Data driven analysis of brain activity and functional connectivity in fMRI

Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von

Silke Dodel
aus München

Göttingen 2002

D7

Referent:	Prof. Dr. Theo Geisel
Koreferent:	Prof. Dr. Ulrich Parlitz

Tag der mündlichen Prüfung: 20. Dezember 2002

Zusammenfassung

In dieser Dissertation werden statistische und graphentheoretische Methoden für die Analyse von fMRT-Daten (fMRT: funktionelle Magnetresonanztomographie) untersucht. Dies beinhaltet sowohl die Vorverarbeitung der Daten als auch die Extrahierung von Aktivitätskomponenten, sowie die Untersuchung von funktioneller Konnektivität und analytische Ableitungen.

Der Hauptaugenmerk liegt dabei auf funktioneller Konnektivität, für deren Analyse wir ein graphentheoretisches Verfahren entwickelt haben, das auf Korrelationsmatrizen basiert. Das Verfahren ist vollständig datengetrieben und benötigt keine vorab definierten Areale (ROI: regions of interest). Darüber hinaus bezieht es sowohl gleichzeitige als auch zeitverschobene Korrelationen mit ein, und identifiziert so zeitliche Beziehungen zwischen funktionellen Einheiten. Der Ansatz ist nicht auf fMRT-Daten beschränkt, sondern kann auch für Daten von anderen bildgebenden Modalitäten, z.B. der Elektroenzephalographie (EEG) und der Magnetenzephalographie (MEG), verwendet werden. Es wird gezeigt, daß es dieses Verfahren erlaubt, interessante Netzwerkstrukturen von funktionellen Einheiten zu extrahieren, die eine Grundlage für die großskalige Modellierung von Hirnfunktionen bilden könnten.

Der graphentheoretische Ansatz basiert auf lokalen Eigenschaften der Daten. Im Gegensatz dazu stehen die multivariaten datengetriebenen Methoden, wie die Hauptkomponentenanalyse (PCA) oder die Analyse der statistisch unabhängigen Komponenten (ICA), die globale Eigenschaften der Daten extrahieren. Diese werden am Anfang dieser Dissertation ebenfalls beschrieben. Die multivariaten Methoden werden analysiert im Hinblick auf ihre Kapazität, identifizierbare Gehirnaktivität von Rauschen zu trennen und werden verglichen mit häufig verwendeten stimulus-getriebenen Verfahren. Darüber hinaus wird die Analyse der Bildvektoren verglichen mit der Analyse der Zeitverläufe und im Fall der Hauptkomponentenanalyse werden analytische Bedingungen für die Gleichheit der beiden Aspekte abgeleitet. Ferner wird die intrinsische Dimension der Daten mit Hilfe der Theorie der Zufallsmatrizen (RMT) abgeschätzt.

Weitere Beiträge in dieser Dissertation sind ein halbautomatischer Algorithmus, der den Bereich des Gehirns aus MRT-Bildern extrahiert, sowie ein Ansatz, um Kopfbewegungen und Pulsationen des Gehirns zu quantifizieren, und eine analytische Ableitung der Stichprobenkorrelation einer endlichen, statistisch unabhängigen, identisch gaußverteilten Stichprobe mit einem festen Referenzvektor unter Benutzung von hochdimensionalen Kugelkoordinaten.

Die Dissertation ist folgendermaßen gegliedert:

Kapitel 1 gibt einen kurzen Überblick über die physikalischen Prinzipien und Techniken, die die Grundlage der Magnetresonanztomographie (MRI) im allgemeinen und der funktionellen Magnetresonanztomographie (fMRI) im speziellen bilden, und gibt eine kurze Einführung in den BOLD-Effekt (BOLD: blood oxygen level dependent).

Kapitel 2 beginnt mit einer Diskussion der statistischen Eigenschaften der Daten. In Abschnitt 2.2 werden Methoden zur Vorverarbeitung wie die Extraktion des Gehirnbereichs aus den Bildern und Ansätze zur Identifizierung von Kopfbewegungen und Gehirnpulsationen abgeleitet. Schliesslich beschreibt Abschnitt 2.3 die Experimente, in denen die Daten, die in dieser Dissertation analysiert werden, gewonnen wurden.

Kapitel 3 führt häufig verwendete stimulus-getriebene Verfahren ein wie Differenzenbilder in Abschnitt 3.1 und stimulus-getriebene Korrelationsanalyse in Abschnitt 3.2. In letzterem Abschnitt wird außerdem die analytische Form der Verteilungsdichte der Stichprobenkorrelation einer endlichen, statistisch unabhängigen, identisch gaußverteilten Stichprobe mit einem festen Stimulusvektor angegeben. Diese wird im Anhang Appendix A mit Hilfe von hochdimensionalen Kugelkoordinaten abgeleitet. Das Ergebnis wird in Abschnitt 3.4 zur Schwellwertbestimmung benutzt, nach der Einführung des allgemeinen linearen Modells (GLM) in Abschnitt 3.3, das dem weit verbreiteten Ansatz des 'statistical parametric mapping' (SPM) zugrunde liegt.

Kapitel 4 diskutiert die Hauptkomponentenanalyse (PCA). Zusätzlich zu den Ergebnissen der Anwendung der Hauptkomponentenanalyse auf fMRT-Daten wird in Abschnitt 4.1 ein Vergleich der zeitlichen und örtlichen Hauptkomponentenanalyse vorgestellt, der auf der Visualisierung des Datenraums und analytischen Bedingungen für Gleichheit beruht. Letztere werden im Anhang Appendix B abgeleitet. In Abschnitt 4.2 stellen wir ferner eine Methode zur Dimensionsreduzierung vor, der die Theorie der Zufallsmatrizen (RMT) zugrunde liegt.

Kapitel 5 führt die Analyse von statistisch unabhängigen Komponenten (ICA) ein, indem ein kurzer Überblick über die Grundlagen dieses Verfahrens gegeben wird. Ferner werden die Ergebnisse von ICA mit denen von PCA verglichen, sowohl im Hinblick auf zeitliche und örtliche Eigenschaften als auch was die Dimensionalität betrifft.

Kapitel 6 entwickelt ein graphentheoretisches Verfahren für die Untersuchung von funktioneller Konnektivität. In dieser Dissertation stellen wir das Verfahren hauptsächlich unter Benutzung von Korrelationsmatrizen vor, es ist aber nicht auf letztere beschränkt. Abschnitt 6.1 diskutiert funktionelle Konnektivität und bisher verwendete Ansätze. In Abschnitt 6.2 wird die Beziehung zwischen Graphen und Korrelationsmatrizen hergestellt, und ein datengetriebener graphentheoretischer Ansatz zur Reduzierung der Korrelationsmatrix entwickelt. Verschiedene Eigenschaften von Untergraphen werden in Abschnitt 6.3 analysiert und ihre Eignung als funktionelle Einheiten wird diskutiert. In Abschnitt 6.4 wird die vorhergehende Methode auf zeitverschobene funktionelle Konnektivität verallgemeinert, indem zeitverschobene Korrelationen und das Konzept von Hypergraphen eingeführt werden. Schliesslich werden in Abschnitt 6.5 andere Maße als die Korrelation für die funktionelle Konnektivität diskutiert und Ergebnisse, die auf der Benutzung dieser Maße beruhen, vorgestellt.

Kapitel 7 faßt die Verfahren und Ergebnisse, die in dieser Dissertation entwickelt und vorgestellt wurden, zusammen und gibt einen Ausblick auf mögliche Weiterentwicklungen.

Abstract

In this thesis the perspectives of statistical and graph-theoretical methods for the analysis of fMRI data are investigated. This involves preprocessing, extraction of components, functional connectivity and analytical derivations.

The main emphasis is on functional connectivity for which we develop a graph theoretical framework based on correlation matrices. The framework is completely data driven and does not rely on previously defined regions of interest. Furthermore, it takes into account both undelayed and delayed correlations, thereby identifying temporal relationships between functional units. The approach is not restricted to fMRI but can be used also in other imaging modalities, including EEG and MEG. It is shown that by using this approach interesting network structures of functional units can be extracted, which could provide a basis for the large scale modeling of brain function.

The graph theoretical approach is based on local properties of the data, in contrast to global multivariate data driven methods such as principal and independent component analysis, which are described earlier in this thesis. These methods are analyzed with respect to their ability to separate identifiable brain activity from noise and compared with commonly used stimulus-driven methods. Furthermore, the analysis based on image vectors is compared with the analysis based on time course vectors, and in the case of PCA analytical conditions for equality of the two aspects are derived. In addition, the intrinsic dimension of the data is estimated by using random matrix theory.

Other contributions include a semi-automatized algorithm to extract the brain area from MR images, an approach to quantify head movements and brain pulsations, and an analytical derivation of the probability density of the sample correlation of a finite Gaussian independent identically distributed (i. i. d.) sample with a fixed reference using high dimensional spherical coordinates.

The thesis is organized as follows:

Chapter 1 gives a short overview of the physical principles and techniques underlying magnetic resonance imaging in general and functional MRI in particular, and briefly introduces the BOLD effect.

Chapter 2 begins with a discussion of the statistical properties of the data. In section 2.2 preprocessing methods are derived such as the extraction of the brain area from images and approaches to identify head movements and brain pulsations. Finally section 2.3

describes the experiments from which the data in this thesis were taken.

Chapter 3 presents common stimulus driven methods, such as difference maps in section 3.1 and stimulus driven correlation analysis in section 3.2. Also in this section the analytical form for the probability density of the sample correlation of a finite i. i. d. Gaussian distributed sample with a fixed stimulus is presented. This is derived in Appendix A using high dimensional spherical coordinates. The result is used in section 3.4 for thresholding, after the presentation of the general linear model in section 3.3, which underlies the popular approach of statistical parametric mapping (SPM).

Chapter 4 discusses principal component analysis (PCA). In addition to showing results of applying PCA to fMRI data in section 4.1, a comparison of spatial and temporal PCA is presented which is based on the visualization of the data space and on analytical conditions for equality. These are derived in Appendix B. In section 4.2 we present a method for dimensionality reduction using random matrix theory.

Chapter 5 introduces independent component analysis (ICA) giving a brief overview of the main approaches in the field and validates the results from ICA against those from PCA, including a comparison of spatial and temporal features as well as aspects of dimensionality.

Chapter 6 establishes a graph theoretical framework for functional connectivity based on correlation matrices. Section 6.1 discusses functional connectivity and approaches commonly used so far. In section 6.2 the relation between graphs and correlation matrices is established, and a data driven graph theoretical method to reduce the size of the correlation matrix presented. Various properties of subgraphs are analyzed in section 6.3 and their suitability as functional units is discussed. In section 6.4 the previous method is extended to delayed functional connectivity by including delayed correlations and introducing the concept of hypergraphs. Finally in section 6.5 the results of using other measures than correlation for functional connectivity is discussed.

Chapter 7 sums up the approaches and results presented in this thesis and discusses further developments.

Notational conventions

Notational convention: Unless otherwise stated matrices are indicated by upper case bold letters ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$), vectors by lower case bold letters ($\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$), and scalars by lower case italic letters (a, b, c) except for one-dimensional random variables which are indicated by upper case italic letters (A, B, C). No notational distinction is made between usual and random vectors.

Notions of Probability: In the notion of probability distributions or probability densities subscripts are omitted. It is understood from the context that $p(x)$ is the probability density or distribution of the random variable X and the probability density or distribution of the random variable Y denoted by $p(y)$ may be a different function. The same applies for the notion of joint and conditional probability distributions or densities.

$\mathbf{A} = \big(a_{ij}\big)_{m\times k}$	$m \times k$ matrix $\mathbf{A}$ having the elements a_{ij} where $i = 1, \ldots, m$ and $j = 1, \ldots, k$
$\mathbf{A_{ij}}$	$(m-1)\times(k-1)$ matrix resulting from an $m\times k$ matrix $\mathbf{A}$ by discarding the ith row and jth column of $\mathbf{A}$
$\mathbf{A_{\{\}j}}$	$m \times (k-1)$ matrix resulting from an $m \times k$ matrix $\mathbf{A}$ by discarding the jth column of $\mathbf{A}$
$\big(\mathbf{a}, \mathbf{b}\big)$	$n \times 2$ matrix with the n-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ as columns
$\mathbf{a} \parallel \mathbf{b}$	the vectors $\mathbf{a}$ and $\mathbf{b}$ are collinear (not necessarily pointing to the same direction)
$diag(d_1, \ldots, d_k)_{m\times k}$	$m \times k$ diagonal matrix having the diagonal elements $d_1, \ldots, d_k$.
$\mathbb{I}_{\mathbf{k}} = \big(\delta_{ij}\big)_{k\times k}$	$k \times k$ identity matrix
$\big(\mathbb{I}_{\mathbf{k}}\big)_{m\times m} = \begin{array}{cc} & \overbrace{\hphantom{1\;\cdot\;\cdot\;1}}^{k} \\ k\left\{\vphantom{\begin{array}{c}1\\ \cdot\\ \cdot\\ 1\end{array}}\right. & \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \end{pmatrix} \end{array}$	$m \times m$ diagonal matrix where the upper left $k \times k$ submatrix is the identity matrix and the $m - k$ lower diagonal elements are zero ($m > k$).

$\mathbf{v} = \big(v_i\big)_k$	$k \times 1$ column vector with the elements v_i where $i = 1, \ldots, k$
$\mathbf{1_k}$	constant k-dimensional vector $\mathbf{1_k} = (\underbrace{1, \ldots, 1}_{k})^T$
δ_{ij}	Kronecker delta $\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$
$[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$	commutator of the matrices $\mathbf{A}$ and $\mathbf{B}$
p(x)	probability distribution of the discrete random variable X or probability density, in case that X is a continuous random variable
EEG	electroencephalography
MEG	magnetoencephalography
EVD	eigenvalue decomposition
fMRI	functional magnetic resonance imaging
fcMRI	functional connectivity magnetic resonance imaging
ICA	independent component analysis
PCA	principal component analysis
ROI	region of interest
SVD	singular value decomposition
a. k. a.	also known as
iff	if and only if
i. i. d.	independent identically distributed
l. h. s.	left hand side
m. m.	mutatis mutandum
r. h. s.	right hand side
w. r. g.	without restricting generality
w. r. t.	with respect to

Contents

Title i

Zusammenfassung iii

Abstract vii

Notational conventions ix

1 Principles of fMRI 1

2 The data set 9
- 2.1 Properties of the data set . . . 9
- 2.2 Data preprocessing . . . 12
- 2.3 The experiments . . . 17

3 Stimulus driven data analysis 19
- 3.1 Difference maps . . . 20
- 3.2 Correlation analysis . . . 21
- 3.3 General Linear Model . . . 26
- 3.4 Thresholding . . . 29

4 Principal component analysis 31
- 4.1 Temporal versus spatial PCA . . . 36
- 4.2 Dimensionality reduction . . . 39

5 Independent component analysis 47
- 5.1 Linear ICA model . . . 47
- 5.2 Prewhitening . . . 48
- 5.3 Criteria for statistical independence . . . 50
  - 5.3.1 Definitions from information theory . . . 50
  - 5.3.2 Mutual information . . . 52
  - 5.3.3 Cumulant diagonalization . . . 52
- 5.4 Results from ICA . . . 53
  - 5.4.1 Actual and imaginary fingertapping . . . 54
  - 5.4.2 Temporal vs. spatial ICA . . . 57

6 Functional connectivity **61**
6.1 Functional connectivity 61
6.2 From correlation matrices to graphs 63
6.2.1 Definitions of correlation 64
6.2.2 Definitions from graph theory 66
6.2.3 Correlation graphs 68
6.2.4 Graph reduction 69
6.3 Subgraphs of $G(\theta)$ as functional units 76
6.3.1 Connectivity components and cliques 76
6.3.2 Revealing subgraph structures from threshold dependencies 82
6.4 Delayed functional connectivity 93
6.4.1 Delayed correlation matrices 93
6.4.2 Hypergraphs 94
6.4.3 De-autocorrelation 95
6.4.4 Delayed intra-correlations 97
6.4.5 Results for hypergraphs from cliques and connectivity components . 98
6.5 Graphs from other matrices 98

7 Discussion **103**

A Mapping probability densities **109**
A.1 Sample correlation 111
A.2 Generalized spherical coordinates 114

B Equality of temporal and spatial PCA **117**
B.1 Equality of the time courses 117
B.2 Equality of the eigenimages 125
B.3 Equality of both time courses and images 126

C Correlation graphs **127**

References **130**

Acknowledgement **140**

Curriculum Vitae **143**

Chapter 1

Principles of fMRI

Functional magnetic resonance imaging (fMRI) in neuroscience refers to imaging of brain activity using MRI. In this chapter we give a short overview about the physical principles and techniques underlying magnetic resonance imaging (MRI) in general and functional MRI in particular. As sources were used mainly [40, 46, 58, 99].

Historical remarks

MRI is based on nuclear magnetic resonance (NMR), the word 'nuclear' omitted because it had negative connotations in the late 1970's, at a time MRI technology was emerging as a viable imaging technique. NMR refers to the phenomenon that when certain nuclei are placed in a magnetic field they absorb energy in the radiofrequency range of the electromagnetic spectrum, and re-emit this energy when the nuclei transfer to their original state. NMR was first observed in the 1930's by Isidor Rabi [62] in molecular and atomic beam experiments. 1946 Felix Bloch [39] and Edward Purcell [38] independently observed it in bulk material and Bloch gave a theoretical description [14]. In 1971 Raymond Damadian [28] used NMR to detect tumor tissue. A backprojecting technique for retrieving spatial information in an NMR experiment was developed in 1973 by Paul Lauterbur [71] who also proposed frequency encoding. In 1975 Richard Ernst [1] performed magnetic resonance imaging using phase and frequency encoding along with the Fourier transform. This is the basis of current MRI techniques. A few years later, in 1977, Raymond Damadian demonstrated MRI of the whole body and Peter Mansfield [81] developed the echo-planar imaging (EPI) technique now frequently used in fMRI. FMRI emerged at the beginning of the 1990's [70, 97, 9, 44, 13] after the discovery that the level of cerebral blood oxygenation can influence the signal intensity of MR images [90, 107].

NMR

NMR is a resonance phenomenon of nuclei with a nonzero spin in a magnetic field. Nuclei with an odd number of protons or neutrons have an angular momentum or spin $\hbar\mathbf{I}$ that is associated with a magnetic moment μ which can be taken as parallel to $\mathbf{I}$, i. e.

$$\mu = \gamma\hbar\,\mathbf{I} \tag{1.1}$$

where the scalar γ is called the *gyromagnetic ratio* and $\hbar$ is the Planck constant. Without an outer magnetic field all spin orientations have the same energy. The application of a magnetic field $\mathbf{B}_0 = B_0\mathbf{e}_z$ which is assumed w. r. g. to be in z-direction leads to Zeeman splitting of the otherwise degenerated energy levels. It produces an interaction energy of the nucleus which can be described by the Hamiltonian

$$\mathfrak{H} = -\mu \cdot \mathbf{B}_0 = -\gamma\hbar\, B_0\, I_z \tag{1.2}$$

where I_z is the spin component in z-direction. The eigenvalues of the Hamiltonian are

$$E = -\gamma\hbar\, B_0\, m \tag{1.3}$$

with $m \in \{-I, -I+1, \ldots, I-1, I\}$ being the eigenvalues of I_z and I the eigenvalue of $\mathbf{I}^2$. Using spectral absorption it is possible to detect the presence of such a set of energy levels. The coupling most commonly used to induce an interaction that causes transitions between the energy levels and hence magnetic resonances is an alternating magnetic field $\mathbf{B}_1$ applied perpendicular to the static field $\mathbf{B}_0$. Let us take w. r. g. $\mathbf{B}_1 = B_1\mathbf{e}_x$ to be in x-direction. The perturbing term in the Hamiltonian is then

$$\mathfrak{H}_{\text{pert}} = -\gamma\hbar B_1\, I_x \cos\omega t \tag{1.4}$$

where I_x is the spin component in x-direction. The allowed transitions can be derived to be those between levels adjacent in energy, giving as resonance condition for ω

$$\hbar\omega = \Delta E = \gamma\hbar B_0 \tag{1.5}$$

The ω satisfying the resonance condition is called the *Larmor frequency*. For protons the energy difference ΔE at a magnetic field strength of 2.0 T is 0.35 μeV which corresponds to a resonance frequency of 85.2 MHz, which lies in the radio frequency range. The HF excitation is therefore often called radio frequency (RF) excitation. Let us now consider for simplicity the case of a proton where according to Equation 1.3 with $I = \frac{1}{2}$ and hence $m = \pm\frac{1}{2}$ there are two energy levels in the presence of a static magnetic field $\mathbf{B}_0$. The proton is the nucleus of ^{1}H which has the highest natural abundance (99.985% [108]) and 0.63 biological abundance (calculated from [108] by [58])) and is therefore the one mostly used in MRI.

$\mathbf{B_1} = 0$:
In thermal equilibrium with only the static magnetic field $\mathbf{B}_0$ present the energy levels are populated according to Boltzmann distribution $p(E)$ as

$$p(E) \propto e^{\frac{E}{kT}} \tag{1.6}$$

where k is the Boltzmann constant and T the temperature. For two energy levels this leads to a population ratio at equilibrium of

$$\frac{N_+}{N_-} \propto e^{-\frac{\Delta E}{kT}} \tag{1.7}$$

where N_+ and N_- are the numbers of nuclei occupying the upper and lower energy level, respectively. This leads to a net magnetization in z-direction, the direction of $\mathbf{B}_0$. The macroscopic magnetization $\mathbf{M}$ satisfies the equation

$$\frac{d\mathbf{M}}{dt} = \gamma\,\mathbf{M} \times \mathbf{B}_0 \tag{1.8}$$

which is true also for the expectation values of the spin magnetic moments.

$\mathbf{B}_1 \neq 0$:
The alternating radio frequency (RF) field $\mathbf{B}_1$ induces transitions from the lower to the higher energy level thereby diminishing the population difference which again leads to a decrease in the longitudinal magnetization and the buildup of transversal magnetization which is due to a phase coherence of the spins. Classically this can be described as tilting the magnetization vector $\mathbf{M}$ such that it has a nonzero transversal component and is hence precessing with larmor frequency around the direction of $\mathbf{B}_0$. This is called *free precession*. If there were no interaction between the spins and the surrounding lattice, as are called the surrounding atoms, after switching off $\mathbf{B}_1$ the precession of $\mathbf{M}$ would simply continue, but because of the interaction the transversal part of $\mathbf{M}$ decays and the longitudinal part relaxates to the equilibrium value $\mathbf{M} = M_0\mathbf{e}_z$. Without further interaction occuring this is called the *free induction decay* (FID).

Relaxation:
The relaxation process of the free induction decay was phenomenologically included to Equation 1.8 by Bloch leading to

$$\frac{dM_x}{dt} = \gamma\,(\mathbf{M} \times \mathbf{B}_0)_x - \frac{M_x}{T_2} \tag{1.9}$$

$$\frac{dM_y}{dt} = \gamma\,(\mathbf{M} \times \mathbf{B}_0)_y - \frac{M_y}{T_2} \tag{1.10}$$

$$\frac{dM_z}{dt} = \gamma\,(\mathbf{M} \times \mathbf{B}_0)_z + \frac{M_0 - M_z}{T_1} \tag{1.11}$$

where T_1 and T_2 are the time constants of the decay of the longitudinal and transversal magnetization, respectively. The longitudinal relaxation is due to spin-lattice interactions and hence called *spin-lattice relaxation*. After switching off the excitation by the RF-field $\mathbf{B}_1$ more transitions from the upper to the lower energy level occur than vice versa leading to an increase of the population difference and hence to an increase of the longitudinal magnetization. The decay of the transversal magnetization is mainly due to spin-spin interactions and hence called *spin-spin relaxation*. It is due to the exchange of energy within the spin ensemble which leaves the population difference of the energy levels unaffected but leads to a dephasing of the spins and hence to a decrease in transversal magnetization. Dephasing due to spin-lattice interactions are included in T_2 as well. It is always $T_2 \leq T_1$. In practical applications often the time constant T_2^* is encountered. This time constant plays an important role in functional neuroimaging as will be seen below. It includes the dephasing effect of inhomogeneities in the static magnetic field $\mathbf{B}_0$ which leads to local differences in the resonance frequency and hence to a change in the phase relations of the spins which results in a signal loss. The relation between T_2 and T_2^* is

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_{2\,inh}} \tag{1.12}$$

where $T_{2\,inh}$ is the relaxation time from the inhomogeneous T_2 effect. The relaxation times T_1 and T_2 depend on factors as the motility of the spins and the lattice (temperature, state of aggregation), their interaction probability (concentration), the magnetic properties and the resonance frequency of the spins (magnetic field). T_2^* is determined by macroscopic and microscopic magnetic field variations, e. g. inhomogeneities of the static magnetic field, differences in susceptibility (particularly at tissue-air-boundaries) or magnetic moments of other particles. Typical values for biological tissue at a magnetic field strength of 2 T are $T_1 = 1000$ ms and $T_2 = 100$ ms. T_2 and T_2^* weighted images can be obtained using spin echo and gradient echo, respectively.

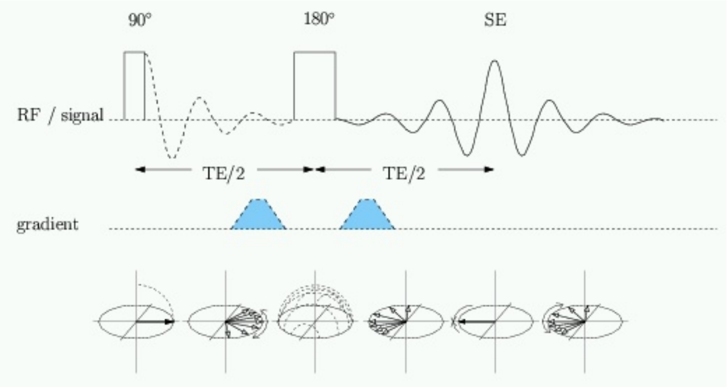


Figure 1.1: Illustration of the spin echo sequence (courtesy of [40], adapted). After the RF pulse, the spins dephase due to local field inhomogeneities. Applying a 180° pulse flips the spins in the directions which are illustrated in the third cartoon on the bottom of the figure. This leads to a refocussing of the spins and a signal of opposite phase to that observed right after the RF pulse. The gradients depicted are used to suppress remains of FID signals resulting from non optimal flip angles.

Spin Echo:
The signal loss due to dephasing effects caused by magnetic field inhomogeneities can be recovered when using a second RF excitation: an appropriate RF pulse applied at a time TE/2 after the initial RF excitation flips the magnetization by 180° and thus leads to refocussing after another time TE/2 (cf. Figure 1.1). The corresponding signal ist called spin echo (SE), its amplitude depends on T_2 because dephasing effects due to spin-spin-interaction can not be compensated. TE is called the echo time. The method of spin echo is used to obtain T_2 weighted images.

Gradient Echo:
T_2^*-weighted images can be obtained by using a so called gradient echo instead of a RF-refocussing echo. Here the refocussing of the spins is achieved by a reversal of the applied

field gradient. Since the local field inhomogeneities remain in their original direction the gradient echo does not cancel out their effect. This results in a strong signal loss, e. g. at tissue-air boundaries, but on the other hand provides sensitivity to local field changes by blood oxygenation and hence can be used to visualize metabolic effects such as occur in consequence of neural activity (see below). The gradient echo sequence is illustrated in Figure 1.2.

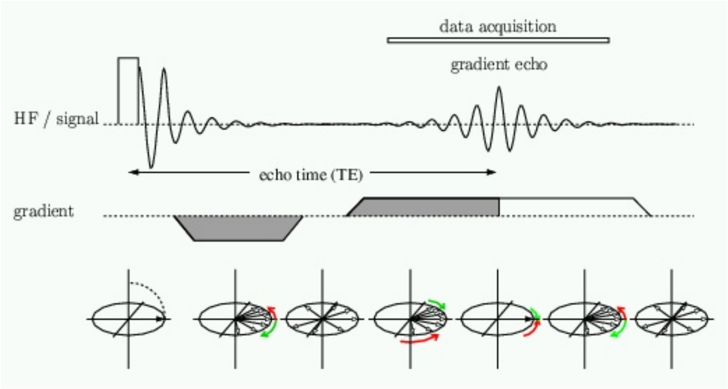


Figure 1.2: Illustration of the gradient echo sequence (courtesy of [40], adapted). After the RF pulse, the spins dephase due to local field inhomogeneities. Reversing the direction of the gradient field reverses the direction of precession and leads to a refocussing of the spins.

Spatial encoding

For spatial encoding additional magnetic fields are used that point in the same direction as the static magnetic field $\mathbf{B}_0$ the strength of which can be varied both temporally and spatially. The spatial gradients of the additional magnetic fields, i. e. the gradients of the magnetic field strength B_z, point to three orthogonal directions and can be combined to constrain the excitation to a volume of arbitrary shape [40]. For simplicity here we describe the spatial encoding of a plane perpendicular to the direction of $\mathbf{B}_0$ which was defined as the z-direction. Spatial encoding is performed at different points in time and can be divided into three steps as follows

1. RF-excitation (slice selection) ⟶ z-direction
2. before data acquisition (phase encoding)
3. during data acquisition (frequency encoding)

$$\left.\begin{array}{l}\text{2.}\\\text{3.}\end{array}\right\}\ x\text{- and } y\text{-direction}$$

1: A RF-excitation which approximates the temporal profile of a $\text{sinc}(t) = \frac{\sin(t)}{t}$ function is used to define a sharp range of excitation frequencies. Together with a static magnetic field having a gradient in z-direction which imposes a spatial modulation of the resonance frequency this confines the excitation to an essentially two dimensional plane section.

2 and 3: Frequency and phase of the transversal magnetization are used to encode the x- and y-direction by modulating the transversal magnetization by magnetic gradient fields, the gradients of which point in x or y direction, respectively. The phase encoding gradient is switched on after the termination of the RF-excitation and leads to a spatially varying phase of the previously phase coherent magnetization. For frequency encoding a gradient field is switched on during the data acquisition time which leads to a spatially dependent change of the precession frequency. For the signal $V(t)$ in the receiver coil during data acquisition we have

$$V(t) \propto \int M_{\perp}(\mathbf{r},t)\, e^{i\,(\omega y t+\phi x)}\, d\mathbf{r} \tag{1.13}$$

where $M_{\perp}(\mathbf{r},t)$ is the modulus of the transversal magnetization, and ω and ϕ are the space dependent precession frequency and phase, respectively, depending linearly on the vector $\mathbf{r} = (x,y)^T$. Since the transversal magnetization outside the slice chosen by the RF-excitation is essentially zero we have considered Equation 1.13 as an integral over two dimensions. Using the notation $\mathbf{k}(t) = (k_\omega, k_\phi)^T = (\omega\, t, \phi)^T$ we can write Equation 1.13 as

$$V(t) \propto M_{\perp}(\mathbf{k},t) = \int M_{\perp}(\mathbf{r},t)\, e^{i\,\langle \mathbf{k}(t),\mathbf{r}\rangle}\, d\mathbf{r} \tag{1.14}$$

Thus for every point in time the signal is proportional to the spatial Fourier transform of the modulus $M_{\perp}(\mathbf{r},t)$ of the transversal magnetization. Since only k_ω is time dependent and k_ϕ remains constant within one period of signal acquisition several cycles of RF-excitation and signal acquisition using various strengths of the phase encoding gradient are necessary to cover the two-dimensional **k**-space. After having sampled the **k**-space the space dependent transversal magnetization $M_{\perp}(\mathbf{r},t)$ is retrieved by a two-dimensional spatial Fourier transform and by this means an image created.

What was described above is essentially the principle underlying the FLASH (fast low angle shot) sequence. Using only one RF-excitation and incrementing the phase encoding gradient by a fixed value one arrives at the echo-planar imaging (EPI) sequence which allows faster image acquisition at the expense of lower spatial resolution. Both sequences are commonly used in fMRI and illustrated in Figure 1.3. The trajectories through k-space are shown in Figure 1.4.

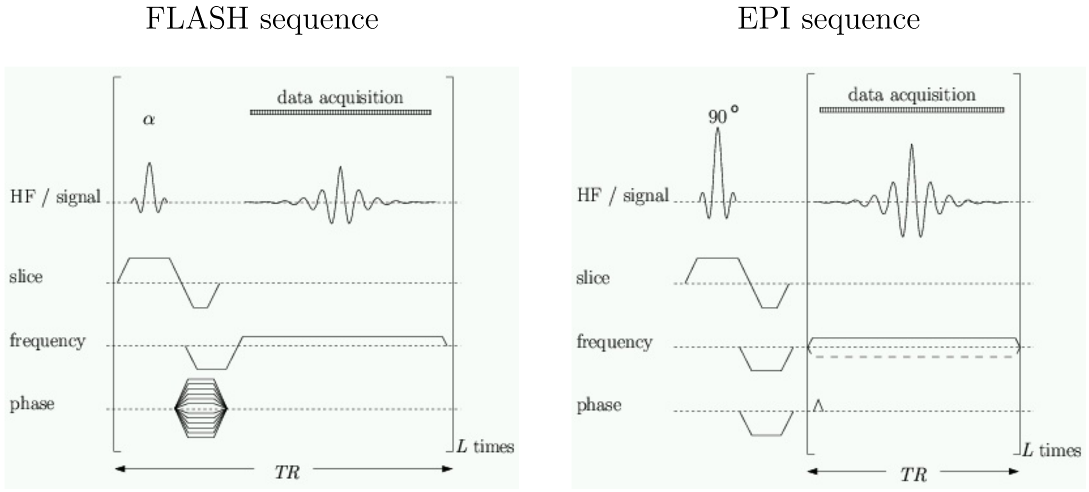
The total NMR signal (the reconstructed image) of a voxel is proportional to

$$M_Z(1-e^{-\frac{TR}{T_1}})e^{-\frac{TE}{T_2}} \tag{1.15}$$

where M_Z is the equilibrium magnetization in z-direction (the direction of the static magnetic field $\mathbf{B}_0$). TR is the repetition time between the RF-excitation pulses and TE the time between echo excitation. M_Z is proportional to the proton density in the voxel.

BOLD

The blood oxygenation level dependent (BOLD) effect on the T_2^* relaxation time can provide MR images with functional sensitivity. The BOLD effect is due to the different



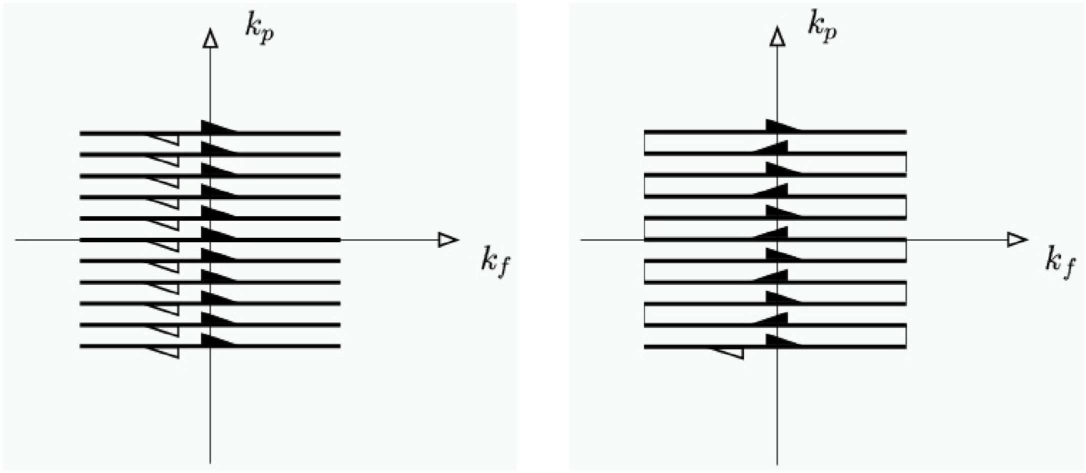
(a) (b)

Figure 1.3: Illustration of the FLASH and EPI sequences (courtesy of [40], adapted). The sequences in brackets are repeated L times during the experiment. In (a) α denotes the flip angle of the macroscopic magnetization $\mathbf{M}$. α is usually chosen to be low to prevent saturation of the signal for short repetition times TR [53]. The gradients used for spatial encoding are denoted by their encoding quantities, slice, frequency, and phase, respectively. The slice and frequency encoding gradient are reversed before data acquisition to obtain a gradient echo.

magnetic properties of oxyhaemoglobin (HbO_2) and deoxyhaemoglobin (Hb). The former is diamagnetic whereas the latter is paramagnetic leading locally to a shorter T_2^* relaxation time and hence to a signal loss in T_2^* weighted images [53]. Haemoglobin therefore can be used as a natural intrinsic contrast agent which is a considerable advantage over functional imaging by Positron Emission Tomography (PET), where slightly radioactive extrinsic contrast agents are required.

The functional T_2^* signal depends on the balance of spatial and temporal alterations in local concentrations of Hb and HbO_2 [106]. PET studies have shown that cortical activity results in oxygenation changes due to an increase in local blood flow with relatively little change in oxygen consumption, so that the venous blood becomes more oxygenated on activation [42, 41]. This leads to a signal increase of T_2^* weighted images in the active region. There is also a T_2 effect of oxygenation, which could be used to distinguish between large and small vessels [53], however the T_2^* effect is larger by a factor of 3-10 [106].

From what was stated above it is clear that fMRI does not measure neural activity directly but rather the signal induced by the related hemodynamics and relative oxygenation. The hemodynamic response to neural activity is nonlinear and yet not fully understood, however recently Logothetis and co-workers [78] by simultaneous electrophysiology and fMRI showed that the fMRI signal is more closely related to local field potentials than to multiunit spiking activity suggesting that it reflects the input and intracortical processing of a given area rather than its spiking output. Further a couple of models relating blood



(a) (b)

Figure 1.4: Trajectories through k-space for (a) FLASH and (b) EPI (courtesy of [40], adapted). The filled vectors denote the sampling direction during data acquisition.

flow and oxygenation changes to brain activations have been proposed [18, 19, 76, 51].

The bold signal in response to brain activity using a single external stimulus reveals that the maximum of the signal occurs at a time delay in the order of 6-8 seconds, whereas the onset of the signal change occurs already after 1.5-2 seconds [46]. An undershoot of the signal relative to the beginning of the stimulus occurs a few seconds after the end of the stimulus. This is attributed to volume changes [77, 19, 45, 69]. The presence of an initial dip in the signal which is believed to be due to blood deoxygenation right after the onset of increased neural activity before the subsequent hyperoxygenation is controversial [66, 47, 17]. Even very short activation elicits considerable response as was shown in [48]. However local image intensity increases will also be dependent on differences in haemodynamic (blood volume, flow and oxygenation) and vessel architecture (radii, orientation, vascular openness).

Chapter 2

The data set

In this chapter in section 2.1 we discuss general statistical properties of fMRI data which are to be taken into account in the choice and adaptation of processing algorithms. In section 2.2 we derive preprocessing procedures that extract the brain area from fMR images as well as approaches to cope for movement and brain pulsation effects. Finally in section 2.3 we describe briefly the paradigms and methods of the experiments from which the data used in this thesis were taken.

2.1 Properties of the data set

The point of view adopted in the following is a statistical one. This means that the data $\mathbf{X}$ are considered as a sample drawn from a presupposed probability distribution $p(\mathbf{X})$, that is generally unknown. However, repetitions of fMRI experiments in different sessions are comparable only to a very limited extent, and hence do not allow for an interpretation as multiple realizations of an underlying probabilistic model. By averaging the results over multiple subjects often a coarse localization of brain activity can be obtained, if the images are appropriately aligned. Statistically, nonzero averages in this context occur on the account of common properties of multiple underlying probabilistic models. The lack of comparability is due to nonstationarities in metabolism, arousal and attention in test subjects, as well as to readjustment of the measurement device between the individual sessions. Because these nonstationarities are considered to be more severe than inhomogeneities within a single trial, in this thesis the focus is on single subject trials also allowing a more detailed analysis of the data.

A number of assumptions about properties of the probability distribution p are to be made in order to be able to retrieve meaningful information. This essentially touches the issues of stationarity, homogeneity, and independence of the data. In this respect two features are particularly striking. First the data are strongly non-stationary in the sense that a large fraction of the time courses of the individual voxels show an either increasing or decreasing trend. Although these trends are easily removable, similar trends may provide evidence that the areas represented by the respective voxels are engaged in related processes (cf. chapter 6). Secondly, there is a clear difference between the statistics of signals originating from the brain and those from extra-cranial regions.

Marginal distributions averaging over both intra- and extra-cranial regions are generally bimodal to such an extent that the extra-cranial voxels of the data images can be removed reliably based on thresholding of appropriate histograms of the data (cf. section 2.2).

Homogeneity of the voxels within the brain is taken as a null-hypothesis, i.e. deviations from homogeneity are interpreted as meaningful information which is then to be separated into various processes presupposedly influencing the data. Thus a main assumption is that the data are generated by a small set of partially separable processes such as respiratory and cardiovascular rhythms, involuntary head movements, various brain-activity related processes, and systematic effects of the measurement procedure. These processes can be grouped into two classes: they either affect the BOLD signal directly or the brain is moving relative to the measurement grid which leads to time-dependent spatial shifts or to crosstalk among voxels. Further image distortions occur at tissue air boundaries and on account of the Fourier reconstruction process from which the images are obtained (cf. chapter 1). In section 2.2 approaches to eliminate movement related effects are discussed that try to establish a unique correspondence of image voxels and positions in the brain based on spatio-temporal continuity assumptions. Since it is reasonable to assume that movement related processes are independent from cerebral information processing they can be also separated using information theoretic methods such as in chapter 5.

Formally the data are characterized by a spatio-temporally discrete set of voxels forming a sequence of volume or slice images. Each image containing m voxels can be considered as a vector in an m-dimensional vector space. Having measured k images the data can be written into an $m \times k$ data matrix $\mathbf{X}$ the columns of which correspond to the images and the rows of which represent the time courses of the signal in the corresponding voxels. Figure 2.1 illustrates this concept. Emphasizing either temporal or spatial aspects of the data one can deal with the row space $\mathbb{R}^k$ or the column space $\mathbb{R}^m$ of $\mathbf{X}$. The intrinsic dimension of the data set is of course independent of the representation, but the statistical properties of the sample vectors may be different. In most cases the number of voxels highly exceed the number of images measured, often by about one order of magnitude, therefore henceforth $m > k$ is assumed.

Which representation of the data is to be favored depends on the questions to be answered and on the assumptions that are made. Spatial localization of brain functions could be determined by focusing on the the set of images, whereas for the identification of distributed activity looking for time courses of interest may be more meaningful. The spatial and temporal aspects of the data are closely interconnected since one cannot change one aspect without affecting the other. However, the relation of the information extracted from the two representations is not obvious and will be discussed in some detail in Appendix B.

From a geometrical point of view the data set is simply a point cloud in an m- or k-dimensional data space the location of the points being determined by the column or row vectors, respectively. The shape of the data cloud is assumed to reflect features of interest in the data, hence, we aim at determining its structure. Statistically the data cloud can be interpreted as a realization of drawing m or k random vectors from an underlying probability distribution. The joint probability density of the data as stated above is unknown and the sample size the data provides is far too small for a reliable estimation. To illustrate this fact assume that we are given data from $k = 128$ images having $m = 1200$ voxels, the signal x of which is in the range of $x \in (0, 200]$ binned at the integer values into $n = 200$ bins. To estimate the joint probability density of the vectors in row space which are k-dimensional we have n^k bins and only mk data points to fill them, which with the values chosen gives a relation of $\frac{mk}{n^k} \approx 10^{-290}$. However assuming a maximum likelihood principle we can hope that the realization provided by the measurement is typical in the sense that the data has high probability to occur, i.e. it lies at the peaks of the underlying probability density. We aim at extracting brain states that are reflected by activity patterns that are interesting in some sense. The term activity pattern here is used for images or time courses or more generally any spatiotemporal subset computed from the data matrix.

Statistical properties when they occur are introduced using a continuous formulation however when computing them from the data we can compute only their discrete finite sample counterparts. Statistical inferences on the data are based on various assumptions about the underlying probability distributions. In the following chapters we will discuss data analysis methods based on geometrical and on probabilistic viewpoints, respectively, relating the two to each other where this is possible.

Since the data are subject to strong noise, which is in part independent for individual voxels, the rank of $\mathbf{X}$ and thus the intrinsic dimension is generally equal to k. The number of relevant dimensions, however, can be expected to be considerably smaller, which is already suggestive from visual inspection of the first principal components (cf. chapter 4), where except for about the first 20 no structure is apparent. Whether or not this observation is justified can be tested by comparison of the statistics of the components with components from a random matrix with independent entries, cf. section 4.2. Other data reduction algorithms which also account for nonorthogonality of the relevant subspaces (cf. chapter 5) or for nonlinearity of relevant submanifolds (such as selforganizing maps, cf. [68]) or nonstationarity (temporally local selforganizing maps) have been shown not to provide any conflicting evidence to this data reduction assumption.

More of a problem is the temporal resolution of fMRI data which is of the order of one second. The relevant information-processing operations in the brain, however, are expected to run on time-scales of less than 100ms. Even if the temporal resolution can be increased in future the nature of the measured signal, i.e. the delayed BOLD-response, amounts to a convolution with a kernel having a width of about six seconds. This leads to a signal-to-noise ratio in temporal resolution of about one hundredth due to covering of the fast cerebral activity by the slow metabolic processes. Therefore operations on a neural time scale will generally not be identifiable, however, the nonlinear nature of the hemodynamics leads to a measurable response also for very short stimuli and for stimuli with an interstimulus time interval of less than six seconds [43]. The variety of physiological interdependences and their influence on the signal suggest that including models thereof probably could significantly improve the analysis of fMRI data, particularly if applied to a concrete system (visual, motor, etc.), where appropriate

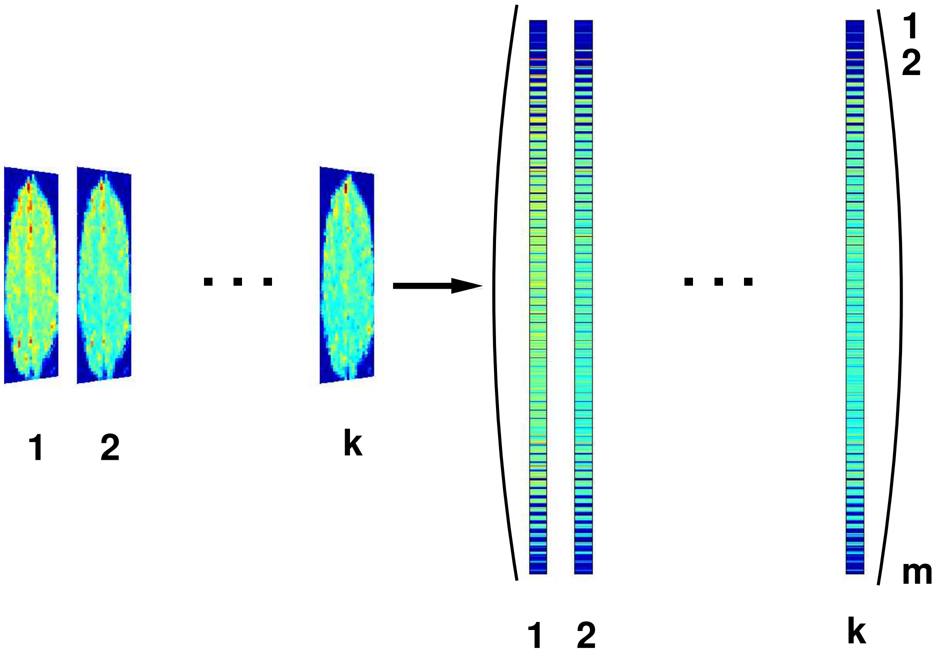


Figure 2.1: Illustration of the data matrix $\mathbf{X}$ in which each column represents an image vector and each row a voxel time course. The left hand side shows the image data measured at k points in time. Rearranging the voxels of each image into a column vector gives the data matrix $\mathbf{X}$ shown on the right hand side.

models exist. However this was beyond the scope of the thesis presented here. Yet, in chapter 6 we will investigate the trace of repeated interactions among brain regions by the analysis of time-delayed correlations which may be identified as effective functional connectivity although this leaves fast interactions unrevealed.

2.2 Data preprocessing

Apart from neural activity there are various other influences on the signal, some of them being unwanted. Head movement and brain pulsation, both influences impossible to prohibit in the experimental setup, can cause leakage of the signal to neighboring voxels. In this section we discuss possibilities of data preprocessing aiming at minimizing these effects.

Brain extraction

For global data driven methods as they are discussed in this thesis each image is considered as a vector in a high dimensional vector space. Using a rectangular image including voxels outside the brain can influence the results of the data driven methods. Further the amount of data may be prohibitive for some application as e. g. computing the temporal correlation matrix, if the brain is not extracted. Here we present a semi-automatized brain extraction method that relies on the temporal statistics of the voxel values.

The idea is that the histogram of the signal of extra-cranial voxels is highly peaked around zero (or another baseline value), since it contains only noise whereas the signal of a voxel in the brain has a broader distribution with a higher mean value. We chose the voxel in the upper left edge of the rectangular (or square) data image as a reference since it is extra-cranial in all data sets we analyzed. The histogram of its values accross time is computed and its standard deviation σ used as a reference. A user defined number p assigns to the brain all voxels for which the mean of their histograms exceeds $p\sigma$. Figure 2.2 illustrates the procedure for data from experiment 2. It would be easy to fully automatize the procedure by fitting a unimodal density to the histogram and varying p such that the fit becomes optimal.

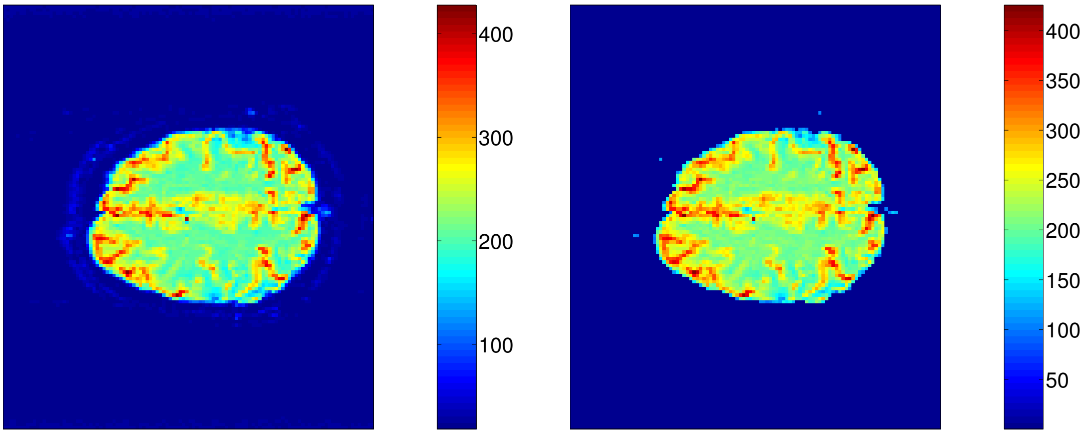
Rigid body transformations

Head movements can be considered as rigid body transformations thereby neglecting possible brain pulsations and physiological signal changes. By defining one image as a reference and shifting the subsequent images by an amount that minimizes a distance measure between the images the effect of head movements can be computationally minimized. One could even expect to increase spatial resolution by this means. If the head is moved by an amount that is smaller than the size of a voxel this shifts the voxel boundaries as well. Imposing a continuity assumption the signal is taken to be essentially constant from one time step to the next. A change in the signal then must be due to a spill-over from neighbouring voxels. Interpolation using the continuity assumption can be done to create an image where the voxels are resized to the finer grid provided by the combination of the shifted and unshifted voxel boundaries.

When only one slice is measured, head movements most probable cause signal from areas adjacent to the slice to spill into the image such that essentially the slice is shifted or rotated from its original location in an uncontrolled way. In principle one could detect the direction of the head movement by modeling it as a shift plus a rotation and fitting the corresponding parameters under the assumption that the movement changes the spatial structure of the image continuously. However the signal from the original slice cannot be retrieved if the head movement does not occur in a direction parallel to the slice. Under the assumption that they occur in-slice Figure 2.3 shows the directions of rigid body head movements occuring on the spatial scale of fractions of the voxel size.

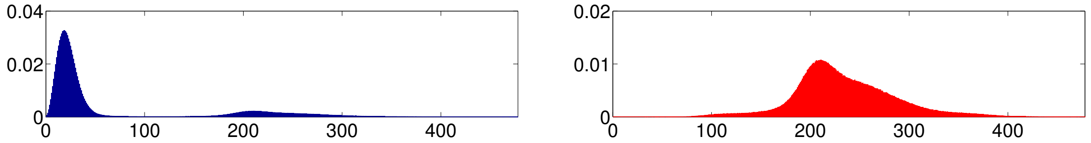
Pulsations

Pulsations in the vicinity of large vessels and at a global scale represent a further source of spill-over of signals to neighbouring voxels. To detect the centers of pulsations the direction of spill-overs can be stored in a vector field the locations of nonzero divergence of which give the centers of pulsations. The direction of spill-over from one image to the next is thereby computed by determining the voxel in a neighborhood of a given voxel the signal of which has the least difference to that of the given voxel. Various weighting



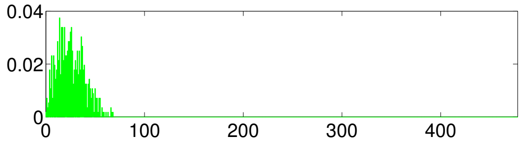
(a) Mean image before brain extraction.

(b) Mean image after brain extraction with $p = 2.5$.

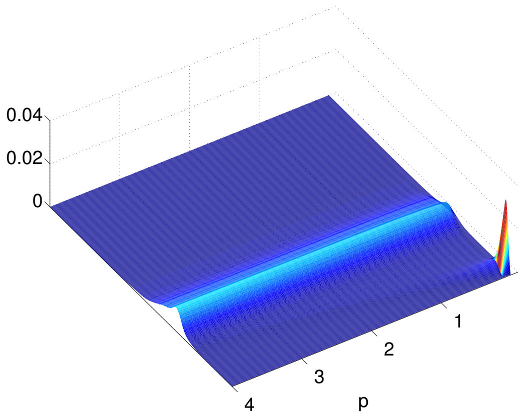


(c) Relative frequency of voxel values before brain extraction.

(d) Relative frequency of voxel values of the extracted brain.



(e) Relative frequency of the values of the voxel in the upper left corner of the image.



(f) Relative frequencies of voxel values after brain extraction for various values of p.

Figure 2.2: The method of brain extraction. Mean images and relative frequencies of the voxel values before and after brain extraction.

techniques can be employed when defining the distance measure. Figure 2.4 shows the result of the approach for data from experiment 2.

Another possibility to cope with pulsations is the use of the Fourier transform in time.

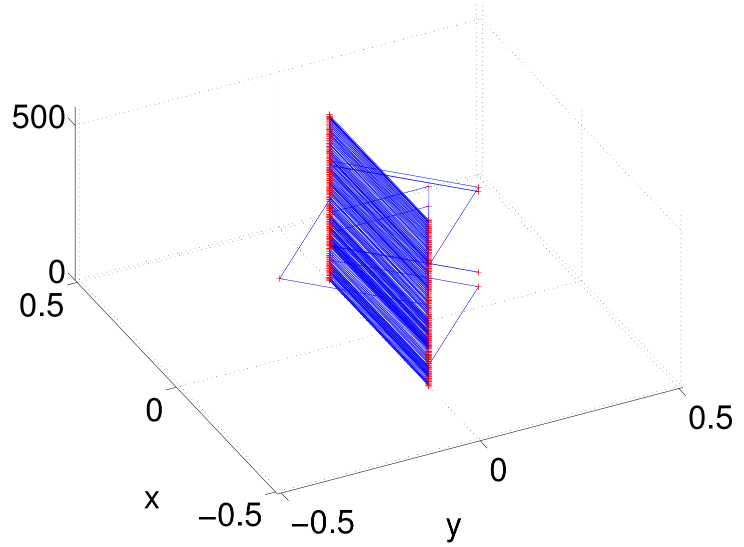


Figure 2.3: Direction of head movements in the units of voxel size for data from experiment 2 under the assumption that the movement is in slice. A strong prevalence for the x-direction can be observed, which corresponds to a forward-backward movement. Note that the amplitude of the movement is very small being of about a fraction of 0.25 of the size of one voxel.

Since it is reasonable to assume that the pulsations are related to heart beat and breathing one can extract the voxels having the most significant signal contributions at the respective frequencies. However this approach requires a temporal resolution which makes it possible to resolve the heart rate in frequency space. Assuming a heart rate of about 1 Hz the corresponding temporal resolution must be 0.5 s.

Spatial and temporal filtering and interpolation

In fMRI often prior to analysis the data are spatially or temporally filtered, sometimes even both. This increases the robustness of the results, however, it further deteriorates the spatial or temporal resolution. Temporally the data are given on a discrete time axis with time bins of size Δt. The acquisition time of a single slice is given by $\frac{\Delta t}{n}$, where n is the number of slices in the volume. Usually when acquiring volume images the odd slices are measured first, followed by the even slice numbers (or vice versa) to minimize interferences of spatially adjacent slices during the measurement.

A promising approach seems to us to spatio-temporally interpolate the data, particularly in the case of volume images where the spatio-temporal resolution usually is low, but the time in which one slice is acquired is high. Furthermore by temporal interpolation the true temporal relation between the slices is preserved. The interpolation becomes spatiotemporal by using information of the neighbouring slices which are rather far apart in temporal respect, but spatially close together.

All preprocessing methods however implicitly rely on the suitability of the underlying

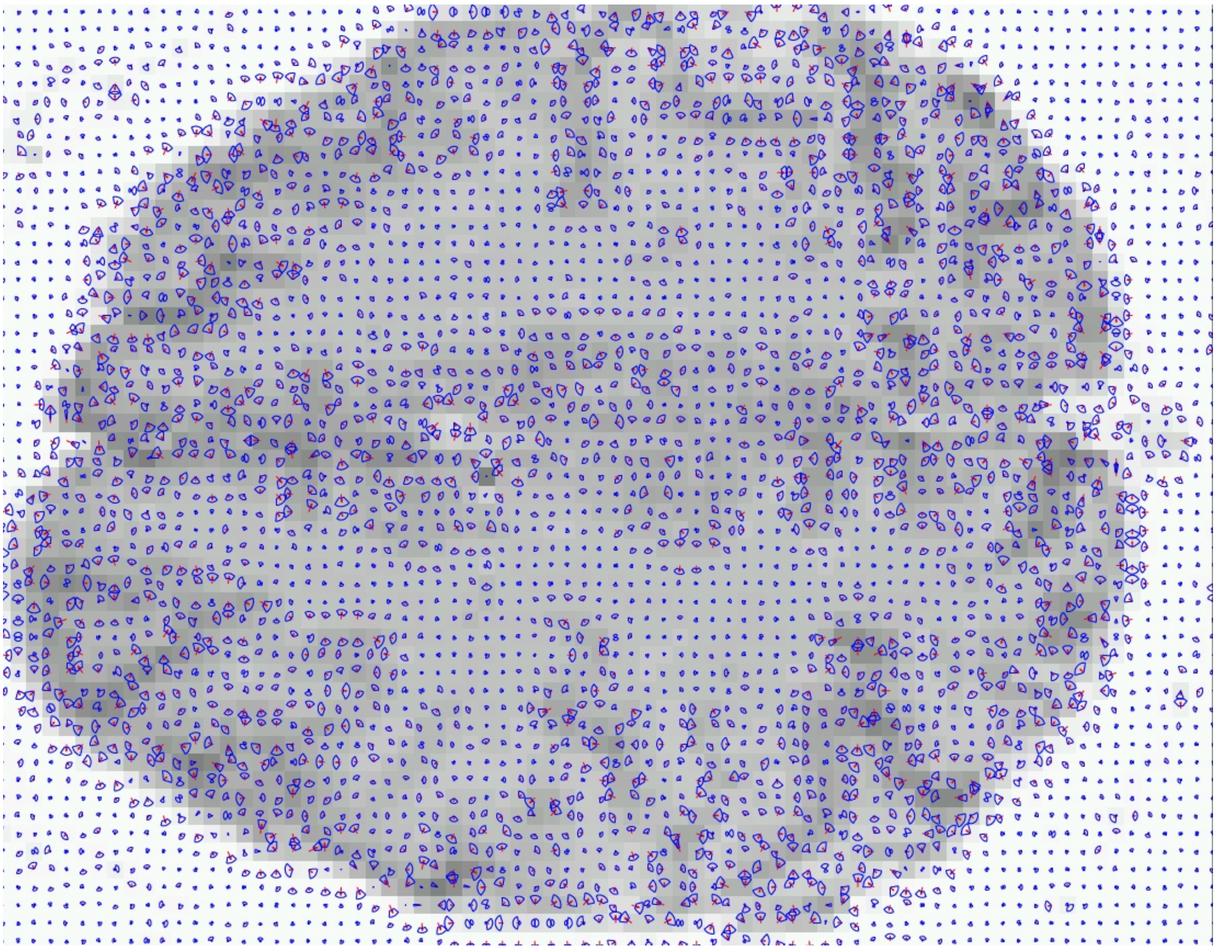


Figure 2.4: Brain pulsations for data from experiment 2 overlaid the mean image of the data. The red arrows indicate the mean vector field of voxel movements and the blue boundaries indicate the standard deviation of the length and the angles of the arrows. The values of the mean vector field are so small that they preclude a reliable computation of the divergence. This could be done, however, using a sliding window and thus computing the mean vector field from a smaller number of images. Further, including to the model an elastic interaction term penalizing shearing strains between neighbouring voxels of the same image would probably also ameliorate the results. From visual inspection it is apparent that the centers of divergence are located along the gyri and sulci of the brain.

model. The latter therefore should be carefully analyzed and the effects of the preprocessing strategies studied before they can be used successfully, i. e. in a controlled rather than ad hoc manner. Particularly useful would be a physical model of fMRI signal and an improved image reconstruction scheme. However investigating the effects of preprocessing was beyond the scope of this thesis and hence no preprocessing was used apart from brain extraction described above.

2.3 The experiments

To illustrate the results of this thesis the data of two experiments were used, which were performed at the Max-Planck-Institute of Biophysical Chemistry by the Biomedical Nuclear Magnetic Resonance group headed by Prof. Frahm.

Both experiments involve a fingertapping paradigm. The subject was visually instructed when to start and stop the fingertapping. In experiment 1 data from a 3-dimensional volume was measured whereas in experiment 2 data from one slice was acquired. Experiment 2 additionally contains data from a “resting brain” period as well as from imaginary fingertapping.

Experiment 1

In experiment 1 volume images consisting of 16 oblique slices containing the motor cortex were acquired in 160 time steps with a temporal resolution of 2 s. The measurement period hence took 320 s. The first 5 images were discarded to ensure that the measurement setup has reached a steady state. After a 80 s rest the subject was asked to tap fingers of both hands for 20 s followed by a 20 s rest and repeat this cycle six times. The activity related to finger tapping occured mainly in slice 3 the data of which was frequently considered separately as well and for which the results are shown in this thesis. Each slice contained voxels from a 128 x 112 matrix. The number of voxels in one volume image was reduced by almost 85% using the brain extraction procedure from section 2.2.

Experiment 2

Experiment 2 consisted of three periods in each of which 560 slice images were acquired with a temporal resolution of 0.5 s. The first 20 images corresponding to the first 10 s were discarded for steady state reasons. In the first period the subject was told to stay awake and refrain from moving. This was hence a so-called “resting brain” measurement where no predefined external stimulus was given. In the second period after a 40 s rest the subject was asked to tap fingers of both hands for 20 s and then rest again for 20 s, repeating this cycle six times. The instructions were given visually on a computer screen and projected into the eye field of the subject. The third part of the experiment was identical to the second except that the subject was not supposed to actually perform the finger tapping but only to imagine it according to the same time course as the actual finger tapping. The comparatively high frequency of 2 Hz with which the images were taken allows a frequency resolution of 1 Hz when Fourier transforming the time courses. The heart rate which occurs approximately at this frequency could hence be resolved.

However the finger tapping frequency of about 3 Hz was beyond this limit, but the finger tapping *cycle* frequency of $\nu = 1/40$ s $= 0.025$ Hz was resolved. Each image contained 128 x 112 voxels and was reduced by an amount of approximately 80% using the brain extraction procedure described above.

Chapter 3

Stimulus driven data analysis

In most fMRI experiments an externally controlled stimulus is used to evoke identifyable brain activity. Since the stimulus related signal is hardly visible from the raw images data analysis methods are used to extract it. In this chapter we discuss common stimulus driven analysis methods such as difference maps in section 3.1 and stimulus driven correlation analysis in section 3.2, where we also present the analytical form for the probability density of the sample correlation of a finite i. i. d. Gaussian distributed sample with a fixed stimulus which we derive in Appendix A. This result is used in section 3.4 dealing with thresholding, after in section 3.3 the general linear model which underlies the popular approach of statistical parametric mapping (SPM) [96] was presented.

We distinguish between directly stimulus related data analysis methods and data driven analysis methods. They differ in the point in time stimulus or other external information is used when applying the method. Stimulus driven methods relate the stimulus directly to the data as e. g. in correlation analysis, whereas data driven analysis methods are independent of a priori stimulus information but use this information a posteriori to identify stimulus related components as e. g. in Principal and Independent Component Analysis (cf. chapters 4 and 5)

Most of the stimuli used in fMRI are on-off type stimuli. The brain images corresponding to the 'on'-state of the stimulus are also referred to as the images of the active state whereas the term resting state is often used for images corresponding to the stimulus 'off'-state. On-off type stimuli can be modeled by a boxcar function which is zero when the stimulus is 'off' and one when the stimulus is 'on'. Many stimulus protocols are periodic which allows to enhance the statistical power by averaging over all periods. A drawback of periodic stimulus protocols is that the brain may adapt to the regularity of the stimulus presentation and react with a decreased response deteriorating the results of averaging. Therefore in the recent years increasingly so called 'event-related' stimuli are used, which can still be modeled by a boxcar function but are randomized in their appearance and duration (cf. Figure 3.1). In this chapter we discuss some of the most common stimulus driven analysis methods and compare their properties.

Probably the oldest and most straightforward analysis method in fMRI is the computation of difference maps (cf. section 3.1), i. e. the mean difference between the images where the stimulus was 'on' and those where the stimulus was 'off'. Obviously this

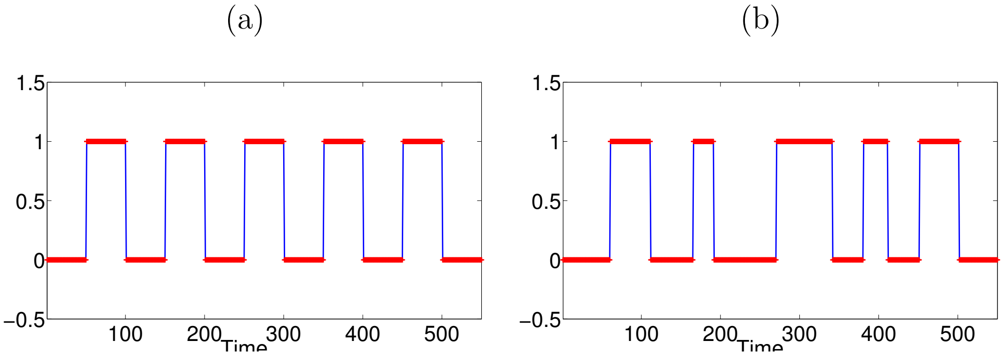


Figure 3.1: (a) Periodic on-off type stimulus protocol modeled by a discretized boxcar function. (b) Discretized boxcar function of an event related on-off type stimulus protocol.

method can be used only for on-off type stimuli (with possibly multiple 'on' states), but not for continuous stimuli.

This restriction does not hold for correlation analysis, which is discussed in section 3.2. It models the BOLD response to the stimulus by a reference function and correlates the data with it. In the simplest case the reference function equals the stimulus protocol but frequently more sophisticated reference functions aiming at mimicking the hemodynamic response are used as well. The General Linear Model introduced in section 3.3 can be considered as an extension of correlation analysis in that it uses multiple reference functions for a parametric fit of the data to extract multiple processes of interest and eliminate nuisance processes at the same time. Thresholding procedures are discussed in section 3.4. They are useful for visualization purposes defining a distinction between active and non-active voxels w. r. t. the given stimulus.

3.1 Difference maps

Difference maps have been used from the beginning of fMRI for on-off type stimuli. They are computed by subtracting the images corresponding to the resting state from the images related to the active state. Figure 3.2 shows difference maps for the data of experiment 2 for various shifts of the stimulus function to account for the hemodynamic delay.

Often, e. g. if the activation is focal, most voxels of an image do not participate at the active state in the sense that there is no significant signal change. Hence the voxel value distribution of an active state image often cannot be distinguished from the voxel value distribution of a resting state image. In the sequel we derive the conditions for which difference maps can be expected to detect active voxels.

Let us denote by M_a the set of voxels participating in the active state and by M_n the set of voxels 'neutral' to the stimulus. Correspondingly p_a and p_n denote the probability densities of the signal values of the voxels in M_a and M_n, respectively. From a statistical point of view the value $D = A - R$ of one pixel in a difference map is the result of a sum of two random variables A and $-R$, where A denotes the value of the signal in the active state and R the value of the signal in the resting state. The probability density of a sum of two independent random variables is proportional to the convolution of the probability densities of the two summands [91], thus

$$p_a(D) \propto p_a(A) * p_a(-R) \tag{3.1}$$

where $p_a(A)$ and $p_a(R)$ are the probability densities of the voxels in M_a during the active state and during the resting state, respectively. Of course, relation 3.1 holds true for the probability density p_n of the signals of non participating voxels as well, where we have $p_n(A) = p_n(R)$. To detect an active voxel the probability density $p_a(D)$ must be clearly distinguishable from $p_n(D)$, e. g. in that the difference of the means is larger than the respective variances hence enabling a distinction between $p_a(D)$ and $p_n(D)$.

By assuming gaussian densities for $p_a(A)$ and $p_a(R)$ as well as for $p_n(A)$ and $p_n(R)$ this statement can be easily quantified. Thus for

$$p_a(A) \ \propto \ e^{\frac{(A-\mu_A)^2}{2\sigma_a^2}}, \qquad p_a(R) \ \propto \ e^{\frac{(R-\mu_R)^2}{2\sigma_a^2}}, \qquad p_n(A) = p_n(R) \ \propto \ e^{\frac{(R-\mu_R)^2}{2\sigma_n^2}} \tag{3.2}$$

using the relation 3.1 results in

$$p_a(D) \ \propto \ e^{\frac{(D-(\mu_A+\mu_R))^2}{4\sigma_a^2}} \qquad p_n(D) \ \propto \ e^{\frac{(D-2\mu_R)^2}{4\sigma_n^2}} \tag{3.3}$$

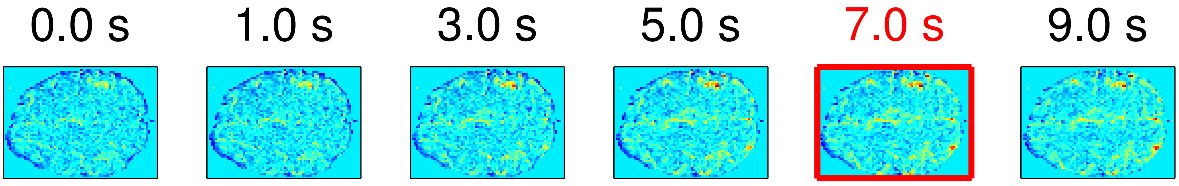
Here for simplicity we have assumed that $p_a(R)$ and $p_a(A)$ have the same variances σ_a^2 and differ only by their means. Further the mean μ_R of the voxels in M_a during the resting state was assumed to be the same as the mean of the neutral voxels. Defining the detectability of active voxels by the condition that the intersection of $p_a(D)$ and $p_n(D)$ is less or equal than the half maximum of the curves leads to the requirement

$$\mu_A - \mu_R \geq \sqrt{4\,\log 2}\,(\sigma_a + \sigma_n) \tag{3.4}$$

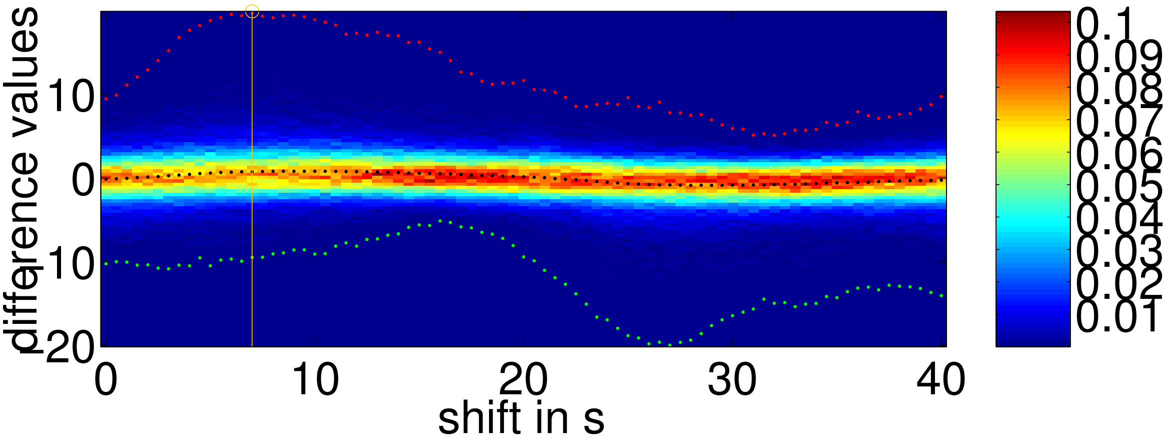
Hereby we have assumed that $\mu_A \geq \mu_R$, which is a reasonable assumption in fMRI, the corresponding condition for $\mu_A \leq \mu_R$ is straightforward. The percent signal change $\frac{\mu_A-\mu_B}{\mu_B}$ for fMRI is of the order of $1 - 10\%$ [43, 78]. For experiment 2 the voxel values are not approximately gaussian distributed (cf. Figure 2.2(d) in chapter 2) and hence Equation 3.4 not applicable. A coarse comparison of the signal of the stimulus related voxels during the presence of the stimulus with the signal when the stimulus is not present gives a percent signal change of about 3%.

3.2 Correlation analysis

What we refer to as correlation analysis is modeling the BOLD response to the stimulus as a function which is then used as a reference for correlating each voxel time course with. In this manner correlation images can be produced similarly to the difference maps discussed in the previous section. The voxels highly correlated with the reference



(a)

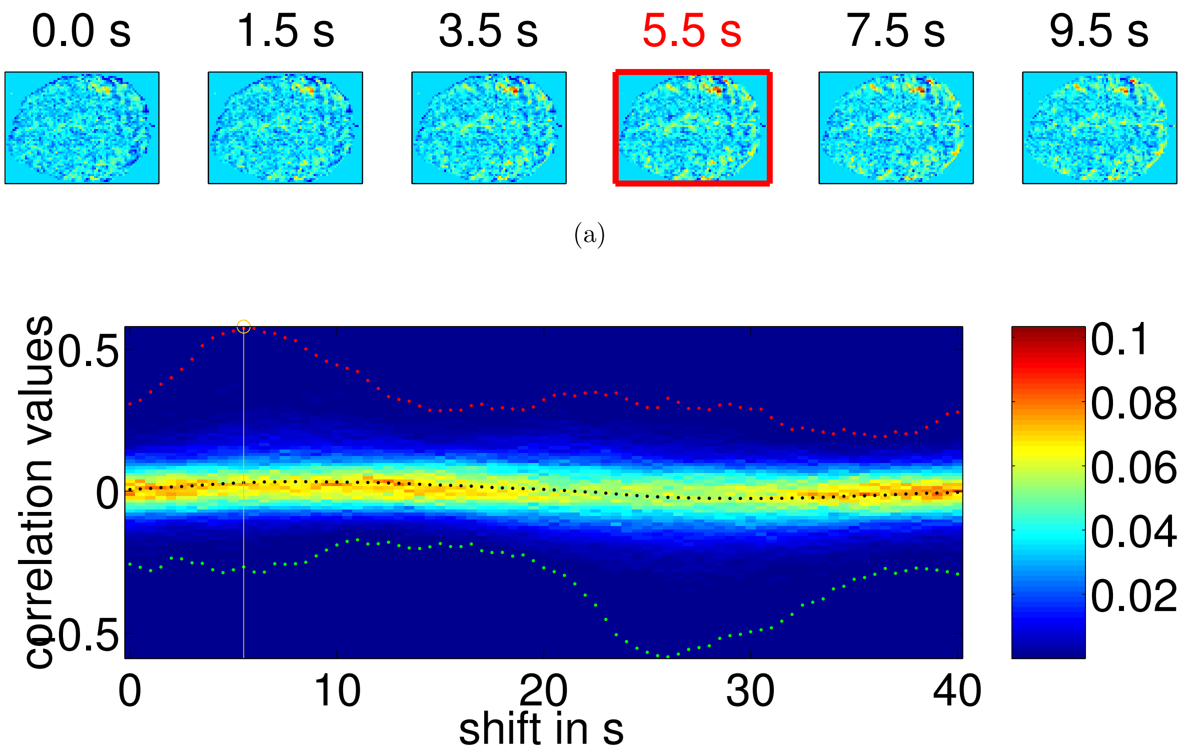


(b)

Figure 3.2: (a) Difference maps for various shifts of the stimulus boxcar function (cf. Figure 3.4). The image with the maximum positive value is indicated by a red frame. It occurs at a shift of 7 s from the onset of the stimulus. The delay is an effect of the hemodynamics (cf. chapter 1).
(b) Color coded relative frequencies of the difference values in dependence of shifts ranging from zero to the wavelength of the boxcar stimulus function. The red dotted curve indicates the maximum difference value of the respective shift, the green dotted curve the minimum difference value, and the black dotted curve denotes the mean difference value. An orange line with a circle is drawn at the shift where the absolute maximum difference value occured.

function are considered as activated by the stimulus. The simplest reference function for on-off type stimuli is a boxcar function, occasionaly shifted in time to account for the hemodynamic delay. Attempts have also be made to model the BOLD response more precisely and to use this as a reference function [51]. Figure 3.3 shows the result of correlating the data of experiment 2 with a boxcar reference function using various shifts.

FMRI data are discrete with an unknown underlying probability distribution and has finite sample size. In this case the sample correlation is used as an estimation for the 'real' correlation. The sample correlation of the ith voxel's time course with the reference



(a)

(b)

Figure 3.3: (a) Correlation maps for various shifts of the stimulus boxcar function (cf. Figure 3.4). The maximum correlation occurs at a shift of 5.5 s (red frame). (b) Color coded relative frequencies of the correlation values for the different shifts. The color code as well as the curves are the same as in Figure 3.2 (b) to enable comparison. Obviously the distribution of the correlation values is slightly broader than the distribution of the difference values and hence the correlation maps are more structured than the difference maps. The curve of the maximum correlation is more peaked around its maximum than the curve of the maximum difference value indicating a higher sensitivity to the shift.

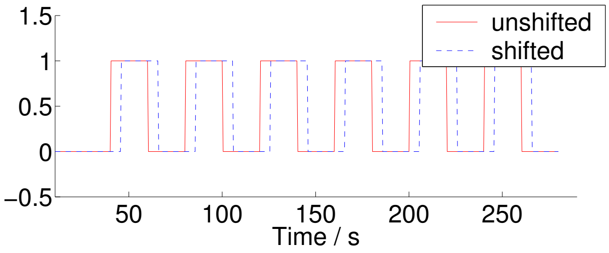


Figure 3.4: The unshifted and shifted stimulus boxcar function used to produce the difference maps in Figure 3.2 and the correlation maps in Figure 3.3

function can be written using vector notation as

$$c_i = \frac{\mathbf{x_i}^T \mathbf{v}}{\|\mathbf{x_i}\| \, \|\mathbf{v}\|} \tag{3.5}$$

with $\mathbf{x_i} = (x_{i1}, \ldots, x_{ik})^T - \left[\frac{1}{k}\sum_{j=1}^{k} x_{ij}\right]\underbrace{(1, \ldots, 1)}_{k}^T$ being the centered time course of the ith voxel, ($i \in \{1, \ldots, m\}$), and $\mathbf{v} = (v_1, \ldots, v_k)^T - \left[\frac{1}{k}\sum_{j=1}^{k} v_j\right](1, \ldots, 1)^T$ denoting the discretized reference function. The second terms of $\mathbf{x_i}$ and $\mathbf{v}$ equal the respective sample means. Note that centering finite samples results in the sample vector being orthogonal to $(1, \ldots, 1)^T$. From this and from Equation 3.5 follows that the sample correlation has a geometrical interpretation as the cosine of the angle between two vectors in $[(1, \ldots, 1)^T]^\perp$, which is the subspace of $\mathbb{R}^k$ that is orthogonal to $(1, \ldots, 1)^T$. Effects of centering are investigated in more detail in section 4.1 in the context of principal component analysis.

The sample correlation c in Equation 3.5 itself can be considered as a random variable whose probability density is dependent on the sample size k. In section A.1 the probability density $p(c)$ of the sample correlation c of a *gaussian* random vector $\mathbf{x} \in \mathbb{R}^k$ having i. i. d. elements with a deterministic vector $\mathbf{v} \in \mathbb{R}^k$ in dependence of k is derived. This is of interest for evaluating the results of correlation analysis, e. g. can it be useful in the context of thresholding (cf. section 3.4). The probability density $p(c)$ of the sample correlation $c \in [-1, 1]$ of k i. i. d. gaussian distributed samples x_i with fixed reference values v_i ($i \in \{1, \ldots, k\}$) computed according to Equation 3.5 is derived to be (cf. section A.1, Equation A.25)

$$p(c) = \begin{cases} \frac{1}{2}\left(\delta(c+1) + \delta(c-1)\right) & k = 2 \\ C_k\left(1 - c^2\right)^{\frac{k-4}{2}} & k \geq 3 \end{cases} \tag{3.6}$$

where $C_k = \frac{\sqrt{\pi}\,\Gamma(\frac{k}{2}-1)}{\Gamma(\frac{k}{2}-\frac{1}{2})}$ is a normalization factor ensuring that the integral over $p(c)$ equals unity. Figure 3.5 shows the probability density $p(c)$ given by Equation 3.6 as a function of the sample size k. Note that for values $k \leq 3$ the density is peaked at ± 1, only for $k \geq 5$ is it unimodal with decreasing width for increasing k.

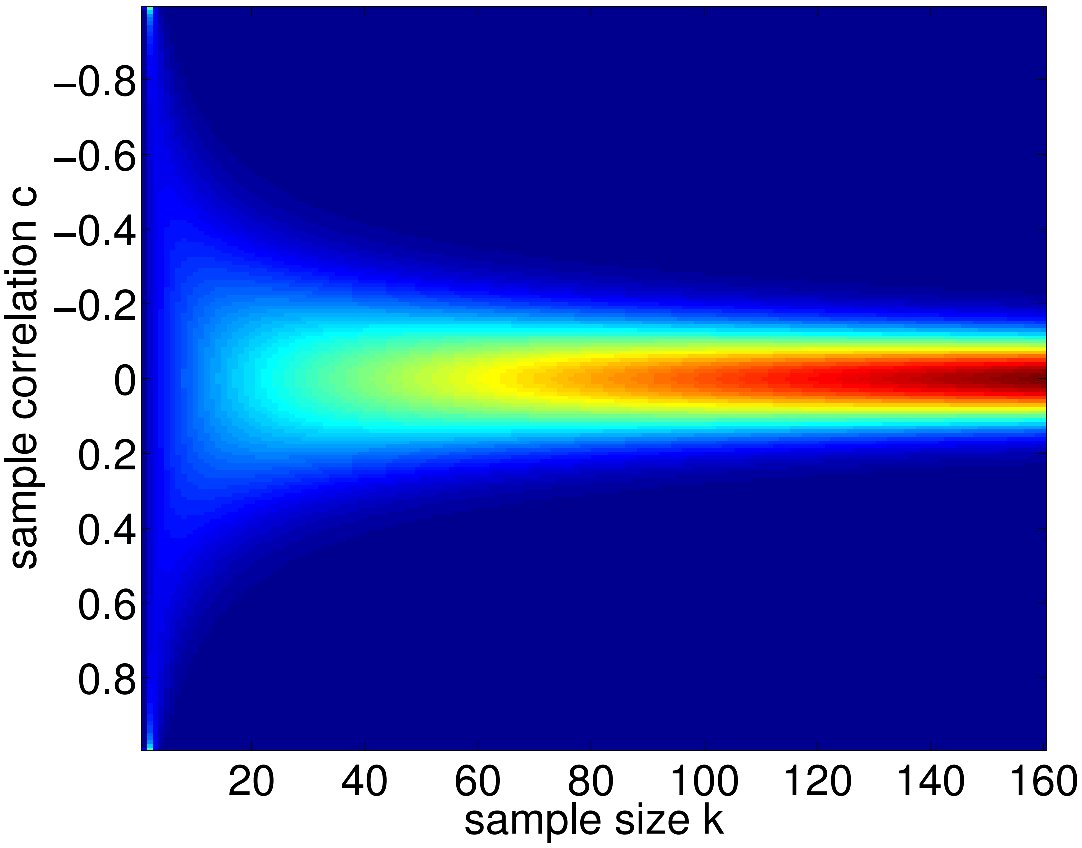


Figure 3.5: Analytical probability density $p(c)$ of the sample correlation c (cf. Equation 3.6) of an i. i. d. gaussian distributed sample with a fixed reference as a function of the sample size k. Note that $k \geq 2$ since for $k = 1$ no correlation is defined. For $k \in \{2, 3\}$, $p(c)$ is peaked around $c = \pm 1$, for $k = 4$, $p(c)$ is a uniform density, and for $k \geq 5$, $p(c)$ is unimodal, becoming increasingly peaked around $c = 0$ for growing k.

3.3 General Linear Model

The General Linear Model (GLM) is a linear multiple regression model which can be used to extract multiple processes of interest from the data and eliminate nuisance processes at the same time. The prior information about the data, i. e. the processes of interest as well as the nuisance processes, are modeled a priori and used as regression variables. An introduction to GLM's can be found in [2], from which here we review the basic facts. The GLM has been widely used in fMRI [96].

For discrete finite data such as in fMRI the GLM can be written as a matrix equation

$$\mathbf{x} = \mathbf{G}\,\beta + \epsilon \tag{3.7}$$

where $\mathbf{x} = (x_1, \ldots, x_k)^T$ is the data vector (e. g. the time course of a voxel in an fMR image), $\mathbf{G} = \left(g_{ij}\right)_{k\times p}$ is the so called *design matrix* containing prior information about the data in the form of the columns which are discretized functions of processes of interest and nuisance processes; $\beta = (\beta_1, \ldots, \beta_p)^T$ is a parameter vector and $\epsilon = (\epsilon_1, \ldots, \epsilon_k)^T$ is a noise vector. The parameter vector β is estimated by the least squares (LS) principle, i. e. by minimizing the cost function

$$S = \|\epsilon\|^2 \tag{3.8}$$

The underlying assumptions thereby are that

1. the model is linear in the parameters. This is already ensured by Equation 3.7.
2. the noise is additive, which is implied by Equation 3.7 as well.
3. the noise has zero mean, is mutually uncorrelated with equal variances σ^2, written as
$$E(\epsilon) = \mathbf{0}, \qquad \mathbf{V}(\epsilon) = \sigma^2\,\mathbb{I}_\mathbf{n} \tag{3.9}$$
where E denotes the mean, $\mathbf{V}$ the covariance matrix and $\mathbb{I}_\mathbf{k}$ the $k\times k$ identity matrix. This assumption is necessary for the estimators to have certain optimal properties (cf. [2]).
4. the design matrix has full rank, so that $\mathbf{G}^T\mathbf{G}$ is positive definite. This is necessary to ensure that the LS solution is unique. It further implies that for the number p of processes, that are included in the design matrix $\mathbf{G}$ is $p \leq k$, which we shall assume henceforth.

Given the model Equation 3.7 and the assumptions 1-4 the LS estimator for β is

$$\hat{\beta} = (\mathbf{G}^T\mathbf{G})^{-1}\,\mathbf{G}^T\,\mathbf{x} \tag{3.10}$$

Further

$$E(\hat{\beta}) \;=\; \beta \tag{3.11}$$

and

$$\mathbf{V}(\hat{\beta}) \;=\; \sigma^2(\mathbf{G}^T\mathbf{G})^{-1} \tag{3.12}$$

thus $\hat{\beta}$ is an unbiased estimator with the covariance matrix $\mathbf{V}$ as in Equation 3.12 [2].

In fMRI the vector $\mathbf{x}$ in Equation 3.7 represents the time course of one voxel. Suppose we have an fMRI experiment with an external stimulus. Then a simple example of a design matrix for $p = 3$ is to set the first column of $\mathbf{G}$ to $\mathbf{1} = (1, \ldots, 1)^T$ to allow for a constant offset, the second column to a discretized stimulus reference function and the third column to a discretized linear function to model the nuisance process of a linear drift which can be considered as a nuisance process when related to the stimulus. Applying Equation 3.10 to each voxel time course results in a $p \times m$ parameter matrix

$$\mathbf{B} = (\mathbf{G}^T\mathbf{G})^{-1}\,\mathbf{G}^T\,\mathbf{X}^T \tag{3.13}$$

where $\mathbf{X}$ is the $m \times k$ data matrix. Each row of $\mathbf{B}$ reflects an image containing the coefficients of the respective column in $\mathbf{G}$. In principle up to k functions can be written into $\mathbf{G}$, however the functions (columns) which can be added are constrained by assumption 4. In our example with $p = 3$ the first row of $\mathbf{B}$ represents the mean image while the second row indicates the amplitudes of the stimulus function for each voxel. A large amplitude characterizes the voxels the time courses of which are reasonably well approximated by the stimulus function and which hence can be assumed as being activated. The third row of $\mathbf{B}$, the amplitude image of the linear drift, often stems from head movements of the subject, which is indicated by negative and positive amplitudes being contiguous in the two-dimensional image. Figure 3.6 shows the results for applying the design matrix described above to the data of experiment 2.
The quality of fit can be estimated by the correlation matrix $\mathbf{K} = \left(k_{ij}\right)_{m\times m}$ of the noise matrix $\mathbf{E} = \mathbf{X} - \mathbf{GB}$. $\mathbf{K}$ should be essentially diagonal if the data is to be explained by the processes modeled in the design matrix $\mathbf{G}$. This can be done by examining, if $\mathbf{K}$ is diagonal dominant, i. e. if $\sum_{i\neq j} |k_{ij}| < |k_{jj}|$. For the example given, $\mathbf{K}$ obviously is not diagonal dominant, thus the three processes modeled are not enough to explain the data. As a first step however the use of a design matrix can be useful to give a coarse grain view on the data structure. For a more detailed view additional analysis methods particularly those which are data driven as are presented in the following chapters are indispensable.

GLM and correlation analysis in comparison

Correlation analysis using a fixed reference function as introduced above is quite similar to the use of a $k \times 2$ design function $\mathbf{G}$ with the first column being $\mathbf{1_k}$ and the second being equal to $\mathbf{v}$ as defined in Equation 3.5. Note that the two columns of $\mathbf{G}$ are orthogonal if $\mathbf{v}$ is centered. Let $\mathbf{b_1}$ and $\mathbf{b_2}$ be the first and second row of $\mathbf{B}$, respectively, i. e. the mean image and the amplitude image related to the stimulus reference function. Then comparing Equation 3.5 and Equation 3.13 under the assumption that $\mathbf{v}$ and $\mathbf{x_i}$ are centered (the latter implies that $\mathbf{b_1}$ is zero) we get

$$c_i = \frac{\|\mathbf{v}\|}{\|\mathbf{x}_i\|}\, b_{2i} \tag{3.14}$$

where c_i is the correlation value computed from Equation 3.5 and b_{2i} is the ith element of $\mathbf{b}_2$. The correlation maps are hence proportional to the parametric map where each entry b_{2i} is weighted essentially by the norm $\|\mathbf{x}_i\|$ of the voxel time course.

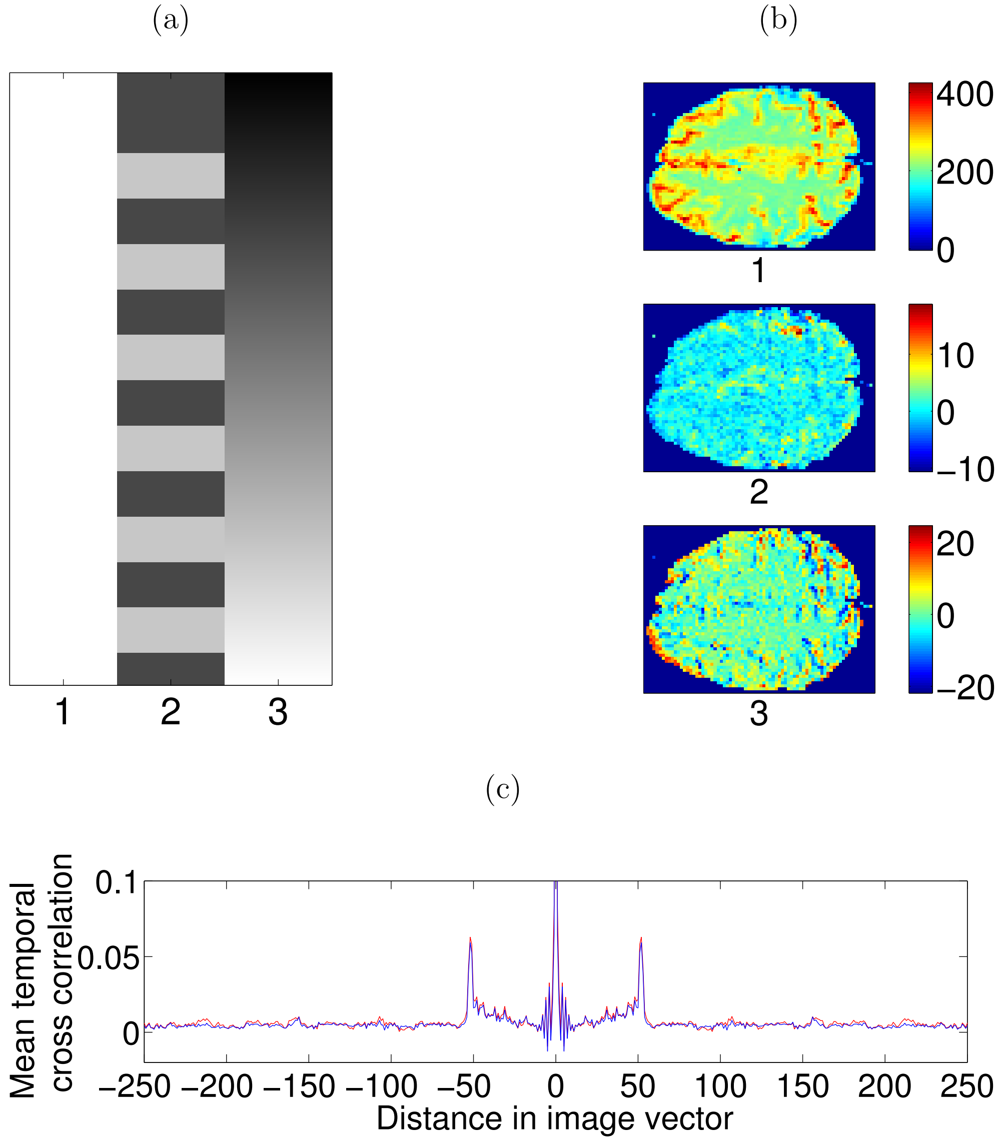


Figure 3.6: Design matrix **G** and corresponding parameter maps with mean residual temporal cross correlation using data from experiment 2.
(a) Design matrix **G** with $p = 3$. First column: constant offset, second column: stimulus protocol shifted by 6 s to account for the hemodynamic delay, third column: linear drift.
(b) Parameter maps corresponding to the weight of the respective columns of **G** if the voxel time course is modeled by a weighted sum of the columns of **G** in LS sense.
(c) Blue curve: Mean temporal cross correlation of the residual images. Red curve: Mean temporal cross correlation of the original data. The curves are appropriately normalized to prevent strong variations due to differences in the number of off-diagonal elements, from which the mean is taken. The peak in the middle has height 1 and is cut at 0.1 to enable better visibility of the curves. The side peaks arise because by considering each image as a vector, some voxels contiguous in the two-dimensional image are not contiguous in the image vector. Obviously the two curves are almost identical. Although the mean temporal correlation is low, the residual cross correlation matrix is not diagonal dominant (the absolute non-diagonal values sum up to about 100 whereas the diagonal values are 1) indicating that the three columns of **G** did not capture all essential processes in the data.

3.4 Thresholding

Thresholding procedures are needed if the voxels are to be partitioned into an activated and a non-activated set. An overview over existing thresholding procedures can be found in [93]. In [43] correlation values are thresholded by using a symmetrized distribution of correlation values as a reference. For GLM common statistical tests, such as t-test, F-test and χ^2-test, have been used [96] along with more sophisticated approaches based on characteristics of random fields [112, 113].

Here we used a thresholding procedure for correlation analysis based on the analytical derivation of the probability density $p(c)$ in Equation 3.6. We compare the histogram of the correlation values in a correlation map with the theoretical probability density $p(c)$ of the correlation values of an i. i. d. gaussian distributed sample of the same size with a fixed reference function. The histogram surface of sample correlations in correlation maps for varying sample size is shown in Figure 3.7 along with the surface of the difference to the analytic probability density $p(c)$. As expected the correlation histograms of the data are broader than the density $p(c)$ the difference being largest at positive correlation values. At the maximum sample size a fraction of about 18% of the positive and about 3% of the negative correlation values in the correlation map made up for the positive difference to the analytic density. However using these values for thresholding leads to spurious activation assignment. Therefore we used a heuristic value of 5% of the fractions for thresholding the correlation map in Figure 3.8(a) leading to the map in Figure 3.8(b). The occurence of spurious activations when using the whole fraction of voxels accounting for the positive difference can be due essentially to two reasons, first, the distribution of voxel values from which the sample correlation was computed, is not gaussian and secondly and more important, the samples are not independent, i. e. the voxel value at a given time depends on the voxel values at previous points in time. Deriving the probability density $p(c)$ of sample correlations for non-independent samples, e. g. using a Markov assumption of different order, hence should be useful in this context.

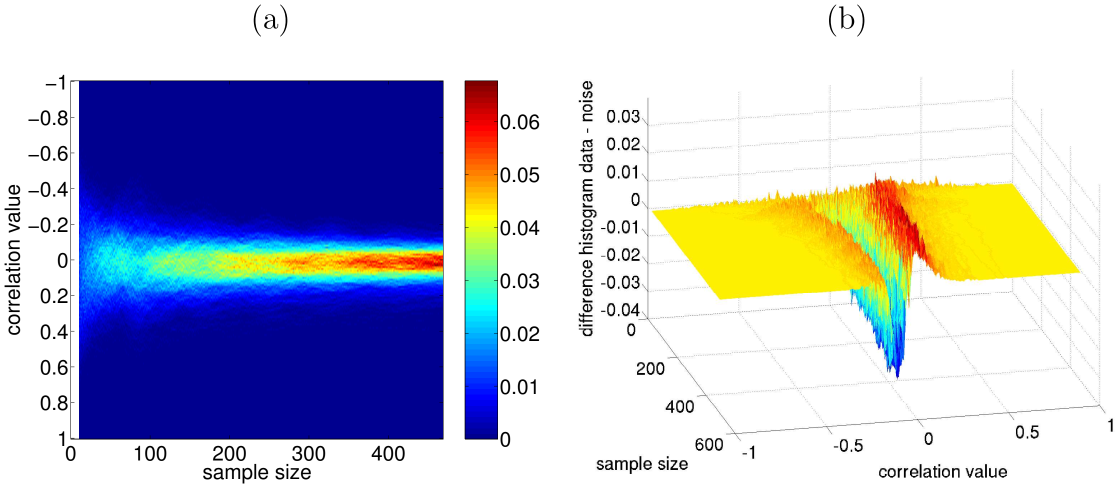


Figure 3.7: Histogram and difference surfaces. (a) Histogram surface of correlation values as a function of sample size for data from experiment 2 with a reference function shifted by 5.5 s. The minimum sample size was 12 since for smaller samples the remaining reference function was constant. (b) Difference surface of the histogram in (a) and the appropriately normalized function $p(c)$ from Figure 3.5. The difference surface has positive parts at higher absolute correlations because the histogram of correlation values of the data was broader than the probability density $p(c)$. The negative part in the center is due to normalization resulting in the volumes enclosed by the positive and the negative fractions of the difference surface and the zero plane being equal.

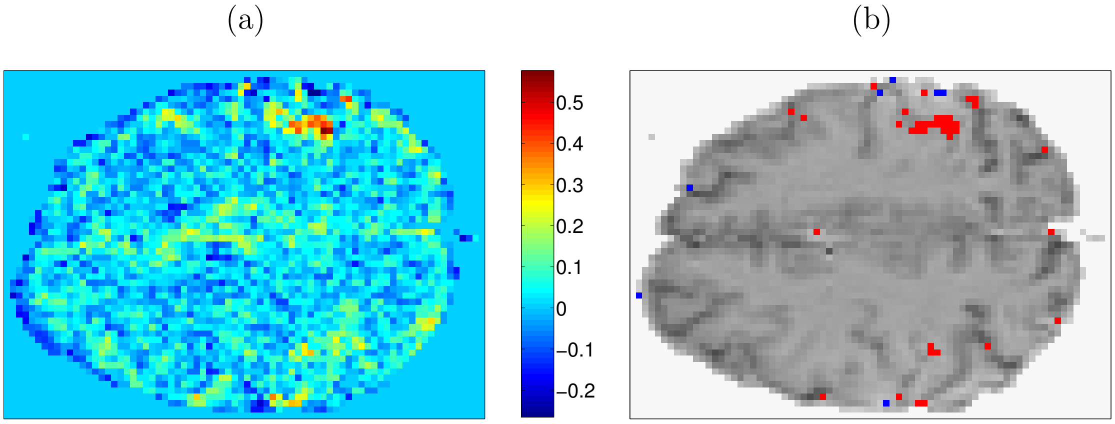


Figure 3.8: Unthresholded and thresholded correlation map. (a) Correlation map at shift 5.5 s taken from Figure 3.3. (b) Mean image with activated voxels obtained by thresholding the correlation map in (a). Red voxels: suprathreshold positive correlation with the reference function. Blue voxels: suprathreshold negative correlation with the reference function. See text for the details of thresholding.

Chapter 4

Principal component analysis

Principal Component Analysis (PCA) is an orthogonal basis transform to the directions of maximal variance of multivariate data. In this chapter in addition to showing results of applying PCA to fMRI data we present a comparison of spatial and temporal PCA, [34], which is based on the visualization of the data space and on analytical conditions for equality, which are derived in Appendix B. Further, we present a method for dimensionality reduction using random matrix theory.

Geometrically PCA is equivalent to fitting a high dimensional ellipsoid to the data. Its half axes then correspond to the directions of maximal variance. Figure 4.1 shows an illustration of the concept of PCA. As one can think of a data set as realization of a multivariate probability density the geometrical and statistical point of view in this case are equivalent. In terms of statistics PCA is a second order procedure. PCA is global in that it determines the structure of the whole data. Where this gives already interesting insight [30], PCA could also serve as a basis to extract local features by applying PCA hierarchically thresholding the extracted features and repeat PCA on the so reduced data set or in a nonlinear fashion as e. g. in [52].

PCA is performed by the eigendecomposition of the covariance matrix $\mathbf{C}$ of the data. The eigenvectors of $\mathbf{C}$ are called *principal vectors* (PV) or *principal axes* and provide a basis with respect to which the data is uncorrelated. Correspondingly the data projected onto the eigenvectors are called *principal components* (PC). The eigenvalues of $\mathbf{C}$ are the variances of the data distribution in the direction of the principal axes.

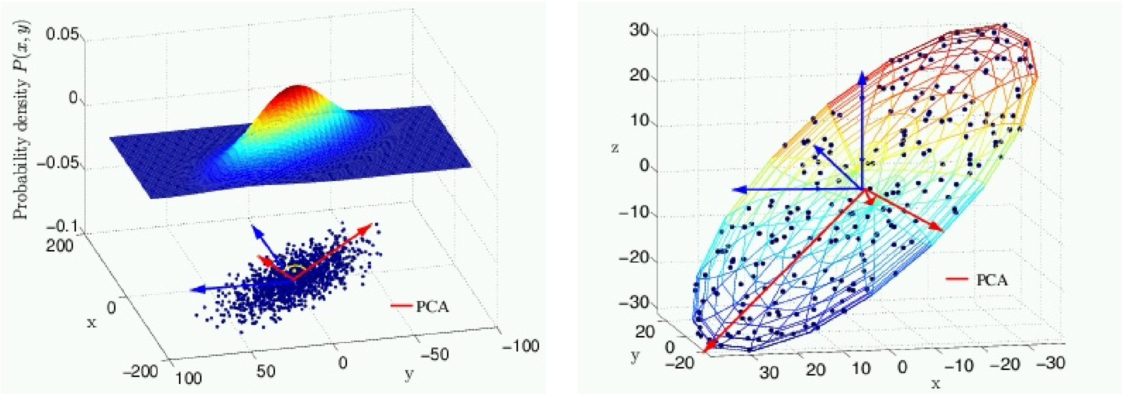
We now shortly review some basic covariance related definitions needed in this chapter. For continuous random variables X and Y their *covariance* is defined as

$$C(X,Y) = \int p(X,Y)\ (X - E(X))\ (Y - E(Y))\ dX\ dY \tag{4.1}$$

where $p(X,Y)$ is the joint probability density of X and Y and

$$E(Z) = \int p(Z)\ Z\ dZ \tag{4.2}$$

is the *expectation value* of Z, $(Z \in \{X,Y\})$, with the marginal probability density $p(Z)$



(a) (b)

Figure 4.1: Illustration of the concept of PCA in two and three dimensions. PCA is an orthogonal data transformation for which the new basis vectors correspond to the (orthogonal) directions of highest variance.
(a) Surface: Bivariate probability density $P(x, y)$. Black dots: Realization of $P(x, y)$ drawn into the same axes to illustrate the relation with $P(x, y)$. Blue arrows: Arbitrary original coordinate system. Red arrows: Orthogonal directions of maximal variance providing a new coordinate system. The different lengths of the vectors correspond to the different variances in the directions they are representing.
(b) Realization (black dots) of a 3-dimensional multivariate density and the ellipse fitted to the data by PCA. Blue arrows: Axes of the original coordinate system. Red arrows: Half axes of the fitted ellipse.

of Z. The *variance* of a random variable Z is defined as

$$V(Z) = \int p(Z)\,(Z - E(Z))^2\,dZ \tag{4.3}$$

and can be interpreted as the covariance of the random variable with itself, hence $V(Z) = C(Z, Z)$. The *correlation* between two random variables X and Y is defined by

$$R(X, Y) = \frac{C(X, Y)}{\sqrt{V(X)\,V(Y)}} \tag{4.4}$$

For discrete random variables the integrals are replaced by discrete sums and the probability densities by probability distributions. Note that R, C, V and E are not functions of X and Y in the usual sense as is suggested by the above notation, rather, due to the definition of random variables (cf. [91]), they are functionals.

If we have random vectors rather than one-dimensional random variables the covariance structure is expressed in the *covariance matrix* $\mathbf{C}$ which contains the covariance of each pairs of elements of the random vector $\mathbf{x} = (X_1, \ldots, X_m)^T$ thus

$$\mathbf{C} = \big(C(X_i, X_j)\big)_{m \times m} \tag{4.5}$$

Obviously $\mathbf{C}$ is symmetric and its diagonal equals the variance of the random vector $\mathbf{x}$ on the account of which $\mathbf{C}$ is sometimes termed the *variance-covariance matrix.*

PCA can be also considered in the context of neural networks [65], where it can be used to adaptively decorrelate the data. Here we concentrate on batch PCA which is performed by an eigenvalue decomposition of $\mathbf{C}$. Since $\mathbf{C}$ is symmetric its eigenvectors are orthogonal. The eigenvalues usually are arranged in descending order, thereby imposing an order on the principal components and principal axes. Commonly the large eigenvalues are interpreted as due to data features of interest while the small eigenvalues are assigned to noise. If this interpretation holds true PCA can be used to improve the signal to noise ratio and to reduce the dimension of the data by projecting it onto the most relevant PV's spanning the so called *signal space*. The data space is hence considered to be the direct sum of the signal space and the *noise space* the latter of which is spanned by the remaining PV's. "In-plane-noise", i. e. noise within the signal space, however, cannot be removed by PCA. The problem of determining the dimension of the signal space is discussed in section 4.2.

As for real data sets the underlying probability density usually is not known, PCA in these cases is performed on the sample covariance matrix computed from the data matrix $\mathbf{X}$. Each column of $\mathbf{X}$ is thereby considered as a realization of a random vector. This is for temporal PCA where is assumed that the processes of interest are characterized by mutually uncorrelated time courses. The other possibility is spatial PCA where the processes of interest are characterized by mutually uncorrelated activity patterns, i. e. images. In this section we consider only temporal PCA, a comparison of temporal and spatial PCA being presented in section 4.1.

The temporal sample covariance matrix of the data writes

$$\mathbf{C_Z} = \frac{1}{k}\mathbf{Z}\mathbf{Z}^T \tag{4.6}$$

where $\mathbf{Z}$ is the $m \times k$ temporally centered data matrix having the elements $z_{ij} = x_{ij} - \frac{1}{k}\sum_{l=1}^{k} x_{il}$. The result of an EVD on $\mathbf{C_Z}$ is equivalent to those obtained by performing an SVD on $\mathbf{Z}$. However, the SVD also provides the activity patterns corresponding to the uncorrelated time courses and hence in this thesis is preferred. The SVD of $\mathbf{Z}$ reads [95]

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{4.7}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $m \times m$ and $k \times k$ orthogonal matrices, respectively, and $\mathbf{D}$ is a $m \times k$ rectangular diagonal matrix the diagonal elements of which are called the *singular values*. Note that apart from the ambiguity regarding the order of the columns of $\mathbf{U}$ and $\mathbf{V}$, which m. m. occurs in EVD as well there is an additional ambiguity in SVD regarding $\mathbf{U}$ and $\mathbf{D}$, in that chosing $\mathbf{U}$ as a rectangular $m \times k$ matrix with orthogonal columns and $\mathbf{D}$ as a $k \times k$ square diagonal matrix leads to the same result in Equation 4.7. However, throughout this thesis we take $\mathbf{U}$ to be square and orthogonal and $\mathbf{D}$ to be rectangular. The last $m - k$ columns of $\mathbf{U}$ then form an orthogonal basis of the nullspace and are not unique, if the nullspace has more than one dimension, i. e. if the multiplicity of zeros in the diagonal of $\mathbf{D}\mathbf{D}^T$ exceeds one. Note that for symmetric (and hence square) matrices SVD is equivalent to EVD. For non-symmetric but diagonalizable matrices, however, the relation between eigenvalues and singluar values is not clear, apart from the fact that the number of nonzero eigenvalues equals the number of nonzero singular values, which again equals the rank of the matrix. Note further that the noise space is different from the nullspace, the latter occuring due to rank deficiency or rectangularity of $\mathbf{X}$ whereas the former is the space of low variance.

From the SVD of $\mathbf{Z}$ in Equation 4.7 follows that the EVD of $\mathbf{C_Z}$ is

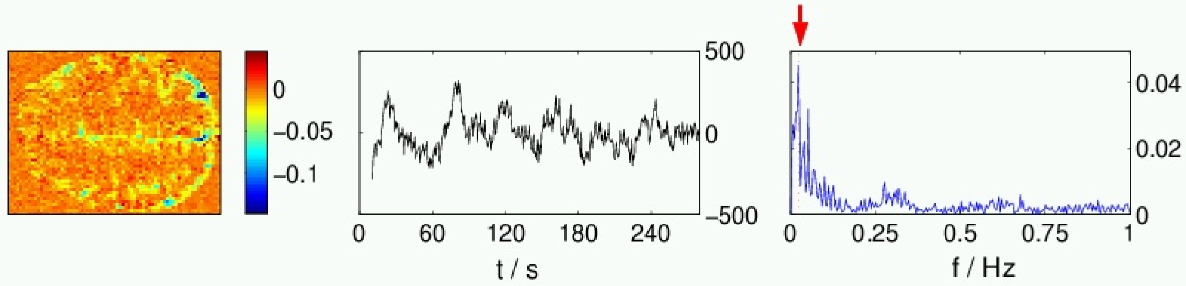
$$\mathbf{C_Z} = \mathbf{U}\mathbf{D}_\mathbf{k}^2\mathbf{U}^T \tag{4.8}$$

where $\mathbf{D}_\mathbf{k}^2 = \frac{1}{k}\mathbf{D}\mathbf{D}^T$. Thus the columns of $\mathbf{U}$ are the principal vectors and the variances of the data in the direction of the PV's are essentially given by the square of the diagonal elements in $\mathbf{D}$. The principal components are obtained by projecting the centered data matrix $\mathbf{Z}$ onto the columns of $\mathbf{U}$. They thus read

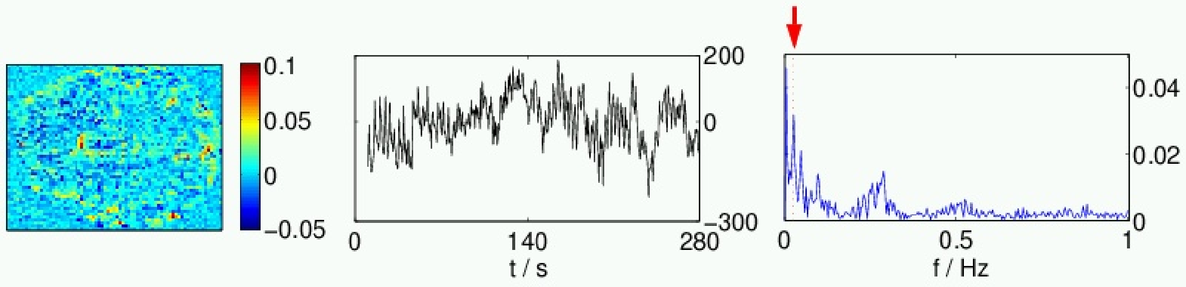
$$\mathbf{U}^T\mathbf{Z} = \mathbf{D}\mathbf{V}^T \tag{4.9}$$

From Equation 4.9 it is obvious that the principal components are essentially equal to the columns of $\mathbf{V}$ (up to scalar factors consisting in the singular values in $\mathbf{D}$, and up to additive constants when using the uncentered data matrix $\mathbf{X}$ rather than $\mathbf{Z}$, these constants disappear due to centering when computing the sample correlation matrix). For discrete finite data the PC's and PV's hence represent the orthogonal matrices of the SVD of the centered data matrix. Note also, that the PC's are k-dimensional since from orthogonality considerations it is not possible to have $m > k$ uncorrelated time courses of length k. We will show below (cf. section 4.1) that due to centering there are only $k-1$ uncorrelated time courses.

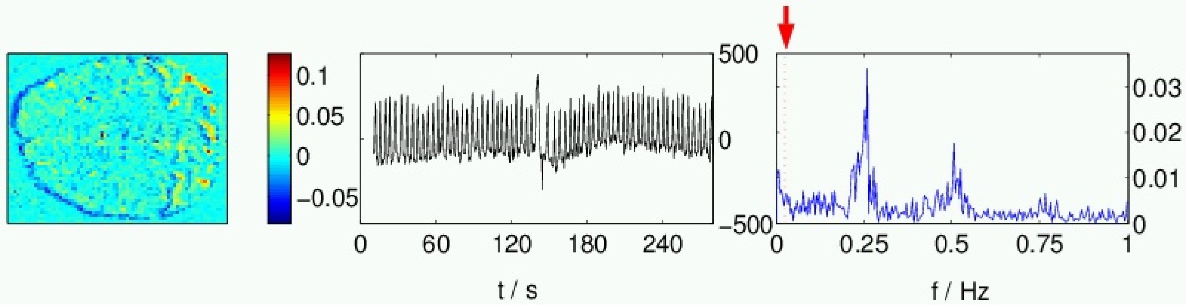
Figure 4.2: (next page) PCA results for data from experiment 2. Left hand side: principal axes (images in this case), middle: principal components (time courses), right hand side: Fourier spectrum of the principal components normalized to the total power. The finger tapping cycle frequency 0.025 Hz is marked by an arrow.
(a) Third PC of both hand finger tapping. A strong peak is visible at the finger tapping cycle frequency and at its first harmonic. The image shows weak activations in the area of the motor cortex and otherwise is dominated by what appear to be blood vessels.
(b) Fourth PC of imaginary finger tapping. An activation is visible in an area which can be attributed to the supplementary motor cortex, which is believed to be involved in planning and memory retrieval of movements. Apart from slower variations the spectrum is peaked at the stimulus cycle frequency, but also shows contributions from higher frequencies which can be attributed to breathing.
(c) Second PC of the resting state. The brain shaped boundaries visible indicate head movement. The direction of the head movement is consistent with the direction determined by rigid body preprocessing in section 2.2. The spectrum shows a broad peak at about 0.25 Hz and smaller peaks at its harmonics. The value of the peak frequency indicates that the head movement occurs due to breathing.
(d) Fourth PC of the resting state. The image contains essentially the sagittal sinus along with another blood vessel in the left frontal region. The time course is dominated by a slow variation. The strong peak at almost zero frequency in the spectrum could be the aliased heart rate.



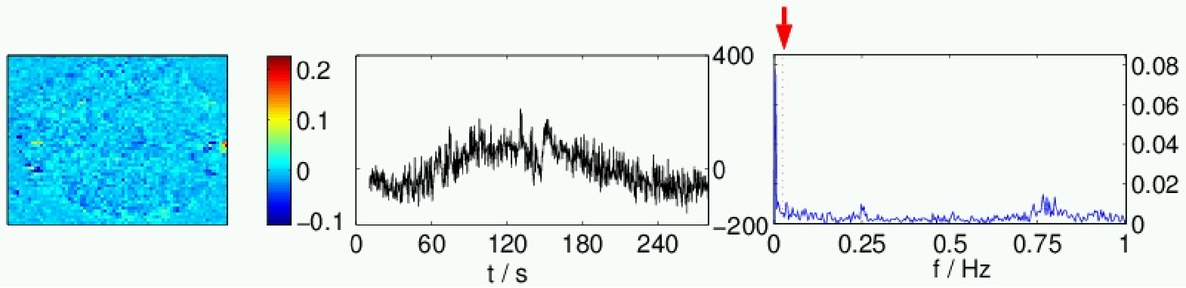
(a) Third PC of both hand finger tapping.



(b) Fourth PC of imaginary finger tapping.



(c) Second PC of the resting state.



(d) Fourth PC of the resting state.

In fMRI for each PC time course we have a PV image which represents the related activity pattern. Figure 4.2 shows principal components from experiment 2 along with the corresponding images and the Fourier spectra of the time courses which help identify periodic components such as the stimulus time course and cardiac as well as respiratory processes. PCA captured essential features in the data, however we will see in chapter 5 that ICA improves the separation. See also Figure 4.3 for a PCA of the data from experiment 1. In the case the stimulus time course is much better reproduced. The details are given in the figure captions.

4.1 Temporal versus spatial PCA

In the previous section we discussed temporal PCA, which yields the uncorrelated time courses underlying the data together with the corresponding images. Sometimes [86] one is interested in the uncorrelated activity patterns, i. e. images, that underly the data. This corresponds to spatial PCA, in which the roles of PV's and PC's are interchanged w. r. t. temporal PCA. Formally the difference between temporal and spatial PCA is due to the different centering of the data matrix. In temporal PCA the data is centered such that the mean of each voxel time course is zero, wheras in spatial PCA centering is performed such that the mean of each image is zero. Hence for the temporally centered data the mean image is zero whereas for the spatially centered data the mean time course is zero. Note the difference between the mean of an image and the mean image. The former is up to normalization equivalent to summing up the elements of a single column of the data matrix $\mathbf{X}$ wheras the latter is computed from the sum of the elements of all rows of $\mathbf{X}$. Similar statements apply for the relation between the mean of a voxel time course and the mean time course.

The elements of the temporally and spatially centered data matrices $\mathbf{Z} = \big(z_{ij}\big)_{m\times k}$ and $\mathbf{Y} = \big(y_{ij}\big)_{m\times k}$ are

$$z_{ij} = x_{ij} - \frac{1}{k}\sum_{r=1}^{k} x_{ir} \qquad \text{and} \qquad y_{ij} = x_{ij} - \frac{1}{m}\sum_{s=1}^{m} x_{sj} \tag{4.10}$$

where $\mathbf{X} = \big(x_{ij}\big)_{m\times k}$ is the uncentered data matrix. In matrix notation Equations 4.10 read

$$\mathbf{Z} = \mathbf{X}\mathbf{P_k} \qquad \text{and} \qquad \mathbf{Y} = \mathbf{P_m}\mathbf{X} \tag{4.11}$$

where $\mathbf{P_n}$ with $n \in \{m, k\}$ is the $n \times n$ projection matrix

$$\mathbf{P_n} = \mathbb{I}_\mathbf{n} - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^T = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdot & \cdot & \cdot & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & \cdot \\ -\frac{1}{n} & \cdot & \cdot & \cdot & \cdot & 1-\frac{1}{n} \end{pmatrix} \tag{4.12}$$

which projects onto the space orthogonal to $\mathbf{1_n} = (1, \ldots, 1)^T$. Obviously $\mathbf{P_n}$ has rank $n-1$ which is also the multiplicity of the eigenvalue 1 and $\mathbf{1_n}$ is the eigenvector to the eigenvalue 0.

The question now arises how the differences in centering affect the results of PCA. Figure 4.3 shows the results for spatial and temporal PCA of fMRI data from experiment 1. There is a striking similarity between the nth temporal and the $(n+1)$th spatial PC's and PV's for the vast majority of the $n \in \{1, \ldots, k\}$. Further the first spatial PC is essentially equal to the mean image of the data. As similarity measure the inner products of the normalized PV's and PC's were computed. Their histogram is shown in Figure 4.4.

To understand the finding of high similarity we analyze the data space, i. e. row and column space of the centered data matrices $\mathbf{Y}$ and $\mathbf{Z}$. A visualization of the data space is given in Figure 4.5. We see that centering results in a dimension loss about one dimension, since for temporally and spatially centered data we have, respectively,

$$\sum_{i=1}^{m} y_{ij} = 0 \qquad \text{and} \qquad \sum_{j=1}^{k} z_{ij} = 0 \tag{4.13}$$

where $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, k\}$. Equations 4.13 define hyperplanes orthogonal to $\mathbf{1_m}$ and $\mathbf{1_k}$, respectively. Thus the rows of the temporally centered data matrix $\mathbf{Z}$ lie on a $k-1$ dimensional hyperplane and hence $\operatorname{rank}(\mathbf{Z}) \leq k-1$. The spatially centered data matrix $\mathbf{Y}$, however, in general is not rank deficient. According to Equation 4.13 the k columns of $\mathbf{Y}$ lie on an $m-1$ dimensional hyperplane, of which the k-dimensional column space is a subspace (recall that we have assumed $m > k$ throughout this thesis). Thus spatially centered data in general are k dimensional whereas temporally centered data are $k-1$ dimensional provided that the uncentered data matrix $\mathbf{X}$ is of full rank k. Note that the columns of $\mathbf{Z}$ and the rows of $\mathbf{Y}$ do not lie on hyperplanes defined by the Equations 4.13 instead they are centered around the origin and span the $k-1$ and k dimensional column and row spaces of $\mathbf{Z}$ and $\mathbf{Y}$, respectively. See Figure 4.5 for an illustration.

Analytical conditions for equality

In Appendix B we derive the analytical conditions for equality of the time courses resulting from temporal and spatial PCA. The conditions involve a number of case differentiations of which here we state only those for the typical and general case, where $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ is of full rank, has no degenerate eigenvalues, and is not diagonal. Then the conditions for equality of the time courses of spatial and temporal PCA are

(A). $\mathbf{P_k}\bar{\mathbf{z}}$ is an eigenvector of $\mathbf{C} = \mathbf{X}^T\mathbf{X}$, $\mathbf{1_k}$ is contained in a subspace spanned by eigenvectors of $\mathbf{C}$ with eigenvalues being either 0 or $\frac{S_C}{k}$, and

  (a) $\bar{\mathbf{z}} \perp \mathbf{1_k}$ or

  (b) $d_p = \frac{1}{k}(S_C - \sum_{l=1}^{k} d_l\, y_l^2)$.

(B). $\bar{\mathbf{z}} \parallel \mathbf{1_k}$ and $\mathbf{1_k}$ is an eigenvector of $\mathbf{C}$.

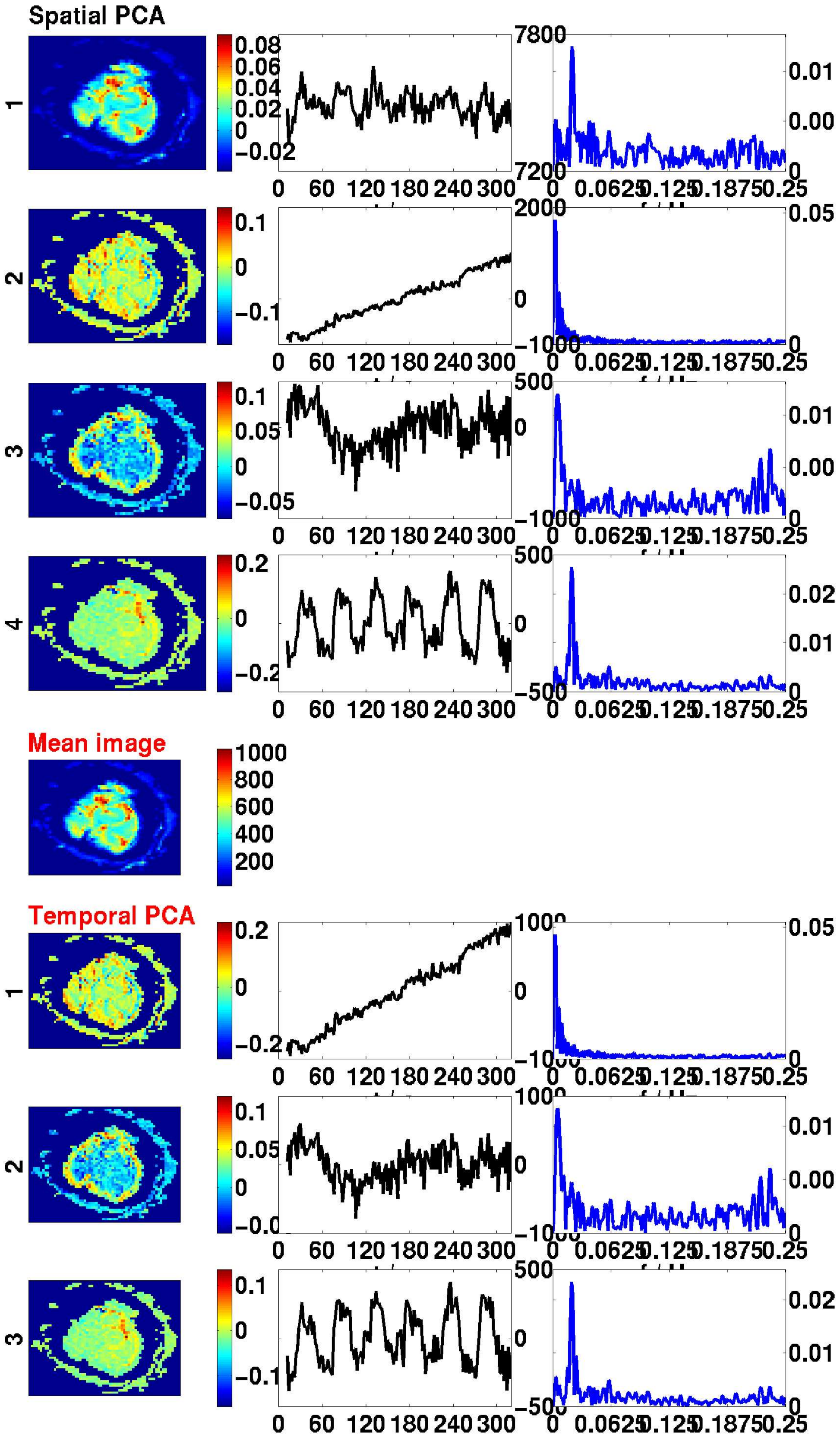
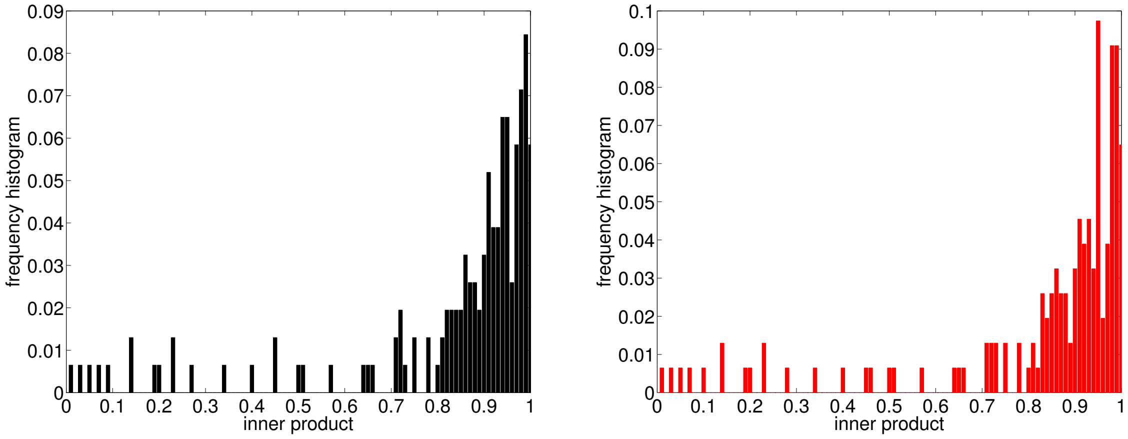


Figure 4.3: Results of spatial and temporal PCA in comparison. The first spatial PC is essentially equal to the mean image, subsequently the similarity occurs between the nth temporal and the $(n + 1)$th spatial components, respectively.



(a) Images

(b) Time courses

Figure 4.4: Frequency histogram of the inner products between the normalized images and time courses, respectively, of temporal and spatial PCA. About 80% of the inner products had values ≥ 0.8, the value 1 standing for equality. Note that in the context of similarity the sign of the inner product is not of interest and hence the absolute values were taken.

Here the y_i are the inner products of $\mathbf{1_k} = (1, \dots, 1)^T$ with the eigenvectors of $\mathbf{C}$ and S_C is the sum of all elements of $\mathbf{C}$. $\bar{\mathbf{z}} = \left(\frac{1}{m}\sum_{i=1}^{m} x_{ij}\right)_{k\times 1}$ is the mean time course of the data. Note that $\mathbf{z} \perp \mathbf{1_k}$ implies that $\bar{\mathbf{z}}$ is centered, i. e. has zero mean. Equivalent conditions must hold for the equality of the images. In Appendix B is shown that the equality of the time courses implies equality of the images and vice versa.

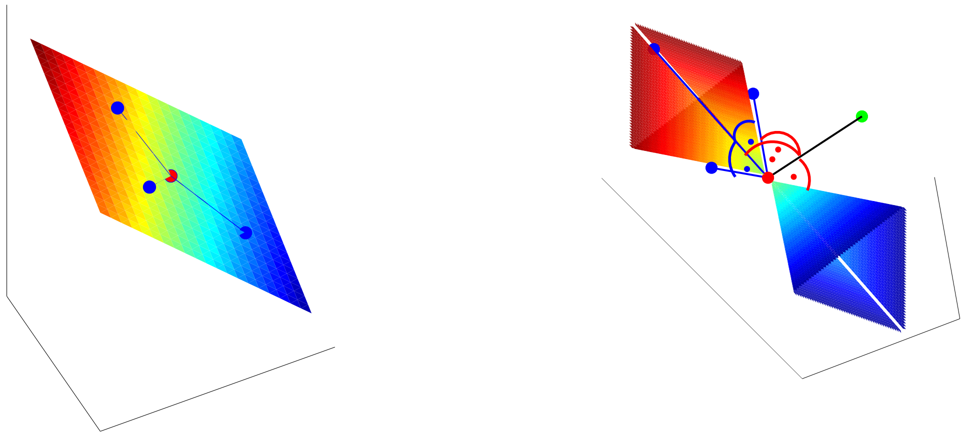
We also investigated the similarity of spatial and temporal PCA numerically. From the results thereof it seems that similarity, however, not equality, of spatial and temporal PCA occurs frequently. However there is a multitude of effects which can cause similarity and hence the single effects as well as their connection is cumbersome to check. On the other hand it could be this multitude of effects that is responsible for the fact that similarity of spatial and temporal PCA is encountered so often.

4.2 Dimensionality reduction using random matrix theory

As stated at the beginning of the chapter PCA is often used to divide the data space into two orthogonal subspaces, the signal space and the noise space. The assumption thereby is that the intrinsic dimension of the data is smaller than the rank of the data matrix $\mathbf{X}$. In this section we use criteria from random matrix theory to determine the intrinsic dimension of the data, which amounts to the dimension of the signal space. The signal space is characterized by large variances. Thus to separate the signal space from the noise space the eigenvalue spectrum of the covariance matrix of $\mathbf{X}$ is analyzed. In this section we use random matrix theory to extract the number of relevant dimensions which make

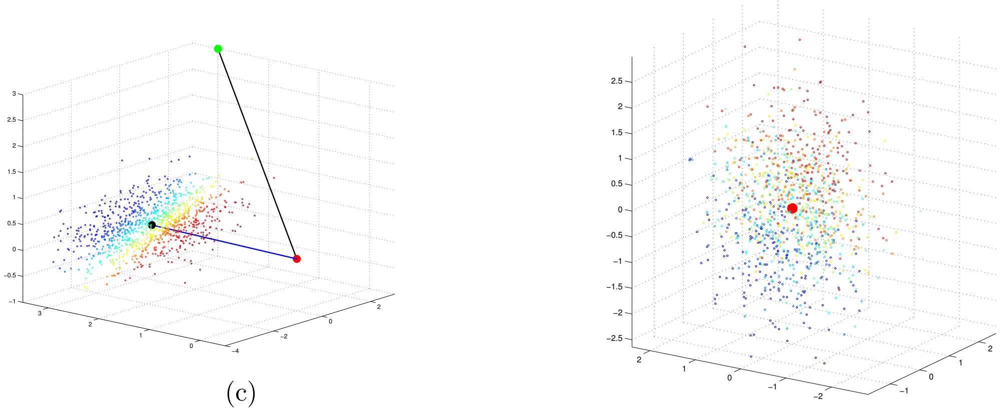
Temporally centered data **Z** Spatially centered data **Y**

Space of images (column space)



(a) (b)

Space of time courses (row space)



(c) (d)

Figure 4.5: Visualization of the data space of the temporally and spatially centered data matrices $\mathbf{Z}$ and $\mathbf{Y}$, respectively. Subfigures (a) and (b) show the column spaces of $\mathbf{Z}$ and $\mathbf{Y}$ whereas subfigures (c) and (d) show the respective row spaces. The red circles indicate the origins, the lines ending in blue circles in (a) and (b) represent the data vectors in the m dimensional space and the lines ending in green circles the vectors $\mathbf{1_m}$ and $\mathbf{1_k}$, respectively. The data in (c) and (d) is represented by coloured circles. For illustrative reasons $k = 3$ was assumed together with an arbitrary value $m > 3$. In this case only the row spaces can be visualized properly, the column spaces represented by coloured surfaces are subspaces of a higher than three dimensional space. However, choosing $m = 3$ and $k = 2$ would not have been very intuitive either, since then the row space of $\mathbf{Z}$ is one-dimensional. The visualizations of the column spaces are included to illustrate the formal similarity of the column space of $\mathbf{Z}$ (a) and the row space of $\mathbf{Y}$ (d) on one hand and the row space of $\mathbf{Z}$ (c) and the column space of $\mathbf{Y}$ (b) on the other. Note that only the data in (a) and (d) are centered around the origin which reflects the fact that the mean image of temporally centered data and the mean time course of spatially centered data equal zero. The data in (b) and (c) lie in $m - 1$ and $k - 1$ dimensional hyperplanes, respectively, defined by Equations 4.13.

up the signal space.

Wigner's semi-circle law

In random matrix theory (cf. e.g. [87]) statements about generic properties of ensembles of random matrices are derived. Most prominent in this context is the Wigner semi-circle law

$$p(\mu) = \frac{2}{\pi}\sqrt{1 - \mu^2} \tag{4.14}$$

which describes the density $p(\mu)$ of eigenvalues μ of an ensemble of random matrices. The eigenvalues thereby are uniformly scaled to the interval $[-1, 1]$. The random matrices are assumed to have independent Gaussian entries [109, 110], but Equation 4.14 applies also to more general cases [101].

Correlation matrices are positive definite or, in case of rank deficiency, positive semi-definite as e. g. the temporal correlation matrix $\mathbf{C_Z}$. However the rank deficiency here is not of importance, particularly, when only one eigenvalues is zero, as is the case for $\mathbf{C_Z}$, thus w. r. g. we restrict our considerations to positive definite matrices. A positive definite matrix has only positive eigenvalues and can be written as $\mathbf{C} = \mathbf{A}^T\mathbf{A}$ with a real matrix $\mathbf{A}$. Because $|\mu|^2$ is an eigenvalue of $\mathbf{C}$ if μ is an eigenvalue of $\mathbf{A}$, the semi-circle law becomes a quarter-circle law

$$p(\lambda) = \frac{4}{\pi}\sqrt{1 - \lambda^2}, \tag{4.15}$$

with $\lambda = |\mu| \geq 0$, provided that Equation 4.14 holds and $\mathbf{A}$ is a symmetric square root of $\mathbf{C} = \mathbf{A}^T\mathbf{A} = \mathbf{A}^2$ satisfying the conditions for Equation 4.14. The correlation matrices $\mathbf{C} = \mathbf{B}^T\mathbf{B}$ are constructed from asymmetric or rectangular matrices $\mathbf{B}$, which are related to the quadratic symmetric matrix $\mathbf{A}$ by an orthogonal projection $\mathbf{R}$ and the eigenvalues of $\mathbf{A}$ equal the singular values of $\mathbf{B}$. The relation between $\mathbf{A}$ and $\mathbf{B}$ reads

$$\mathbf{A} = \mathbf{RY} = \mathbf{VU}^T\mathbf{B} \tag{4.16}$$

where $\mathbf{B} = \mathbf{UDV}^T$ is the SVD, which in contrast to the previous sections here is assumed to consist of an $m \times k$ matrix $\mathbf{U}$, the columns of which are orthogonal and a square $k \times k$ diagonal matrix $\mathbf{D}$. The orthogonal projection $\mathbf{R}$ can be interpreted as an orthogonal transform of the m-dimensional column vectors in $\mathbf{B}$ to a basis where the spurious $m - k$ elements in the vectors are zero such that they can be written as k-dimensional vectors. From Equation 4.16 it is easily seen that the singular values of $\mathbf{B}$ equal the eigenvalues of $\mathbf{A}$ for which random matrix theory is applicable.

For real data we expect $\mathbf{A}$ to be only partly random. When testing whether Equation 4.15 holds, eigenvalues of large modulus appear to be responsible for deviations from Equation 4.14: if they are included in the rescaling of the spectrum Equation 4.15 fails completely. If a number of large eigenvalues is excluded Equation 4.14 is reproduced reasonably well, but the excluded eigenvalues become outliers. We are arguing here that these outliers form the extractable information from an underlying noise and are hence to be assigned to the signal space. We assume thus Equation 4.15 as a null-hypothesis and will now review criteria for the detection of significant deviations from random matrix theory.

The density $p(\lambda)$ in Equation 4.15 allows only for qualitative statements, because histogram techniques are required, which are bound to be inaccurate here. A more robust quantity is the integrated density of eigenvalues (a.k.a. integrated density of states; remind that we are considering positive matrices)

$$P(\Lambda) = \int_0^{\Lambda} p(\lambda) d\lambda \tag{4.17}$$

which in the case of Equation 4.15 becomes

$$P(\Lambda) = \frac{2}{\pi}\left(\Lambda\sqrt{1-\Lambda^2} + \arcsin(\Lambda)\right) \tag{4.18}$$

Equation 4.17 allows for more robust results, because it can be directly compared with the empirically obtained ordered sequence of eigenvalues by [87]

$$P^{-1}\left(\frac{n}{N}\right) = \lambda_n \tag{4.19}$$

where λ_n are the eigenvalues and $n \in \{1, \ldots, N\}$ with N being the total number of eigenvalues.

Fluctuations

There is a number of results on large deviations from the theoretically predicted curves, which allow to determine the significance of the deviations from random matrix theory. A reasonable estimate is [88] that for a matrix with entries in the interval $[0, 1]$ the probability of λ_s to deviate from its median by $\Delta\lambda$ is at most $4\exp\left(-\frac{(\Delta\lambda)^2}{16s^2}\right)$, where $s \in \{1, \ldots, N\}$.

Thus e. g. deviations of the first eigenvalue for more than 10 occur with probabixlity of less than 0.01. Taking into account that the mean and median of the largest eigenvalue of a matrix with entries from $\{0, 1\}$ are of order $n/2$ we should expect deviations deviations of less than 13% for $N = 150$, such that the 100% deviation in the data is apparently significant. Although the above bound is less tight when taking further eigenvalues into account it is not implied that these are of larger variance, rather they may be expected to vary less under certain conditions [88].

Fitting the eigenvalue distribution

In order to quantify the deviation from random matrix theory, we define the discrete function $N_C(\lambda_n) = n$ for all eigenvalues $\{\lambda_1, \dots \lambda_N\}$ of $\mathbf{C}$. Note that the curve of $N_C(\lambda_n)$ is obtained by interchanging the axes when plotting the eigenvalues λ_n in descending order. To compare N_C and Equation 4.18 the latter is generalized to

$$n(\Lambda) = \frac{2c}{\pi}\left(\left(\frac{\Lambda - a}{b}\right)\sqrt{1 - \left(\frac{\Lambda - a}{b}\right)^2} + \arcsin\left(\frac{\Lambda - a}{b}\right)\right) \tag{4.20}$$

by introducing a shift a and scale factors b and c. To fit the scales is necessary, because any straight-forward scaling would depend on the large eigenvalues which are expected not to obey the Wigner law. The parameters a, b, and c are optimized by gradient descent with respect to the energy function

$$E_{n_0}(a, b, c) = \sum_{n=n_0}^{n_1} \left(n(\lambda_n) - N_C(\lambda_n)\right)^2,$$

where $n_1 < N$ in order to exclude errors that occur due to the inclusion of small eigenvalues, and $n_0 \in \{1, \dots, n_{cut}\}$ with $n_{\text{cut}} < n_1$ such that a, b, and c still can be reliably optimized. Thus $1 \leq n_0 \leq n_{cut} < n_1 < N$, e. g. we may choose $N - n_1 = 40$ for $N = 150$. The results presented below show that the interesting region in the sense of large deviations is found for $n_0 \leq 30$ such that we can use $n_1 - n_{\text{cut}} = 60$, i. e. $n_{\text{cut}} = 50$. For each n_0 optimal parameters a, b, and c are determined and the individual errors

$$E_{n,n_0} = \left(n(\lambda_n) - N_C(\lambda_n)\right)^2$$

for each eigenvalue are determined for $1 \leq n_0 \leq n_1$, for $n_0 \leq n \leq n_1$, and as an extrapolation also for $n < n_1$. Several regions of E_{n,n_0} can be distinguished the boundaries of which are shown in Figure 4.6. We find a region of valid random matrix theory (region (1) in Figure 4.6) which is bounded by two types of critical values: If $n_0 < n_{0,\text{crit}}$ where $5 \leq n_{0,\text{crit}} \leq 10$ is a data-dependent value, the fit of the parameters a, b, and c is very poor, i.e. the 5 to 10 largest eigenvalues clearly do not obey the relations implied by random matrix theory (region (2) in Figure 4.6). This leads also to a failure of the first few n to satisfy the conditions for applicability of the fit, due to the arcsin-function (region (3) in Figure 4.6). On the other hand, for the first 20 to 25 eigenvalues, there is a significant elevation of the error (region (4) in Figure 4.6) which is more or less independent of $n_0 > n_{0,\text{crit}}$, i. e. n_{crit} is around 20.

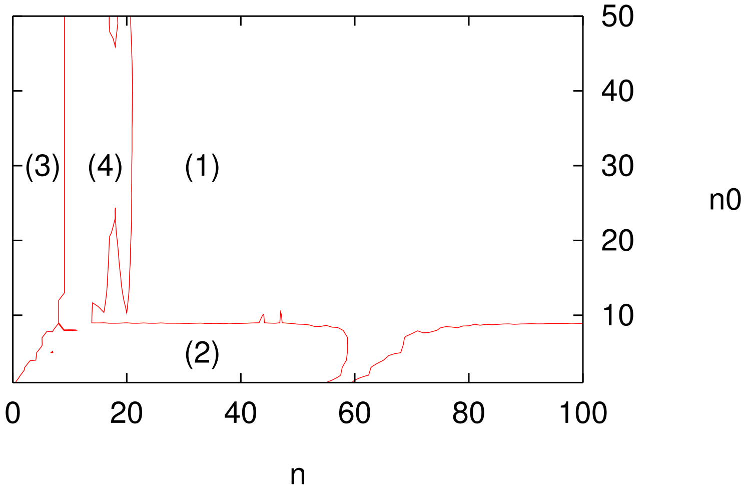


Figure 4.6: Contour plot of E_{n,n_0} for data from experiment 1, layer 9, eigenvalues from spatial PCA. The contour line is given by the lowest bound of the error level The regions (1) - (4) refer to the situations described in the text.

We may hence conclude that a most 25 eigenvectors are sufficient to represent the non-random contributions in the data.

The theory of random matrices provides a number of other results which might be of relevance here. E. g. Wigner's surmise on the distribution $P(s)$ of spacings s between neighboring eigenvalues

$$P(s) = As \exp\left(-Bs^2\right)$$

with constants A and B, might be checked, but would require density estimation or smoothing which leads to results of little robustness in the present case. But, for the purpose of justifying the dimension reduction of the data to about the first 20 principal vectors, the above procedure seems to be sufficient.

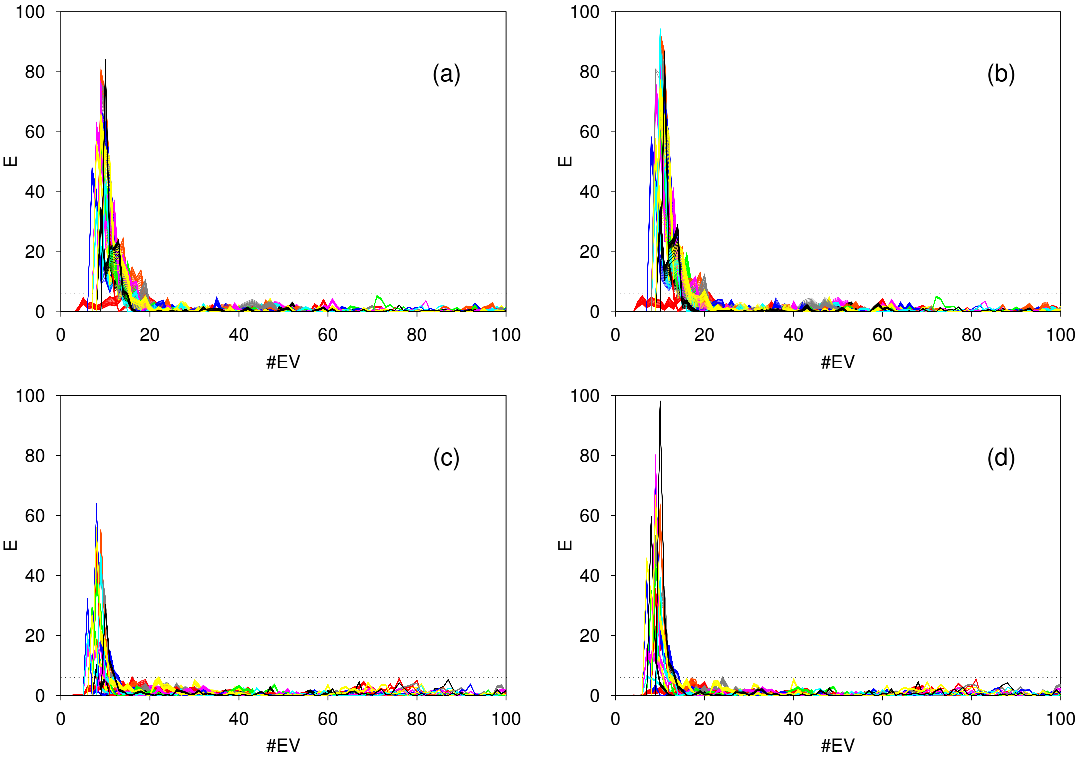


Figure 4.7: The error values E_{n,n_0} as a function of n for data from experiment 1. The traces for all layers and for $n_0 \in \{10, \ldots, 30\}$ are plotted on top of each other. The dotted horizontal line indicates a maximal noise level which was chosen identical for all four plots. Note that the traces for layer 1 do not significantly surpass the noise threshold determined over all layers, which is probably due to the number of brain voxels in the layer being small. It is apparent that where the brain has been extracted prior to the analysis (cf. chapter 2) more components ((a) 20 temporal, (b) 21 spatial) of potential relevance are separable from the noisy background than in the full image including both brain and extracranial regions ((c) 13 temporal, (d) 13 spatial). The values left of 10 on the x-axis could not be obtained by the fit used here. The terms temporal and spatial refer to the underlying eigenvalues resulting from temporal and spatial PCA.

Chapter 5

Independent component analysis

Independent component analysis (ICA) is an approach which aims to extract from the data the presumably statistically independent underlying processes. ICA is a special case of Blind Source Separation (BSS) where the aim is to separate the underlying sources from a mixed signal according to various criteria, the one of ICA being statistical independence. In contrast to principal component analysis (PCA) it is an approach which accounts for the statistical properties higher than second order. In the case of multivariate gaussian distributed data, uncorrelated components, i. e. components independent up to the second order, are statistically independent and hence in this case PCA is equivalent to ICA. In a similar way as the principal components the independent components provide a basis of the data space, however in contrast to PCA the basis is in general not orthogonal. In this chapter we discuss the main approaches in the field of ICA and present a comparison of spatial and temporal ICA, at the same time analyzing the effect of prior dimensionality reduction of the data. An overview over various approaches to ICA can also be found in [4, 59]. Historical reviews are provided in [23] and [57]. Results of separating fMRI data into spatially independent components can also be found in [105].

5.1 Linear ICA model

The basic assumption of independent component analysis is, that the data is a linear superposition (also called mixing in the following) of statistically independent processes and can hence be written as

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

where $\mathbf{x}(t) \in \mathbb{R}^m$ is a vector of measurements, $\mathbf{A}$ is a time-independent $m \times n$ mixing matrix, $\mathbf{s}(t) \in \mathbb{R}^n$ is a random vector the components of which are statistically independent and t indicates time. Also ICA using a nonlinear model has been considered as e. g. in [60]. Often the elements of $\mathbf{x}$ are called sensors and the elements of $\mathbf{s}$ sources, for ICA has been developed in the field of signal processing [27, 23].

Statistical independence of the elements of $\mathbf{s}(t) = (s_1(t), \ldots, s_n(t))^T$ means that

$$p(\mathbf{s}(t)) = \prod_{i=1}^{n} p(s_i(t)) \tag{5.2}$$

where $p(\mathbf{s}(t))$ is the joint probability density of $\mathbf{s}$ and $p(s_i(t))$ is the marginal probability density of the elements $s_i(t)$, $i \in \{1, \dots, n\}$. All that is known a priori in this ICA model is the measurement vector $\mathbf{x}(t)$, whereas the matrix $\mathbf{A}$, the number n of columns of $\mathbf{A}$, the random vector $\mathbf{s}(t)$ as well as the probability densities $p(\mathbf{s}(t))$ and $p(s_i(t))$ are unknown.

For most applications the assumption of $\mathbf{s}(t)$ being non-stationary and hence the probability density $p(\mathbf{s}(t))$ being time-dependent is not tractable, since the measurement vector must be used to infer informations about $p(\mathbf{s}(t))$. In the following is thus assumed that $\mathbf{s}$ is stationary and hence $p(\mathbf{s})$ and the $p(s_i)$ are time independent. As for n it is often assumed that $n = m$ and $\mathbf{A}$ is invertible. Approaches where the number of sources is different from the number of signals can be found in [5, 72, 22].

For the solution of Equation 5.1 two main classes of approaches can be distinguished, information theoretic and cumulant based. Information theoretic approaches are used e. g. in [61, 6]. The information theoretic quantities derived from $p(\mathbf{s})$ require implicit assumptions about $p(\mathbf{s})$ whereas cumulant based approaches, such as in [63, 73], are equivalent to estimating $p(\mathbf{s})$ up to a certain order. On the basis of the chosen criterion a so called *contrast* or *objective function* $F(\mathbf{x}, \mathbf{A})$ is defined which is to be extremized w. r. t. the elements of $\mathbf{A}$.

In the context of fMRI measurements t is discrete and there is only a finite number of measurements. Hence we can write the model in Equation 5.1 as

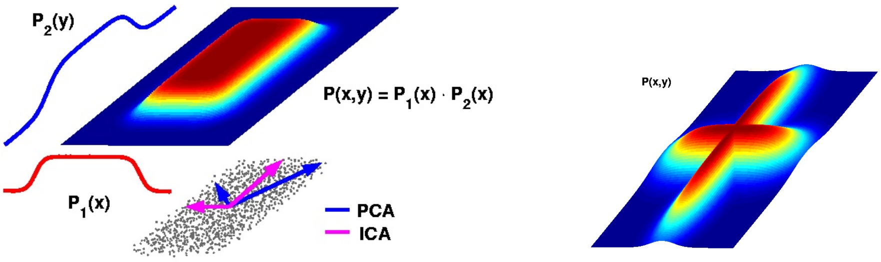
$$\mathbf{X} = \mathbf{AS} \tag{5.3}$$

where $\mathbf{X}$ is the $m \times k$ data matrix, $\mathbf{A}$ is as in Equation 5.1 and $\mathbf{S}$ is a $n \times k$ random matrix the columns of which consist of statistically independent elements. Note that $\mathbf{A}$ and $\mathbf{S}$ are defined only up to permutation and scaling, since $\mathbf{A}' = \mathbf{AP\Lambda^{-1}}$ multiplied by $\mathbf{S}' = \mathbf{\Lambda PS}$, where $\mathbf{\Lambda}$ is a (diagonal) scaling matrix and $\mathbf{P}$ a permutation matrix, leaves the data matrix $\mathbf{X}$ invariant. In contrast to PCA for ICA there is no inherent order of the components. An illustration of ICA in contrast to PCA is given in Figure 5.1(a). However, there are cases where the model assumption is not valid. Such a case is illustrated in Figure 5.1(b) by the example of a bivariate probability density which cannot be composed of two independent probability densities.

5.2 Prewhitening

The first step in ICA usually consists in a prewhitening of the data. This is done in order to remove second order statistical dependencies from the data, since in ICA one is interested in the higher order statistical properties. In addition to decorrelation, the data is rescaled to unit variance, which has the advantage that in this case the mixing matrix is orthogonal. Prewhitening can be achieved by PCA. If $\mathbf{U\Lambda U}^T$ denotes the eigendecomposition of the covariance matrix $\mathbf{C} = E(\mathbf{xx}^T)$ of the data then the prewhitened data reads

$$\mathbf{y} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{x} \tag{5.4}$$



(b)

(a)

Figure 5.1: Illustration of ICA of a bivariate data distribution (a) and an example where the ICA model is not valid (b).
(a) The surface depicts the joint probability density p of a 2-dimensional random vector $\mathbf{y}$. The point cloud represents the data (realizations of p). Each coordinate y_i of a realization is a linear superposition of the realizations of two statistically independent random variables s_1 and s_2, since the joint probability density p was created by multiplying the two marginal probability densities $p_1(s_1)$ and $p_2(s_2)$ shown along the two axes. PCA reveals the orthogonal directions with respect to which the data is uncorrelated (blue arrows). ICA reveals the directions with respect to which the data is statistically independent (red arrows). Referring to our data the coordinates of each point would reflect an image (consisting of only two voxels) and the PCA vectors would indicate the images which have uncorrelated time courses whereas the ICA vectors indicate the images having statistically independent time courses.
(b) The bivariate probability density $P(x, y)$ shown in (b) cannot be the result of a product of two marginal probability densities.

The covariance matrix of the whitened data $\mathbf{y}$ equals the identity matrix $\mathbb{I}$ and thus the data distribution is spherical. It is assumed here that the covariance matrix has full rank, which is not the case for finite sample size. This was discussed in the context of PCA in section 4.1, however, since we performed ICA on a subspace of the data space (cf. section 5.4) this is not critical here.

For ICA is now used the fact that prewhitening is only defined up to an orthogonal transformation. If $\mathbf{R}$ is an orthogonal matrix, left multiplying $\mathbf{x}$ by $\mathbf{R}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T$ prewhitens the data as well as can be easily seen by inserting this term into Equation 5.4 and computing the covariance matrix. Thus in ICA for prewhitened data the indeterminacy up to an orthogonal transform is used to determine the "right" transform for source separation according to a chosen criterion. For prewhitened data Equation 5.3 is written as

$$\mathbf{Y} = \mathbf{R}\mathbf{S} \tag{5.5}$$

with the prewhitened data matrix $\mathbf{Y}$ and an orthogonal mixing matrix $\mathbf{R}$. Now we need a criterion w. r. t. which **R** is to be determined. This is dealt with in section 5.3, but before that, in the next section some useful basics of information theory are given. ICA model is not valid. The bivariate probability density $P(x, y)$ shown cannot be the result of a product of two marginal probability densities.

5.3 Criteria for statistical independence

In this section we discuss some of the approaches derived from the two main criteria for statistical independence used in ICA, namely minimizing mutual information and joint-diagonalizing higher order cumulants. First we need some information theoretical definitions.

5.3.1 Definitions from information theory

Information theory is of great importance in ICA, since the aim of ICA can be understood as separating the information flow of the underlying sources. Information theory provides useful criteria for statistical independence such as the widely used mutual information. Also other BSS related criteria such as the interestingness of a probability distribution in the sense of quantifying its non-gaussianity can be derived from information theory. In this section some information theoretical definitions necessary to understand most of the criteria used in ICA are given and some of their properties shortly sketched. Figure 5.2 shows an illustration of the relation of the quantities defined below.

Statistically independent random variables thus do not exchange information in the sense of information theory. In information theory the term information is equivalent to the term entropy in statistics [20].

Definition: (Entropy, Joint Entropy, Conditional Entropy) [104]
The *entropy* of a discrete random variable X is defined by

$$H(X) = -\sum_{x\in\Theta_X} p(x)\log p(x) \tag{5.6}$$

where $p(x)$ is the probability distribution of X and Θ_X is the set of values that X can assume. The *joint entropy* $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X,Y) = -\sum_{x\in\Theta_X}\sum_{y\in\Theta_Y} p(x,y)\log p(x,y) \tag{5.7}$$

The *conditional entropy* is defined by

$$\begin{aligned} H(Y|X) &= \sum_{x\in\Theta_X} p(x)H(Y|X=x) \\ &= -\sum_{x\in\Theta_X} p(x)\sum_{y\in\Theta_Y} p(y|x)\log p(x|y) \\ &= -\sum_{x\in\Theta_X}\sum_{y\in\Theta_Y} p(x,y)\log p(y|x) \end{aligned} \tag{5.8}$$

It is the remaining uncertainty of Y if X is known. The relation of the joint entropy and the conditional entropy is governed by the chain rule [104]

$$H(X,Y) = H(X) + H(Y|X) \tag{5.9}$$

Definition: (Relative Entropy or Kullback Leibler (KL) Distance, Negentropy) [104]
The *relative entropy* or *Kullback Leibler* (KL) *distance* between two probability distributions $p(x)$ and $q(x)$ is defined as

$$D(p \parallel q) = \sum_{x \in \Theta_X} p(x) \log \frac{p(x)}{q(x)} \tag{5.10}$$

The Kullback Leibler distance is a measure of discrepancy between two distributions, however it is not a metric. The KL distance is always non-negative and zero iff $p = q$ almost everywhere, but it is not symmetric and in general does not satisfy the triangle inequality. It is invariant under invertible transformations, if the densities p and q are defined on the same space. The KL distance $D(p_{\mu\sigma^2} \| N(\mu,\sigma^2))$ between an arbitrary density p_{μ,σ^2} having mean μ and variance σ^2 and the Gaussian density $N(\mu,\sigma^2)$ is called *negentropy* [61].

We now introduce mutual information, which can be interpreted as a symmetrized version of the KL distance. Mutual information is a measure of the amount of information that one random variable contains about another, and hence the reduction in the uncertainty of one random variable due to the knowledge of the other [104]. Mutual information is an important quantity for ICA as it equals zero, iff the random variables from which it is computed are statistically independent.

Definition: (Mutual information) [104]
The *mutual information* between two random variables X and Y with the joint probability distribution $p(x,y)$ and the marginal probability distributions $p(x)$ and $p(y)$ is defined as

$$I(X,Y) = \sum_{x \in \Theta_X} \sum_{y \in \Theta_Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{5.11}$$

It can be interpreted as the uncertainty of X that is due to the uncertainty of Y or vice versa. The mutual information $I(X,Y)$ is the KBD between the joint distribtuion and the product distribution $p(x)p(y)$

$$I(X,Y) = D(p(x,y) \| p(x)p(y)) \tag{5.12}$$

An important relation between statistical independence and entropy is given by the *independence bound on entropy* [104]:

Let $X_1, \ldots, X_n$ be random variables with the joint probability density $p(x_1, \ldots, x_n)$. Then

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i) \tag{5.13}$$

with equality iff the X_i are statistically independent. From this can be derived that

$$I(X_1, \ldots, X_n) \geq 0 \tag{5.14}$$

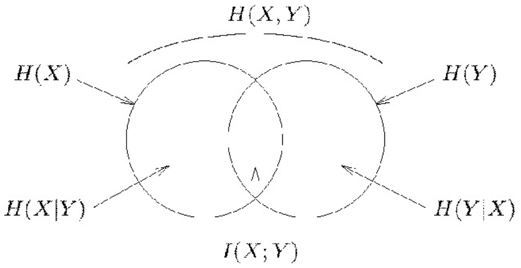


Figure 5.2: Venn diagram illustrating the relations among the important information theoretical quantities joint entropy $H(X,Y)$, entropy $H(X)$, $H(Y)$, conditional entropy $H(X|Y)$, $H(Y|X)$, and mutual information $I(X,Y)$ (from [54]).

5.3.2 Mutual information

The previous section introduced the mutual information as the information one random variable contains about the other. Intuitively it is obvious that statistically independent random variables should not contain any information about each other and hence the mutual information should be zero, which is indeed true (cf. Equation 5.13 and Equation 5.14), statistical independence and zero mutual information being even equivalent. Under the assumption of the linear model Equation 5.1 the mutual information between the elements of the measurement vector $\mathbf{x}$ equals the KL distance of the joint probability density of the sources and the factorized probability density of the measurements. Minimizing this w. r. t. to the elements of the mixing matrix $\mathbf{A}$ leads to the "true" mixing matrix for the linear model.

Approaches related to information theory are given e. g. in [23] where the negentropy is related to mutual information. Maximizing negentropy minimizes mutual information. In [61] ICA is performed by maximizing an approximation of negentropy up to the fourth order as objective function. In [27] a maximum likelihood approach is performed which can be shown to be equivalent to minimizing mutual information. Another approach is the so called "infomax" principle [74, 75, 6] which occurs in the development of information-theoretic unsupervised learning rules for neural networks. There the aim is to maximize the mutual information between inputs and outputs in a neural network. In [89] is shown that under certain conditions the infomax principle is equivalent to minimizing mutual information.

5.3.3 Cumulant diagonalization

Another class of approaches to ICA is based on cumulant diagonalization. These approaches rely on the fact that for statistically independent random variables the cumulant tensors are diagonal, and at least for bivariate random variables the reverse is also true, provided that the joint distribution is determined by its moments ([82], p. 36). The n-th order cumulant $\kappa^{(n)}$ of a one dimensional random variable Y is given by the n-th derivative of the cumulant generating function $K_Y(\tau)$ at $\tau = 0$ [3]:

$$K_Y(\tau) = \ln \int e^{i\tau y} p(y) dy \tag{5.15}$$

$$\kappa^{(n)} = (-i)^n \frac{\partial^n K_Y}{\partial \tau^n}\bigg|_{\tau=0} \tag{5.16}$$

with p being the probability density of the random variable Y. The first and second order cumulants correspond to the first and second order moments, respectively. For higher dimensional random variables the nth order cumulant $\kappa^{(n)}$ is a contravariant tensor of the order n [82]. The algorithm in [63] aims at diagonalizing the fourth order cumulant tensor. The goal is to determine an orthogonal matrix $\mathbf{R}$ such that the fourth order cumulant tensor of $\mathbf{R}^T\mathbf{y}$ is diagonal, where $\mathbf{y}$ is the prewhitened data vector in Equation 5.4. In [63] an approach is taken which is based on the following two properties: First it can be shown that $\mathbf{R}$ jointly diagonalizes so called 'cumulant matrices', matrices resulting from a weighted sum over two dimensions of the fourth order cumulant tensor with arbitrary weights. Second, an orthogonal data transformation does not affect the total sum of the squared elements of the fourth order cumulant tensor, so the latter can be diagonalized by maximizing the sum of squares of its diagonal elements with respect to $\mathbf{R}$. This is essentially the approach taken in [23]. The approach in [63] maximizes the sum of of squares of the cumulant tensor elements having identical first and second indices. This is equivalent to jointly diagonalizing a set of n^2 cumulant matrices. This set can again be reduced to the n 'most significant' cumulant matrices by diagonalizing an $n^2 \times n^2$ matrix. While having the appealing relation to diagonalization as in PCA and showing good results in separation the diagonalization of an $n^2 \times n^2$ matrix restricts the applicability of the approach to a small number of sources.

The third order cumulant tensor could have been used as well in the above approach, although then the algorithm would fail in the important case of symmetrical source distributions, where the third order cumulant is zero. However recently an approach for simultaneous diagonalization of the third and fourth order cumulant tensor [73] has been proposed which seems to work reasonably well.

5.4 Results from ICA

We used the cumulant based approach in [63], the algorithm of which is publicly available, to perform ICA on fMRI data. As the diagonalization of an $n^2 \times n^2$ matrix is prohibitive for high dimensional fMRI data, a reduction down to about 30 dimensions of the data is necessary which can be achieved by projecting the data onto the first principal axes. This is justified since one can assume that the higher principal components mainly reflect Gaussian noise. Gaussian components are a nuisance in ICA anyway, because their higher order cumulants are zero, which makes the diagonalization of the fourth order cumulant tensor difficult. Further we showed in section 4.2 of chapter 4 that the dimension of the signal space of the data for the experiments considered is most probable less than 30 except for one case where it is 32.

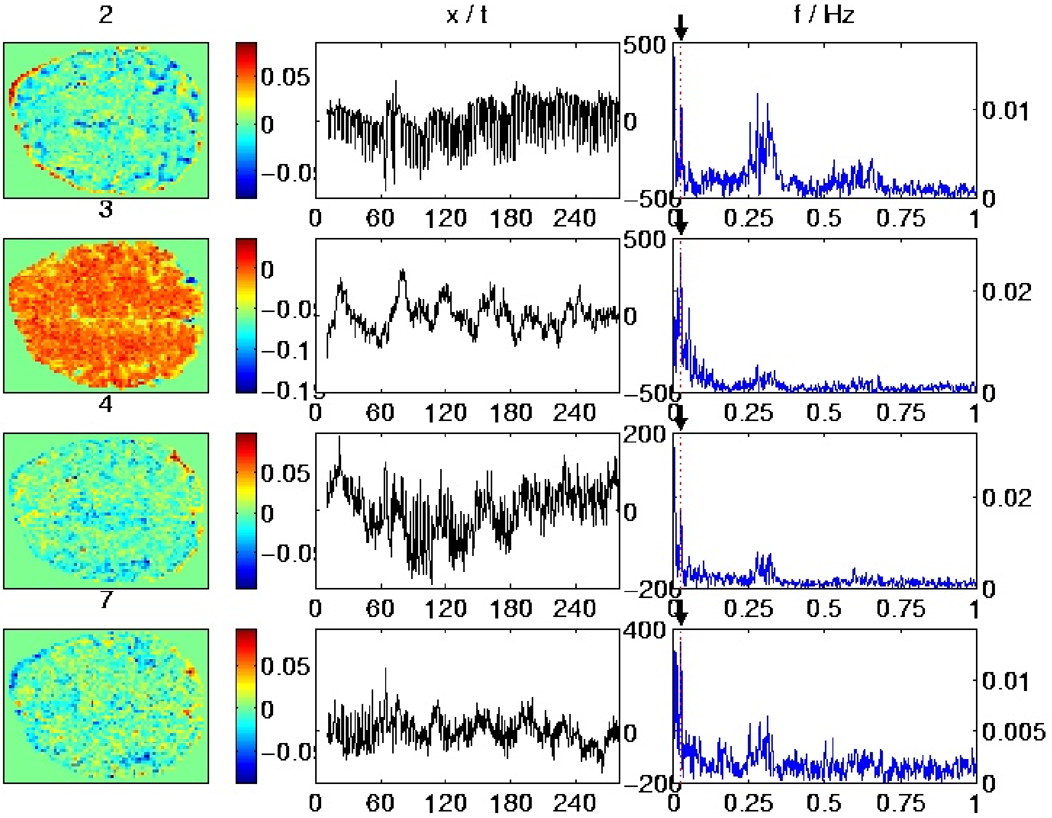


Figure 5.3: PCA of actual finger tapping.
Left hand side: Images corresponding to the PC's 2, 3, 4 and 7. These PC's were the ones with the highest peaks at the finger tapping cycle frequency of 0.025 Hz.
Middle: PC's (time courses) indicating the strength with which the PC is present in the measured image at the respective point in time.
Right hand side: Square root of the power spectrum of the PC's time course normalized to enable comparison. The finger tapping cycle frequency is indicated by arrows and dotted lines.

5.4.1 Actual and imaginary fingertapping

Successfull application of ICA can be characterized by the merging of features that occur in several principal components into one independent component [32, 33]. In PCA on data from experiment 2 the frequency of 0.025 Hz related to the stimulus cycle shows up in the Fourier spectrum of several components indicating suboptimal separation. Figure 5.3 shows the four PC's in which the stimulus cycle frequency was most prominent. The "brain-like" pattern in PC 2 most likely indicates head movements due to breathing an assumption which is also supported by the broad peak at the frequency around 0.3 Hz. The stimulus time course is best reproduced by the third PC, but with a weak activation in the motor cortex. Note that another weak activation is visible in an area which can be identified as the supplementary motor cortex, which is active also in imagined finger tapping (cf. Figure 5.5). The most prominent activation

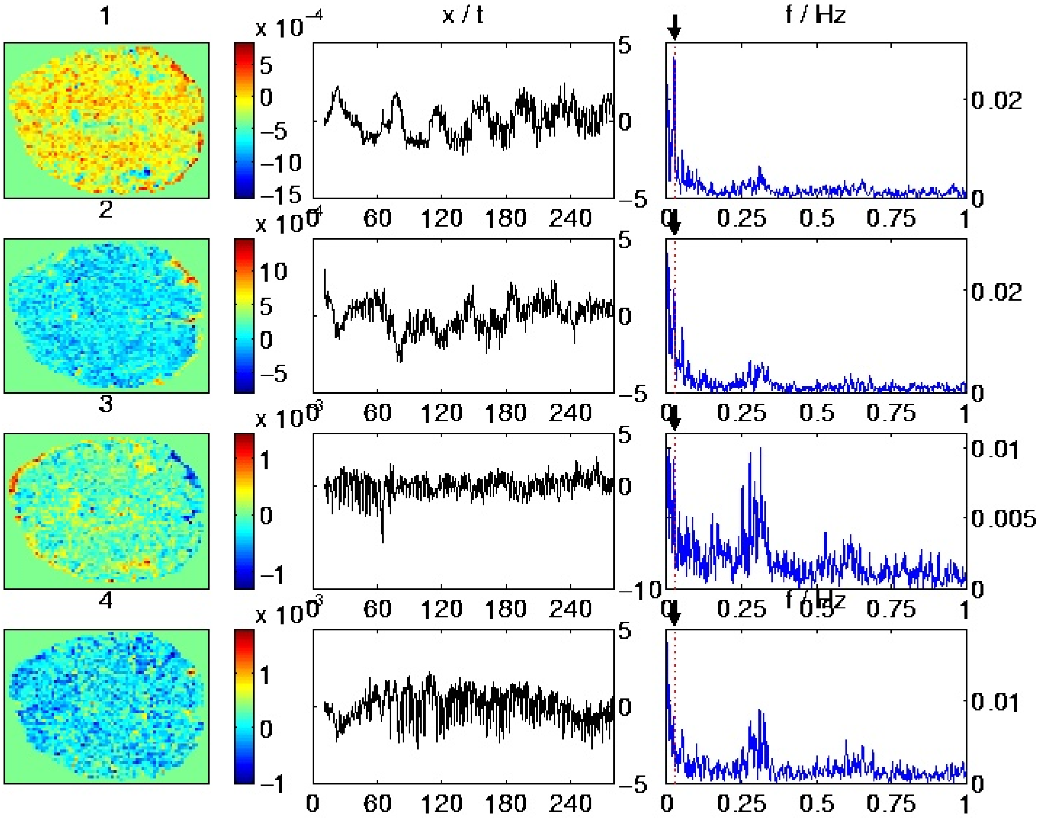


Figure 5.4: ICA of actual finger tapping.
Independent Components obtained by ICA of the projection of the data onto the PC's numbered 2, 3, 4 and 7. For simplicity the IC's are numbered as well although no intrinsic order exists for independent components. Obviously the first IC contains most of the finger tapping cycle frequency and also an activated area in the motor cortex is visible. This indicates that indeed by ICA the stimulus related activity has been separated into one component. However the second independent component contains stimulus related activity as well however not as pronounced as in the first component. From the activity distribution which apparently includes the supplementary motor cortex and some blood vessel this could indicate a secondary stimulus related process which was separated from the first.

of the motor cortex is present in PC 7. The corresponding independent components are shown in Figure 5.4. As a result of using the stimulus to chose the components for dimension reduction performing ICA leads to a far better, however not total separation of the stimulus. The first independent component is clearly stimulus related, however exhibiting an increasing noise level towards the end of the time course, and shows a localized activation in the primary motor cortex. The activation is essentially unilateral which could be due to the slice only partially including the contralateral part of the primary motor cortex. In the second independent component the relatedness to the stimulus is less pronounced and also the activity distribution does not show activation of the primary, but of the supplementary motor cortex, which is also present on imaginary fingertapping (Figure 5.5, [33]). Also the activation pattern of the second independent component is similar to the one of the third principal component which is the one most closely related to the stimulus and includes an activated region near the scull which is most probably a blood vessel that could be engaged in the blood supply in response to the stimulus. A similar, slightly less concentrated result is obtained by projecting the data onto up to the first 30 principal components regardless of their relatedness to the stimulus.

Figure 5.5 shows the stimulus related principal components of the imaginary fingerapping. The results of performing an ICA on the data projected thereon are shown in Figure 5.6. In this case ICA did not lead to a visible improvement which could indicate that in a sense the data from imaginary finger tapping are more gaussian than the data from actual fingertapping. It could be speculated that the imagination of fingertapping is a more complex task than actual fingertapping involving more independent processes the sum of which tend to be more gaussian distributed, stressening the central limit theorem.

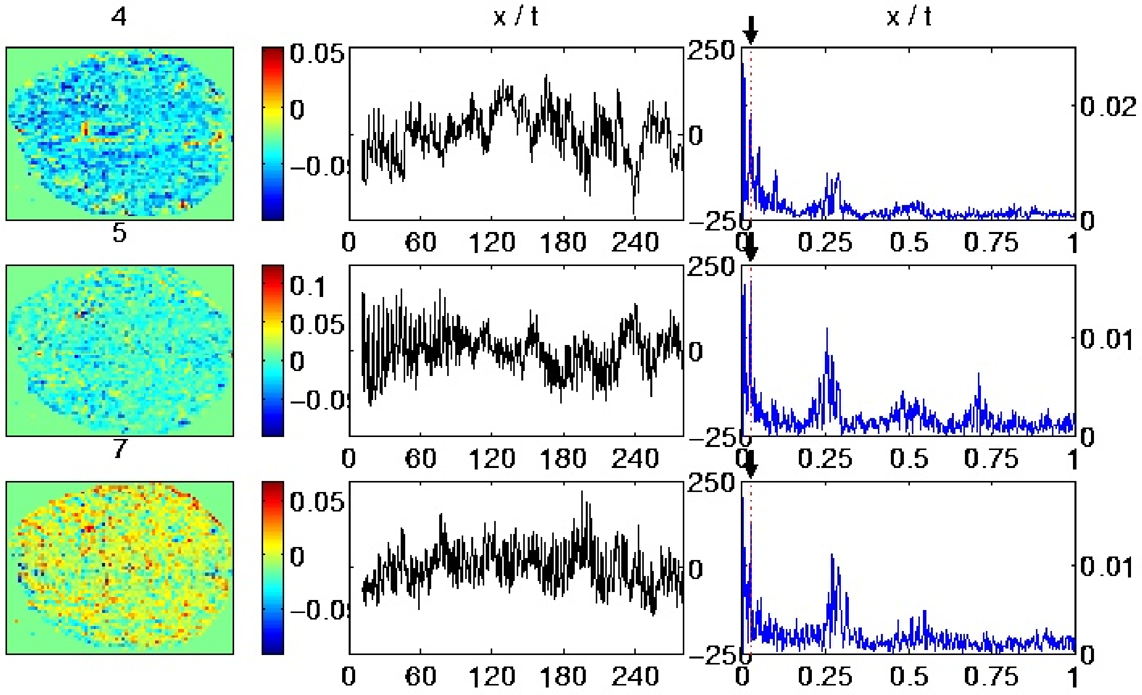


Figure 5.5: PCA of imagined finger tapping.
PC's 4, 5 and 7 with the highest peaks at the tapping cycle frequency. In the fourth PC an activated area which can be identified as the supplementary motor cortex is visible. The other two PC's have prominent peaks at 0.025 Hz, however they exhibit less spatial structure. In the fifth PC the sagittal sinus seems to be involved as well a small area in the left forebrain. The seventh PC shows more of a distributed activation with the same small area present.

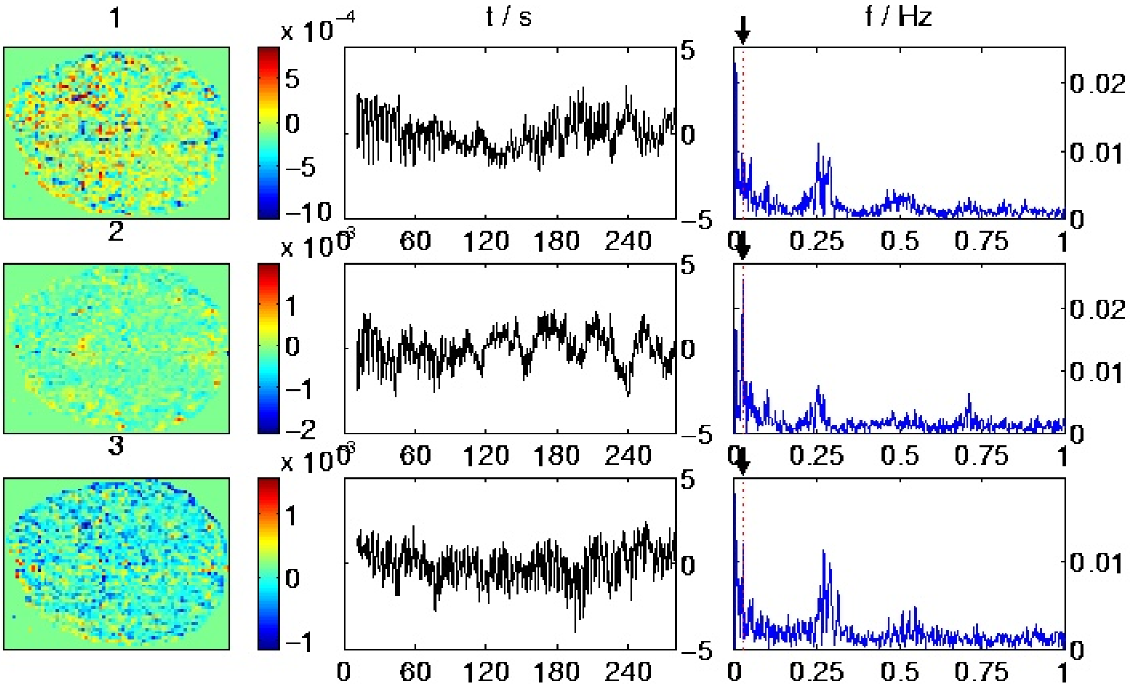


Figure 5.6: ICA of imagined finger tapping.
Independent Components obtained by ICA of the projection of the data onto the PC's numbered 4, 5 and 7. The second IC contains most of the finger tapping cycle frequency, however the activated areas are not so clearly visible as in the fourth PC in Figure 5.5. It seems that in this case ICA did not yield an improvement over PCA though the peak at 0.025 Hz is very pronounced in the second IC compared with the others.

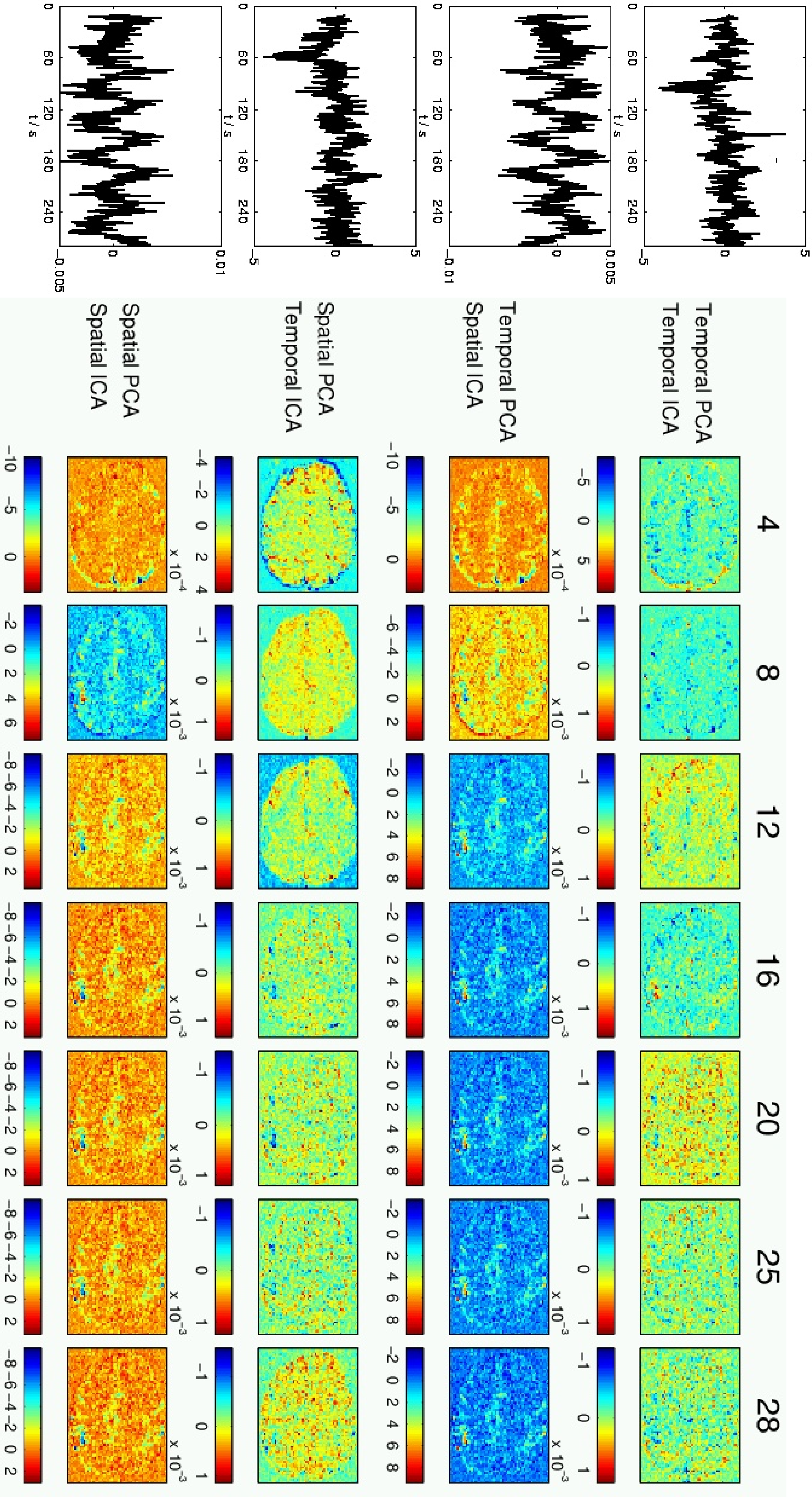
5.4.2 Temporal vs. spatial ICA

So far in our analysis using PCA and ICA we have assumed that the interesting factors are characterized by a time course which is uncorrelated to or independent of time courses resulting from other influences on the data. However, one could also assume that the factors of interest are characterized by a certain independent or uncorrelated activity distribution. These factors could then be revealed by spatial PCA and ICA. In Figure 5.7 we show a comparison of temporal and spatial ICA for various numbers of principal axes used for preprojection together with the respective time courses plotted on top of each other. Only the stimulus related component is shown. The time courses of the spatial ICA reflect the stimulus more closely than the ones from temporal ICA and also the images show more structure than the ones for temporal ICA. Thus in contrast to PCA where temporal and spatial results are very similar, for ICA differences occur, which provide evidence in favor of spatial ICA [31, 35].

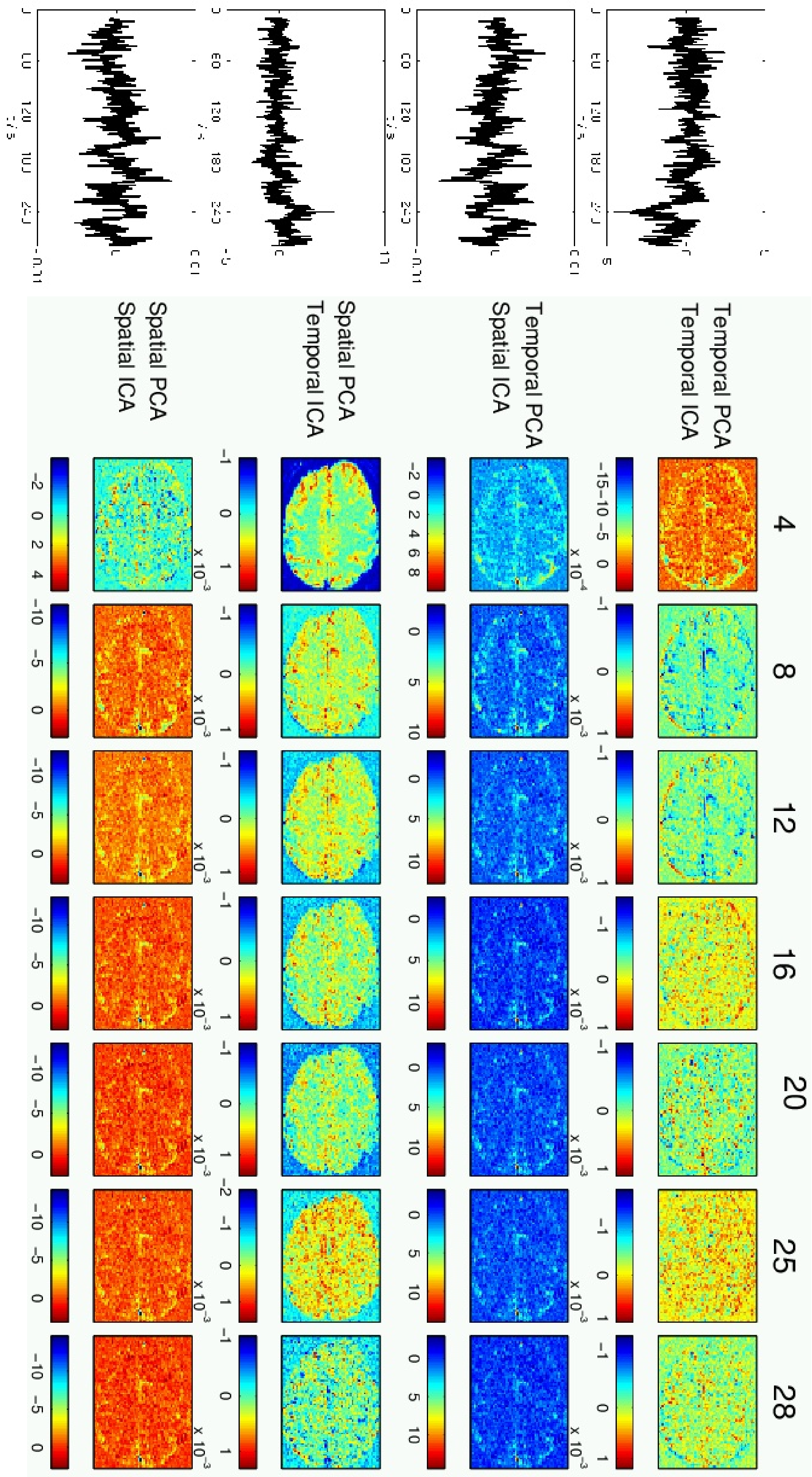
For completeness all four possible combinations of temporal and spatial PCA and ICA, respectively, were performed, although the mixed combinations are hardly interpretable from a theoretical point of view. The optimum number of 12 for PV's to preproject the data onto that is indicated by the figure could reflect the fact that the stimulus has a strong influence which is increasingly shaded by including more than 12 PV's as would be suggested from the results in the previous chapter. Note, however, that although the third PC is most closely stimulus related the optimum number of PV's suggest that the stimulus related influences were distributed over the first 12 PC's and are best recovered by projecting on them and separating the stimulus from other processes by ICA.

Figure 5.7: (next two pages) (a) Actual and (b) imaginary fingertapping in experiment 2. Results of temporal and spatial ICA for various numbers of PV's onto which the data were projected prior to performing ICA. The time courses resulting from ICA are plotted on top of each other. The stimulus related components were extracted by Fourier analysis. Spatial ICA leads to considerably better results than temporal ICA. The time courses of spatial ICA better reflect the stimulus time course than the time courses of temporal ICA and the images of spatial ICA show focal activations in the motor and supplementary motor cortex, whereas the images of temporal ICA are rather noisy. The optimum number of PV's for preprojecting seems to be around 12. Note that in the images of spatial ICA of the actual fingerapping the area on the left occipital cortex which was strongest in the stimulus related PCA image of PC 3 in Figure 5.3 and could be interpreted as a blood vessel has disappeared in the ICA images for a preprojecting number of PV's ≥ 8. For the imaginary fingertapping results this area disappeared more slowly. Further, the influence of the sagittal sinus was not separated from the supplementary motor cortex activity. The result that spatial ICA outperforms temporal ICA was found consistently over the data sets.

(a)


Temporal PCA
Temporal ICA
Temporal PCA
Spatial ICA
Spatial PCA
Temporal ICA
Spatial PCA
Spatial ICA
4
8
12
16
20
25
28
t / s
0
60
120
180
240
−5
0
5
−0.01
0
0.005
0.01
−0.005
x 10⁻⁴
x 10⁻³

(b)


4
8
12
16
20
25
28
Temporal PCA
Temporal ICA
Temporal PCA
Spatial ICA
Spatial PCA
Temporal ICA
Spatial PCA
Spatial ICA
x 10^-3
x 10^-4
60
120
180
240

Chapter 6

Functional connectivity and graph theory

In this chapter we derive a graph theoretical approach to functional connectivity which is based on correlation matrices. The structure of the correlation matrices is revealed by analyzing the structure of a graph which is extracted from the correlation matrix and the subgraphs of which are determined and identified with functional units. Delayed correlation matrices along with the concept of hypergraphs are used to include time delayed, possibly causal functional connectivity.

6.1 Functional connectivity

Functional connectivity between brain regions affects cross-correlations of time series of voxel activities in fMRI. In order to reveal relevant activity the latter must be separated from vegetative processes, artefacts and noise. On a global scale components of brain activity can be identified automatically by independent component analysis, but this technique does not clarify the mutual relationships between the activated regions. For this purpose we propose a graph-theoretical clustering method which is based on the assumption that pairs of voxels could be considered as functionally connected if the temporal cross-correlation of their signal is high. The method proposed here is completely data driven and involves the identification of correlational structures among voxels using general graph theoretical properties. Further, we include time-delayed correlations and the concept of hypergraphs to identify delayed functional connectivity between brain regions.

Cross-correlations have been widely used as measure of functional connectivity in fMRI. Functional connectivity MRI, in short fcMRI, is usually based on resting state images, however there is evidence that the functional connectivity and task activation MR signal changes arise from the same BOLD-related origins [92] and further that functional connectivity of regions unrelated to task activation seems to be unaffected by the presence of activated areas [7]. These findings also suggest that the distinction between activation and rest in functional brain imaging is mainly gradual as should be expected, since the concept of a resting brain is somewhat artificial in a living brain.

The first approaches to reveal functional connectivity in the context of fMRI appeared in [8, 12], where ROI's were defined by conventional methods during task activation and regions functionally connected to them were determined by correlations among the voxels inside and outside the ROI's during a resting state measurement. This approach is used frequently to date, e. g. [80, 26], sometimes slightly varied in the choice of the ROI as e. g. in [103]. Functional connectivity is often computed by low-pass filtering the time course. This restricts the considerations to low frequency fluctuations which are believed to arise from spontaneous BOLD oscillations [92]. A combination of correlation measures and frequency content of the voxels done in [25] found evidence that low frequency oscillations (< 0.1 Hz) contributed to more than 90% of the correlation coefficient. However in the present context we did not impose any constraints or preprocessing on the time courses, which in the light of [92, 7] appears reasonable. Our framework does not rely on any assumptions about preprocessing or the absence thereof and hence filtering could be easily included.

Further approaches to determine functional connectivity include hypothesis building [56], multidimensional scaling using definitions of functional connectivity as its metric [50], partial least squares [85], prediction [49], image based funcional connectivity using random field theory [21, 111], structural equation modeling [83, 84], modeling by replicator dynamics [79] and various types of clustering [11, 10, 24]. Approaches to relate functional connectivity to anatomical connectivity have been made e. g. in [67] using diffusion tensor imaging and in [102] based on anatomical connectivity matrices. An approach related to the one presented here has been independently suggested in [115] where complete graphs from a region of interest (ROI) defined by conventional stimulus-driven methods are determined and extended to complete graphs including voxels outside the ROI.

In our approach we do not rely on predefined ROI's, but use the correlation structure of the voxel time courses as a whole. Relating the correlation structure to graphs, we analyze subgraphs based on various criteria to define functionally connected areas. The basic idea of our approach is to consider the voxels of an image as the vertices of a graph and the temporal correlation matrix of their time courses as the weight matrix of the edges between the vertices. For zero delay the correlation matrix is symmetric and the corresponding graph is undirected. An unweighted graph can be extracted from the weighted graph by deleting all edges that have weights below a certain threshold. Subgraphs of the so extracted graph can be identified with undelayed functionally connected units. We compare various subgraph definitions with respect to their suitability for definiting functional units.

Delayed correlation matrices can be used to identify delayed, i. e. possibly causal, functional connectivity. Since delayed correlation matrices in general are not symmetric, the corresponding weighted graphs are directed. It is not appropriate for determining delayed functional connectivity to simply extend the subgraph formalism to directed graphs, since delayed functional connectivity is not independent of the previous definition of functional units. On the other hand, however, for the evaluation of the approach, it could be of interest to investigate the structure of the directed graph and define the functional units from there. A consistent definition should lead to the same results. To analyze delayed functional connectivity we use the concept of hypergraphs where the hypervertices consist of the previously defined functional units and the hyperedges are the multiple directed edges between the voxels of each pair of functional units. The weights of the hyperedges are taken from the delayed correlation matrices. Figure 6.1 illustrates the concept of functional connectivity.

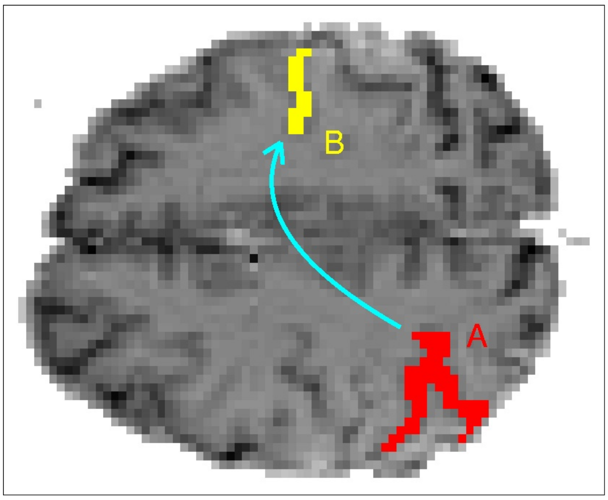


Figure 6.1: Illustration of the concept of functional connectivity. The areas A and B are considered as functional units (to be determined w. r. t. a similarity measure of the voxels e. g. subgraphs of a graph extracted from the undelayed correlation matrix (see text)). A possibly causal relationship between A and B is indicated by an arrow. By possibly causal is meant that the behaviour of area B is similar in a certain sense to the behaviour of area A with a time delay, so that it cannot be precluded that area A exhibits an influence on area B. The aim in this chapter is to both determine functional units and their mutual relationships by defining a time delayed similarity measure between the units.

Outline of the approach:

- Identify functionally connected voxels by the unshifted ($\tau = 0$) cross-correlation of their time courses (section 6.2).
- Use graph theory to cluster functionally connected voxels into cortical areas (section 6.3)
- Use time-delayed ($\tau \neq 0$) cross-correlations and a more general graph theoretical framework (hypergraphs) to identify possibly causal relationships between clusters (section 6.4).

6.2 From correlation matrices to graphs

In this section the approach of constructing graphs from correlation matrices is presented. In the following two subsections we review some definitions of correlation and from graph theory before deriving the correlation graph concept in more detail. Part of this work was published in [36, 37]

6.2.1 Definitions of correlation

Correlation for continuous stochastic processes and for deterministic functions

FMRI data can be considered as a stochastic process, i. e. as a function whose values are random variables [114]. A *stochastic process* $X(t)$ with a time variable t is determined by its probability densities

$$p_{t_1,\dots,t_n}(x_1,\dots,x_n) = p(\{X(t_1)=x_1,\dots,X(t_n)=x_n\}) \tag{6.1}$$

($n \in \mathbb{N},\ x_i \in \mathbb{R}$). The probability densities must satisfy a *symmetry condition* and a *compatibility condition*, according to which the probability densities must be invariant under simultaneous permutations of the indices of t and x and, respectively, integrating the probability density $p_{t_1,\dots,t_n}(x_1,\dots,x_n)$ over the variable x_n must result in the probability density $p_{t_1,\dots,t_{n-1}}(x_1,\dots,x_{n-1})$ (cf. [114]).

Since for a given experiment we only have the experimental data to make inferences about the probability densities, it is reasonable as a first approach to assume that the underlying stochastic processes are stationary. *Stationarity in the strong sense* means that the finite-dimensional probability densities $p_{t_1,\dots,t_n}$ remain the same if the whole group of points $t_1,\dots,t_n$ is shifted along the time axis by a fixed amount τ [114]. A weaker assumption is *stationarity in the wide sense*, which requires that the mean of the process is constant over time and the covariance is a function of the time difference τ only, which amounts to stationarity up to second order. Thereby the mean and the covariance of stationary processes are defined as

$$\begin{aligned} \mu(t) &= \int x\, p_t(x)\, dx && (6.2)\\ \Gamma(t_1,t_2) &= \int\!\!\int (x_1(t_1)-\mu(t_1))(x_2(t_2)-\mu(t_2))\, p_{s,t}(x_1,x_2)\, dx_1 dx_2 && (6.3) \end{aligned}$$

The correlation function of a stochastic process is defined as [94]

$$\rho(t_1,t_2) = \frac{\Gamma(t_1,t_2)}{\sqrt{\sigma^2(t_1)\sigma^2(t_2)}} \tag{6.4}$$

where

$$\sigma^2(t) = \int (x(t)-\mu_x(t))^2\, p_t(x)\, dx \tag{6.5}$$

is the variance. In the sequel we assume wide sense stationarity, and hence the correlation function reads

$$\rho(\tau) = \frac{\Gamma(\tau)}{\sigma^2} \tag{6.6}$$

where τ is a time difference. For $p_{s,t}(x_1,x_2)=\delta(\tau)$ and $\mu(t)=0$, Equation 6.3 is equivalent to the correlation of two deterministic functions $g, h \in L^2$, which is defined as [95]

$$C_{gh}(\tau) = N_{gh}\int g(t)h(t+\tau)dt \tag{6.7}$$

where $N_{fg} = \left(\int g^2(t)dt \int h^2(t)dt \right)^{-\frac{1}{2}}$ is a normalization factor which makes $C_{gg}(0) = 1$. The correlation theorem [95] states that the Fourier transform of the correlation $C_{gh}(\tau)$ is equal to multiplying the Fourier transform of one function with the complex conjugate of the Fourier transform of the other, thus

$$C_{gh}(\tau) \propto \int G(f)\, \bar{H}(f)\, e^{-if\tau}\, df \tag{6.8}$$

where $G(f)$ and $H(f)$ are the Fourier transforms of $g(t)$ and $h(t)$, respectively, and $\bar{H}(f)$ is the complex conjugate of $H(f)$. Obviously from Equation 6.8 follows immediately the Wiener-Khinchin theorem, which states that the autocorrelation of a function and the squared modulus of its Fourier transform act as a Fourier pair [95].

Correlation for discrete and finite samples of stochastic processes

Since fMRI data is discrete in time and for a given experiment we have only a finite number of measurements the definition of correlation has to be adapted accordingly. The discreteness can be handled by substituting the integral by a summation and the probability density by a discrete probability. The finiteness of the data however requires some considerations about normalization.

Let $\mathbf{x} = (x_1, \ldots, x_k)^T$ and $\mathbf{y} = (y_1, \ldots, y_k)^T$ be the signal time courses of two voxels. The delayed sample correlation is determined by

$$\rho_{XY}(\tau) = \frac{N}{k-\tau}\Big(\langle \mathbf{x}(\tau), \mathbf{y}(0) \rangle - (k-\tau)M \Big) \tag{6.9}$$

where N is a normalization factor yet to be defined, M is given in Equation 6.12. and the delayed and undelayed time courses $\mathbf{x}(\tau)$ and $\mathbf{y}(0)$, respectively, are given by (cf. Figure 6.2)

$$\mathbf{x}(\tau) = (x_1, \ldots, x_{k-\tau})^T \in \mathbb{R}^{k-\tau} \tag{6.10}$$

$$\mathbf{y}(0) = (y_\tau, \ldots, y_k)^T \in \mathbb{R}^{k-\tau} \tag{6.11}$$

unless Equation 6.9 is determined by using the Fourier transform. In this case the vectors $\mathbf{x}$ and $\mathbf{y}$ are to be supplemented by zero padding to vectors in $\mathbb{R}^{2k}$ since otherwise the use of periodic boundary conditions in the context of the Fourier transform leads to artifacts. The term M is the product of the mean values of $\mathbf{x}(\tau)$ and $\mathbf{y}(0)$:

$$M = \frac{1}{(k-\tau)^2} \sum_{i=1}^{k-\tau} x_i \sum_{i=1}^{k} y_{i+\tau} \tag{6.12}$$

For simplicity in the following we assume that the mean values of $\mathbf{x}(\tau)$ and $\mathbf{y}(0)$ are zero and hence we disregard M. The delayed sample correlation in Equation 6.9 is computed only from $k-\tau$ elements of $\mathbf{x}(\tau)$ and $\mathbf{y}(0)$ as is illustrated in Figure 6.2.

As for the normalization factor N there are three more or less natural choices:

- $N = \|\mathbf{x}(0)\| \|\mathbf{y}(0)\|$

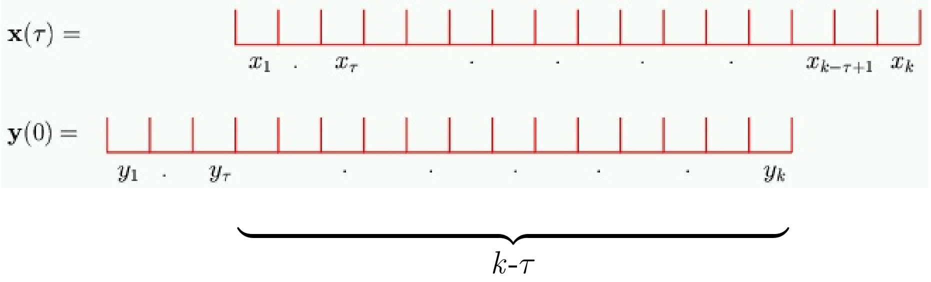


Figure 6.2: Illustration of the shifted and unshifted time courses $\mathbf{x}(\tau)$ and $\mathbf{y}(0)$, respectively, and the $k - \tau$ elements in the common subspace $\mathbb{R}^{k-\tau}$, from which the sample correlation in Equation 6.9 is computed.

- $N = \|\mathbf{x}(\tau)\| \|\mathbf{y}(0)\|$
- $N = (k - \tau)^{\frac{1}{2}} \|\mathbf{x}(\tau)\| \|\mathbf{y}(0)\|$

The first choice of N makes Equation 6.9 equivalent, up to centering, to a discretized version of Equation 6.7. For the second choice of N Equation 6.9 is the sample correlation of two vectors in $\mathbb{R}^{k-\tau}$ (note that $\mathbf{y}(0) \in \mathbb{R}^k$ for the first choice of N wheras for the second and third $\mathbf{y}(0) \in \mathbb{R}^{k-\tau}$). For $\tau = 0$ the second choice of the normalization is equivalent to the first. The third choice of N takes into account that the variance of the sample correlation for i. i. d. gaussian distributed finite samples scales with the square root of the sample size and hence this choice of N aims at supressing the variations due to small sample size. In the following for small values of τ we will use the second choice for N as it has the convenient geometrical interpretation as the cosine of the angle between two vectors in $\mathbb{R}^{k-\tau}$. When considering the dependence on τ we will use the correlation theorem, i. e. the discretized Fourier transform, in the case of which the vectors are in $\mathbb{R}^{2k}$ and the first normalization is employed.

6.2.2 Definitions from graph theory

The temporal correlation matrix $\mathbf{R}(\tau)$ along with a threshold θ_τ can be used to relate fMR images to undirected ($\tau = 0$) and directed ($\tau \neq 0$) graphs. In this section some basic graph theoretical definitions are reviewed which will be needed later on.

Definition (Graph) [29]
A *graph* is a pair $G = (V, E)$ of sets satisfying $E \subseteq V \times V$. The elements of V are the *vertices* of the graph G, the elements of E are its *edges*. The number of vertices of the graph G is its *order* denoted by $ord(S)$, the number of edges is its *size* denoted by $sz(G)$. A graph G is *undirected* if $(v_i, v_j) \in E \iff (v_j, v_i) \in E$ for each pair of vertices v_i and v_j. The *adjacency matrix* $\mathbf{A} = \left(a_{ij}\right)_{m \times m}$ of a graph G having order m is defined by

$$a_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{6.13}$$

A graph G is *weighted*, if each edge $e \in E$ is associated to a weight $w \in \mathbb{R}$. This case can be considered as a graph G for which the elements of the adjacency matrix $\mathbf{A}$ assume more than two values. For an undirected graph the matrix $\mathbf{A}$ is symmetric.

In the sequel we will define functional units by subgraphs of the graph extracted from the fMRI data.

Definition: (Subgraph, maximum spanning tree)
A *subgraph* $S = (V_S, E_S)$ of a graph $G = (V, E)$ is a subset $S \subseteq G$ which is a graph itself. A *spanning subgraph* is a subgraph for which $V_S = V$. A *tree* is a graph containing no cycles, i. e. no edge sequences $\{(v_{i_1}, v_{i_2}), (v_{i_2}, v_{i_3}), \ldots, (v_{i_{n-1}}, v_{i_n})\}$ with $i_n = i_1$. Trees of order n have size $n-1$. A set of disjoint trees is a *forest*. A tree of order n is a *spanning tree*. A *maximum spanning tree* is defined for a weighted graph as the spanning tree having the highest possible weights of all spanning trees of the graph.

An important concept in the context of subgraphs is their graph theoretical connectivity. Functional units should be subgraphs with a high graph theoretical connectivity but since the edge weights are noisy, the subgraph definition should not be too sensitive to noise. Connectivity in a graph theoretical context is defined as follows [15, 29, 64]

Definition: (Connectivity)
A *walk* W in a graph G is an alternating sequence of vertices and edges, say $[v_{i_1}, (v_{i_1}, v_{i_2}), v_{i_2}, \ldots, (v_{i_{n-1}}, v_{i_n}), v_{i_n}]$. A walk is called a *path* if all its vertices are distinct and a *trail*, if all its edges are distinct. A graph is *connected* if there is a walk between every pair of its vertices. A graph G is *k-vertex-connected*, if at least k vertices must be deleted to render the graph disconnected, and it is *l-edge-connected*, if at least l edges must be deleted to make it disconnected. The *vertex-connectivity* and *edge-connectivity* of a graph G are denoted by $\kappa(G)$ and $\lambda(G)$, respectively. If $\kappa(G) = 1$ the graph is said to have a *cut vertex*, if $\lambda(G) = 1$ the graph has a *bridge*. In these cases there exist a vertex or an edge, respectively, the deletion of which leads to a disconnected graph. The number of edges of a given vertex is called the *degree* of the vertex. The *minimum degree* of the graph is denoted by $\delta(G)$. Vertex-connectivity, edge-connectivity and minimum degree satisfy the following relation

$$\kappa(G) \leq \lambda(G) \leq \delta(G) \tag{6.14}$$

Intuitively it is clear that the vertex-connectivity must be less or equal than the edge-connectivity, since with every vertex deleted at least one edge is deleted as well. Figure 6.3 shows a graph for which the inequality 6.14 strictly holds.

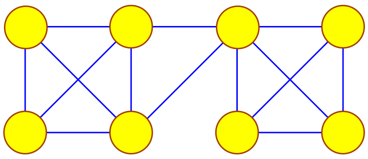


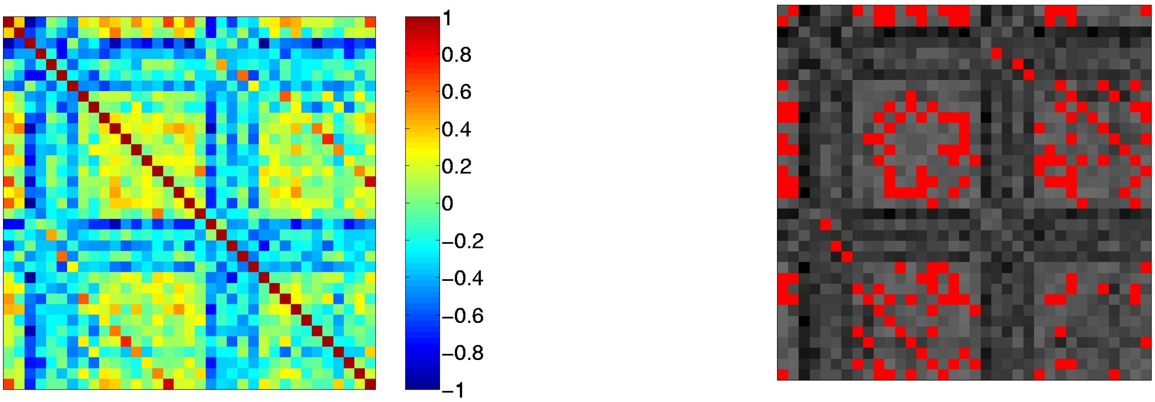
Figure 6.3: Example of a graph for which $\kappa(G) < \lambda(G) < \delta(G)$. $\kappa(G) = 1$, $\lambda(G) = 2$, $\delta(G) = 3$.

6.2.3 Correlation graphs

Using Equation 6.9 along with the second normalization the undelayed correlation matrix for an fMRI experiment where m is the number of voxels and k is the number of images reads

$$\mathbf{R}(0) = \left(\frac{\langle \mathbf{x_i}(0), \mathbf{x_j}(0) \rangle}{\|\mathbf{x_i}(0)\| \|\mathbf{x_j}(0)\|} \right)_{m \times m} \tag{6.15}$$

with $\mathbf{x_i}(0) = (x_{i1}, \ldots, x_{ik})^T \in \mathbb{R}^k$ being the centered time course of voxel i. Obviously $\mathbf{R}(0)$ is symmetric and its diagonal elements equal unity.



(a) (b)

Figure 6.4: Illustration of the graph extraction leading to the graph $G(\theta)$. (a) Simplified example of an unshifted correlation matrix $\mathbf{R}(0)$. It is symmetric and has unit diagonal elements. (b) The same matrix as in (a) but with the elements exceeding θ marked in red and the diagonal elements set to zero as they do not contribute to the graph structure of $G(\theta)$.

(c)

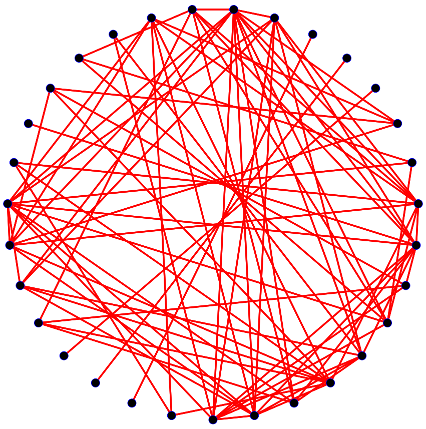


Figure 6.4: (continued) (c) Graph $G(\theta)$ extracted from the matrix $\mathbf{R}(0)$ in Figure 6.4(b) by taking only the red elements of $\mathbf{R}(0)$ as edges together with the associated vertices. The structure of $G(\theta)$ is to be analyzed to define the functional units.

The correlation matrix $\mathbf{R}(0)$ can be considered as the continuous adjacency matrix of a weighted graph the vertices of which correspond to the voxels. In the following we will use the terms voxels and vertices interchangeably, also, when referring to the correlation between two voxels it is understood that is meant the temporal cross-correlation of their time courses.

Instead of investigating the whole weighted graph we use a threshold $\theta \in (-1, 1)$ to extract an unweighted graph $G(\theta)$ from $\mathbf{R}(0)$. Figure 6.4 illustrates this process. The graph $G(\theta)$ has the vertex and edge sets

$$\begin{aligned} V &= \{i \mid i \in \{1, \ldots, m\} \;\wedge\; \exists j \neq i : (i, j) \in E\} && (6.16) \\ E &= \{(i, j) \mid r_{ij}(0) \geq \theta\} && (6.17) \end{aligned}$$

where m is the number of voxels in the image and the second constraint for the elements in V discards isolated vertices. Strictly speaking each voxel is connected to itself since the diagonal elements of $\mathbf{R}(0)$ equal unity and hence always exceed the threshold $\theta \in (-1, 1)$, but this self-connection is of interest only in the delayed case $\mathbf{R}(\tau \neq 0)$, where the diagonal is not always unity. In the undelayed case a voxel connected only to itself is discarded, so that the unweighted correlation graphs $G(\theta)$ do not contain isolated vertices.

To cluster the voxels into functional units we investigate the structure of $G(\theta)$ and determine subgraphs of it according to criteria suitable for the definition of functional units. But before that we discuss methods of reducing the order of $G(\theta)$ by defining a submatrix $\mathbf{R}(0)$ from which the graph $G(\theta)$ is then extracted.

6.2.4 Graph reduction

The reason for not using all voxels is that for small values of the threshold θ the computation of some subgraphs becomes untractable in terms of computation time and that with all voxels included some of the subgraphs are not of interest, anyway. In this paragraph we analyze general properties of $G(\theta)$ in dependence of the threshold θ to extract voxels of interest thereby reducing the size of the correlation matrix $\mathbf{R}(0)$.

Threshold dependencies of the following properties are fast to compute and suited to indicate the coarse structure of $G(\theta)$:

1. The order $ord(G(\theta))$
2. The size $sz(G(\theta))$ relative to the maximum size of a graph of the same order (i. e. normalized by $\binom{ord(G(\theta))}{2}$)
3. The degree distribution

By contrasting the graph structure of $G(\theta)$ against the structure of graphs that are random in a sense to be defined below, a value θ_0 is determined and a submatrix $\mathbf{R}_{\theta_0}(0)$ of $\mathbf{R}(0)$ extracted, which contains only the rows and columns corresponding to the voxels in $G(\theta_0)$. Graph extraction for the reduced graph is then performed as described in subsection 6.2.3. For $\theta \geq \theta_0$ the reduced extracted graphs $G(\theta)$ are equal to the unreduced, only for $\theta < \theta_0$

they differ, because no further voxels can occur with decreasing θ. Each of the contrast graph definitions below leaves invariant different properties of the matrices from which $G(\theta)$ is defined. Graphs based on the following matrices were investigated to find an appropriate contrast:

1. Surrogate data from a random permutation of all elements in the data matrix $\mathbf{X}$. This leaves the distribution of the values in $\mathbf{X}$ constant but changes the values in the correlation matrix $\mathbf{R}(0)$.
2. Surrogate data from a random permutation of the elements in each *row* of the data matrix $\mathbf{X}$. This in addition leaves the distribution of the values in each row of $\mathbf{X}$ invariant and otherwise has the same properties as the previous contrast.
3. Surrogate data from permutation of the elements in each *column* of the data matrix $\mathbf{X}$. This leaves the distribution of the values of each column of $\mathbf{X}$ invariant and otherwise has the same properties as the first contrast.
4. Random permutation of the entries in the correlation matrix $\mathbf{R}(0)$. This leaves the distribution of the values in $\mathbf{R}(0)$ invariant and corresponds to a random graph as it is usually defined in random graph theory. However, it is not ensured that the resulting graphs are correlation graphs, i. e. that permuting the entries of $\mathbf{R}(0)$ results again in an correlation matrix.
5. Restriction of the permutations of the entries of $\mathbf{R}(0)$ in 4 to those that result again in a correlation matrix, i. e. contrasting against correlation graphs.

It turns out that the properties of 1-3 are essentially equal so that in this paragraph we show only the properties of contrast 1. However, easy to compute and suited to determine a value θ_0 for graph reduction seems to be only the graph from a random permutation of the elements of $\mathbf{R}(0)$, i. e. contrast 4. This graph will be called random graph in this paragraph, if no other specifications of random graphs are given. Further the graph from 1 will be called the surrogate graph and the graph from the original correlation matrix $\mathbf{R}(0)$ will be referred to as data graph. The data underlying the graphs in this section are from experiment 1.

In contrast 5 we aimed at taking into account the special structure of our graphs as “correlation graphs” due to their extraction from correlation matrices. Therefore contrast 5 would be of particular interest, however, to estimate the proportion of correlation graphs from the number of random graphs with a given weight distribution is cumbersome. A short sketch of the steps to determine the correlation graphs from a given distribution of correlation values is presented in Appendix C. Also it can be expected that if the distribution of correlation values is broad and includes high correlations, the fraction of the set of all graphs from permuting the elements in $\mathbf{R}(0)$ of correlation graphs may be very small such that it has to be discussed whether the occurence of a correlation graph of this type may be considered as significant by itself.

Figure 6.5 shows in dependence of θ the graph order of the data graph, the surrogate graph, and the random graph relative to their maximum possible order. The graph from

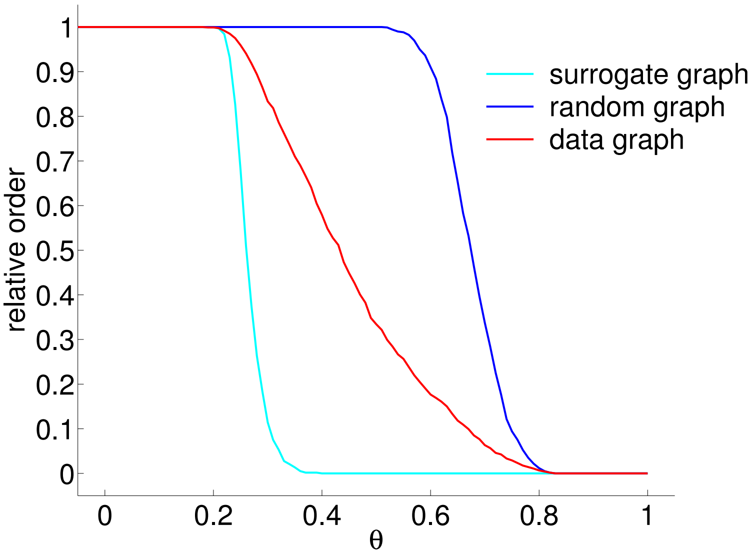


Figure 6.5: Order relative to the maximum possible graph order in dependence of θ. Cyan curve: surrogate data from a random permutation of all elements in $\mathbf{X}$. Equal results were found for permuting the rows and columns of $\mathbf{X}$ seperately. Blue curve: random graph from a random permutation of the elements of the correlation matrix. Red curve: Extraction graph $G(\theta)$ from the correlation matrix $\mathbf{R}(0)$ of the data.

surrogate data exhibits a relatively sharp threshold for values $\theta \in (0.2, 0.4)$ from all vertices to almost none. This is probably due to the presence of only low correlations in the correlation matrix as it is expected for surrogate data. The random graph exhibits a more steep descent than the data graph from all vertices present to none, but for higher values $\theta \in (0.5, 0.81)$ than the surrogate graph. A steep threshold from all vertices to none is also what is expected for random graphs [98]. The surrogate graph hence seems to behave more like a random graph which is also indicated by the further properties investigated below. This could be interpreted as a hint that graphs from correlation matrices with low values may most often be correlation graphs and the distinction between random graphs and correlation graphs for low value correlation matrices may not be of interest.

Note that the size of the data graph equals the size of the random graph for every value of θ, since the size of an extracted graph depend only on the values in the correlation matrix $\mathbf{R}(0)$, not on their configuration in the matrix. The fact that the order of the data graph decays more slowly than the order of the random graph indicates that the former is more dense than the latter. This can also be seen in Figure 6.6, where the size of the graphs relative to their maximum size for a given θ in dependence of θ is shown. The relative size is given by

$$sz_{rel}(G(\theta)) = \frac{sz(G(\theta))}{\binom{ord(G(\theta))}{2}} \tag{6.18}$$

The maximum of the ratio $\frac{ord(R(\theta))}{ord(G(\theta))}$ between the order of the random graph $R(\theta)$ and the

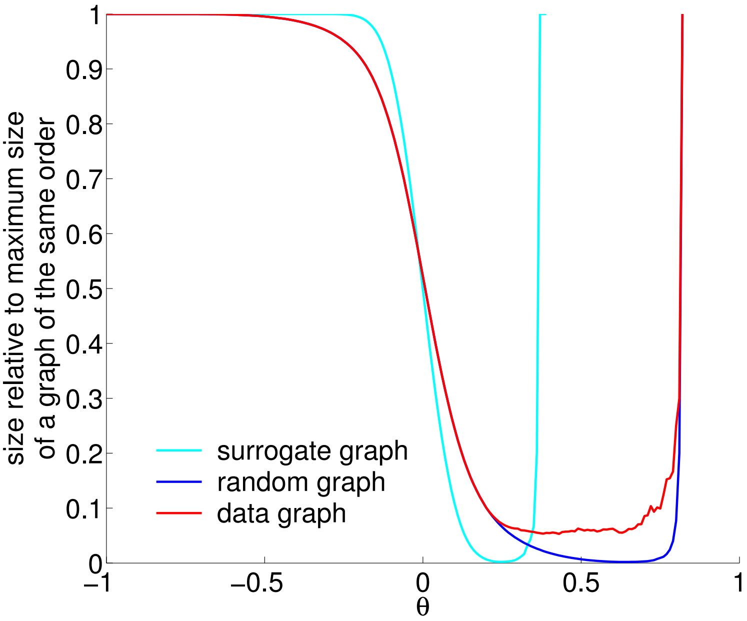
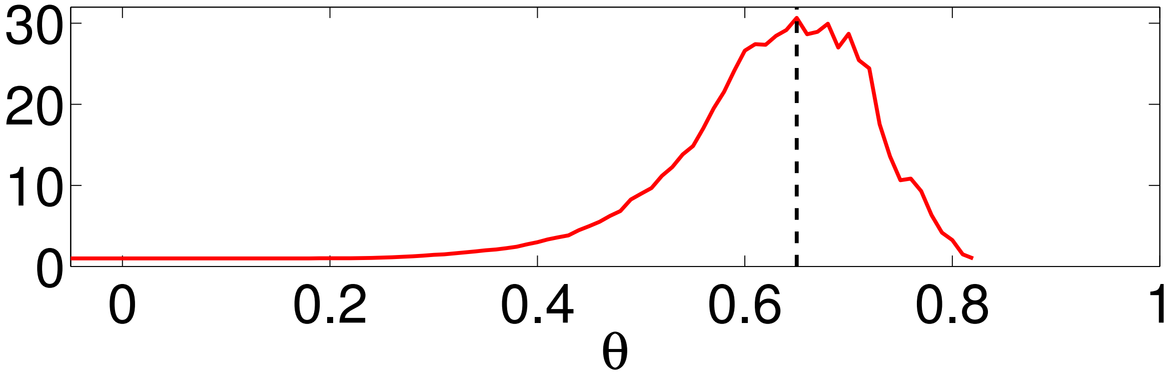


Figure 6.6: Size relative to the maximum size of a graph of the same order, cf. Equation 6.18. Not that here the range of θ is $[-1, 1]$. Cyan curve: Surrogate graph. Blue curve: Random graph. Red curve: Data graph. The high values at the tails occur because there the graphs consist only of very few vertices and edges and hence are complete or close to complete. Note that the difference between the data graph and the random graph occurs only in the range $\theta \in (0.25, 0.8)$.

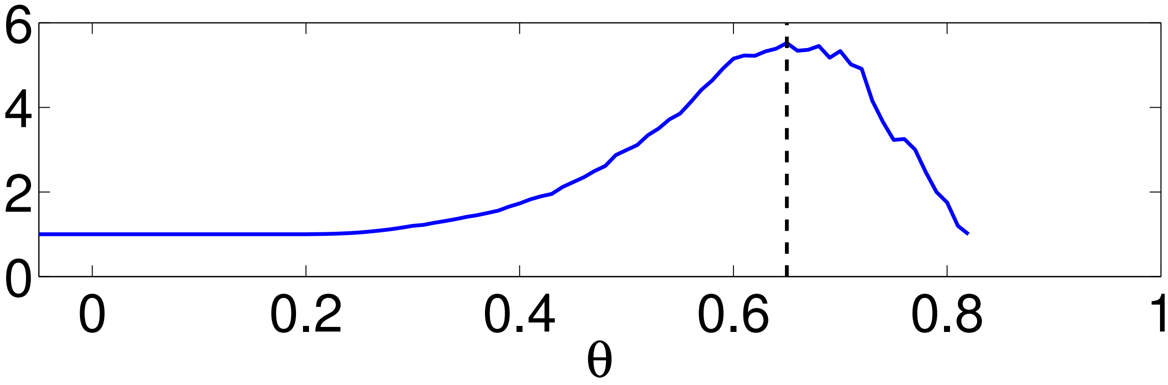
order of the data graph $G(\theta)$ can be used to define θ_0, since it indicates in a sense the distance of the data graph from the random graph, without of course formally being a metric. Figure 6.7 shows the ratio of the orders and the ratio of the relative sizes of the random graph and the data graph. The maximum occurs at a value of $\theta_0 = 0.65$ which was then used for graph reduction. The ratio of the relative sizes gives a more pronounced maximum than the ratio of the relative order at the same value θ_0. This is because both curves are equivalent, since the sizes of the graphs are equal and the ratio of their relative sizes hence given by

$$\frac{sz_{rel}(R(\theta))}{sz_{rel}(G(\theta))} = \frac{\binom{ord(R(\theta))}{2}}{\binom{ord(G(\theta))}{2}} \tag{6.19}$$

Finally we investigated the degree distribution of the various graphs in dependence of θ. They are shown in Figure 6.8. The degree distribution of the data graph is significantly different from the degree distribution of the random graph and of the surrogate graph. The degree distribution of the surrogate graph exhibits a steeper threshold for decreasing θ than the degree distribution of the random graph, but apart from this is similar. For both graph types the vertices for a given value of θ have essentially the same degree. The degree distribution of the data graph in contrast indicates a diverse structure of the graph, where vertices with a high degree coexist with vertices having a low degree for each value of θ except for values near zero. The latter is also confirmed when computing the difference and the ratio of the degree distributions of the data graph and the random



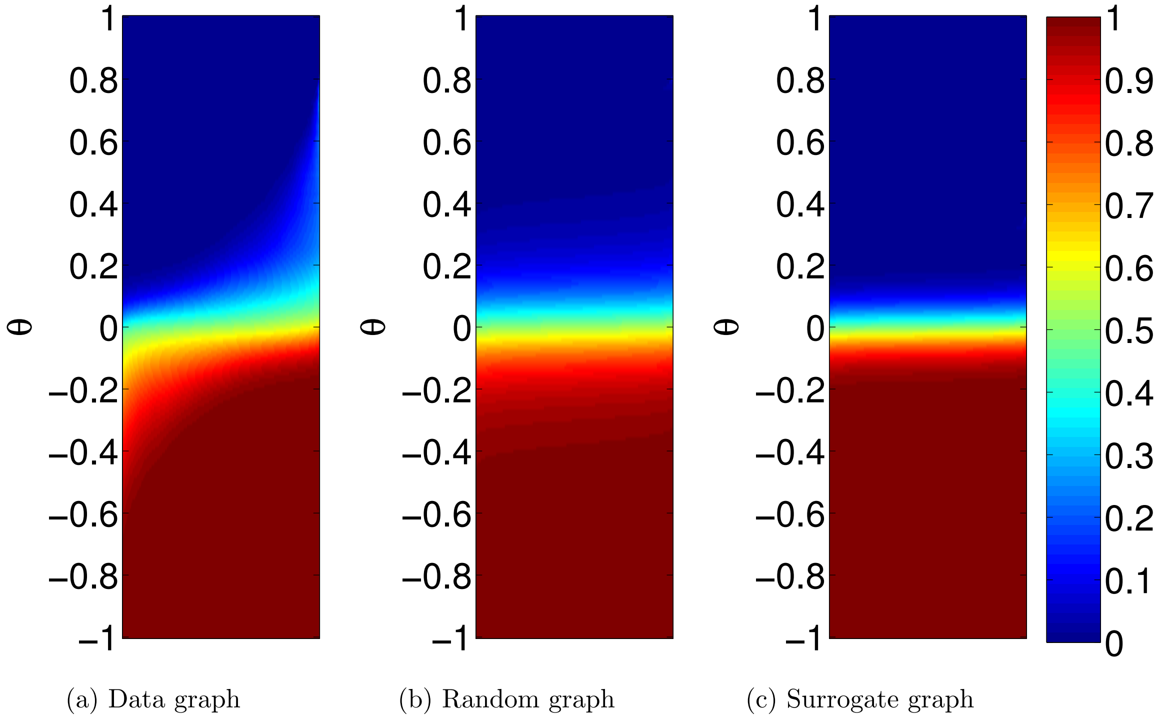
(a) Ratio of the relative sizes



(b) Ratio of the orders

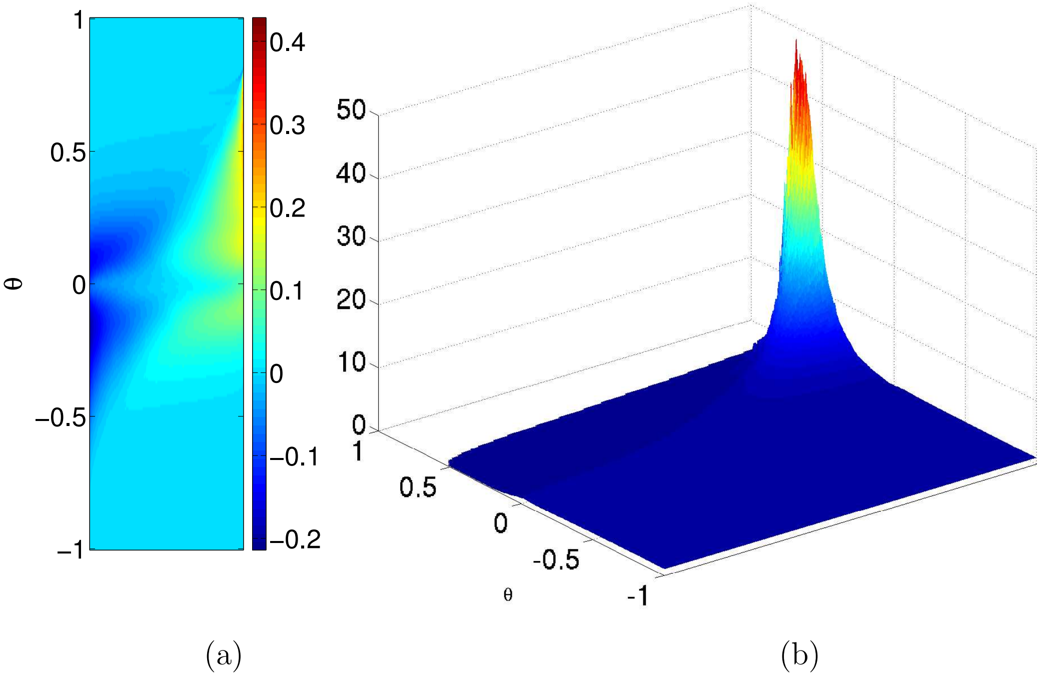
Figure 6.7: Ratio of the relative sizes and ratio of the orders of the data graph and the random graph. The black dotted lines indicate the location of the maximum at a value $\theta_0 = 0.65$ which was then used for graph reduction.

graph as is shown in Figure 6.9. Furthermore the maximum of the ratio of the degree distributions occurs at a value of $\theta = 0.68$ which is similar to the value $\theta = 0.65$ which was found for the maximum of the order ratio in Figure 6.7. The latter value was used for graph reduction since the measure of relative order is more robust than the relative degree distribution. However to investigate all properties such as order, size and degree distribution is useful to get information about the coarse structure of the graph $G(\theta)$. Some subgraphs of the reduced graph shall now be determined to analyze the structure in more detail and as a basis for defining functionally connected units.



(a) Data graph (b) Random graph (c) Surrogate graph

Figure 6.8: Degree distribution of $G(\theta)$ in dependence of θ for a data graph, a random graph, and a surrogate graph. Each vertical line in color code denotes the relative degree of a vertex, i. e. the proportion of vertices in $G(\theta)$ to which the vertex is connected, in dependence of θ. The vertices are ordered appropriately to show the structure of the degree distribution. The voxel indices on the abscissa are not labeled, since the relation to individual voxels is not of interest here. Obviously the degree distribution of the data graph significantly differs from the degree distribution of the random graph and the surrogate graph. In the latter two each vertex has essentially the same degree for a given value of θ whereas in the data graph for values of about $|\theta| \geq 0.1$ there is a strong diversity of degrees leading to a more structured graph.



(a) (b)

Figure 6.9: (a) Difference and (b) ratio of the relative degree distributions showed in Figure 6.8 of the data graph and the random graph. Near $\theta = 0$ the difference map shows that the degree distributions of the data graph are almost the same as for the random graph. The maximum of the ratio occurs at a value of $\theta = 0.68$ which is similar to the value of $\theta = 0.65$ at which the maximum ratio shown in Figure 6.7 between the orders of the respective graphs occured and which was used for graph reduction.

6.3 Subgraphs of $G(\theta)$ as functional units

6.3.1 Connectivity components and cliques

The concepts of connectivity components and cliques form the two extremes of subgraph definitions in that connectivity components have the minimum and cliques the maximum possible connectivity w. r. t. the number of voxels. See Figure 6.10 for an illustration.

Definition: (Connectivity components and cliques)
A *connectivity component* S of a graph $G = (V, E)$ is a maximal connected subgraph, where the term maximal is defined in the sense of inclusion. A *clique* C of a graph G is a *complete*, i. e. fully connected, subgraph of G. A *maximal* clique is a clique which is not contained in any other clique. A *maximum* clique is a maximal clique of largest size.

Connectivity components S form a partition of a graph G. Cliques C are possibly overlapping subgraphs of the connectivity components S, i. e.

$$\bigcup C = \dot{\bigcup} S = G \tag{6.20}$$

The edge-connectivity l of a clique C is related to the order of C by $l(C) = \operatorname{ord}(C) - 1$. Note however that the edge-connectivity for a small clique can be lower than the edge-connectivity of a larger graph which is not fully connected.

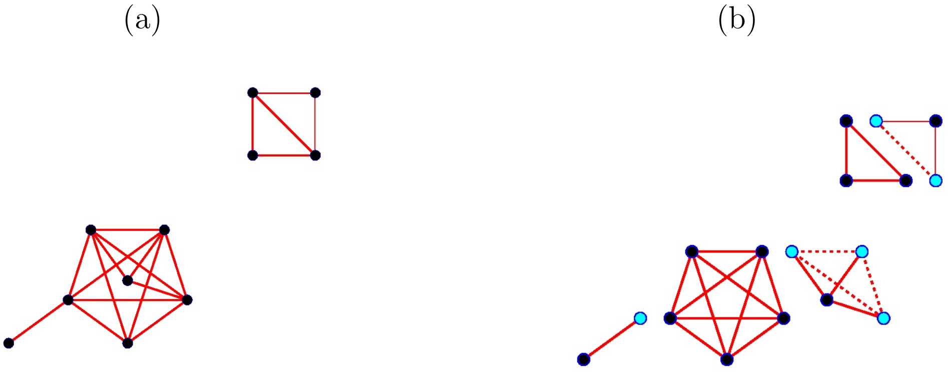


Figure 6.10: Illustration of the concept of connectivity components and cliques. (a) Graph with two connectivity components. (b) The cliques of the graph in (a) drawn seperately for better visibility. Light blue vertices and dashed lines indicate vertices and edges that are shared by more than one clique.

Some of the mutually exclusive properties of cliques and connectivity components are listed in the table below, the desirable of them in terms of suitability for defining functional units are written in italic style:

connectivity components	cliques
minimal connectivity	*all-to-all connectivity*
structure loss by a giant component emerging for lower values of θ	*revealing structure details*
robust to noise	noise prone
disjoint	largely intersecting

The properties of connectivity components and cliques make intermediate cluster definitions desirable, which preferably combine the advantages of both. The term intermediate hereby refers to the connectivity constraints which are to be chosen between all-to-all and plain connectedness. A tuning parameter between the two extremes could be the edge-connectivity defined in subsection 6.2.2. However as for the cliques where for a given graph $G(\theta)$ we extracted cliques of all orders, extracting all clusters with a given l-connectivity is an NP-hard problem. In subsection 6.3.2 we analyze threshold dependencies of cliques and connectivity components. It turns out that the data graph is dominated by two large connectivity components exhibiting a complex edge structure as can be seen from so called overlap matrices of cliques. The stability of the two large connectivity components along with the complexity of their clique structure, makes the definition of an intermediate subgraph cumbersome, if not unnecessary. For the time being we therefore stick with cliques and connectivity components.

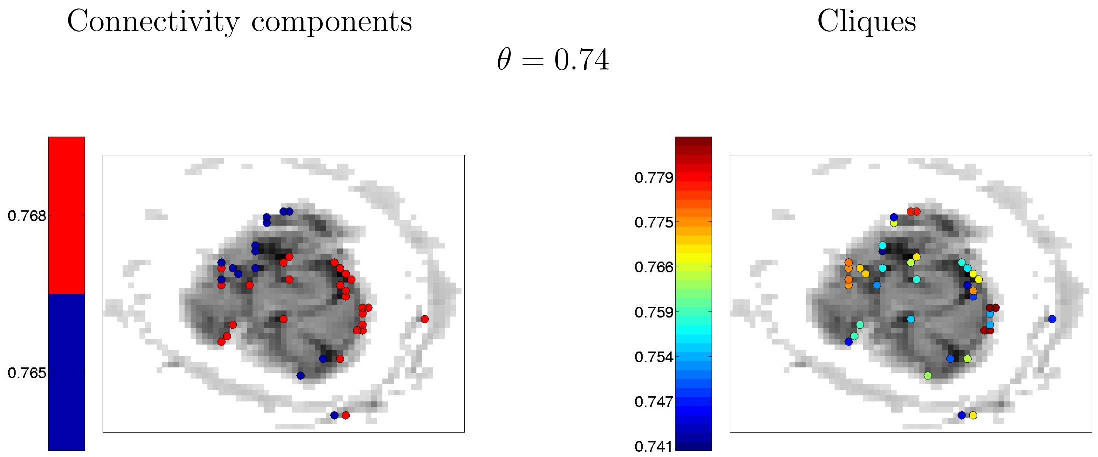
Connectivity components and cliques for data from experiment 1

Figure 6.11 and Figure 6.12 show the connectivity components and cliques of the reduced graph for data from experiment 1 for the values $\theta = 0.74$ and $\theta = 0.65$, respectively, together with the time courses and the overlap matrix of the cliques. Each element in the overlap matrix indicates the fraction of vertices a clique has in common with another clique relative to the number of its own vertices. In the paragraph **overlap matrices, p. 87** we consider overlap matrices of cliques in more detail. The time courses were taken as the mean of the time courses of the voxels which represent the vertices of the subgraphs. Prior to taking the mean the time courses were centered and linearly detrended to better show their structure.

A pronounced linear trend was present for most of the cliques and connectivity components. Apart therefrom mainly two interesting temporal features occured. One is the stimulus related time course which is found at $\theta = 0.65$ mostly for small subgraphs, such as the first three connectivity components in Figure 6.12(c) which correspond to trivial cliques, i. e. cliques of order 2, but also in the fifth component, which has order 6. The stimulus related time courses of the subgraphs exhibit phase shifts with respect to each other indicating differences in stimulus processing. Further some of the subgraphs such as the first two connectivity components only partly participate at the stimulus time course. This could be due to head movements or again due to processing differences. There are more stimulus related cliques than stimulus related connectivity components reflecting the coarser structure of connectivity components in comparison to cliques and the fact that cliques are subgraphs of connectivity components. When considering various thresholds it appears that stimulus related subgraphs are found up to a value of about $\theta = 0.71$.

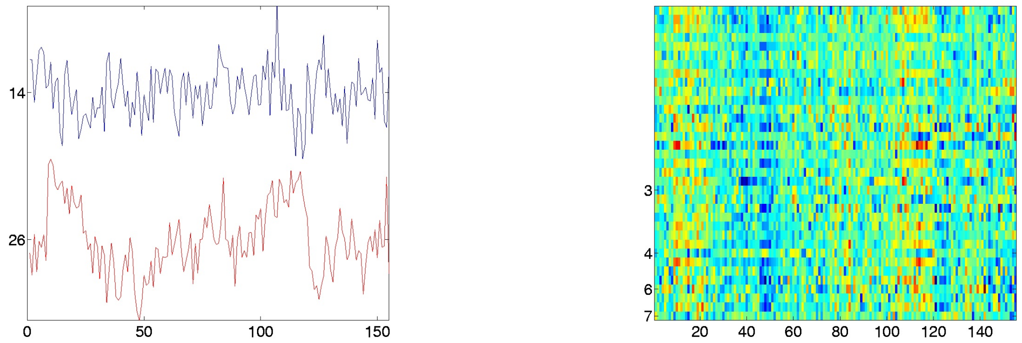
The other prominent temporal process is a slower apparently periodic time course the maxima of which roughly coincide with the maximum of the first and the fifth stimulus cycle as it occurs in the time course of the second connectivity component for $\theta = 0.74$ in Figure 6.11(c) and in the time courses of numerous cliques for $\theta = 0.74$ as well as for $\theta = 0.65$, shown in in Figure 6.11(d) and Figure 6.12(d), respectively. The overlap matrices in Figure 6.11(e) and in Figure 6.12(e) show a cluster of comparatively large large overlapping cliques that exhibit this time course. Considering the time courses of the voxels in these cliques in more detail reveals that the prominent temporal process is mostly due to a few voxels contiguous to the stimulus-related region. The appearance of this process could hence be a signature of head movements of brain pulsations, but also of an only partly stimulus related process.

In Figure 6.13 stimulus related connectivity components and cliques for $\theta = 0.65$ are shown separately. For periodic stimuli as in the experiment considered, the relatedness to the stimulus could be found by using the Fourier transform of the time courses and extracting those with the highest contributions at the stimulus cycle frequency as it was done in chapter 4 and chapter 5. However, to include also time courses only partly related to the stimulus for this figure the stimulus related subgraphs were chosen by hand. For $\theta = 0.65$ only trivial cliques, i. e. cliques of order 2, are related to the stimulus. Partial relatedness occurs e. g. in the fifth and the eighth clique time course in Figure 6.13(d). This partial relatedness however is not similar to the second prominent temporal process described above, indicating different underlying processes. This is confirmed by the observation that the voxels exhibiting a time course according to the second prominent temporal process belong to a different connectivity component than the stimulus related voxels (for $\theta > 0.41$, where there is more than one component). In Figure 6.12(b) some single voxels occur that are located far away from the voxels of the connectivity components in Figure 6.12(a). They are part of only partially stimulus related cliques. More detailed analysis yields that their time courses are not related to the stimulus.



(a) Connectivity components overlaid the mean image data from experiment 1. The colors correspond to the mean correlation among the voxels of the connectivity components.

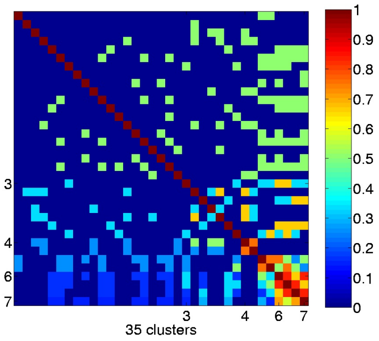
(b) Cliques overlaid the mean image data. Note that due to overlap not all vertices of the cliques are visible in their proper color. Small cliques are drawn on top of large cliques.



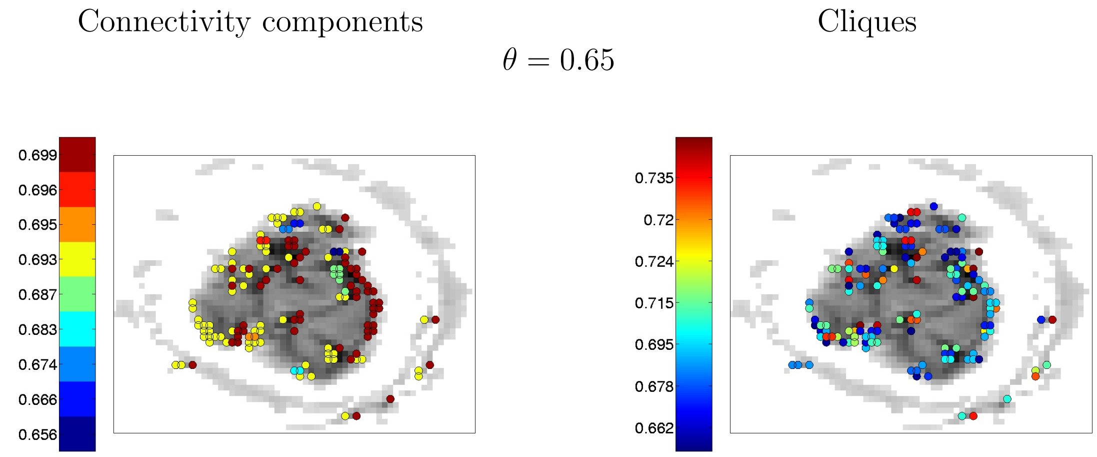
(c) Mean time courses of the connectivity components in the corresponding color. The numbers on the ordinate indicate the order of the connectivity component.

(d) Color coded mean time courses of the cliques ordered from small to large cliques. The numbers on the ordinate indicate the order of the cliques for the time courses thereunder.

Figure 6.11: Connectivity components and cliques for $\theta = 0.74$ together with their mean time courses and the overlap matrix. (e) The (asymmetric) overlap matrix of the cliques. Most of the cliques are trivial and only partly overlapping except for a cluster of strongly overlapping larger cliques. For a further discussion of the results shown, see text.

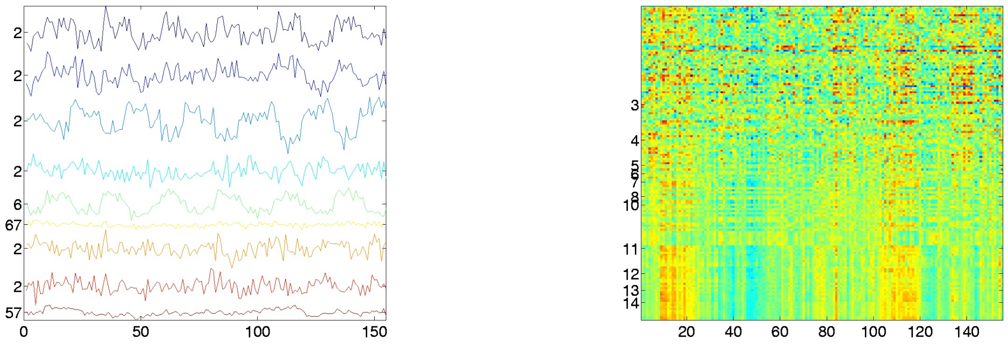


(e) Overlap matrix of the cliques.



(a) Connectivity components overlaid the mean image data from experiment 1.

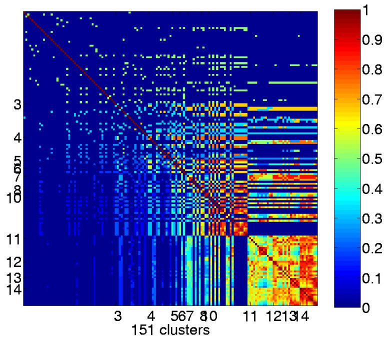
(b) Cliques overlaid the mean image data.



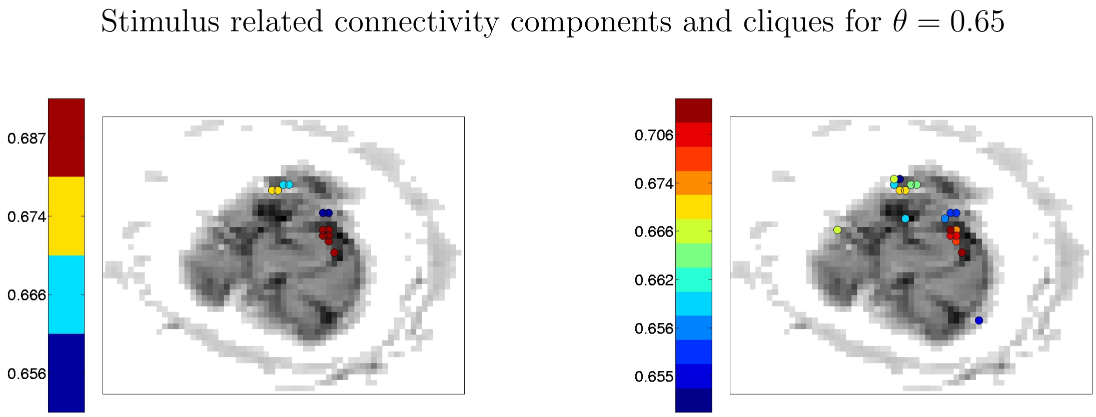
(c) Mean time courses of the connectivity components in the corresponding color.

(d) Color coded mean time courses of the cliques.

Figure 6.12: Connectivity components and cliques for $\theta = 0.65$ together with their mean time courses and the overlap matrix. (e) The (asymmetric) overlap matrix of the cliques. For a further discussion of the results shown, see text.

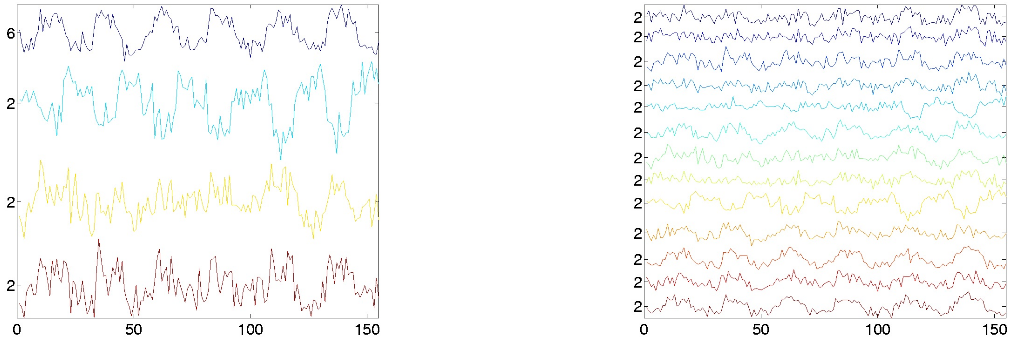


(e) Overlap matrix of the cliques.


Stimulus related connectivity components and cliques for $\theta = 0.65$
0.687
0.674
0.666
0.656
0.706
0.674
0.666
0.662
0.656
0.655

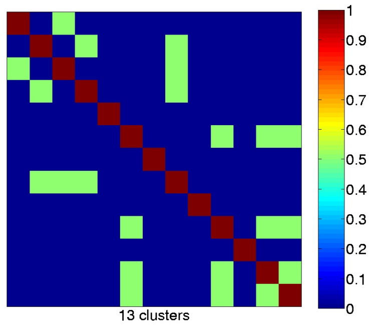
(a) Stimulus related connectivity components.

(b) Stimulus related cliques.



(c) Mean time courses of the connectivity components.

(d) Mean time courses of the cliques.



(e) Overlap matrix of the cliques.

Figure 6.13: Stimulus related connectivity components and cliques for $\theta = 0.65$. The stimlus related connectivity components are located at areas which are also found by correlation analysis, PCA or ICA. The cliques include voxels lying far away from these areas and the individual time course of which does not show stimulus relatedness (cf. text). Only cliques 5, 7 and 9 do not overlap with any other stimulus related cliques.

6.3.2 Revealing subgraph structures from threshold dependencies

In a similar way as in the previous section threshold dependencies of various graph properties were used to reduce the graph, also threshold dependencies of subgraph properties further reveal their structure. In this paragraph we determine the threshold dependencies of properties of data subgraphs and contrast them to the threshold dependencies of subgraphs from random graphs. The properties discussed in this paragraph are the following:

1. Number of subgraphs.
2. Mean order of the subgraphs.
3. Mean size of the subgraphs.
4. Mean correlation of the subgraphs, where the intra correlation of a subgraph is given by the mean intra correlation of its voxels including only the edges that are present for the given threshold.
5. Histogram of the subgraph order.
6. Tree- and pseudotree structures relating the subgraphs for different values of θ.
7. Overlap matrices of cliques.
8. Spanning trees

Figure 6.14 shows the properties 1-4 of connectivity components and cliques from data of experiment 1 in comparison to the respective subgraphs from random graphs.

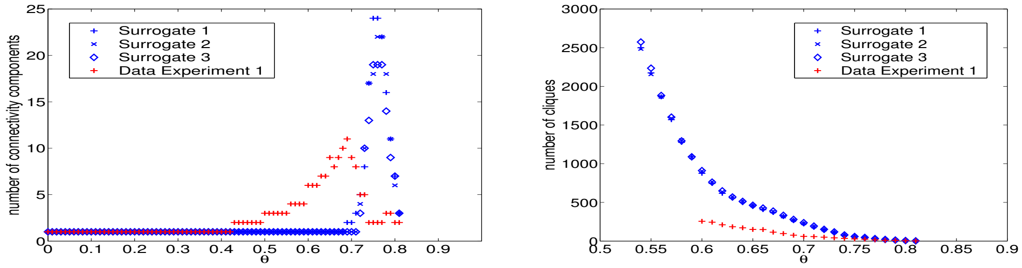
Number, mean order, size, and correlation of the connectivity components

For high values of θ there are many though small connectivity components from the random graphs whereas for the data graphs the number of connectivity components reaches its maximum for lower values of θ (cf. Figure 6.14(a),(c)). Comparing Figure 6.14(a) and Figure 6.14(c) it seems that the curves in Figure 6.14(c) are strongly related to the derivative of the curves in Figure 6.14(a).

The lower mean size of the data graph connectivity components in Figure 6.14(e) is consistent with their higher number in Figure 6.14(a). For the random graph connectivity components the mean size is given by the total number of edges for a given value of θ and hence is equal to the integrated histogram of edges in the graphs.

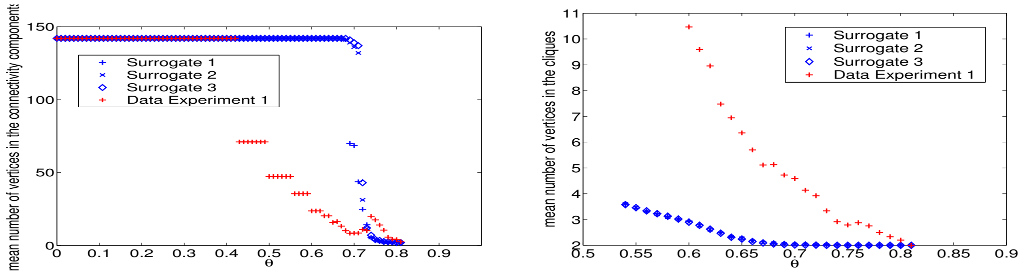
The seeming equality of the mean size in the connectivity components of random graphs and data graph for high values of θ is an effect of the low resolution of the plot in this area, the true differences there being of the order of 10, where the data graph connectivity components have a larger mean size than the random graph connectivity components.

Number of connectivity components and cliques



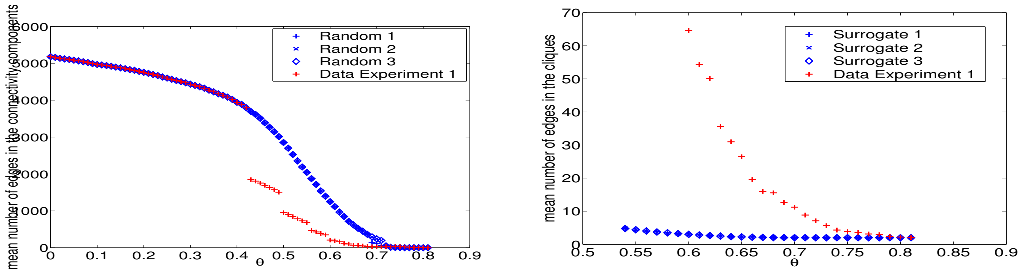
(a) (b)

Mean order of connectivity components and cliques



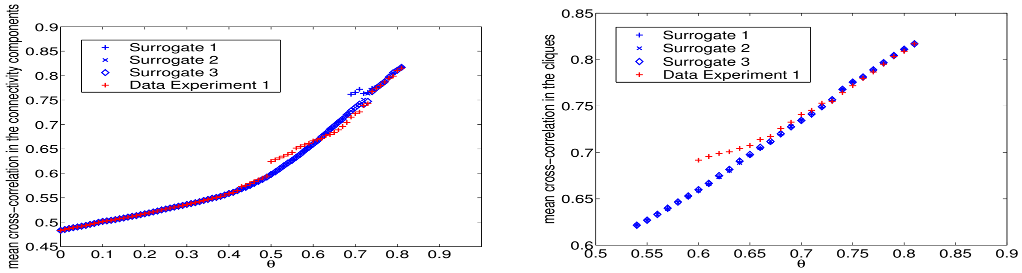
(c) (d)

Mean size of connectivity components and cliques



(e) (f)

Mean correlation within the connectivity components and cliques



(g) (h)

Figure 6.14: Properties of connectivity components and cliques of the data graph and random graphs. The red curves denote the properties of the data graph and the blue curves with various markers show the results for three realizations of random graphs. Note that the range of θ is different for the connectivity components and cliques since computing cliques for smaller values of θ was computationally untractable. The properties are discussed in the text.

The mean correlation of the data connectivity components in Figure 6.14(g) is slightly higher than that of the random graph connectivity components except for the interval $\theta \in (0.6, 0.7)$. This might be due to the fact that in this region the connectivity components are rather small and hence a low xcorrelation value has more effect on the mean correlation of the connectivity component.

Number, mean order, size, and correlation of the cliques

The number of random graph cliques is higher than that of data graph cliques and they are of lower order than the data graph cliques as is expected for random graphs. The order of the random graph cliques does not exceed the trivial order of 2 for $\theta > 0.6$. The mean number of vertices n_v and the mean number n_e of edges for cliques are related through $n_v \propto \sum_{c \in cliques} ord(c)$ and $n_e \propto \sum_{c \in cliques} \binom{ord(c)}{2}$. The mean correlation of data cliques slightly exceeds the mean correlation of cliques from random graphs for $\theta \leq 0.73$, the differences becoming significantly more pronounced for $\theta \leq 0.68$.

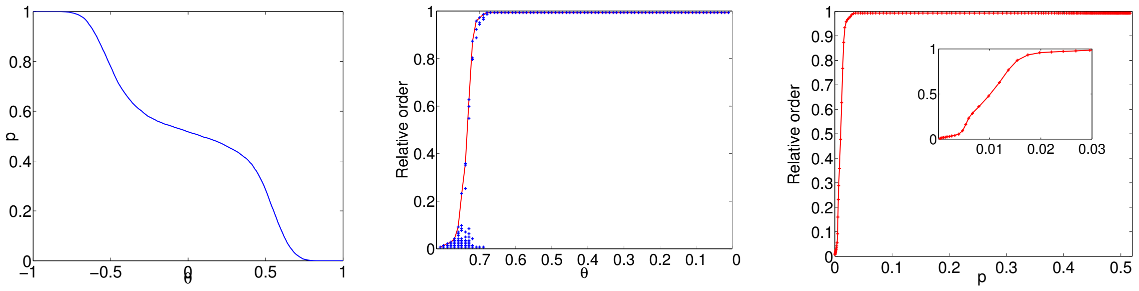
Histogram of the subgraph order

Color coded histograms of the orders of connectivity components and cliques are shown in Figure 6.16. An interesting feature is exhibited by the connectivity components of the data graph. For $\theta \geq 0.43$ there are essentially two large connectivity components. One of them is stable in its order for $\theta \in [0.66, 0.43]$, which means that in this range it becomes denser and does not acquire new vertices. The voxels of the two components for $\theta = 0.43$, i. e. just above their merging into one giant component for $\theta \leq 0.42$, are shown in Figure 6.17. The stable component is the one exhibiting the second prominent temporal process referred to in the previous paragraph. The voxels of the two large components are frequently contiguous to each other and are grouped around tissue boundaries. Since at tissue boundaries the signal change due to head movement is expected to be particularly high it could be that this is the main cause for high correlation.

The order of the connectivity components of random graphs exhibit a phase transition at $\theta \approx 0.7$ from small orders to essentially one giant component. This is consistent with random graph theory, where for $c = \frac{p}{n}$ with n being the graph order and p the edge probability there is a component of the order ac with a constant a if $a = 1 - e^{-ac}$. The relation between edge probability in random graph theory and the threshold θ is

$$p = \frac{sz(G(\theta))}{sz(G(-1))} \tag{6.21}$$

where $G(\theta)$ is the graph extracted from the reduced correlation matrix and $G(-1)$ is the complete graph occuring for $\theta = -1$. Figure 6.15 shows the relation 6.21 for the correlation matrix used. The two bumps are due to the reduction of the correlation matrix yielding a bimodal histogram of the correlation values.



(a) Edge probability p vs. threshold θ.

(b) Relative order of random graph connectivity components vs. θ.

(c) Relative order of random graph connectivity components vs. p.

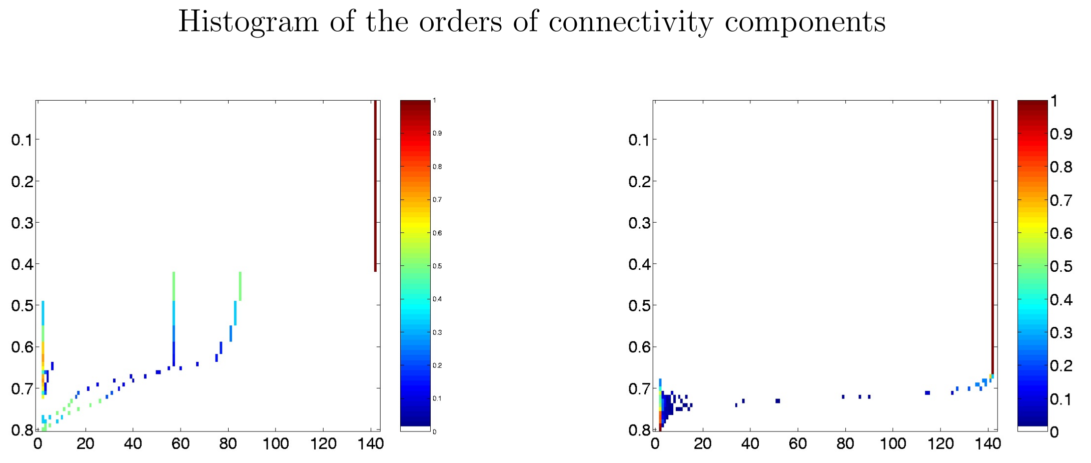
Figure 6.15: Threshold θ, edge probability p, and relative orders of the random graph connectivity components in dependence of θ and p, respectively.
(a) Relation between θ and edge probability p. The values $\theta \leq 0$ are shown for completeness, though they were not used for graph extraction. The bumps are due to graph reduction and the resulting bimodality of the histogram of correlation values in the reduced correlation matrix.
(b) Nonzero histogram values of the orders of connectivity components of random graphs as shown in Figure 6.16(b). The orders are normalized to the maximum order. The red curve represents an interpolation of the values.
(c) The same interpolation curve as in Figure 6.15(b) but in dependence of p instead of θ. For $p \approx 0.0098$ the relative order of the connectivity component is 0.5 and for $p \approx 0.0296$ the relative order is 0.99. This is consistent with results from random graph theory (see text).

The color coded histograms of the clique orders in Figure 6.16(c) and Figure 6.16(d) show that for the data graph at high values of θ the clique orders are mainly trivial, the highest clique order increasing roughly linearly with decreasing θ, approaching more and more a uniform distribution. For random graphs for $\theta > 0.71$ there are only trivial cliques. At $\theta \approx 0.61$ the mode of the order histogram switches from 2 to 3, the maximum clique order does not exceed 4 for the range $\theta \geq 0.6$ considered here.

Trees and pseudotrees

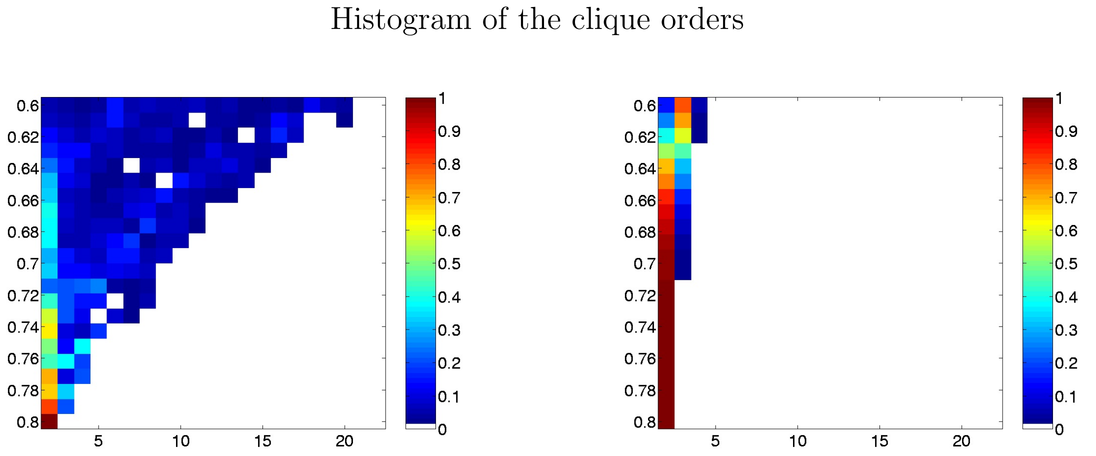
Tree- and pseudotree structures are obtained, when relating the connectivity components and cliques for different values of θ. This is shown in Figure 6.18. A tree structure emerges for connectivity components which obey the inclusion relation

$$S(\theta_i) \subseteq S(\theta_j) \qquad \vee \qquad S(\theta_i) \cap S(\theta_j) = \emptyset \qquad \forall\, \theta_i > \theta_j \tag{6.22}$$


Histogram of the orders of connectivity components

(a) Data connectivity components

(b) Connectivity components from random graphs


Histogram of the clique orders

(c) Data cliques

(d) Cliques from random graphs

Figure 6.16: Color coded histograms of the order of connectivity components and cliques. The histograms are normalized to the total number of subgraphs for a given value of θ such that the color codes the relative frequency of subgraphs of a particular order. The random graph histograms are the result of averaging over three realizations. The properties are discussed in the text.

where $S(\theta_i)$ and $S(\theta_j)$ are connectivity components of the extracted graphs $G(\theta_i)$ and $G(\theta_j)$, respectively. For a clique $C_r(\theta_i)$, $r \in \{1, \ldots, N_{\theta_i}\}$, the inclusion relation is

$$\forall\, \theta_i > \theta_j \;\exists\, s \in \{1, \ldots, N_{\theta_j}\} : \; C_r(\theta_i) \subseteq C_s(\theta_j) \tag{6.23}$$

where N_{θ_j} is the number of cliques of $G(\theta_j)$. Thus a given clique may be contained in more than one clique of an extraction graph of lower threshold. Therefore in general no tree structure can be defined for cliques, however, a pseudotree structure as in Figure 6.18(b) can be employed to visualize the clique structure for various thresholds. Note that because of their frequently occuring overlap the branches of the pseudotree may not be

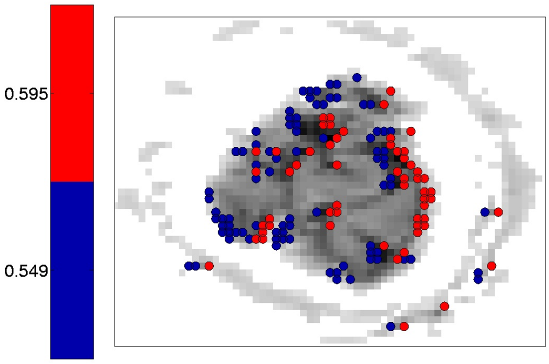


Figure 6.17: Voxels of the two large connectivity components of which the data graph consists for $\theta = 0.43$ before merging into one large connectivity component for $\theta < 0.43$. The order of the blue connectivity component equals 85, the order of the red connectivity component equals 57. The blue connectivity component contains the stimulus related voxels whereas the mean time course of the red connectivity component is dominated by the second prominent temporal process also underlying the red connectivity component in Figure 6.11(a) and (c).

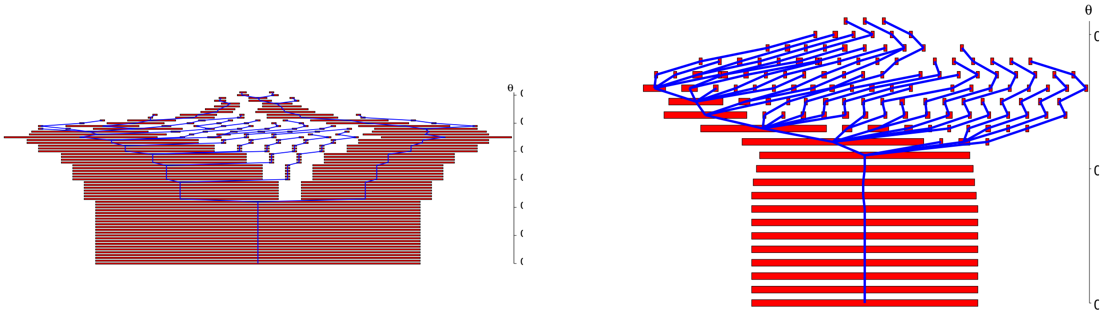
independent (see caption).

Overlap matrices

For a given value of θ overlap matrices of cliques are helpful to give insight to the clique structure. Figure 6.19 shows the overlap matrices for different values of θ, thereby showing the evolution of the clique structure in dependence of θ. The entries of the overlap matrices are of the form $\frac{ord(\,C_i(\theta) \cap C_j(\theta)\,)}{ord(C_i\theta)}$, where $C_i(\theta)$ and $C_j(\theta)$ are the ith and jth clique of $G(\theta)$ ordered by size and by mean intra-correlation. The normalization makes the overlap matrices non-symmetric. Values near one can occur only for large clique orders, their number of vertices sampling the interval $[0, 1]$ more closely. For $\theta \geq 0.66$ the overlap matrices are characterized by a number of overlapping larger cliques and a bunch of small cliques some of which are appended to the larger ones. For $\theta \leq 0.64$ the overlap matrix starts to be dominated by the two large connectivity components present in the graph. The stability of one connectivity component that was seen in Figure 6.16(a) shows up in a cluster of cliques that becomes denser for decreasing θ. The other connectivity component is of larger order and exhibits smaller cliques. Therefore the two connectivity components show up as essentially separated structures in the overlap matrix. Figure 6.20 shows the clique overlap matrices for the two large connectivity components separately. The differences in the edge density of the connectivity components is obvious. Both overlap matrices, however, show a complex structure of more or less intersecting cliques, reflecting the complex edge structure of the underlying graph.

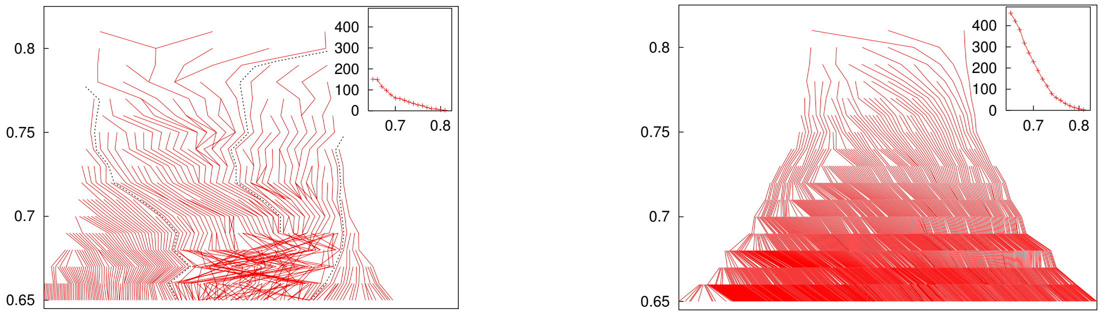
Spanning trees

Spanning trees were computed successively starting from the whole graph at $\theta = 0$ and removing the edge with the lowest weight after each time a maximum spanning tree



(a) Tree of data connectivity components

(b) Tree of random graph connectivity components



(c) Pseudotree of data cliques

(d) Pseudotree of random graph cliques

Figure 6.18: Trees and pseudotrees relating connectivity components and cliques for different values of θ. The width of the horizontal bars at the trees in (a) and (b) indicate the order of the respective component. The tree structure of the data connectivity components reflects the graph structure consisting of essentially two components, whereas the tree of the random graph connectivity components exhibits branches only for very small components. The pseudotrees in (c) and (d) relate cliques for which one clique is included in the other. Since a given clique can be included in multiple cliques of a lower value θ in general no tree structure is obtained. The dotted black lines in the pseudotree of the data cliques indicates the boundaries of the two main components. The pseudotree of the data cliques has more merging branches than the pseudotree of the random graph cliques. Most of the latter are trivial and scattered through the graphs and hence remain relatively stable for varying θ. The random graph pseudotree is thus almost a tree, in contrast to the data graph pseudotree, which is clearly not a tree.

was computed. This leads to a tree structure similar to the trees of the connectivity components, the difference, however is, that the maximum spanning tree structure reflects the edge structure, a stable branch indicating a high number of edges with large weights. Figure 6.21 shows the spanning tree structure of the data graph and a random graph. The ordinate shows both, the iteration number and the corresponding minimum θ which was present in the spanning tree. The random graph spanning tree structure is essentially composed of one large and many small spanning trees whereas the spanning tree structure of the data graph reflects the two large connectivity components which are stable with only a few small components being separated. The left branch of the data tree also reflects the stable connectivity component that does not acquire new vertices for decreasing θ (cf. Figure 6.16(a)).

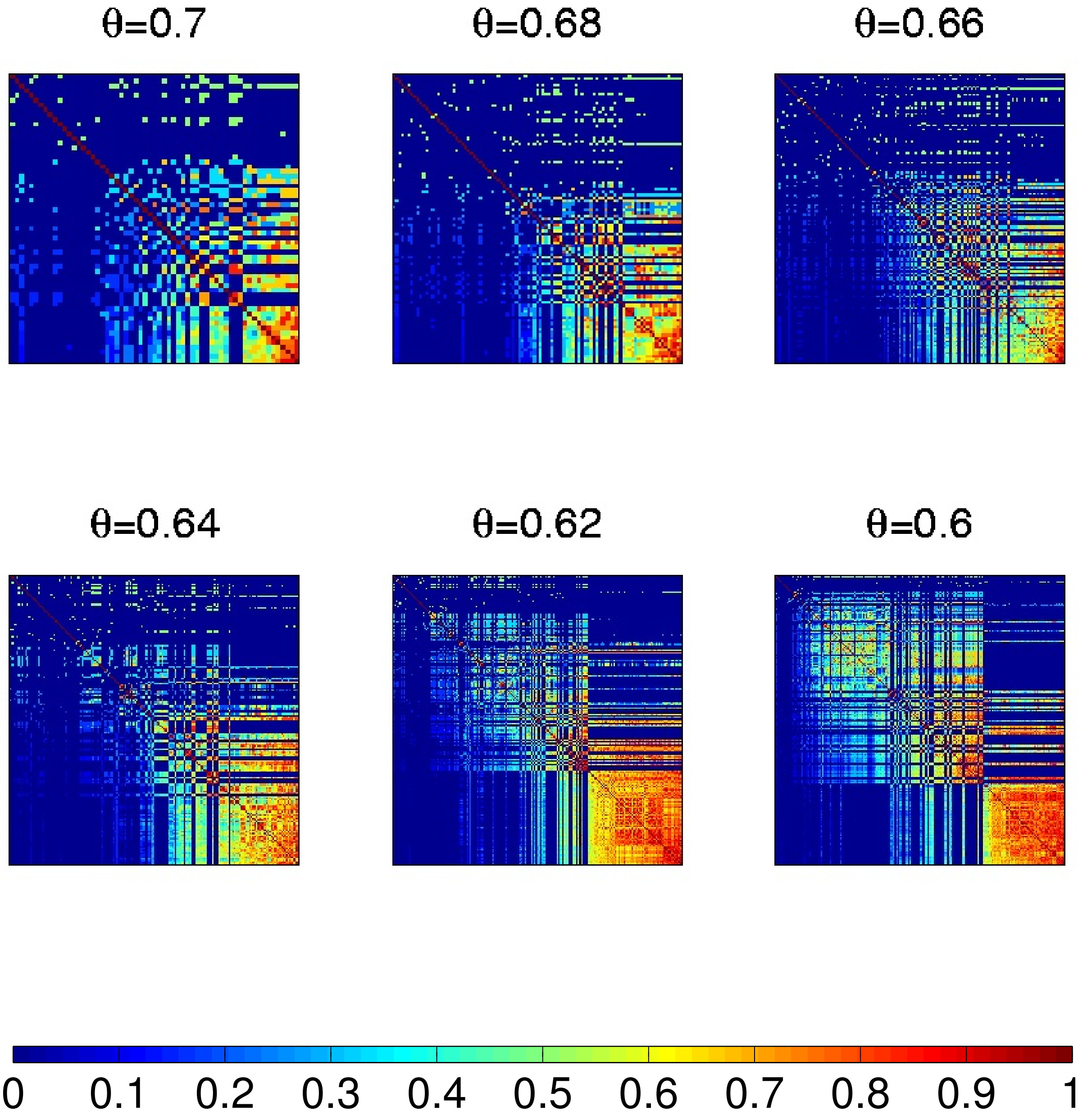


Figure 6.19: Overlap matrices for various values of θ in decreasing order. In each matrix the clique order is increasing from left to right and from top to bottom. Further, cliques having the same order are sorted by their mean intra correlation. A growing cluster of largely overlapping high order cliques is apparent as an increasingly red square on the bottom right of the matrices. Note that the number of cliques in the matrices increases for decreasing θ, such that the overlapping cluster of overlapping may appear smaller although it is larger in absolute size. Appart from the shrinking proportion of small cliques the overlap matrices also exhibit the signature of the two large connectivity components discussed in the text. For decreasing θ a second square appears for cliques of slightly lower order. This second square belongs to the "unstable" large connectivity component which grows for decreasing θ as was concluded also from Figure 6.16(a). Only in the stable connectivity component cliques of the highest order for a given θ occur, whereas the other connectivity component exhibits a greater variety of clique orders, which is shown also in Figure 6.20.

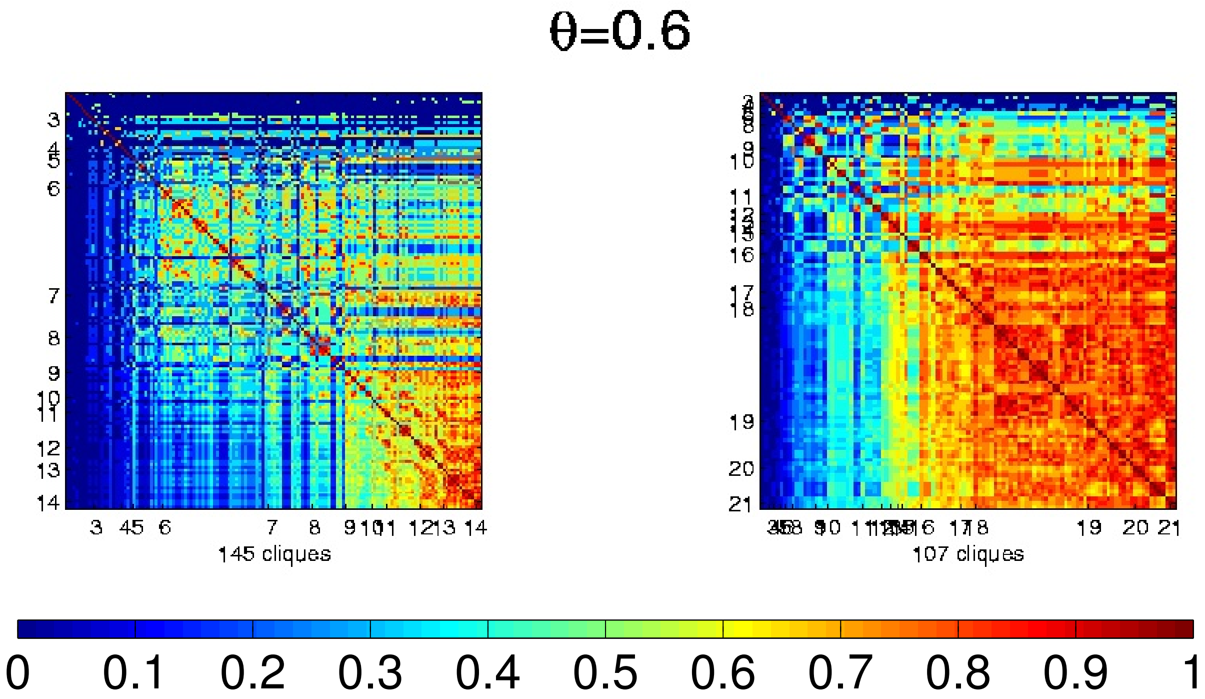
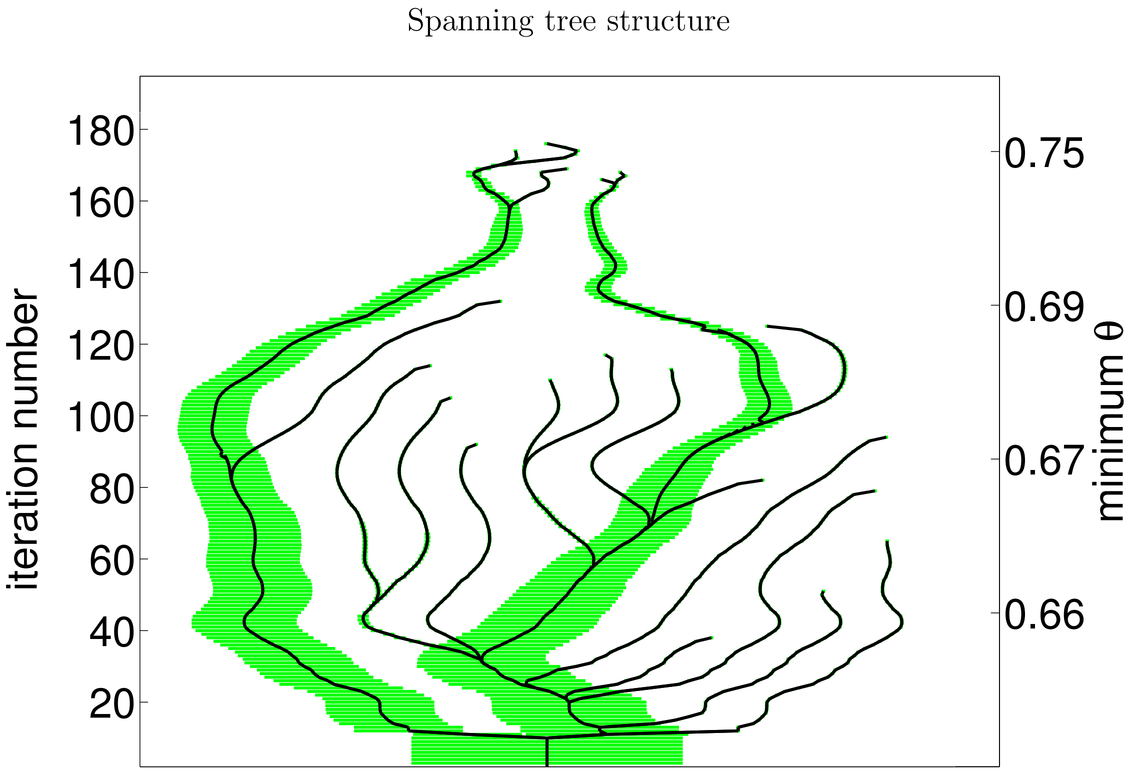
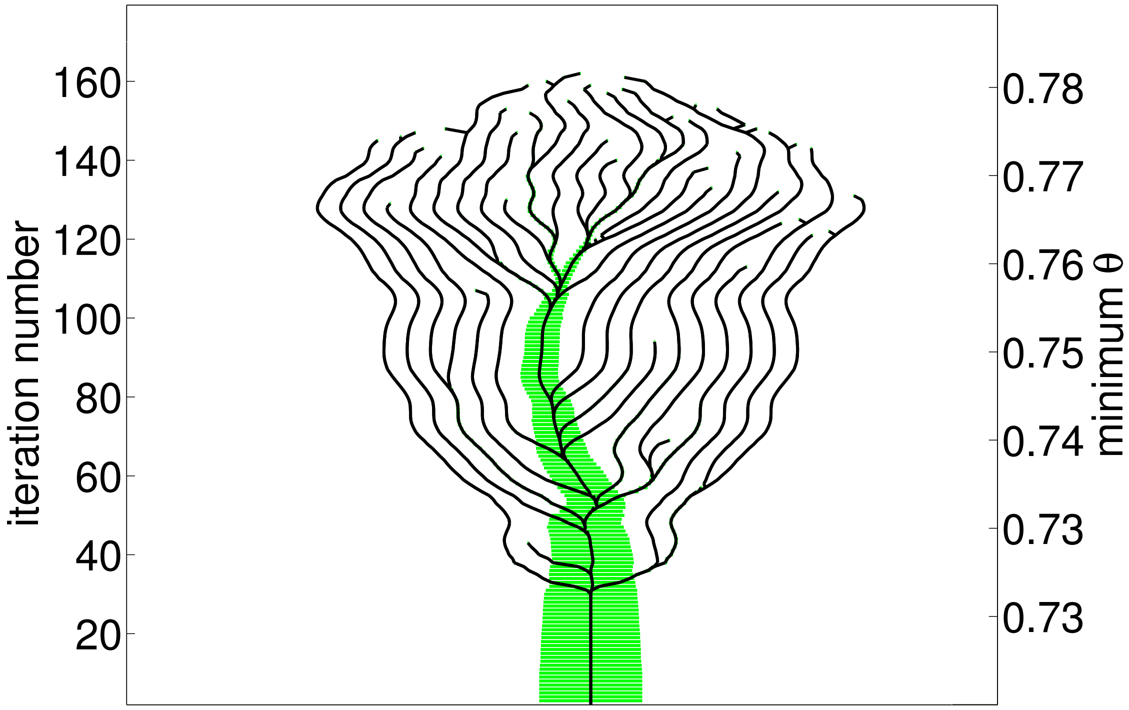


Figure 6.20: Overlap matrices of the cliques for $\theta = 0.6$ in the two large connectivity components. The numbers at the matrices denote the first occurence of the respective clique order in the matrix. The large cluster which was visible also in the overlap matrix of the total graph dominates the structure of the stable connectivity component, whereas the other shows a cluster of larger cliques as well, however less overlapping and of smaller clique orders than the former. Further this connectivity component has a second cluster with cliques of the order of about 6 which is loosely overlapping with the cluster of larger cliques.



(a) Spanning tree structure of the data graph.



(b) Spanning tree structure of the random graph.

Figure 6.21: Spanning tree structure of the data graph and the random graph. The horizontal bars indicate the order of the spanning tree.

6.4 Delayed functional connectivity

In the previous section we derived subgraphs from $G(\theta)$ the underlying voxels of which can be considered as functional units. Assuming the (possibly transient) existence of functional units in the brain, it is of interest to determine their mutual temporal relationships, in particular possibly causal influences of one unit on the other. Causality, however, is hard to prove, therefore usually properties are considered which would appear in this case and are taken as a hint of its presence. A useful property in this context is delayed correlation. If the temporally shifted signal time course of one functional unit has a high correlation with the signal time course of another, it can be conjectured that the two exhibit delayed functional connectivity. This is true in general of course only for a linear interaction between the units, which we will assume henceforth as a first approximation.

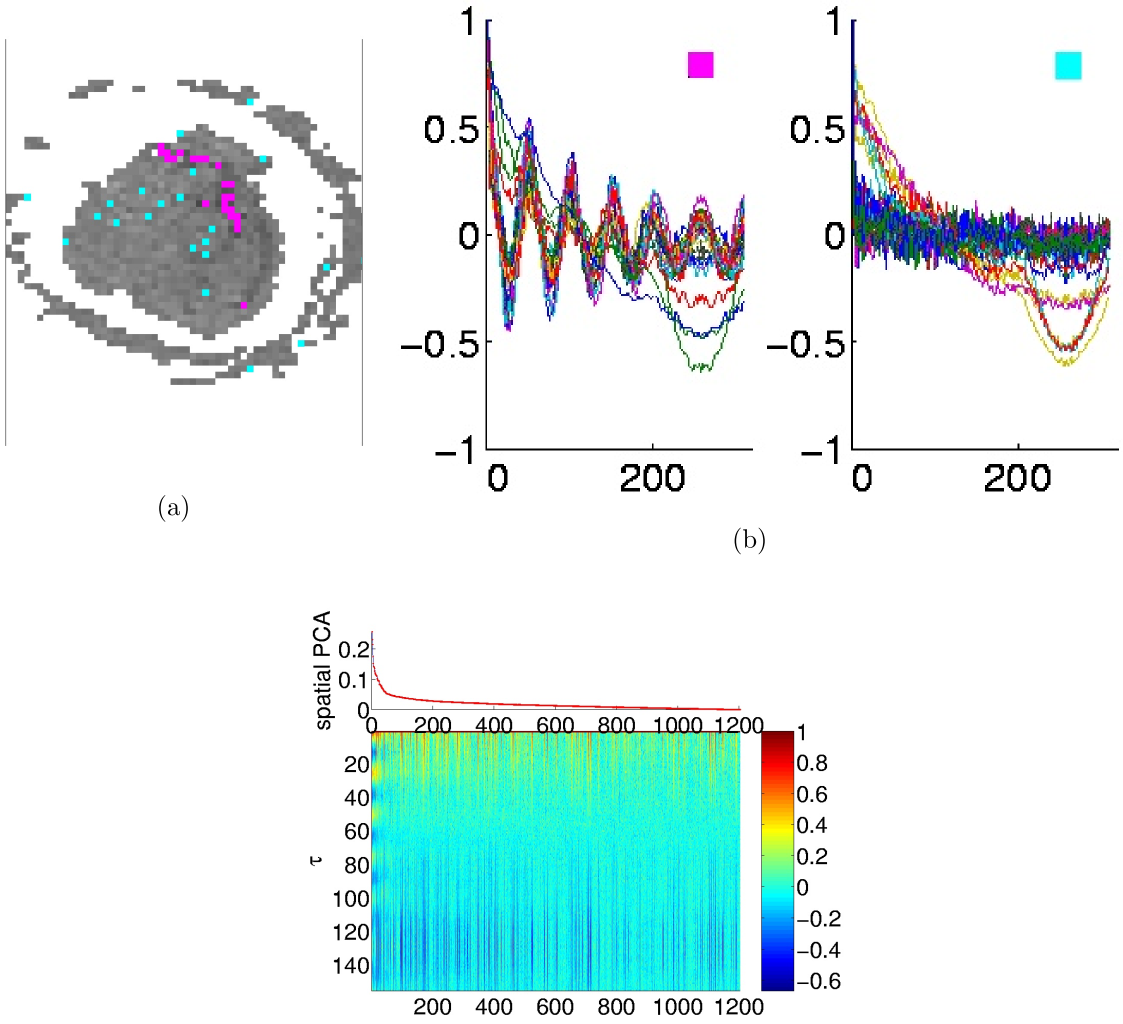
6.4.1 Delayed correlation matrices

Analogously to definition 6.15 for small values of τ we use the definition of the delayed correlation matrix of the data that reads

$$\mathbf{R}(\tau) = \left(\frac{\langle \mathbf{x_i}(\tau), \mathbf{x_j}(0) \rangle}{\|\mathbf{x_i}(\tau)\| \|\mathbf{x_j}(0)\|} \right)_{m \times m} \tag{6.24}$$

with $\mathbf{x_i}(\tau) = (x_{i1}, \ldots, x_{i,k-\tau})^T \in \mathbb{R}^{k-\tau}$ and $\mathbf{x_j}(0) = (x_{i,\tau+1}, \ldots, x_{i,k})^T \in \mathbb{R}^{k-\tau}$ being the centered time course of voxel i temporally shifted by an amount of τ and the unshifted centered time course of voxel j, respectively. Each element in $\mathbf{R}(\tau)$ is thus the sample correlation of two random processes of sample size $k - \tau$ (cf. paragraph on p.65). Obviously $\mathbf{R}(\tau)$ in general is not symmetric. The diagonal is not unity as in the undelayed case, but in most cases decreases in absolute value with increasing delay.

Figure 6.22(a) shows in color code and ordered according to the value of the most stimulus related spatial PC the evolution of the diagonal elements of $\mathbf{R}(\tau)$ in dependence of τ computed by using the correlation theorem. In Figure 6.22(b),(c), some of the autocorrelation curves along with the respective location of the voxels in the brain are depicted. Essentially three classes of curves occur. One is a damped oscillation corresponding to the autocorrelation of the signal time courses of stimulus related voxels, reflecting the periodic nature of the stimulus. The second is a curve peaked at zero where the autocorrelation is trivially one and for values $\tau > 0$ it essentially fluctuates around zero. The third curve exhibit a slow decay and an undershoot for large values of θ. This class of curves correspond to the autocorrelation of a linear signal, the undershoot being a finite size effect which occurs when Fourier transforming a discretized linear function that is padded with zeroes at the end. From Figure 6.22 it can be seen that only a small part of the autocorrelation of the time courses of the data from experiment 1 is stimulus related whereas the large majority falls to about equal parts into the other classes. The voxels with long lasting autocorrelations are most likely located in more or less homogeneous regions whereas the voxels the autocorrelation of which decays immediately probably are located at tissue boundaries.



(a)

(b)

(c)

Figure 6.22: Shifted autocorrelation in of selected voxels in dependence of τ. (a) Selected voxels in magenta are chosen from stimulus related areas wheras voxels in cyan are randomly chosen to display the contrast. (b) Shifted autocorrelation in dependence of τ for the respective voxels sets. The curves are appropriately normalized to suppress fluctuations for large values of τ due to small sample size. (c) Color coded autocorrelation curves for voxels ordered according their absolute values in the most stimulus related spatial PC. See text for futher information.

6.4.2 Hypergraphs

Since for $\tau \neq 0$ the correlation matrix $\mathbf{R}(\tau)$ is in general is not symmetric the equivalent weighted graph is directed. Delayed functional connectivity defined on the basis of correlation is characterized by a high undelayed ($\tau = 0$) correlation within the functional units and a high delayed ($\tau \neq 0$) correlation between the functional units. Thus we need information from both, $\mathbf{R}(0)$ and $\mathbf{R}(\tau \neq 0)$. Therefore instead of using a straightforward generalization of the definition of subgraphs as in the undelayed case we use the concept of hypergraphs. There are several definitions of hypergraphs. We take the following

Definition: (Hypergraph)
A *hypergraph* $H = (S_V, S_E)$ is a family S_V of sets of vertices $v \in V$ and a family $S_E = S_V \times S_V$ of sets of edges $e \in E$. The elements of S_V and S_E are called *hypervertices* and *hyperedges*, respectively.

The relation to the correlation matrices $\mathbf{R}(0)$ and $\mathbf{R}(\tau)$ can be set up in defining $H(\theta_0, \theta_\tau)$ as a hypergraph the hypervertices of which are subgraphs of the graph $G(\theta_0)$ extracted from $\mathbf{R}(0)$ and the hyperedges of which are the directed edges the weights in $\mathbf{R}(\tau)$ of which exceed the threshold θ_τ. E. g. a *clique hypergraph* in this context is a hypergraph the hypervertices of which consist in cliques. An illustration of a clique hypergraph is given in Figure 6.23. The hypergraphs from other subgraphs, such as connectivity components, are defined correspondingly.

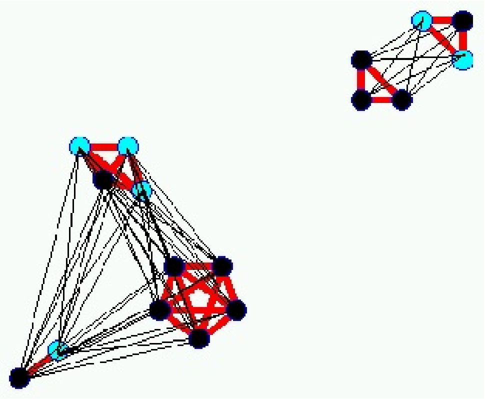


Figure 6.23: Illustration of the hypergraph concept. The cliques drawn in color form the hypervertices and the black lines between two vertices of different cliques the hyperedges.

6.4.3 De-autocorrelation

Figure 6.24 shows the "pathways" correlation between two voxels can be mediated. To preclude that the delayed correlations were essentially due to high autocorrelations we also de-autocorrelated the data. Thereby the autocorrelation of each time course was removed by projecting the shifted time course vector $\mathbf{v}(\tau)$ to an orthogonal of the unshifted time course vector $\mathbf{v}(0)$ by

$$\mathbf{v}_{de}(\tau) = \big(\mathbb{I} - \frac{\mathbf{v}(0)\mathbf{v}(0)^T}{\|\mathbf{v}(0)\|^2}\big)\,\mathbf{v}(\tau) \tag{6.25}$$

where $\mathbf{v}_{de}$ is the de-autocorrelated time course. In terms of Figure 6.24 by de-autocorrelation essentially one pathway is cut constraining the correlation to be mediated directly.

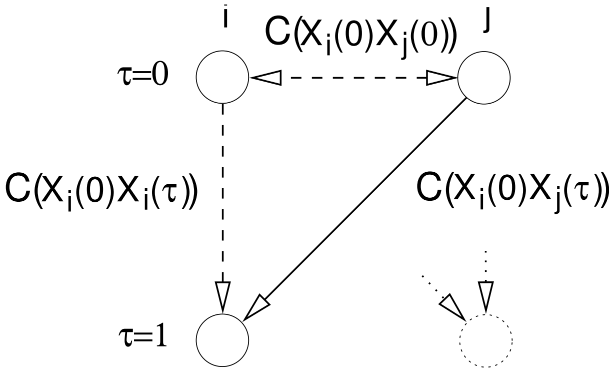
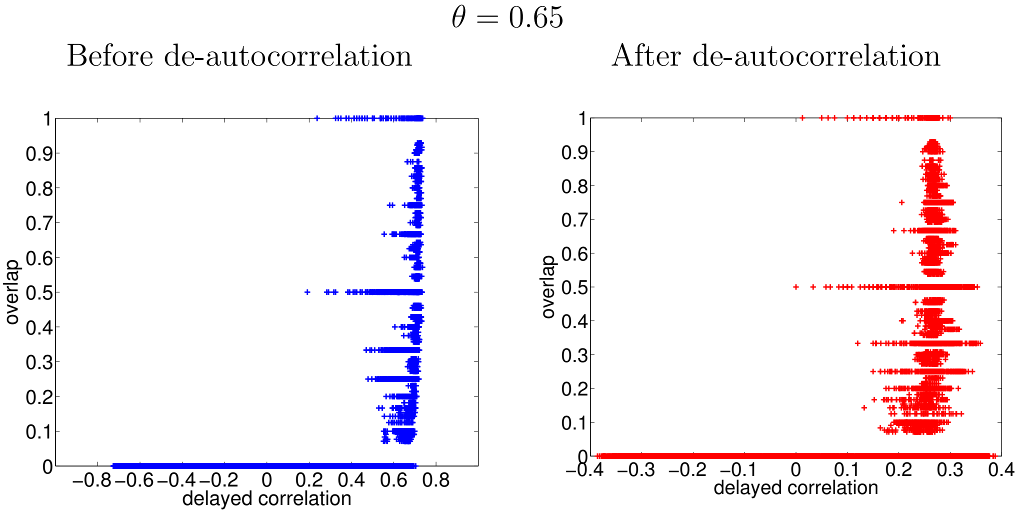


Figure 6.24: Correlation “pathways” between two voxels. Delayed correlation between voxel j and voxel i can occur directly (solid arrows) or be due to high non-delayed cross-correlation and a high autocorrelation (dashed arrows). De-autocorrelation precludes the latter by “cutting” the autocorrelation pathway $C(X_i(0)X_j(\tau))$.

By de-autocorrelation the absolute delayed correlations are decreased, however for de-autocorrelated data the delayed correlation is slightly more independent from a possible overlap of the hypervertices as they occur in clique hypergraphs. Figure 6.25 shows the clique overlap vs. the mean delayed correlation before and after de-autocorrelation for data from experiment 1 and $\theta = 0.65$. Before de-autocorrelation the largest delayed correlations are exhibited by overlapping cliques whereas after de-autocorrelation the largest delayed correlations occur for non-overlapping cliques.

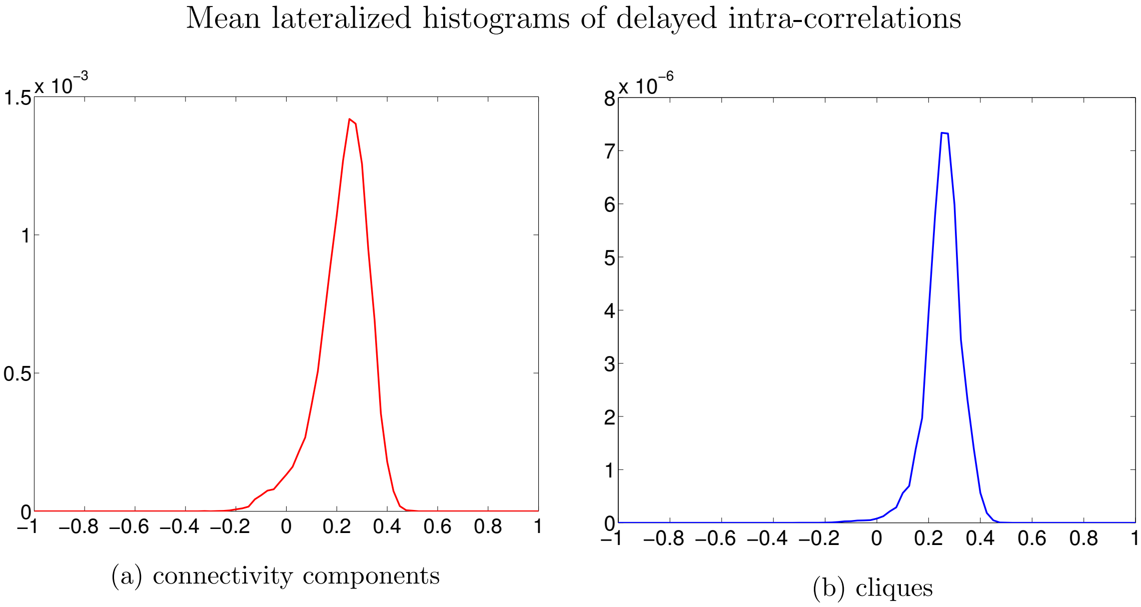


(a) (b)

Figure 6.25: Clique overlap vs. mean delayed correlation for non-deautocorrelated and deautocorrelated data. Note the different ranges on the abscissae. After de-autocorrelation the maximum delayed correlation is exhibited by non-overlapping cliques. The horizontal peaks at $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{2}{3}$, etc. are due to overlap values of smaller cliques, which occur more frequently than the larger ones.

6.4.4 Delayed intra-correlations

In addition to the delayed correlations among the voxels of the hypervertices, we considered also the delayed intra-correlations among the voxels within the hypervertices. It turned out that for connectivity components and cliques as hypervertices the delayed intra-correlations within a given hypervertex were essentially either all positive or all negative, even for de-autocorrelated data. The effect was strongest for large θ but persisted also for decreasing θ, even to a high degree for connectivity components. The latter is a signature of a certain homogeneity of the connectivity components which by their weak connectivity constraints is not obvious. Figure 6.26 shows the mean lateralized histograms of delayed intra-connectivity within connectivity components and cliques for data from experiment 1 and a threshold $\theta = 0.65$. Further information as to how the lateralized histogram was computed can be found in the caption.



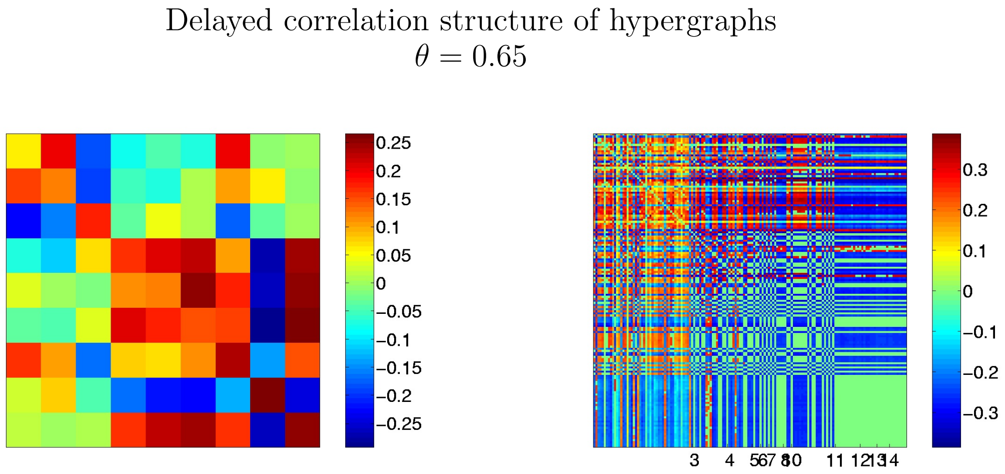
(a) connectivity components

(b) cliques

Figure 6.26: Mean lateralized histograms of the delayed intra-correlations in connectivity components and cliques for de-autocorrelated data. Lateralizing in this context means that the histogram of the delayed intra-correlations of each connectivity component or clique which had a negative mean value was flipped around the ordinate before taking the mean of all histograms. The reason for doing so was to check if the visual impression that all histograms had essentially either positive or negative support was true, which is confirmed, if the lateralized histogram has only positive support. Here in both cases the lateralization is almost complete which is particularly surprising for the connectivity components considering the weak connectivity constraints they are based on. Note that the delayed correlations between all voxel pairs in a given subgraph were counted, even if in case of the connectivity components for the given threshold they were not connected by an edge. Further the histograms are normalized to the sum of the total number of voxel pairs considered, which in the case of cliques amounts to a weighting by the number of cliques in which a given voxel pair is present.

6.4.5 Results for hypergraphs from cliques and connectivity components

We used clique hypergraphs and connectivity component hypergraphs to determine the delayed functional connectivity for $\tau = 1$. To be consistent with the threshold $\theta = 0.65$ for which the connectivity components and cliques were shown in Figure 6.12 of subsection 6.3.1 we use the same threshold in this section. Figure 6.27 shows the matrices of the mean delayed correlations between connectivity component hypergraphs and clique hypergraphs, respectively. The values in the matrices are are not thresholded except to cope for the clique overlap (cf. caption). Using a threshold of $\theta_1 = 0.19$ for connectivity components and of $\theta_1 = 0.35$ for cliques a network structure emerges that is shown in Figure 6.28, which reflects the temporal relationships between the functional units.



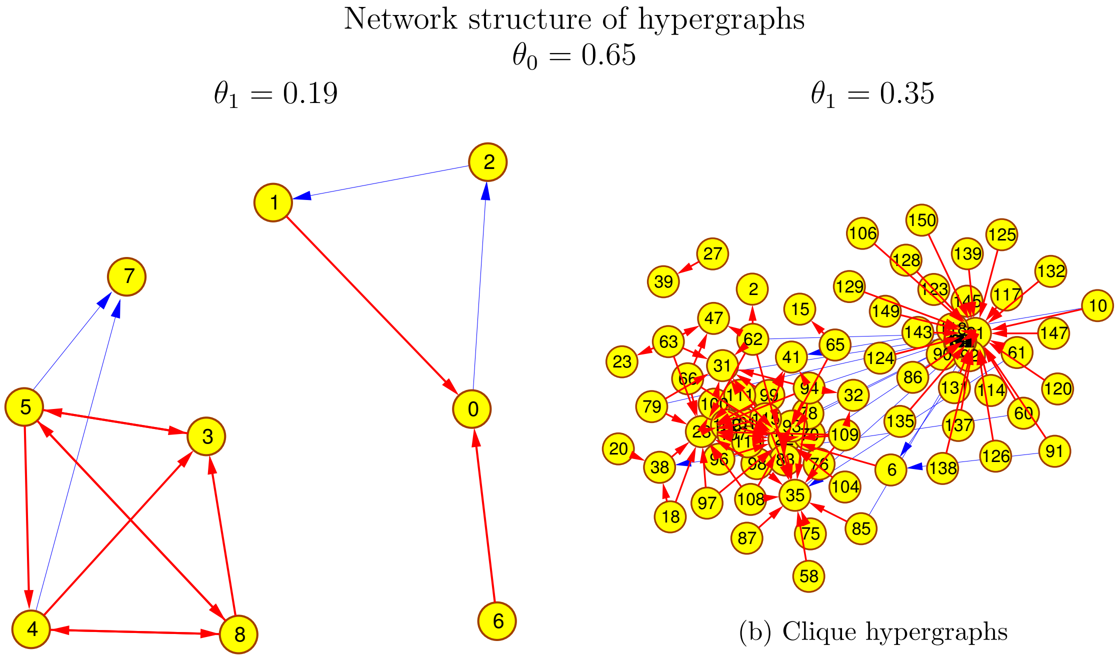
(a) Connectivity component hypergraphs

(b) Clique hypergraphs

Figure 6.27: Mean delayed correlation for connectivity component and clique hypergraphs. The hypervertices are the connectivity components and cliques determined in subsection 6.3.1 and for the given value of $\theta = 0.65$ are shown in Figure 6.12. Here the mean delayed correlation between the respective hypervertices is depicted. In (b) the delayed correlation between overlapping cliques was set to zero. The numbers at the bottom of the matrix indicate the first occurence of the respective clique order. Positive delayed correlations are found for clique orders up to 4. The negative delayed correlations between large cliques reflect probably the fact that they belong to different connectivity components. Note that the delayed correlation values for the cliques are larger than the ones for connectivity components.

6.5 Graphs from other matrices

In the previous sections we investigated functional connectivity on the basis of graphs from correlation matrices. Instead of using correlation as a measure of functional connectivity, we could also have used other measures as e. g. the covariance or the euclidean distance

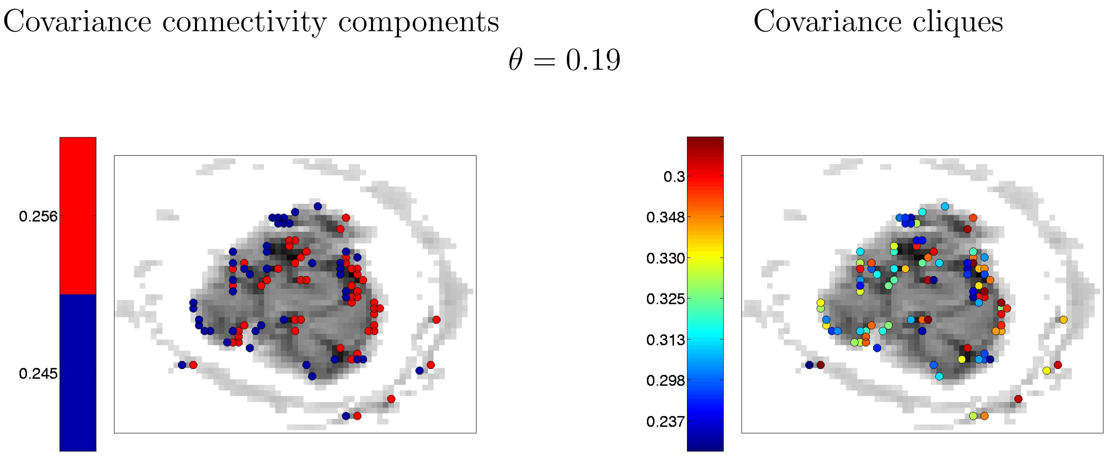


(a) Connectivity component hypergraphs

(b) Clique hypergraphs

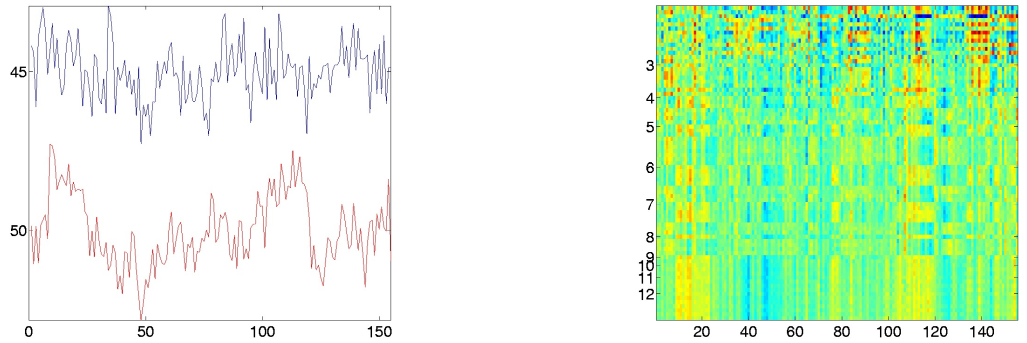
Figure 6.28: Network structure for connectivity component hypergraphs (a) and clique hypergraphs (b). The clique hypergraphs are divided reflecting the two connectivity components which are negatively correlated. A striking feature of the clique network is the emergence of centers from or to which most of the connections point.

of the time course vectors in $\mathbb{R}^k$ as well as information theoretic measures. Figure 6.30 shows the histograms of the correlation values, covariances and euclidean distances in comparison. To follow the same approach as in the previous sections the covariances and euclidean distances would be rescaled to the interval $[-1, 1]$, however, in case of the euclidean distances which to not reflect negative correlations or covariances, negative values could be included by weighting the euclidean distances with the sign of the inner product of the time course vectors. The covariances exhibit a very narrow distribution with long tails. From the shape of the covariance histogram can be expected that the subgraphs do not exhibit much structure for different values of θ, if the latter are small. This expectation is indeed confirmed by the results of using rescaled covariances instead of correlations in the graph extraction and subgraph derivation approach of the previous sections. Graph reduction here was performed at an optimal value of $\theta = 0.22$ for the rescaled covariances. Figure 6.29 shows the results for covariance connectivity components and cliques for $\theta = 0.19$. Only two large connectivity components were found for $\theta \leq 0.22$, which resemble the two large connectivity components found when using correlations. A striking difference are exhibited by the covariance cliques for which the overlap matrix in the overlapping regions is almost homogeneous. Remember that the different colors essentially are due to the normalization of the overlap. Thus as expected the covariance graphs are much less structured than the correlation graphs.



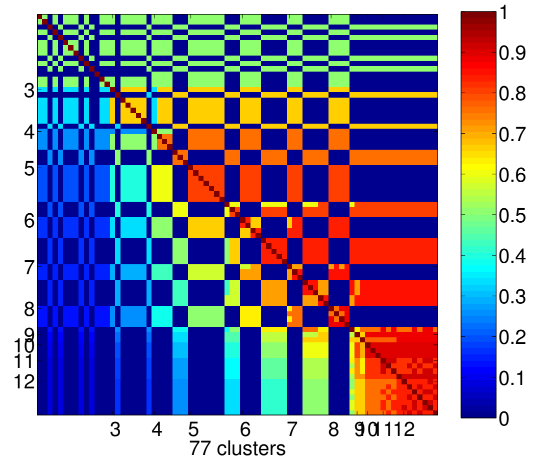
(a) Connectivity components overlaid the mean image data from experiment 1.

(b) Cliques overlaid the mean image data.



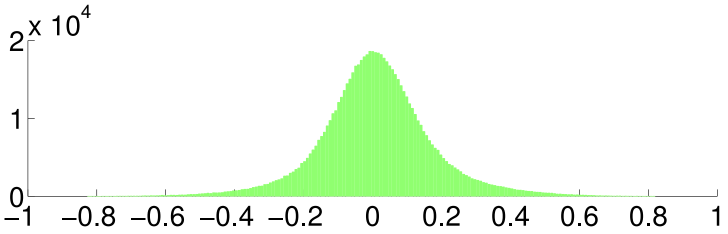
(c) Mean time courses of the connectivity components.

(d) Color coded mean time courses of the cliques

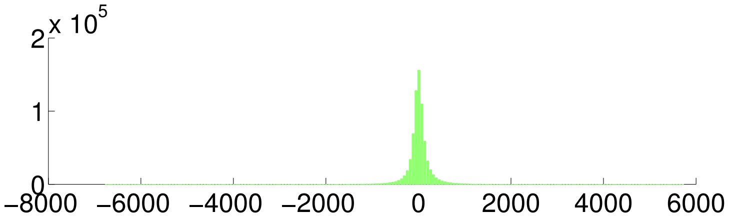


(e) Overlap matrix of the cliques.

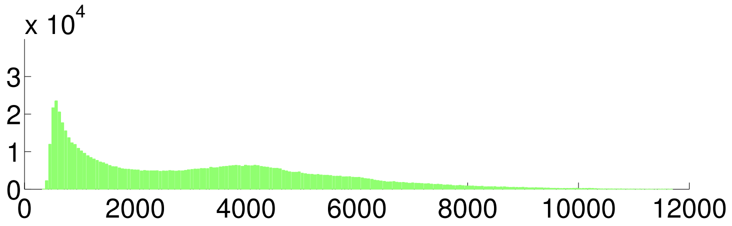
Figure 6.29: Covariance connectivity components and cliques for $\theta = 0.19$ together with their mean time courses and the overlap matrix. See text for a discussion of the results shown.



(a) Histogram of correlation values.



(b) Histogram of covariances.



(c) Histogram of euclidean distances.

Figure 6.30: Histogramms of correlation values, covariances and euclidean distances of data from experiment 1.

Chapter 7

Discussion

In this thesis we have discussed data driven approaches which are only recently finding their way into fMRI, the vast majority of the data being analyzed by conventional stimulus- or task driven approaches. Where these methods are useful to give a hint as to which brain areas are involved in the processing of a stimulus or the performance of a task, they may mask by the way of modeling the stimulus time course processes that are related to the stimulus but do not have the same time course. In contrast, data driven methods are based on generic criteria that do not depend on the particular data set. If a stimulus or task can be characterized by such a criterion, corresponding features will be appear in the results. Along with the stimulus, additional processes are identifyable.

Data driven methods produce a variety of potentially interesting aspects of the data, e. g. PCA (although being similar to correlation analysis) identifies a number of components related to the stimulus whereas correlation analysis yields a single activity distribution only. Hence, additional criteria need to be employed to chose those that are most appropriate or interesting. If a stimulus has been present it provides of course the most natural criterion. For periodical stimuli the stimulus related component can easily be identified by the power spectrum of the respective time course. This is of course true also for periodically changing non-stimulus-related factors, e. g. many physiological influences. By this means we identified stimulus-related as well as non-stimulus-related components such as blood vessels, head movements and breathing. Note, however, that the employment of the stimulus to chose the most relevant result is not equivalent to a stimulus driven approach. The stimulus in data driven methods is used only to chose from a class of results, whereas in stimulus-driven methods it is used to generate the class itself. The criteria used in the data driven approaches involve the properties of the data including possible interactions among themselves, whereas in stimulus driven approaches the data are compared with an external reference.

Analysis of real data must rely on background knowledge, while being based on general criteria at the same time. In the preprocessing section we have quantified the extend of known possible disturbances in the data, such as head movements and pulsations. Diminishing the effect of such influences in the data makes way for the application of more genereal approaches.

Stimulus-driven approaches are often based on correlating the data with a fixed reference based on the stimulus time course. We derived analytically the probability density of the sample correlation of a finite sample with a fixed reference. For i. i. d. Gaussian distributed samples this probability density assumes a closed form and can be used for thresholding.

Principal and Independent Component Analysis

In this thesis we dealt with two multivariate global data driven methods, principal and independent component analysis. Both of them can be considered from a statistical and a geometrical point of view. In principle multivariate data can be thought of as a realization of a probability distribution in a high-dimensional vector space. Usually, however, the underlying probability distribution is unknown such that the only information on it is contained in the data. In high dimensions an estimation of the underlying probability distribution is a hopeless venture, so we must rely on the assumption that the given data structure is typical in the sense that it occurs with high probability. Under this assumption the geometry of the data in the high dimensional space conveys information about the coarse structure of the probability distribution.

For finite samples the measurements can be stored in a data matrix. Due to noise, this matrix usually will have full rank, however, the intrinsic dimension, i. e. the dimensions into which the data extend considerably, may be smaller. The number of the relevant dimensions in this thesis was determined by constrasting the variances occuring against those that would be expected from random matrix theory. For approaches to determine possible fractal dimensions, such as delayed emdedding or box-counting the sampling of the data was too coarse.

Principal component analysis provides an orthogonal transform to the basis of the most relevant direction in the sense of extent, or in terms of statistics, variance, of the data. The criterion in PCA is uncorrelatedness, resulting in the uncorrelated components making up the data. There are two manners to characterize the data - by their time courses or by their images. Both characterizations are in a sense dual since for finite data they span the row and column space of the data matrix, respectively. We first performed temporal PCA as it seemed a reasonable assumption to us that the various processes occuring during an fMRI experiment are characterized by uncorrelated time courses. When determining the uncorrelated images with spatial PCA, however, they were visually indistinguishable from the images corresponding to uncorrelated time courses, except for the first image which was equivalent to the mean image. From a formal perspective the difference between temporal and spatial PCA consists in the centering of the data matrix. Centering the columns of the data matrix also affects the rows and vice versa. In the data space the centered vectors are grouped around the origin such that their sum equals zero. The vectors in the corresponding other space (row space of the data matrix if the columns were centered, or reverse) lie on a hyperplane orthogonal to a vector with all entries being identical. We have analytically derived the conditions for equality of spatial and temporal PCA. Combining the particular conditions for column and row vectors leads to multiple possibilities for equality and provides thus an explanation for the numerically observed abundance of similarity.

Independent component analysis provides an extension of PCA in that also statistical properties of higher than second order are involved. It provides a transform of the data space to a basis of statistically independent directions which need not be orthogonal. The validity of the underlying model, i. e. that the data consists of a linear superposition of statistically independent components, however, is usually impossible to know. Therefore applying ICA will result in a basis that minimizes statistical independence according to a certain criterion. There is a vast number of ICA algorithms using various criteria, which often are derived from information-theoretic properties, particularly based on the minimizations of mutual information. A more geometrical approach is the diagonalization of sample cumulants, which we have used here, since in contrast to the former methods it does not rely on implicit assumptions about source distributions.

In many cases we found ICA to be useful for additional separation of features that could not be separated by PCA alone. However, the strong assumptions contained in the ICA model are not always fulfilled and usually impossible to verify. We therefore propose to use ICA rather with care and use additional criteria to validate the results such as e. g. by comparing whether by ICA factors that were present in more than one PC have merged. The merging could be quantified e. g. by the mutual inner products of the vectors resulting from PCA and ICA.

Comparing the results of spatial and temporal ICA evidence is found that spatial ICA performs clearly better than temporal ICA. An explanation could be that the fraction of stimulus related images from the total number of images is larger than the fraction of the stimulus related voxel time courses from the total number of time courses. Hence the sampling of the image is better than the sampling of the time course. This is at least expected in the case of a periodic stimulus and focal activation. This interpretation could be checked by varying the number of voxels in the image and holding the number of stimulus related voxels constant. Also in case of the cumulant based approaches a geometrical interpretation in a similar way as was performed in the previous chapter for temporal and spatial PCA could be developed.

Functional Connectivity and Graph theory

As a main part of this thesis we developed an approach to determine functional connectivity in a graph theoretical framework, and applied it to data from fMRI. The approach extracts a network of functional units in which the connections represent their mutual temporal relationships. As paradigm for functional connectivity is used the temporal correlation among the time courses of the voxels. A high undelayed correlation is interpreted as the voxels reacting to a common input in a similar way, so that it appears justified to consider them as a functional unit. A high delayed correlation is a measure of similar, but temporally shifted behaviour and can be interpreted as a hint of causality.

The framework is completely data driven and does not rely on any stimulus based previously defined regions of interest. The pairwise correlations of each two voxel time courses constitute the correlation matrix, which from a graph theoretical point of view can be considered as the adjacency matrix of a weighted graph, which in case of delayed correlation matrices are directed. For undelayed correlations the correlation matrix is symmetric with unit elements on the diagonal. Delayed correlations lead to a correlation matrix which in general is not symmetric, the elements on the diagonal being the delayed autocorrelations of the voxel time courses, the off-diagonal elements the delayed cross-correlations.

Delayed connectivity among functional units can be characterized by high undelayed intra-correlation, i. e. a high correlation among the voxel time courses within the functional units, along with high delayed cross-correlation between the functional units.

To define functional units we analyzed the graph structure of the undelayed correlation matrix by imposing a threshold θ thereby extracting an unweighted undirected graph. The subgraphs of the extracted graph were determined according to certain graph theoretical connectivity criteria and defined as functional units. The relations among the functional units were then determined on the basis of the delayed autocorrelation matrix. Formally this was done by using the generalized graph concept of hypergraphs, where the hypervertices consisted of the previously defined subgraphs and the (directed) hyperedges of the edge-set between the vertices of the subgraphs. The hyperedges are weighted by their delayed cross-correlation. A second threshold θ_1 was imposed to extract the most significant hyperedges.

Prior to the graph extraction from the undelayed correlation matrix the latter was reduced to include only voxels that were significant in a certain sense. This so called graph reduction was performed by contrasting various graph properties of the extracted data graph against the same properties determined for random graphs. Analyzing these properties in dependence of θ, a threshold θ_0 was derived, for which the data graph was "least random" in the sense that its properties differed most from those of a random graph. The threshold θ_0 was used to include only voxels with at least one undelayed cross-correlation $\geq \theta_0$ thereby reducing the size of the correlation matrix. The various graph properties considered led to consistent thresholds θ_0. The extracted graphs the structure of which was analyzed by determining their subgraphs were then based on the reduced correlation matrix. Apart from θ_0 there are essentially two parameters involved in the presented approach, θ and θ_τ which threshold the undelayed and delayed correlations, respectively. Because longer delays are likely to be irrelevant in the data sets considered we concentrated on the case $\tau = 1$, which for the data shown corresponds to a shift of 2 s.

The subgraphs used to define functional units were characterized by their edge-connectivity. The two extremal subgraph definitions in this respect, connectivity components and cliques, which are characterized by plain and complete connectedness, respectively, were analyzed w. r. t. their suitability as functional units. Neither of them was optimally suited, e. g. connectivity components exhibiting structure loss for decreasing threshold and cliques being sensitive to noise, both extremal definitions had beneficial properties which were usually not present in the other making the definition of intermediate subgraphs desirable that preferably inherit the advantages of both. However both extreme cases agreee in a number of respects upon the structural details in the data, thus, they still provide useful insight. Analyzing the evolution of the subgraph structures in dependence of θ revealed that for the data considered the extracted graph is essentially composed of two large disjoint components each of which exhibited a complex structure of intersecting cliques, making an intermediate subgraph definition cumbersome. Additional information, which could be provided by the delayed correlation, is hence needed to separate intermediate subgraphs from the complex graph structure.

However, taking the time delayed correlations into account using the hypergraph approach reveals interesting structures for both types of hypervertices, connectivity components and cliques, as well. De-autocorrelation was used to enhance the significance of the delayed correlation. The delayed intra-correlations of hypervertices exhibited a strong homogeneity in that essentially all delayed intra-correlations within a hypervertex had the same sign. This was true even for connectivity components from the weak connectivity constraints of which that could not be expected. By thresholding the hyperedges the network structure of the hypergraph are visualized. In doing so a striking fact is the occurence of central cliques from or to which most of the connections point. This indicates brain areas which can be expected to assume a central role in the ongoing information processing or locations of systematic physiological effects.

In addition to correlations also other quantities could be used as a paradigm of functional connectivity. Natural quantities to consider in this context are the covariance and the euclidean distance of the time courses. The covariances depend on the absolute values which affect the rescaling that is necessary for thresholding. The results based on the covariance matrix were similar to the results from the correlation matrix in the voxels that were extracted, but due to the narrow distribution of the rescaled covariance values the corresponding graphs were less structured. As for the euclidean distance to include a feature corresponding to negative correlations a sign could be imposed on the euclidean distance, e. g. by including the sign of the inner product of the time courses. Also measures including higher-order statistics are of interest, but are more promising for larger data samples.

The presented framework can serve as a starting point for multiple generalizations, e. g. imposing dynamics on the extracted network. Of spectial interest in this context is the approach in [102] where networks built from anatomical connectivity matrices were functionally and structurally compared with networks generated by an evolutionary approach using various information theoretic measures. Imposing dynamics which are based on models approximating the underlying processes, such as the hemodynamic response in fMRI, the framework could be used to build large scale models of brain function. Also including multiple delays with an appropriate weighting into the hyperedges could be useful, this however being more promising for data having higher temporal resolution.

Extracting networks on the basis of hypergraphs reflecting the structure of undelayed and delayed correlations in the data establishes a novel approach to determine temporally delayed functional connectivity from fMRI data. It could be argued that functional connectivity reflecting interactions of areas at small time scales cannot be resolved by present fMRI devices, but will appear as coactivation. This being undoubtedly true, there are however examples where causal interactions can be expected to occur at timescales of seconds such as e. g. on the emergence of 3-D vision when looking at stereoscopic images or at higher cognitive tasks such as mental computations. The presented approach is quite general and therefore applicable also to data from other imaging modalities, particularly for such with high temporal resolution as well, allowing to detect functional connectivity also on short time scales, such as e. g. in EEG and MEG.

Appendix A

Mapping probability densities

When a function is applied to a random variable it is of interest how the corresponding probability density is mapped. An application related to this problem is histogram equalization where the goal is to find a function which makes the mapped probability density of a random variable as close as possible to a uniform density. In this section the rules relating the probability density of the mapped random variable to the probability density of the original random variable are derived for the one dimensional case as well as for random vectors of higher dimensions.

For the one dimensional case the relation between original and mapped probability density is proven in [91] to be

$$p_y(f(x)) = \left| \frac{df}{dx} \right|^{-1} p_x(x) \tag{A.1}$$

where $f : \mathbb{R} \longrightarrow \mathbb{R}$, $y = f(x)$ is a strictly monotone function and p_y, p_x are the probability densities of x and y, respectively. Equation A.1 can be generalized to piecewise strictly monotone functions as well (cf. [91]).

Proceeding to higher dimensions let $\mathbf{f} : \mathbb{R}^N \supset U \longrightarrow \mathbb{R}^M$ be a function which maps a random vector $\mathbf{x} \in U$ to a vector $\mathbf{y} \in \mathbb{R}^M$. Since $\mathbf{x}$ is a random vector so is $\mathbf{y}$. We wish to determine the probability density $p_{\mathbf{y}}(\mathbf{y})$ in terms of the probability density $p_{\mathbf{x}}(\mathbf{x})$. We first state the results in the cases $N > M$, $N = M$ and $N < M$ as well as for the special case $1 = M < N$ which is of importance in section A.1. Then a proof is given for the result in the case $N > M$ which can be easily extended to include the other cases as well. In any case we assume that $\mathbf{f}$ is invertible.

Case $N > M$:

$$p_{\mathbf{y}}(\mathbf{y}) = \int_{\mathbf{g}(\mathbf{z})} p_{\mathbf{x}}(g(\mathbf{z}))\, |\mathbf{J}_{\mathbf{g}(\mathbf{z})}|\, dx^{N-M} \tag{A.2}$$

Here $\mathbf{z} = (x_1, \ldots, x_{N-M}, y_1, \ldots, y_M)$ and $g(\mathbf{z})$ is the inverse mapping $f^{-1}(\mathbf{y})$ resolved with respect to $\{x_{N-M+1}, \cdots, x_N\}$.$|\mathbf{J}_{\mathbf{g}(\mathbf{z})}|$ is the Jacobian of $\mathbf{g}(\mathbf{z})$.

Case $N = M$:

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{f}(\mathbf{y}))\, |\mathbf{J}_{\mathbf{f}}^{-1}(\mathbf{y})| \tag{A.3}$$

Equation A.3 implies that coordinate transforms for which $|\mathbf{J}_\mathbf{f}^{-1}(\mathbf{y})| = 1$, as e. g. orthogonal transforms, leave the probability densities invariant.

Case $N < M$:
This case can be solved by interchanging the parts of $\mathbf{x}$ and $\mathbf{y}$ in Equation A.2. This leaves us with an integral equation for $p_\mathbf{y}$. However this case is not very common in practical applications and is mentioned here only for the sake of completeness.

Case $1 = M < N$:

$$p_y(y) = \int_{g(\mathbf{z})} p_\mathbf{x}(g(\mathbf{z})) \left|\frac{\partial g}{\partial y}\right| dx^{N-1} \tag{A.4}$$

where $\mathbf{z} = (x_1, \dots, x_{N-1}, y)$ and $g(\mathbf{z})$ is the inverse mapping $f^{-1}(\mathbf{y})$ resolved with respect to x_N (again it is assumed that f is such that this is possible).

Proof of Equation A.2:

Let $G \subseteq \mathbb{R}^M$ be the range of f. Obviously we have

$$\int_G p_\mathbf{y}(\mathbf{y})\, dy^M = \int_U p_\mathbf{x}(\mathbf{x})\, dx^N = 1 \tag{A.5}$$

Find an appropriate coordinate transform such that the integral over U in Equation A.5 can be splitted into two integrals one of which is over G. Consider

$$f(\mathbf{x}) = \mathbf{y} \tag{A.6}$$

for a fixed vector of $\mathbf{y} = (y_1, \dots, y_M)$. The vectors $\mathbf{x}$ for which Equation A.6 hold represent a subset of U and we have $U = \dot{\bigcup}\, \mathbf{f}^{-1}(\mathbf{y})$. Then we can write

$$\int_G p_\mathbf{y}(\mathbf{y})\, dy^M = \int_G \int_{\mathbf{f}^{-1}(\mathbf{y})} p_\mathbf{x}(\mathbf{g}_\mathbf{y}(\mathbf{z}))\, |J_{\mathbf{g}_\mathbf{y}(\mathbf{z})}|\, dz^N \tag{A.7}$$

where $\mathbf{g}_\mathbf{y}(\mathbf{z})$ is a parametrization of the set $\mathbf{f}^{-1}(\mathbf{y}) \subset U$ for fixed $\mathbf{y}$. As we assumed that $\mathbf{f}$ is invertible we can w. r. g. solve Equation A.6 w. r. t. the set $\{x_{N-M+1}, \dots, x_N\}$ leaving us with the following coordinate transform $\mathbf{g}(\mathbf{z})$

$$\mathbf{g}_i(\mathbf{z}) = \begin{cases} x_i & 1 \le i \le N - M \\ x_i(\mathbf{z}) & N - M + 1 \le i \le N \end{cases} \tag{A.8}$$

where $\mathbf{z} = (x_1, \dots, x_{N-M}, y_1, \dots, y_M)^T$ and $x_i(\mathbf{z})$ is the solution of $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ w. r. t. x_i for

$N-M+1 \leq i \leq N$. The Jacobian matrix of $\mathbf{g}$ in Equation A.8 is

$$\mathbf{J_g} = \begin{array}{r} \left(\begin{array}{cccccccc}
1 & 0 & \cdots & & 0 & 0 & \cdots & 0 \\
0 & 1 & & & \cdot & \cdot & & \cdot \\
\cdot & & \ddots & & \cdot & \cdot & & \cdot \\
\cdot & & & \ddots & \cdot & \cdot & & \cdot \\
\cdot & & & 1 & 0 & \cdot & & \cdot \\
0 & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 \\
\frac{\partial x_{N-M+1}(\mathbf{z})}{\partial x_{N-M+1}} & \cdots & \cdots & \cdots & \frac{\partial x_{N-M+1}(\mathbf{z})}{\partial x_N} & \frac{\partial x_{N-M+1}(\mathbf{z})}{\partial y_1} & \cdots & \frac{\partial x_{N-M+1}(\mathbf{z})}{\partial y_M} \\
\cdot & & & & & & & \cdot \\
\cdot & & & & & & & \cdot \\
\frac{\partial x_N(\mathbf{z})}{\partial x_{N-M+1}} & \cdots & \cdots & \cdots & \frac{\partial x_N(\mathbf{z})}{\partial x_N} & \frac{\partial x_N(\mathbf{z})}{\partial y_1} & \cdots & \frac{\partial x_N(\mathbf{z})}{\partial y_M}
\end{array}\right) \begin{array}{l} \\ \\ \\ \\ \\ \\ \left.\begin{array}{l} \\ \\ \\ \\ \\ \end{array}\right\} M \end{array} \\ \underbrace{\hphantom{\frac{\partial x_N(\mathbf{z})}{\partial y_1} \cdots \frac{\partial x_N(\mathbf{z})}{\partial y_M}}}_{M} \hphantom{\left(\right\} M} \end{array} \tag{A.9}$$

the determinant of which equals the determinant of the lower right submatrix consisting of the last M rows and columns of $\mathbf{J_g}$ indicated by the curly braces.

Substituting Equation A.8 into Equation A.7 yields

$$\int_G p_{\mathbf{y}}(\mathbf{y})\, dy^M = \int_G \int_{\mathbf{z}} p_{\mathbf{x}}(g(\mathbf{z}))\, |\mathbf{J_g}|\, dx^{N-M} dy^M \tag{A.10}$$

and hence

$$p_{\mathbf{y}}(\mathbf{y}) = \int_{\mathbf{z}} p_{\mathbf{x}}(g(\mathbf{z}))\, |\mathbf{J_M}|\, dx^{N-M} \tag{A.11}$$

■

A.1 The probability density of the sample correlation

The sample correlation c between a random vector $\mathbf{x}$ and a fixed vector $\mathbf{y}$ can be considered as the image of a function of the random vector $\mathbf{x}$ which is describing a surface in the N-dimensional real space (we identify the function c with its image $c(\mathbf{x})$ if there is no danger of confusion):

$$\begin{array}{cccc} c: & \mathbb{R}^N & \longrightarrow & [-1,1] \subset \mathbb{R} \\ & (x_1, \ldots, x_N)^T & \longrightarrow & \frac{\mathbf{y}^T\mathbf{x}}{\|\mathbf{y}\|\,\|\mathbf{x}\|} = \frac{\sum_{i=1}^N y_i x_i}{\left(\sum_{j=1}^N y_j^2 \sum_{i=1}^N x_i^2\right)^{\frac{1}{2}}} \end{array} \tag{A.12}$$

Here as in Equation 3.5 is assumed that $\mathbf{x}$ and $\mathbf{y}$ are centered

$$\sum_{i=1}^N x_i = \sum_{i=1}^N y_i = 0 \tag{A.13}$$

This constraint results in a loss of one dimension the effects of which are examined in section 4.1 in chapter 4. Equation A.13 constraints the surface to lie in an $(N-1)$-dimensional hyperplane of $\mathbb{R}^N$. With the appropriate orthogonal transform we can write $\mathbf{x}$ and $\mathbf{y}$ as

$(N-1)$-dimensional vectors. This leaves the probability density $p_{\mathbf{x}}(\mathbf{x})$ and hence also $p_c(c)$ invariant as is shown in Appendix A, Equation A.3. In the following we assume that this transformation has been applied. Now let $\mathbf{A}$ be an orthogonal matrix which maps $\mathbf{y}$ onto the $(N-1)$-st basis vector. We thus have $\tilde{\mathbf{y}} = \mathbf{A}\mathbf{y} = (0, \ldots, 0, \|\mathbf{y}\|)^T$. Then c simplifies to

$$c = \frac{z_{N-1}}{\sqrt{z_{N-1}^2 + s^2}} \tag{A.14}$$

where $\mathbf{z} = (z_1, \ldots, z_{N-1})^T = \mathbf{A}\mathbf{x}$ and $s^2 = \sum_{i=1}^{N-2} z_i^2$. Note that in the new coordinate frame c is not the sample correlation between $\mathbf{z}$ and $\hat{\mathbf{y}}$ since they are not centered any more. Again as $\mathbf{A}$ is orthogonal, it leaves the probability densities $p_{\mathbf{x}}(\mathbf{x})$ and $p_c(c)$ unaffected. In the previous section was shown how probability densities are transformed when their corresponding random vectors are mapped by invertible functions. Equation A.14 as a function of z_{N-1} is monotone and therefore can be inverted leading to

$$z_{N-1}(c) = \begin{cases} \sqrt{\frac{s^2\, c^2}{1-c^2}} & c \geq 0 \\ -\sqrt{\frac{s^2\, c^2}{1-c^2}} & c < 0 \end{cases} \qquad \text{with} \qquad \frac{\partial z_{N-1}}{\partial c} = \sqrt{\frac{s^2}{(1-c^2)^3}} \tag{A.15}$$

Equation A.4 represents the general formula for the transformed probability density for a function like c given by Equation A.12. Inserting Equation A.15 into Equation A.4 thus yields the probability density $p_c(c)$ in dependence of the probability density $p_{\mathbf{x}}(\mathbf{z})$

$$p_c(c) = \frac{1}{(1-c^2)^{\frac{3}{2}}} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-2} p_{\mathbf{x}}(z_1, \ldots, z_{N-2}, z_{N-1}(c)) \sqrt{\sum_{i=1}^{N-2} z_i^2}\; dz^{N-2} \tag{A.16}$$

In the following subsection we will explicitly consider the dependence of $p_c(c)$ on the sample size N for i. i. d. gaussian samples.

Independent identically Gaussian distributed samples

In this subsection we derive the sample size dependence of the correlation probability density for i. i. d. gaussian distributed samples. Consider N identically gaussian distributed samples with zero mean and variance σ^2. In this case Equation A.16 reads

$$p_c(c) = \frac{(2\pi)^{-\frac{N-1}{2}}}{\sigma^{N-1}(1-c^2)^{\frac{3}{2}}} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-2} e^{\frac{-\sum_{i=1}^{N-2} z_i^2}{2\,\sigma^2\,(1-c^2)}} \sqrt{\sum_{i=1}^{N-2} z_i^2}\; dz^{N-2} \tag{A.17}$$

Since the distribution is spherical the integral is solved most easily by changing to generalized spherical coordinates (cf. section A.2). Using these Equation A.17 becomes

$$p_c(c) = \frac{(2\pi)^{-\frac{N-1}{2}+1}}{\sigma^{N-1}\,(1-c^2)^{\frac{3}{2}}} \int_0^{\infty} e^{\frac{-r^2}{2\,\sigma^2\,(1-c^2)}}\, r^{N-2}\, dr \prod_{k=2}^{N-3} \int_0^{\pi} \sin^{k-1}\vartheta\, d\vartheta \tag{A.18}$$

This holds for $N \geq 4$. The cases $N < 4$ are dealt with in Equation A.26-Equation A.28.

In the following we derive a simplified expression for Equation A.18. Integrating over r we get ([16])

$$\int_0^\infty e^{\frac{-r^2}{2\,\sigma^2\,(1-c^2)}}\, r^{N-2}\, dr = 2^{\frac{N-1}{2}-1}\sigma^{N-1}\,(1-c^2))^{\frac{N-1}{2}}\,\Gamma(\frac{N-1}{2}) \tag{A.19}$$

where Γ refers to the Gamma function. The integrals over $\sin^{k-1}\vartheta$ are ([55])

$$\int_0^\pi \sin^{k-1}\vartheta\, d\vartheta = B(\frac{1}{2},\frac{k}{2}) \tag{A.20}$$

where

$$B(x,y) = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)} \tag{A.21}$$

refers to the Beta function alias Euler's integral of the first kind. The explicit formula for Equation A.20 reads

$$\int_0^\pi \sin^{k-1}\vartheta\, d\vartheta = \begin{cases} \frac{2^{k-1}\,((\frac{k}{2}-1)!)^2}{(k-1)!} & k \text{ even} \\ \frac{(k-1)!}{2^{k-1}\,((\frac{k-1}{2})!)^2}\,\pi & k \text{ odd} \end{cases} \tag{A.22}$$

Multiplying Equation A.22 from for $k = 2$ to $N-3$ we get

$$\prod_{k=2}^{N-3}\int_0^\pi \sin^{k-1}\vartheta_k\, d\vartheta_k = \begin{cases} \frac{\pi^{\frac{N-4}{2}}}{(\frac{N-4}{2})!} & N \text{ even} \\ \frac{2^{N-4}\,(\frac{N-5}{2})!\,\pi^{\frac{N-5}{2}}}{(N-4)!} & N \text{ odd} \end{cases} \tag{A.23}$$

Inserting Equation A.19 and Equation A.23 into Equation A.18 yields

$$p_c(c) = \frac{(2\pi)^{-\frac{N-1}{2}+1}}{\sigma^{N-1}\,(1-c^2)^{\frac{3}{2}}}\;\underbrace{2^{\frac{N-1}{2}-1}\sigma^{N-1}\,(1-c^2))^{\frac{N-1}{2}}\,\Gamma\Big(\frac{N-1}{2}\Big)}_{\text{Equation A.19}}\cdot\underbrace{\begin{cases} \frac{\pi^{\frac{N-4}{2}}}{(\frac{N-4}{2})!} \\ \frac{2^{N-4}\,(\frac{N-5}{2})!\,\pi^{\frac{N-5}{2}}}{(N-4)!} \end{cases}}_{\text{Equation A.23}} \begin{matrix} N \text{ even} \\ N \text{ odd} \end{matrix} \tag{A.24}$$

which can be simplified to the final result for $N \geq 4$

$$p_c(c) = (1-c^2)^{\frac{N-4}{2}} \cdot \begin{cases} \frac{(N-3)!}{\left(\left(\frac{N-4}{2}\right)!\right)^2\,2^{N-3}} & N \text{ even} \\ \frac{2^{N-5}\,(N-3)\left(\left(\frac{N-5}{2}\right)!\right)^2}{\pi\,(N-4)!} & N \text{ odd} \end{cases} \tag{A.25}$$

where the terms on the right of the curly brace are normalization factors ensuring that the integral over $p_c(c)$ equals unity.

The cases $N < 4$:

$N = 3$:

For $N = 3$ we have $N - 2 = 1$ independent coordinates. In this case Equation A.17 reads

$$\begin{aligned} p_c(c) &= \frac{(2\pi)^{-1}}{\sigma^2(1-c^2)^{\frac{3}{2}}} \int_{-\infty}^{\infty} e^{\frac{z^2}{2\,\sigma^2\,(1-c^2)}} \sqrt{z}\, dz \\ &= \frac{1}{\pi\,\sqrt{1-c^2}} \end{aligned} \tag{A.26}$$

$N = 2$:

For $N = 2$ the sample correlation reduces to

$$c = \frac{z}{\sqrt{z}} = \operatorname{sign}(z) \qquad \text{for } z \neq 0 \tag{A.27}$$

For any symmetric probability density of z this results in

$$p_c(c) = \frac{1}{2}\left(\delta(c+1) + \delta(c-1)\right) \tag{A.28}$$

$N = 1$:

For $N = 1$ centering makes all random variables equal to zero for which no correlation is defined.

The shape of $p_c(c)$ exhibits a “phase transition” at $N = 4$. For $N \in \{2, 3\}$ the probability density $p_c(c)$ is strongly concentrated on $c = 1$ and $c = -1$, in fact it is infinite at these points. For $N = 4$ we have a uniform density $p_c(c)$ whilst for $N > 4$, $p_c(c)$ is unimodal with its mode at zero becoming sharper and sharper for increasing N. An illustration of $p_c(c)$ in dependence of N for i. i. d. gaussian samples is given in Figure 3.5 in section 3.2.

Another interesting feature of $p_c(c)$ is that it is independent of the variance σ^2 and also independent of the original fixed vector $\mathbf{y}$ with which the random vector $\mathbf{x}$ was correlated.

A.2 Generalized spherical coordinates

In three dimensions changing from cartesian to spherical coordinates involves the well known equations

$$\begin{aligned} x_1 &= r\,\cos\phi\,\sin\vartheta \\ x_2 &= r\,\sin\phi\,\sin\vartheta \\ x_3 &= r\,\cos\vartheta \end{aligned} \tag{A.29}$$

with the ranges $r \in [0, \infty)$, $\phi \in [-\pi, \pi)$, and $\vartheta \in [0, \pi)$. Going on with the principle underlying the construction of spherical coordinates for three dimensions we arrive at the generalized spherical coordinates for N dimensions

$$\begin{aligned}
x_1 &= r \prod_{k=1}^{N-1} \sin\vartheta_k \\
x_i &= r \cos\vartheta_{i-1} \prod_{k=i}^{N-1} \sin\vartheta_k \qquad\qquad 2 \le i \le N-1 \\
x_N &= r \cos\vartheta_{N-1}
\end{aligned} \tag{A.30}$$

where for consistency the parts of x_1 and x_2 in 3-dimensional spherical coordinates were interchanged and the replacements $\phi = \vartheta_1$ and $\vartheta = \vartheta_2$ made. The ranges here are $r \in [0, \infty)$, $\vartheta_1 \in [-\pi, \pi)$, and $\vartheta_i \in [0, \pi)$ for $2 \le i \le N-1$. The Jacobian matrix for the transform to spherical coordinates in matrix form reads

$$\mathbf{J} = \begin{pmatrix}
\prod_{k=1}^{N-1} \sin\vartheta_k & r \cos\vartheta_1 \prod_{k=2}^{N-1} \sin\vartheta_k & \cdot\ \cdot & \cdot\ \cdot & \cdot\ \cdot & r \cos\vartheta_{N-1} \prod_{k=1}^{N-2} \sin\vartheta_k \\
\cos\vartheta_1 \prod_{k=2}^{N-1} \sin\vartheta_k & -r \prod_{k=1}^{N-1} \sin\vartheta_k & & & & r\cos\vartheta_{N-1} \cos\vartheta_1 \prod_{k=2}^{N-2} \sin\vartheta_k \\
\cos\vartheta_2 \prod_{k=3}^{N-1} \sin\vartheta_k & 0 & \cdot & & & r\cos\vartheta_{N-1} \cos\vartheta_2 \prod_{k=3}^{N-2} \sin\vartheta_k \\
\cdot & \cdot & & \cdot & & \cdot \\
\cdot & \cdot & \cdot & & \cdot & \cdot \\
\cdot & \cdot & & \cdot & & \cdot \\
\cos\vartheta_{N-1} & 0 & \cdot & \cdot & 0 & -r \sin\vartheta_{N-1}
\end{pmatrix} \tag{A.31}$$

The determinant of J is

$$|\mathbf{J^{(N)}}| = (-r)^{N-1} \prod_{k=2}^{N-1} \sin^{k-1}\vartheta_k \tag{A.32}$$

Proof of Equation A.32:

Equation A.32 is valid for $N \in \{1, 2, 3\}$. For $N = 2$ the coordinates are known as polar coordinates and for $N = 3$ as spherical coordinates. The proof for $N > 3$ is by induction. We thus want to show that

$$|\mathbf{J^{(N)}}| = -r \sin^{N-2}\vartheta_{N-1}\, |\mathbf{J^{(N-1)}}| \tag{A.33}$$

Using the Laplace expansion along the last row we get

$$|\mathbf{J^{(N)}}| = (-1)^{N+1} \cos\vartheta_{N-1}\, |\mathbf{J}^{\mathbf{(N)}}_{N1}| - r \sin^{N}\vartheta_{N-1}\, |\mathbf{J^{(N-1)}}| \tag{A.34}$$

where we used that $\det(\lambda\,\mathbf{A}) = \lambda^N \det(\mathbf{A})$ for a $N \times N$ matrix $\mathbf{A}$ and $\lambda \in \mathbb{R}$. For Equation A.33 to hold the following equality needs to be shown

$$|\mathbf{J}^{(\mathbf{N})}_{N1}| = (-1)^{N+2n}\, r\, \sin^{N-2}\vartheta_{N-1}\, \cos\vartheta_{N-1}\, |\mathbf{J}^{(\mathbf{N-1})}| \tag{A.35}$$

where $n \in \mathbb{Z}$. The following two relations can be seen from Equation A.31 or shown formally from Equation A.9:

$$(\mathbf{J}^{(\mathbf{N})}_{N1})_{j\,N-1} = r\, \cos\vartheta_{N-1}\, (\mathbf{J}^{(\mathbf{N-1})})_{j\,1} \qquad 1 \leq j \leq N-1 \tag{A.36}$$

and

$$(\mathbf{J}^{(\mathbf{N})}_{N1})_{j\,k} = \sin\vartheta_{N-1}(\mathbf{J}^{(\mathbf{N-1})})_{j\,k+1} \qquad 1 \leq j \leq N-1, \quad 1 \leq k \leq N-2 \tag{A.37}$$

That is the last row of $\mathbf{J}^{(\mathbf{N})}_{N1}$ equals the first row of $\mathbf{J}^{(\mathbf{N-1})}$ multiplied by the factor $r\,\cos\vartheta_{N-1}$ and the remainders of the respective matrices are equal up to the factor $\sin\vartheta_{N-1}$. These factors are such that Equation A.35 holds true for $n = -1$.

■

Appendix B

Equality of temporal and spatial PCA

B.1 Equality of the time courses

In this section we provide detailed conditions for equality of the time courses of temporal and spatial PCA.

Let $\mathbf{C} = \mathbf{X}^T\mathbf{X} \neq \mathbf{0}$ have the eigenvalues $d_i \geq 0$ with $i \in \{1, \dots, k\}$. $\bar{\mathbf{z}}^T = \left(\frac{1}{m}\sum_{r=1}^{m} x_{ri}\right)_{1\times m}$ is the mean time course of the data, i. e. the mean row of $\mathbf{X}$. $\mathbf{P_k}$ is the projection 4.12 which projects the row vectors of $\mathbf{X}$ to the subspace orthogonal to $\mathbf{1_k}$. Further, y_i are the inner products of $\mathbf{1_k} = (1, \dots, 1)^T$ with the eigenvectors of $\mathbf{C}$ and S_C is the sum of all elements of $\mathbf{C}$. Finally, if $\mathbf{1_k}$ and $\mathbf{P_k}\bar{\mathbf{z}}$ are eigenvectors of $\mathbf{C}$, then d_1 and d_p denote the respective eigenvalues.

Proposition: The time courses obtained by temporal and spatial PCA are equal, iff any of the following conditions hold:

(A). $\mathbf{P_k}\bar{\mathbf{z}}$ is an eigenvector of $\mathbf{C}$, $\mathbf{1_k}$ is contained in a subspace spanned by eigenvectors of $\mathbf{C}$ with eigenvalues being either 0 or $\frac{S_C}{k}$, and

- (a) $\bar{\mathbf{z}} \perp \mathbf{1_k}$ or
- (b) $d_p = \frac{1}{k}(S_C - \sum_{l=1}^{k} d_l\, y_l^2)$.

(B). $\bar{\mathbf{z}} \parallel \mathbf{1_k}$ and $\mathbf{1_k}$ is an eigenvector of $\mathbf{C}$.

Before proving the proposition a few remarks are in order. If the subspace mentioned in (A) is one-dimensional, i. e. if $\mathbf{1_k}$ is itself an eigenvector, the requirements are more specifically that $d_1 = 0$ and $\bar{\mathbf{z}} \perp \mathbf{1_k}$ or $d_p = \frac{S_C}{k}$, or alternatively that $d_1 = \frac{S_C}{k}$ and $\bar{\mathbf{z}} \perp \mathbf{1_k}$ or $d_p = 0$. Further, note that in (B) there are no constraints on the eigenvalues such that (B) can be checked much easier than (A). In particular for the data sets used here, spatial and temporal PCA produce similar time courses because (B) is approximately satisfied, which can be considered to reflect a homogeneity property of the data. On the other hand, condition (Ab) appears to be a rather unlikely special case and (Aa) will not be relevant for the present data set, because some coordinates of $\bar{\mathbf{z}}$ would be negative if this case applied.

The conditions in the proposition are sufficient and necessary, i. e. the equality of the two sets of time courses is equivalent to the disjunction of the conditions. The proof will proceed as follows. First the obvious necessary and sufficient condition for equality of the time courses B.1 will be condensed into the simpler condition B.9. Then by a lemma a further reduction is achieved, cf. Equation B.10, before necessity and sufficieny are shown for Equation B.9.

PROOF: The PC time courses are equal (i. e. $\mathbf{V_Z} = \mathbf{V_Y}$ in section 4.1) if and only if

$$[\mathbf{C}_{\mathbf{Z}}^{(t)}\,,\,\mathbf{C_Y}] = 0 \tag{B.1}$$

where $[\cdot,\cdot]$ is the commutator and

$$\begin{aligned}
\mathbf{C}_{\mathbf{Z}}^{(\mathbf{t})} &= \mathbf{Z}^T\mathbf{Z}\\
&\overset{\text{Equation 4.11}}{=} \mathbf{P_k}\mathbf{X}^T\mathbf{X}\mathbf{P_k}\\
&\overset{\text{Equation 4.12}}{=} (\mathbb{I}_\mathbf{k} - \frac{1}{k}\mathbf{1_k}\mathbf{1_k}^T)\mathbf{X}^T\mathbf{X}(\mathbb{I}_\mathbf{k} - \frac{1}{k}\mathbf{1_k}\mathbf{1_k}^T)\\
&= \mathbf{C} - \frac{1}{k}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\mathbf{C} - \frac{1}{k}\mathbf{C}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T + \|\bar{\mathbf{s}}\|^2\mathbf{1_k}\mathbf{1}_\mathbf{k}^T \qquad \text{(B.2)}\\
\mathbf{C_Y} &= \mathbf{Y}^T\mathbf{Y}\\
&\overset{\text{Equation 4.11}}{=} \mathbf{X}^T\mathbf{P_m}\mathbf{X}\\
&\overset{\text{Equation 4.12}}{=} \mathbf{X}^T(\mathbb{I}_\mathbf{m} - \frac{1}{m}\mathbf{1_m}\mathbf{1}_\mathbf{m}^T)\mathbf{X}\\
&= \mathbf{C} - m\,\bar{\mathbf{z}}\bar{\mathbf{z}}^T \qquad \text{(B.3)}
\end{aligned}$$

where $\bar{\mathbf{s}} = \left(\frac{1}{k}\sum_{r=1}^{k} x_{ir}\right)_{m\times 1}$ is the mean image (mean column of $\mathbf{X}$). $\|\bar{\mathbf{s}}\|^2$ can be written as

$$\|\bar{\mathbf{s}}\|^2 = \sum_{i=1}^{m}(\frac{1}{k}\sum_{l=1}^{k} x_{il})^2 = \frac{1}{k^2}\sum_{l=1}^{k}\sum_{t=1}^{k}\sum_{i=1}^{m} x_{il}x_{it} = \frac{S_C}{k^2} \tag{B.4}$$

Since $\mathbf{C}_{\mathbf{Z}}^{(\mathbf{t})}$ and $\mathbf{C_Y}$ are both symmetric, condition B.1 is equivalent to the condition that the product $\mathbf{C}_{\mathbf{Z}}^T\mathbf{C_Y}$ is symmetric as well: If the matrices $\mathbf{A}$ and $\mathbf{B}$ are symmetric, we have $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T = \mathbf{BA}$ and hence $\mathbf{AB} = \mathbf{BA} \iff (\mathbf{AB})^T = \mathbf{AB}$. Using Equation B.2 and Equation B.3 the symmetry condition reads

$$\begin{aligned}
\mathbf{C}_{\mathbf{Z}}^{(\mathbf{t})}\mathbf{C_Y} &= \left(\mathbf{C} - \frac{1}{k}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\mathbf{C} - \frac{1}{k}\mathbf{C}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T + \frac{S_C}{k^2}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\right)\left(\mathbf{C} - m\,\bar{\mathbf{z}}\bar{\mathbf{z}}^T\right)\\
&= \underbrace{\mathbf{C}^2}_{\text{symm.}} - \underbrace{\frac{1}{k}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\mathbf{C}^2}_{\text{a.r.eq.}} - \underbrace{\frac{1}{k}\mathbf{C}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\mathbf{C}}_{\text{symm.}} + \underbrace{\frac{S_C}{k^2}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\mathbf{C}}_{\text{a.r.eq.}}\\
&\quad - \underbrace{m\,\mathbf{C}\bar{\mathbf{z}}\bar{\mathbf{z}}^T}_{(1)} + \underbrace{\frac{m}{k}\,\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\mathbf{C}\bar{\mathbf{z}}\bar{\mathbf{z}}^T}_{\text{a.r.eq.}} + \underbrace{\frac{m}{k}\,\mathbf{C}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\bar{\mathbf{z}}\bar{\mathbf{z}}^T}_{(2)} - m\underbrace{\frac{S_C}{k^2}\mathbf{1_k}\mathbf{1}_\mathbf{k}^T\bar{\mathbf{z}}\bar{\mathbf{z}}^T}_{\text{a.r.eq.}}\\
&= (\mathbf{C}_{\mathbf{Z}}^{(\mathbf{t})}\mathbf{C_Y})^T \qquad \text{(B.5)}
\end{aligned}$$

Here a. r. eq. stands for “all rows equal” which is a property of the so labeled matrices. Obviously the sum of a. r. eq. matrices is an a. r. eq. matrix as well and can be symmetric only if it is constant, i. e. the matrix elements are all equal, which is meant henceforth when referring to a matrix as being constant. A $k \times k$ matrix with identical rows can be written as $\mathbf{1_k a}^T$ where $\mathbf{a}$ is the row vector of the a. r. eq. matrix. For the a. r. eq. matrices in Equation B.5 the corresponding vector $\mathbf{a}$ reads

$$\mathbf{a} = \frac{1}{k^2}(\mathbf{C} - m\bar{\mathbf{z}}\bar{\mathbf{z}}^T)(S_C \mathbb{I}_\mathbf{k} - k\mathbf{C})\mathbf{1_k} \tag{B.6}$$

Furthermore the sum of (1) and (2) can be written as

$$\begin{aligned} (1) + (2) &= \mathbf{b}\bar{\mathbf{z}}^T & \text{(B.7)} \\ \mathbf{b} &= -m\mathbf{C}\mathbf{P_k}\bar{\mathbf{z}} & \text{(B.8)} \end{aligned}$$

where, again, $\mathbf{P_k}\bar{\mathbf{z}}$ is the part of $\bar{\mathbf{z}}$ which is orthogonal to $\mathbf{1_k}$.
Thus, Equation B.1 can be replaced by the equivalent condition

$$\mathbf{b}\bar{\mathbf{z}}^T + \mathbf{1_k a}^T = \bar{\mathbf{z}}\mathbf{b}^T + \mathbf{a}\mathbf{1_k}^T \tag{B.9}$$

A property of the set of the four vectors occurring in Equation B.9 is stated in the following

Lemma: Equation B.9 is equivalent to the relation

$$\big(\bar{\mathbf{z}}, \mathbf{a}\big) = \big(\mathbf{b}, \mathbf{1_k}\big)\ \mathbf{K} \tag{B.10}$$

where $\mathbf{K}$ is a symmetric 2×2 matrix and $\big(\bar{\mathbf{z}}, \mathbf{a}\big)$ and $\big(\mathbf{b}, \mathbf{1_k}\big)$ are the $k \times 2$ matrices with the columns $\bar{\mathbf{z}}$, $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{1_k}$, respectively.

Note that condition B.10 is equivalent to the equality of the signed areas of the parallelograms spanned by $\mathbf{a}$ and $\mathbf{1_k}$ and $\mathbf{b}$ and $\bar{\mathbf{z}}$, respectively, as this relates to a conservation law of symmetric transformations.

Proof of the lemma: First we show that Equation B.9 implies $span(\mathbf{b}, \mathbf{1_k}) = span(\bar{\mathbf{z}}, \mathbf{a})$, i. e. $\mathbf{b}$ and $\mathbf{1_k}$ span the same subspace as $\bar{\mathbf{z}}$ and $\mathbf{a}$. Then it is demonstrated how Equation B.10 follows from Equation B.9 and finally the reverse implication is proven.
For the span of two pairs of vectors $\mathbf{b}, \mathbf{1_k} \in \mathbb{R}^k$ and $\bar{\mathbf{z}}, \mathbf{a} \in \mathbb{R}^k$, $k \geq 4$, with $\mathbf{b} \nparallel \mathbf{1_k}$ and $\bar{\mathbf{z}} \nparallel \mathbf{a}$, there are the following possibilities

(a). $span(\mathbf{b}, \mathbf{1_k}) \cap span(\bar{\mathbf{z}}, \mathbf{a}) = \{\mathbf{0}\}$.

(b). $span(\mathbf{b}, \mathbf{1_k}) \cap span(\bar{\mathbf{z}}, \mathbf{a}) = span(\mathbf{v})$ for a vector $\mathbf{v} \in \mathbb{R}^k$.

(c). $span(\mathbf{b}, \mathbf{1_k}) = span(\bar{\mathbf{z}}, \mathbf{a})$.

The matrix $\mathbf{b}\bar{\mathbf{z}}^T + \mathbf{1_k a}^T$ in Equation B.9 projects onto the span of $\mathbf{b}$ and $\mathbf{1_k}$. Hence the rank of $\mathbf{b}\bar{\mathbf{z}}^T + \mathbf{1_k a}^T$ is ≤ 2. The same holds true for the rank of $\bar{\mathbf{z}}\mathbf{b}^T + \mathbf{a}\mathbf{1_k}^T$, which projects onto the span of $\bar{\mathbf{z}}$ and $\mathbf{a}$. It is easy to see that $\mathbf{b}\bar{\mathbf{z}}^T + \mathbf{1_k a}^T$ has rank 1 iff $\mathbf{b} \parallel \mathbf{1_k}$, which is true m. m. also for $\bar{\mathbf{z}}\mathbf{b}^T + \mathbf{a}\mathbf{1_k}^T$. In the case of rank 1 (c) is equivalent to (b) and holds trivially with $\mathbf{K}$ in Equation B.10 being a scalar. Now w. r. g. let us assume $\mathrm{rank}(\mathbf{b}\bar{\mathbf{z}}^T + \mathbf{1_k}\mathbf{a}^T) = \mathrm{rank}(\bar{\mathbf{z}}\mathbf{b}^T + \mathbf{a}\mathbf{1_k}^T) = 2$ and write the vectors $\mathbf{a}$ and $\bar{\mathbf{z}}$ in terms of $\mathbf{b}$ and $\mathbf{1_k}$ as

$$\begin{aligned} \bar{\mathbf{z}} &= \beta_1\mathbf{b} + \beta_2\mathbf{1_k} + \bar{\mathbf{z}}^\perp & \text{(B.11)} \\ \mathbf{a} &= \alpha_1\mathbf{b} + \alpha_2\mathbf{1_k} + \mathbf{a}^\perp & \text{(B.12)} \end{aligned}$$

where $\mathbf{a}^\perp, \bar{\mathbf{z}}^\perp \perp \mathit{span}(\mathbf{b}, \mathbf{1_k})$ are the parts orthogonal to the span of $\mathbf{b}$ and $\mathbf{1_k}$. In matrix form Equation B.11 and Equation B.12 read

$$\big(\bar{\mathbf{z}}, \mathbf{a}\big) = \big(\mathbf{b}, \mathbf{1_k}\big)\mathbf{K} + \big(\bar{\mathbf{z}}^\perp, \mathbf{a}^\perp\big) \tag{B.13}$$

where

$$\mathbf{K} = \begin{pmatrix} \beta_1 & \alpha_1 \\ \beta_2 & \alpha_2 \end{pmatrix} \tag{B.14}$$

Multiplying both sides of Equation B.9 from the right by an arbitrary vector $\mathbf{w} = x_1\mathbf{b} + x_2\mathbf{1_k} + \mathbf{w}^\perp$ and writing the result in a similar form as Equation B.13 yields

$$\big(\mathbf{b}, \mathbf{1_k}\big)(\mathbf{A}\mathbf{x} + \mathbf{x}^\perp) = \big(\bar{\mathbf{z}}, \mathbf{a}\big)\mathbf{B}\mathbf{x} \tag{B.15}$$

where $\mathbf{x} = (x_1, x_2)^T$ and

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} \langle \bar{\mathbf{z}}, \mathbf{b}\rangle & \langle \bar{\mathbf{z}}, \mathbf{1_k}\rangle \\ \langle \mathbf{a}, \mathbf{b}\rangle & \langle \mathbf{a}, \mathbf{1_k}\rangle \end{pmatrix} & \text{(B.16)} \\ \mathbf{B} &= \begin{pmatrix} \|\mathbf{b}\|^2 & \langle \mathbf{1_k}, \mathbf{b}\rangle \\ \langle \mathbf{1_k}, \mathbf{b}\rangle & \|\mathbf{1_k}\|^2 \end{pmatrix} & \text{(B.17)} \\ \mathbf{x}^\perp &= \begin{pmatrix} \langle \bar{\mathbf{z}}^\perp, \mathbf{w}^\perp\rangle \\ \langle \mathbf{a}^\perp, \mathbf{w}^\perp\rangle \end{pmatrix} & \text{(B.18)} \end{aligned}$$

Inserting Equations B.11 and B.12 into Equation B.16 yields

$$\mathbf{A} = \mathbf{K}^T\mathbf{B} \tag{B.19}$$

which by inserting Equations B.19 and B.13 into Equation B.15 leads to the result

$$\big(\mathbf{b}, \mathbf{1_k}\big)(\mathbf{K}^T\mathbf{B}\mathbf{x} + \mathbf{x}^\perp) = \big(\mathbf{b}, \mathbf{1_k}\big)\mathbf{K}\mathbf{B}\mathbf{x} + \big(\bar{\mathbf{z}}^\perp, \mathbf{a}^\perp\big)\mathbf{B}\mathbf{x} \tag{B.20}$$

Since the vectors on the left hand side are contained in $\mathit{span}(\mathbf{b}, \mathbf{1_k})$, from the choice of $\bar{\mathbf{z}}^\perp$ and $\mathbf{a}^\perp$ to be perpendicular to $\mathbf{b}$ and $\mathbf{1_k}$ it must be

$$\big(\bar{\mathbf{z}}^\perp, \mathbf{a}^\perp\big)\mathbf{B}\mathbf{x} = \mathbf{0} \tag{B.21}$$

which is fulfilled if

$$\|\mathbf{1_k}\|^2\|\mathbf{b}\|^2 = |\langle \mathbf{b}, \mathbf{1_k}\rangle|^2 \tag{B.22}$$

or

$$\big(\bar{\mathbf{z}}^\perp, \mathbf{a}^\perp\big) = \big(\mathbf{0}, \mathbf{0}\big) \tag{B.23}$$

Equation B.22 is equivalent to $\mathbf{b} \parallel \mathbf{1_k}$ and rank$(\mathbf{b}\bar{\mathbf{z}}^T + \mathbf{1_k}\mathbf{a}^T)$= 1. Since we assumed rank$(\mathbf{b}\bar{\mathbf{z}}^T + \mathbf{1_k}\mathbf{a}^T)$= 2, Equation B.23 remains and, hence, from this and Equation B.21,

$$\mathbf{x}^\perp = \mathbf{0} \tag{B.24}$$

Inserting Equation B.24 and Equation B.23 into Equation B.20 leads to the requirement that

$$\mathbf{K} = \mathbf{K}^T \tag{B.25}$$

i. e. $\mathbf{K}$ must be symmetric. In the following we will write

$$\mathbf{K} = \begin{pmatrix} \beta & \gamma \\ \gamma & \alpha \end{pmatrix} \tag{B.26}$$

For the reverse implication of the lemma, we start from Equation B.10 with symmetric $\mathbf{K}$, i. e. we have

$$\bar{\mathbf{z}} = \beta\mathbf{b} + \gamma\mathbf{1_k} \tag{B.27}$$

$$\mathbf{a} = \gamma\mathbf{b} + \alpha\mathbf{1_k} \tag{B.28}$$

Multiplying Equation B.27 from the right by $\mathbf{b}^T$ and the transpose of Equation B.27 from the left by $\mathbf{b}$ we obtain

$$\bar{\mathbf{z}}\mathbf{b}^T = \beta\mathbf{b}\mathbf{b}^T + \gamma\mathbf{1_k}\mathbf{b}^T \tag{B.29}$$

$$\mathbf{b}\bar{\mathbf{z}}^T = \beta\mathbf{b}\mathbf{b}^T + \gamma\mathbf{b}\mathbf{1_k}^T \tag{B.30}$$

such that

$$\bar{\mathbf{z}}\mathbf{b}^T - \mathbf{b}\bar{\mathbf{z}}^T = \gamma(\mathbf{1_k}\mathbf{b}^T - \mathbf{b}\mathbf{1_k}^T) \tag{B.31}$$

Analogously, we find from Equation B.28

$$\mathbf{1_k}\mathbf{a}^T - \mathbf{a}\mathbf{1_k}^T = \gamma(\mathbf{1_k}\mathbf{b}^T - \mathbf{b}\mathbf{1_k}^T) \tag{B.32}$$

Condition B.9 then follows by equating Equation B.31 and Equation B.32 .

■

PROOF OF THE PROPOSITION (NECESSITY): We now show that the above conditions (A) and (B) can be obtained from Equation B.10.

In the following we will need the basic fact that rescaling a matrix $\mathbf{C}$ and adding a multiple of $\mathbb{I}$ does not change the eigenvectors of $\mathbf{C}$. Namely, let $\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ be the EVD of $\mathbf{C}$ then $\forall\ \rho, \sigma \in \mathbb{R}$, $(\rho\mathbf{C} - \sigma\mathbb{I}) = \mathbf{V}(\rho\mathbf{D} + \sigma\mathbb{I})\mathbf{V}^T$ is the EVD of $(\rho\mathbf{C} + \sigma\mathbb{I})$. Note that this assertion can be proven analogously for nonsymmetric matrices and for $\rho, \sigma \in \mathbb{C}$.

As was shown in the above lemma the relation between $\mathbf{b}$, $\mathbf{1_k}$ and $\bar{\mathbf{z}}$, $\mathbf{a}$ can be written as

$$\bar{\mathbf{z}} = \beta\mathbf{b} + \gamma\mathbf{1_k} \tag{B.33}$$

$$\mathbf{a} = \gamma\mathbf{b} + \alpha\mathbf{1_k} \tag{B.34}$$

where the scalars β, γ and α are as in Equation B.26. Further we can write

$$\bar{\mathbf{z}} = \mathbf{P_k}\bar{\mathbf{z}} + \mu\mathbf{1_k} \tag{B.35}$$

with $\mu = \frac{\langle \bar{\mathbf{z}}, \mathbf{1_k} \rangle}{k}$. We restate the equations for $\mathbf{b}$ and $\mathbf{a}$, Equation B.6 and Equation B.8:

$$\begin{aligned} \mathbf{a} &= \frac{1}{k^2}(\mathbf{C} - m\bar{\mathbf{z}}\bar{\mathbf{z}}^T)(S_C\mathbb{I}_\mathbf{k} - k\mathbf{C})\mathbf{1_k} \\ \mathbf{b} &= -m\mathbf{C}\mathbf{P_k}\bar{\mathbf{z}} \end{aligned}$$

Inserting Equation B.6, Equation B.8, and Equation B.35 into the Equations B.33 and B.34 yields

$$(\gamma - \mu)\mathbf{1_k} = (m\beta\mathbf{C} + \mathbb{I}_\mathbf{k})\mathbf{P_k}\bar{\mathbf{z}} \tag{B.36}$$

$$[\frac{1}{k^2}\mathbf{C}(S_C\mathbb{I}_\mathbf{k} - k\mathbf{C}) - (\rho\mu + \alpha)\mathbb{I}_\mathbf{k}]\,\mathbf{1_k} = (\rho\mathbb{I}_\mathbf{k} - m\gamma\mathbf{C})\mathbf{P_k}\bar{\mathbf{z}} \tag{B.37}$$

where we used the abbreviation $\rho = m\bar{\mathbf{z}}^T(-\frac{1}{k}\mathbf{C} + \frac{S_C}{k^2}\mathbb{I}_\mathbf{k})\mathbf{1_k}$.

We first show that we must have $\gamma = \mu$. If $\mathbf{P_k}\bar{\mathbf{z}} = \mathbf{0}$ this is obvious. Thus let us assume $\mathbf{P_k}\bar{\mathbf{z}} \neq \mathbf{0}$. Under the assumption that $\gamma \neq \mu$ inserting Equation B.36 into Equation B.37 yields

$$\frac{1}{\gamma - \mu}[\frac{1}{k^2}\mathbf{C}(S_C\mathbb{I}_\mathbf{k} - k\mathbf{C}) - (\rho\mu + \alpha)\mathbb{I}_\mathbf{k}](m\beta\mathbf{C} + \mathbb{I}_\mathbf{k})]\,\mathbf{P_k}\bar{\mathbf{z}} = (\rho\mathbb{I}_\mathbf{k} - m\gamma\mathbf{C})\mathbf{P_k}\bar{\mathbf{z}} \tag{B.38}$$

According to the above notice this can hold only if $\mathbf{P_k}\bar{\mathbf{z}}$ is an eigenvector of $\mathbf{C}$ or if $\exists\,\sigma \geq 0$ so that $\mathbf{C} = \sigma\mathbb{I}_\mathbf{k}$. In these cases Equation B.36 reads

$$(\gamma - \mu)\mathbf{1_k} = (m\beta d_p + 1)\mathbf{P_k}\bar{\mathbf{z}} \tag{B.39}$$

where d_p is the eigenvalue to $\mathbf{P_k}\bar{\mathbf{z}}$ or $d_p = \sigma$ as above. For nonzero scalar factors $(\gamma - \mu)$ and $(m\beta d_p + 1)$ Equation B.39 implies $\mathbf{1_k} \parallel \mathbf{P_k}\bar{\mathbf{z}}$ which cannot hold for $\mathbf{P_k}\bar{\mathbf{z}} \neq \mathbf{0}$ since the matrix $\mathbf{P_k}$ projects onto the subspace orthogonal to $\mathbf{1_k}$. Therefore in any case $\gamma = \mu$, thus Equation B.36 becomes

$$\mathbf{0} = (m\beta\mathbf{C} + \mathbb{I}_\mathbf{k})\mathbf{P_k}\bar{\mathbf{z}} \tag{B.40}$$

Equation B.40 requires any of the following three statements to hold:

I. $\mathbf{P_k}\bar{\mathbf{z}}$ is an eigenvector of $\mathbf{C}$.

II. $\mathbf{P_k}\bar{\mathbf{z}} = \mathbf{0}$.

III. $\mathbf{C} = \sigma\mathbb{I}_\mathbf{k}$ with $\beta = -m\sigma^{-1}$.

III can be checked directly with Equation B.5 which is satisfied in this case if $\mathbf{1_k}$ and $\bar{\mathbf{z}}$ are orthogonal or parallel. Note that the second part of III follows from the first and is no additional requirement. Thus III produces special cases of the first two conditions. A third possibility, namely that $2k = \beta m S_C$, can be realized only for $(m\beta)^{-1} = 0$ since $S_C = k/\beta m$. $\mathbf{C} = \mathbf{0}$ was excluded, however, in the assumptions of the proposition.
The proof of necessity is completed by inserting I-II into Equation B.37 which reveals the following additional constraints which will be derived below.

(A-I) (a) $\mathbf{1_k}$ is an eigenvector of $\mathbf{C}$ to the eigenvalue $d_1 = \frac{S_C}{k}$ and furthermore $\mu = 0$ or $d_p = 0$. Or

(b) $\mathbf{1_k}$ is an eigenvector of $\mathbf{C}$ to the eigenvalue $d_1 = 0$ and furthermore $\mu = 0$ or $d_p = \frac{S_C}{k}$. Or

(c) $\mathbf{1_k}$ is contained in a subspace spanned by eigenvectors to eigenvalues $\in \{0, \frac{S_C}{k}\}$ and furthermore $\mu = 0$ or $d_p = \frac{1}{k}(S_C - \sum_{l=1}^{k} d_l\, y_l^2)$

(A-II) $\mathbf{1_k}$ is an eigenvector of $\mathbf{C}$. In this case there are no constraints of the corresponding eigenvalue.

(A-Ia) and (A-Ib) are special cases of (A-Ic) and will, hence, not be further considered. (A-Ic) together with I and (A-II) together with II make up the two conditions of the proposition. It remains to show that (A-Ic) and (A-II) follow from I and II.

(A-I). If $\mathbf{P_k\bar{z}}$ is an eigenvector of $\mathbf{C}$, Equation B.37 can be written as

$$[\frac{1}{k^2}\mathbf{C}(S_C\mathbb{I}_\mathbf{k} - k\mathbf{C}) - (\rho\mu + \alpha)\mathbb{I}_\mathbf{k}]\mathbf{1_k} = (\rho - m\mu d_p)\mathbf{P_k\bar{z}} \tag{B.41}$$

where d_p is the eigenvalue to $\mathbf{P_k\bar{z}}$. Inserting the EVD $\mathbf{C} = \mathbf{VDV}^T$ into Equation B.41 and multiplying by $\mathbf{V}^T$ from the left yields

$$\mathbf{\Lambda w} = \mathbf{0} \tag{B.42}$$

with

$$\mathbf{\Lambda} = \frac{1}{k^2}\mathbf{D}(S_C\mathbb{I}_\mathbf{k} - k\mathbf{D}) - (\rho\mu + \alpha)\mathbb{I}_\mathbf{k} \tag{B.43}$$

$$\mathbf{w} = \mathbf{y} - \frac{(\rho - m\mu d_p)}{\lambda_p}\|\mathbf{P_k\bar{z}}\|\,\mathbf{e_p} \tag{B.44}$$

where $\mathbf{e_p} = (0, \dots, 0, \underset{p}{1}, 0, \dots, 0)^T$, $\mathbf{y} = \big(\, y_i \,\big)_{k\times 1} = \mathbf{V}^T\mathbf{1_k}$, and $\mathbf{\Lambda}$ is a diagonal matrix with the diagonal elements

$$\lambda_i = \frac{1}{k^2}d_i(S_C - kd_i) - (\rho\mu + \alpha) \tag{B.45}$$

for $i \in \{1, \dots, k\}$. Note that $y_p = 0$ because $\mathbf{1_k} \perp \mathbf{P_k\bar{z}}$ and we assumed that $\mathbf{P_k\bar{z}}$ is an eigenvector of $\mathbf{C}$. A necessary and sufficient condition for Equation B.42 is that both

$$(\rho - m\mu d_p)\|\mathbf{P_k\bar{z}}\| = \mathbf{0} \tag{B.46}$$

and

$$\exists\, i, \quad i \neq p, \qquad \text{with} \qquad \lambda_i = 0 \tag{B.47}$$

hold. It is easy to see that Equation B.46 must hold independently of the value of λ_p, thus also in case of $\lambda_p = 0$. However if $\lambda_p = 0$ is the only zero diagonal element of $\mathbf{\Lambda}$ then $\mathbf{\Lambda y} = (\rho - m\mu d_p)\|\mathbf{P_k\bar{z}}\|\mathbf{e_p}$ must hold for item B.41 to be true, but since $y_p = 0$ and $\exists\, i \neq p \,:\, y_i \neq 0$ and $\big(\mathbf{e_p}\big)_i = 0\; \forall\, i \neq p$ this is impossible to fulfill. Hence the requirement B.47 is needed.

The number of nonzero elements in $\mathbf{y}$ determines the number of diagonal elements of $\mathbf{\Lambda}$ that must be zero. This leads to a system of quadratic equations of the form B.45 for the eigenvalues d_i of $\mathbf{C}$. Since the equations are quadratic there are two possible solutions for each d_i for which $\lambda_i = 0$ should hold, namely $d_i = \frac{1}{2}\big(S_C \pm \sqrt{S^2 - 4(\alpha - \mu\rho)}\big)$. Note that ρ and α depend on the d_i as well. Taking this into account the possible solutions become $d_i = 0$ and $d_i = \frac{S_C}{k}$ which is shown below. Since the number of nonzero elements in $\mathbf{y}$ equal the number of eigenvectors that are needed to cast $\mathbf{1_k}$ this proves the first part of A-Ic. The second part of A-Ic follows from item B.46.

Let us now prove that $\lambda_i = 0$ resolved w. r. t. d_i has only the two solutions $d_i = 0$ or $d_i = \frac{S_C}{k}$. We can express α and ρ in terms of $\mathbf{y}$ as

$$\begin{aligned} \alpha &= \frac{1}{k}\Big(\frac{S_C}{k^2} + m\mu^2\Big)\sum_{l=1}^{k} d_l y_l^2 - \frac{1}{k^2}\sum_{l=1}^{k} d_l^2 y_l^2 - \frac{mS_C}{k}\mu^2 && \text{(B.48)} \\ \rho &= \frac{m}{k}\mu\,\Big(S_C - \sum_{l=1}^{k} d_l\, y_l^2\Big) && \text{(B.49)} \end{aligned}$$

Inserting Equation B.48 and Equation B.49 into item B.45 and setting $\lambda_i = 0$ yields

$$0 = \frac{1}{k^2} d_i(S_C - kd_i) - \frac{1}{k^3}\sum_{l=1}^{k} d_l\,(S_C - kd_l)\, y_l^2 \tag{B.50}$$

Obviously $d_i = 0$ and $d_i = \frac{S_C}{k}$ are solutions of Equation B.50 also in the case where the latter is a system of equations and the d_i assume one of the two possible values $\in \{0, \frac{S_C}{k}\}$, but not necessarily all d_i the same. It can be proven that $\{0, \frac{S_C}{k}\}$ are the only possible solutions of Equation B.50 by considering N equations of the form Equation B.50 as a matrix equation

$$\mathbf{A}\mathbf{r} = \mathbf{0} \tag{B.51}$$

with the matrix $\mathbf{A} = \big(\frac{1}{k^2}(\delta_{ij} - \frac{1}{k}y_j^2)\big)_{N\times N}$ and the vector $\mathbf{r} = \big(\,d_i(S_C - kd_i)\,\big)_N$. Using the Laplace expansion and complete induction leads to

$$\det \mathbf{A} = \frac{1}{k^{2N}}\Big(1 + \frac{\|\mathbf{y}\|^2}{k}\Big) \neq 0 \tag{B.52}$$

and hence $\mathbf{r} = \mathbf{0}$ is the only solution of item B.51, thus it must be $d_i \in \{0, \frac{S_C}{k}\}$ for all i for which $\lambda_i = 0$ must hold.

Inserting Equation B.49 into item B.46 yields

$$m\mu\,\Big(\frac{S_C}{k} - \frac{1}{k}\sum_{l=1}^{k} d_l\, y_l^2 - d_p\Big)\,\|\mathbf{P_k}\bar{\mathbf{z}}\| = 0 \tag{B.53}$$

and since we assumed $\|\mathbf{P_k}\bar{\mathbf{z}}\| \neq 0$ we must have $\mu = 0$ or

$$d_p = \frac{1}{k}\,(S_C - \sum_{i=1}^{k}\, d_l\; y_l^2) \tag{B.54}$$

(A-II). follows immediately from Equation B.37 by inserting II. Because the right hand side of Equation B.37 is zero, $\mathbf{1_k}$ is an eigenvector of $\mathbf{C}$. Inserting Equation B.48 and Equation B.49 we further find that the eigenvalue d_1 equals $\frac{S_C}{k}$, which follows already from $\bar{\mathbf{z}}$ being parallel to $\mathbf{1_k}$, i. e. no further restriction to $\mathbf{C}$ is imposed.

Proof of the proposition (sufficiency): For the proof of the sufficiency of the conditions we can check directly that Equation B.5 holds. In case I, if $\bar{\mathbf{z}} = \mathbf{P_k}\bar{\mathbf{z}}$ is an eigenvector of $\mathbf{C}$, the first two "a. r. eq"-terms cancel out because of the specific eigenvalues of the subspace which contains $\mathbf{1_k}$, while the remaining terms are symmetric or even zero. If the condition (Ab) holds the first two of the not obviously symmetric terms in Equation B.5 cancel out as before. Out of those remaining, namely $m\left(-\mathbf{C} + \frac{1}{k}\mathbf{1_k}\mathbf{1_k}^T\mathbf{C} + \mathbf{C}\mathbf{1_k}\mathbf{1_k}^T - \frac{S_C}{k^2}\mathbf{1_k}\mathbf{1_k}^T\right)\bar{\mathbf{z}}\bar{\mathbf{z}}^T$ the second and the fourth cancel in the same way, while the other two sum up to zero because of the condition $d_p = \frac{1}{k}(S_C - \sum_{l=1}^{k}\, d_l\, y_l^2)$. In case II, where $\bar{\mathbf{z}}$ is parallel to $\mathbf{1_k}$ and both are eigenvectors of $\mathbf{C}$ it is obvious that all terms in Equation B.5 are symmetric.

■

B.2 Equality of the eigenimages

Since the interpretation of the rows of $\mathbf{X}$ as time courses was not explicitely used in the above considerations, we may state conditions for images analogously as a corollary of the proposition in the previous section. Although for the dimensionalities occurring in the presently analyzed data sets, $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ can be assumed to have full rank k, such an assumption has not been used in the proof. Therefore, the fact that the matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ is $m \times m$, but has the same rank as $\mathbf{C}$ is not critical. Thus we can adapt the conditions of the proposition as follows.

(C). $\mathbf{P_m}\bar{\mathbf{s}}$ is an eigenvector of $\mathbf{K}$, $\mathbf{1_m}$ is contained in a subspace spanned by eigenvectors of $\mathbf{K}$ with eigenvalues being either 0 or $\frac{S_K}{m}$, and

  (a) $\bar{\mathbf{s}} \perp \mathbf{1_m}$ or

  (b) $d_p^K = \frac{1}{m}(S_K - \sum_{l=1}^{m}\, d_l^K\, {y_l^K}^2)$.

(D). $\bar{\mathbf{s}} \parallel \mathbf{1_m}$ and $\mathbf{1_m}$ is an eigenvector of $\mathbf{K}$.

where d_l^K is an eigenvalue of $\mathbf{K}$, y_l^K denotes the scalar product of $\mathbf{1_m}$ and the l-th eigenvector of $\mathbf{K}$, and $\bar{\mathbf{s}}^T = \left(\frac{1}{k}\sum_{r=1}^{k} x_{ir}\right)$.

B.3 Equality of both time courses and images

It remains to be studied what consequences follow from the requirement that both time courses and images obtained as eigenvectors by temporal and spatial PCA being equal. Concerning the propositions in section B.1 and section B.2 both [(A) $\vee$ (B)] and [(C) $\vee$ (D)] must be satisfied.
Obviously, the sums of the components of $\bar{\mathbf{z}}$ and $\bar{\mathbf{s}}$, respectively, are related via the sum over all entries of $\mathbf{X}$

$$S_X := \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} x_{ij} = \frac{1}{k} \sum_{j=1}^{k} \bar{z}_j = \frac{1}{m} \sum_{i=1}^{m} \bar{s}_i \tag{B.55}$$

Therefore, the criteria $\bar{\mathbf{z}} \parallel \mathbf{1_k}$ and $\bar{\mathbf{s}} \perp \mathbf{1_m}$ are incompatible unless $S_X = 0$, since for $\bar{\mathbf{s}} \perp \mathbf{1_m}$ the components of $\bar{\mathbf{s}}$ sum up to zero. The same holds for $\bar{\mathbf{s}} \parallel \mathbf{1_m}$ and $\bar{\mathbf{z}} \perp \mathbf{1_k}$.
In the following we show that both the criteria (Aa) and (Ca) as well as (B) and (D) are in fact equivalent. If, according to (B) both $\bar{\mathbf{z}} \parallel \mathbf{1_k}$ and $\mathbf{1_k}$ is eigenvector of $\mathbf{C}$ we find that, because of $\mathbf{X}\mathbf{1_k} = \bar{\mathbf{s}}$, $\mathbf{1_k}$ is an eigenvector of $\mathbf{C}$ iff $\bar{\mathbf{s}}$ is an eigenvector of $\mathbf{K}$. Since $\bar{\mathbf{z}} \parallel \mathbf{1_k}$ was incompatible with $\bar{\mathbf{s}} \perp \mathbf{1_m}$, (B) is equivalent to (D). Equivalently, because $\mathbf{X}^T\mathbf{1_m} = \bar{\mathbf{z}}$, $\mathbf{1_m}$ is an eigenvector of $\mathbf{K}$, iff $\bar{\mathbf{z}}$ is an eigenvector of $\mathbf{C}$, thus (Aa) and (Ca) are equivalent.

Finally we do not assume that $\bar{\mathbf{z}}$ and $\mathbf{1_k}$ are parallel or orthogonal, i. e. we consider condition (Ab). We first note that if $\mathbf{P_k}\bar{\mathbf{z}}$ is an eigenvector of $\mathbf{C}$ then $\mathbf{P_m^K}\mathbf{1_m}$ is an eigenvector of $\mathbf{K}$ with eigenvalue d_p, where $\mathbf{P_m^K} = \mathbf{K} - \frac{1}{k}\bar{\mathbf{s}}\bar{\mathbf{s}}^T$. In the same way, the representation of $\mathbf{1_k}$ in terms of the eigenvalues of $\mathbf{C}$ transfers to a decomposition of $\bar{\mathbf{s}}$ which together with the condition on d_p in (Ab) implies that the eigenvalues of $\mathbf{K}$ obey (Cb). Combining these results it can be easily shown that $\mathbf{P_m}\bar{\mathbf{s}}$ is an eigenvector of $\mathbf{K}$.

Thus we have shown that for the cases (Aa) and (Ca) as well as for (B) and (D) the equality of the eigenimages for spatial and temporal PCA implies the equality of the eigen time courses and vice versa.

Appendix C

Correlation graphs

Correlation graphs occur in chapter 6, where correlation matrices are identified with adjacency matrices of weighted graphs, from which again by thresholding unweighted graphs are extracted. To analyze the significance of properties of the extracted graphs it would be of particular interest to contrast them against properties of correlation graphs, which are extracted from a random permutation of the elements of the correlation matrix. However, not all random permutations conserve the property of being a correlation matrix. To estimate the proportion correlation graphs from the number of graphs from random permutations is cumbersome. In this section we present a short sketch of the steps to determine the number of correlation graphs from a given distribution of correlation values.

First, it is convenient to map the correlation values into euclidean distances. This is possible with the interpretation of the correlation as the inner product of centered and normalized vectors. These vectors lie on a $k-1$-dimensional hypersphere and form the edges of a tetrahedron. Basic geometrical considerations then give the relation

$$d = \sqrt{2\,(1-c)} \tag{C.1}$$

where $d \in [0,2]$ is the euclidean distance and $c \in [-1,1]$ is the correlation value. The problem of how many permutations of elements of a given correlation matrix result in a correlation matrix again is thus equivalent to the question of how many high-dimensional tetrahedra can be built of a set of given distances $d_{ij},\ i,j \in \{1,\ldots,k\}$. Thereby we are not interested in isomorphic tetrahedra, since they are equivalent to isomorphic correlation graphs, which are not useful as contrasts. The maximum possible number of permutations leading to non-isomorphic tetrahedra is $\left(\binom{k}{2}\right)!/k!$. The question is now, for a given distribution of d_{ij}, how many permutations lead again to tetrahedra.

The Cayley-Menger determinant [100]

$$V_{k-1}^2\left(d_{ij}\right) = \frac{(-1)^k}{2^{k-1}\left((k-1)!\right)^2} \begin{vmatrix} 0 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & \cdots & d_{1k}^2 \\ 1 & d_{12}^2 & 0 & d_{23}^2 & \cdots & d_{2k}^2 \\ 1 & d_{13}^2 & d_{23}^2 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 & d_{k-1,k}^2 \\ 1 & d_{1k}^2 & d_{2k}^2 & \cdots & d_{k-1,k}^2 & 0 \end{vmatrix} \tag{C.2}$$

allows to calculate the squared volume of a $(k-1)$-dimensional tetrahedron from the mutual distances d_{ij} of its k corner points. Heron's formula appears as special case for $k=3$. The result of Equation C.2 is negative if any of the distances do not obey the triangular equation $d_{ij}+d_{jm} \geq d_{im}$, i.e. when no volume can be defined by the respective distances. Thus the number of tetrahedra that can be built by permutation is equal to the number of permutations for which the Cayley-Menger determinant is positive. Unless Equation C.2 is not significantly simplified, this is only a restatement of the problem, however, it could be used to determine a contrast correlation graph numerically, by permuting the elements of the correlation matrix and checking, if the corresponding Cayley-Menger matrix is positive.

Suppose that we are given a sample d_{ij} for which Equation C.2 is positive. Then for each triple of distances, d_{ir}, d_{jr}, d_{ij}, two of them being in the same row, and two of them being in the same column of the Cayley-Menger matrix, the triangular inequality must hold, i. e. we must have

$$|d_{ir}-d_{jr}| \leq d_{ij} \leq d_{ir}+d_{jr} \tag{C.3}$$

The requirement of the particular arrangement of the triple of distances in the Cayley-Menger matrix ensures that there are three corners to which the distances belong.

Each permutation can be written in terms of successive transpositions. From Equation C.3 we know that, if Equation C.2 is positive, two arbitrary distances d_{ab} and d_{ef} lie in the intervals

$$\begin{aligned} d_{ab} &\in [M_{ab}, m_{ab}] = I_{ab} && \text{(C.4)} \\ d_{ef} &\in [M_{ef}, m_{ef}] = I_{ef} && \text{(C.5)} \end{aligned}$$

with $M_{ab} = \max_r\{|d_{ar}-d_{br}|\}$ and $m_{ab} = \min_r\{d_{ar}+d_{br}\}$ and the same notations m. m. for d_{ef}. If $d_{ab}, d_{ef} \in I_{ab} \cap I_{ef}$, the triangle inequalities hold also after transposition. If $\forall\, i,j\ :\ \ d_{ij} \in [M,m] \wedge [M,m] \neq \emptyset$ with $M = \max_{a,b}\{M_{ab}\}$ and $m = \min_{a,b}\{m_{ab}\}$, all transpositions and hence permutations lead again to tetrahedra. Let $d = \min_{i,j}\{d_{ij}\}$ and $D = \max_{i,j}\{d_{ij}\}$. Then

$$D - d \leq M \qquad \text{and} \qquad m \leq 2d \tag{C.6}$$

From $[M,m] \neq \emptyset$ we must have $M < m$ and hence

$$D > 3d \tag{C.7}$$

Thus if the minimum distance d is smaller than one third of the maximum distance D all permutations lead again to tetrahedra. It is not clear, however, how the proportion of correlation graphs is affected, if Equation C.7 is almost fulfilled, as it is the case for the graphs of the data considered in chapter chapter 6. One could, however, conjecture that if $D - 3d \approx 0$ almost all graphs with the given weight distribution are correlation graphs and it is hence reasonable to use random graphs as contrast.

A rough estimate of the number of tetrahedra can be derived from a probabilistic point of view, i. e. if we do not restrict ourselves on the given realization of the d_{ij}, but consider their underlying probability density $p(x)$ with $p(x) = 0$ for $x < 0$, from which they are assumed to be taken as i. i. d. samples.
We are interested in the probability that k sampled distances d_s, $s \in \{1, \ldots, k\}$, satisfy the triangle inequality, i. e. that Equation C.3 holds. From Equations Equation C.5 we have, that Equation C.3 holds if

$$d_s \in [M, m] \;\; \forall d_s \tag{C.8}$$

thus, if d_s lie in the interval $[M, m]$, where $M = \max\limits_{s_1,s_2}\{|d_{s_1} - d_{s_2}|\}$ and $m = \min\limits_{s_1,s_2}\{d_{s_1} + d_{s_2}\}$. m and M are distributed according to the maximum and minimum of a set of random variables. With the respective integrated distributions P this writes

$$\begin{aligned} P_{\text{min}}(m) &= 1 - (1 - P(m))^k && \text{(C.9)} \\ P_{\text{max}}(M) &= P(M)^k && \text{(C.10)} \end{aligned}$$

where $P_{\text{min}}(m)$ is the probability, that the minimum of a set of random variables is less or equal m, and the definition of $P_{\text{max}}(M)$ reads correspondingly for the maximum. m is a sum and M a difference of random variables. The sum of two independent random variables is proportional to the convolution of the individual probability densities, and hence we have

$$\begin{aligned} P(m) &\propto \int_{-\infty}^{m} \int_{-\infty}^{\infty} p(y)p(x-y)dydx && \text{(C.11)} \\ P(M) &\propto \int_{-\infty}^{M} \int_{-\infty}^{\infty} p(y)p(|x+y|)dydx && \text{(C.12)} \end{aligned}$$

Thus the probability for a value d_s to be in the interval $[M, m]$ can be written as

$$W(M, m) = \int_{M}^{m} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x)dP_{\text{max}}(M)dP_{\text{min}}(m) \tag{C.13}$$

Now we have to take into account that the relation C.8 has to be satisfied for all rows and columns of the (symmetrical) matrix $\left(d_{ij}\right)_{k\times k}$. We arrive thus at the result

$$W = \prod_{n=1}^{k} W(M_n, m_n). \tag{C.14}$$

which is the probability of k samples of size k forming a tetrahedron (assuming that the validity of all triangular inequalities is a sufficient condition for the distances forming a tetrahedron). Note that we have made assumptions of independence, which in general do not hold, such as independence of one sample of the next and independence of m, M from the samples. Thus W can only be an approximation for the distribution of tetrahedra.

References

[1] R.R. Ernst A. Kumar, D. Welti. Nmr fourier zeugmatography. *J. Magn. Reson.*, 18:69–83, 1975. 1

[2] K. Ord A. Stuart. *Classical inference and the linear model*, volume 2 of *Kendall's Advanced Theory of Statistics*. Halsted Press, 1994. 26, 27

[3] K. Ord A. Stuart. *Distribution theory*, volume 1 of *Kendall's Advanced Theory of Statistics*. Halsted Press, 1994. 53

[4] Erkki Oja Aapo Hyvärinen, Juha Karhunen. *Independent Component Analysis*. John Wiley & Sons, Inc., 2001. 47

[5] S. Amari. Natural gradient learning for over- and under-complete bases in ica. *Neural Computation*, 11(8):1875–1883, 1999. 48

[6] Terrence J. Sejnowski Anthony J. Bell. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7:1129–1159, 1995. 48, 52

[7] K. Arfanakis, D. Cordes, V. M. Haughton, C. H. Moritz, M. A. Quigley, and M. E. Meyerand. Combining independent component analysis and correlation analysis to probe interregional connectivity in fmri task activation datasets. *Magnetic Resonance Imaging*, 18(8):921–930, 2000. 61, 62

[8] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. S. Hyde. Processing strategies for time-course data sets in functoinal mri of the human brain. *Magn. Reson. Med.*, 30:161, 1993. 62

[9] P. A. Bandettini, E. C. Wong, R. S. Hinks, R. S. Tikofsky, and J. S. Hyde. Time course epi of human brain function during task activation. *Magn Reson Med*, 25:390–397, 1992. 1

[10] R. Baumgartner. Quantification in functional magnetic resonance imaging: Fuzzy clustering vs. correlation analysis. *Magnetic Resonance Imaging*, 16(2):115–125, 1998. 62

[11] A. Baune. Dynamical cluster analysis of cortical fmri activation. *Neuroimage*, 9(5):477–489, 1999. 62

[12] Bharat Biswal, F. Zerrin Yetkin, Victor M. Haughton, and James S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34:537–541, 1995. 62

[13] A. M. Blamire, S. Ogawa, K. Ugurbil, D. Rothman, G. McCarthy, J. M. Ellermann, F. Hyder, Z. Rattner, and R. G. Shulman. Dynamic mapping of the human visual cortex by high-speed magnetic resonance imaging. *Proc Natl Acad Sci USA*, 89:11069–11073, 1992. 1

[14] F. Bloch. Nuclear induction. *Phys. Rev.*, 70:460–474, 1946. 1

[15] B. Bollobas. *Modern Graph Theory*. Number 184 in Graduate Texts in Mathematics. Springer, 1998. 67

[16] I. N. Bronstein, K.A. Semendjajew, G. Musiol, and H. Mühlig. *Taschenbuch der Mathematik*. Verlag Harri Deutsch, 1993. 113

[17] R. B. Buxton. The elusive initial dip. *NeuroImage*, 13:953–958, 2001. 8

[18] R. B. Buxton and L. R. Frank. A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *Journal of Cerebral Blood Flow & Metabolism*, 17(1):64–72, 1997. 8

[19] R. B. Buxton, E. C. Wong, and L. R. Frank. Dynamics of blood flow and oxygenation changes during brain activation - the balloon model. *Magnetic Resonance in Medicine*, 39(6):855–864, 1998. 8

[20] W. Weaver C. E. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, 1963. 50

[21] J. Cao and K. Worsley. The geometry of correlation fields with an application to functional connectivity of the brain. *Annals of Applied Probability*, 9(4):1021–1057, 1999. 62

[22] A. Cichocki, J. Karhunen, W. Kasprzak, and R. Vigario. Neural networks for blind separation with unknown number of sources. *Neurocomputing*, 24(1-3):55–93, 1999. 48

[23] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. 47, 52, 53

[24] D. Cordes, V. Haughton, J. D. Carew, K. Arfanakis, and K. Maravilla. Hierarchical clustering to measure connectivity in fmri resting-state data. *Magnetic Resonance Imaging*, 20(4):305–317, 2002. 62

[25] D. Cordes, V. M. Haughton, K. Arfanakis, J. D. Carew, P. A. Turski, C. H. Moritz, M. A. Quigley, and M. E. Meyerand. Frequencies contributing to functional connectivity in the cerebral cortex in ”resting-state” data. *American Journal of Neuroradiology*, 22(7):1326–1333, 2001. 62

[26] D. Cordes, V. M. Haughton, K. Arfanakis, G. J. Wendt, P. A. Turski, C. H. Moritz, M. A. Quigley, and M. E. Meyerand. Mapping functionally related regions of brain with functional connectivity mr imaging. *American Journal of Neuroradiology*, 21(9):1636–1644, 2000. 62

[27] C. Jutten D. T. Pham, P. Garat. Separation of a mixture of independent sources through a maximum likelihood approach. *Signal Processing VI, Theories and Applications*, pages 771–774, 1992. 47, 52

[28] R. Damadian. Tumor detection by nuclear magnetic resonance. *Science*, 171:1151–1153, 1971. 1

[29] R. Diestel. *Graph Theory*. Springer, 2 edition, 2000. 66, 67

[30] S. Dodel, J. M. Herrmann, and T. Geisel. Components of brain activity - data analysis for fmri. In *Proc. ICANN 99*, volume 2, pages 1023–1028, 1999. 31

[31] S. Dodel, J. M. Herrmann, and T. Geisel. Comparison of temporal and spatial ica in fmri data analysis. *Proc. ICA 2000, Helsinki, Finland*, pages 543–547, 2000. 57

[32] S. Dodel, J. M. Herrmann, and T. Geisel. Localization of brain activity - blind separation for fmri data. *Neurocomputing*, 32-33:701–708, 2000. 54

[33] S. Dodel, J. M. Herrmann, and T. Geisel. *Emergent Neural Computational Architectures Based on Neuroscience - Towards Neuroscience-Inspired Computing. Lecture Notes in Computer Science 2036.*, chapter Stimulus-Independent Data Analysis for fMRI., pages 39–53. Springer, London, 2001. 54, 56

[34] S. Dodel, J. M. Herrmann, and T. Geisel. Is brain activity spatially or temporally correlated? In *HBM 2001 Proceedings*, page 110, 2001. 31

[35] S. Dodel, J. M. Herrmann, and T. Geisel. Temporal versus spatial pca and ica in data analysis for fmri. In *28th Göttingen Neurobiology Conference*, page 897. Thieme Verlag, 2001. 57

[36] S. Dodel, J. M. Herrmann, and T. Geisel. Functional connectivity by cross-correlation clustering. *Neurocomputing*, 44-46:1065–1070, 2002. 63

[37] S. Dodel, J. M. Herrmann, and T. Geisel. Functional connectivity clusters from correlation graphs. In *HBM 2002 Proceedings*, 2002. 63

[38] R.V. Pound E.M. Purcell, H.C. Torrey. Resonance absorption by nuclear magnetic moments in a solid. *Phys. Rev.*, 69:37–38, 1946. 1

[39] M. Packard F. Bloch, W.W. Hansen. Nuclear induction. *Phys. Rev.*, 69:127, 1946. 1

[40] Jürgen Finsterbusch. *Linienselektive Bildgebungsverfahren in der Magnetresonanz-Tomografie*. PhD thesis, Georg-August-Universität zu Göttingen, 1999. 1, 4, 5, 7, 8

[41] P. T. Fox and M. E. Raichle. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proc. Natl. Acad. Sci. USA*, 83(83):1140–1144, 1986. 7

[42] P. T. Fox, M. E. Raichle, M. A. Mintun, and C. Dence. Nonoxidative glucose consumption during focal physiologic neural activity. *Science*, 241:462–464, 1988. 7

[43] J. Frahm. Magnetic resonance functional neuroimaging: New insights into the human brain. *Current Science*, 76:735–743, 1999. 11, 21, 29

[44] J. Frahm, H. Bruhn, K. D. Merboldt, and Hnicke W. Dynamic mri of human brain oxygenation during rest and photic stimulation. *J Magn Reson Imaging*, 2:501–505, 1992. 1

[45] J. Frahm, G. Kruger, K. D. Merboldt, and A. Kleinschmidt. A. dynamic uncoupling and recoupling of perfusion and oxidative metabolism during focal brain activation in man. *Magn. Reson. Med.*, 35:143–148, 1996. 8

[46] Jens Frahm, Peter Fransson, and Gunnar Krüger. *Modern Techniques in Neuroscience Research*, chapter Magnetic Resonance Imaging of Human Brain Function, pages 1055–1082. Springer, Berlin, 1999. 1, 8

[47] P. Fransson, G. Krüger, K. Merboldt, and J. Frahm. Temporal characteristics of oxygenation-sensitive mri responses to visual activation in humans. *Magn. Reson. Med.*, 41:436–441, 1999. 8

[48] P. Fransson, G. Krüger, K. D. Merboldt, and Frahm J. Temporal and spatial mri responses to subsecond visual activation. *Magnetic Resonance Imaging*, 17(1):1–7, 1999. 8

[49] K. J. Friston, C. Buechel, G. R. Fink, J. Morris, E. Rolls, and R. J. Dolan. Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, 6(3):218–229, 1997. 62

[50] K. J. Friston, C. D. Frith, P. Fletcher, P. F. Liddle, and R. S. J. Frackowiak. Functional topography - multidimensional scaling and functional connectivity in the brain. *Cerebral Cortex*, 6(2):156–164, 1996. 62

[51] K. J. Friston, A. Mechelli, R. Turner, and C. J. Price. Nonlinear responses in fmri: The balloon model, volterra kernels, and other hemodynamics. *NeuroImage*, 12:466–477, 2000. 8, 22

[52] K. J. Friston, J. Phillips, D. Chawla, and C. Büchel. Nonlinear pca: characterizing interactions between modes of brain activity. *Phil. Trans. R. Soc. Lond. B*, 355:135–146, 2000. 31

[53] David G. Gadian. *NMR and its applications to living systems*. Oxford University Press, 1995. 7

[54] John Gill. Ee 376a: Information theory, lecture notes. http://www.stanford.edu/class/ee376a/handouts/lect02.pdf (Nov. 2002). URL: `http://www.stanford.edu/class/ee376a/handouts/lect02.pdf`. 52

[55] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, Inc., 1994. 113

[56] M. Hampson, B. S. Peterson, P. Skudlarski, J. C. Gatenby, and J. C. Gore. Detection of functional connectivity using temporal correlations in mr images. *Human Brain Mapping*, 15(4):247–262, 2002. 62

[57] Simon Haykin. *Blind deconvolution*, volume 2 of *Unsupervised adaptive filtering*. John Wiley & Sons, Inc., 2000. 47

[58] Joseph P. Hornak. The basics of mri. http://www.cis.rit.edu/htbooks/mri (Oct. 2002). URL: `http://www.cis.rit.edu/htbooks/mri`. 1, 2

[59] A. Hyvärinen. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. 47

[60] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. 47

[61] Aapo Hyvärinen. Independent component analysis by minimization of mutual information. Technical Report A46, Helsinki University of Technology, Finland, 1997. 48, 51, 52

[62] S. Millman I. Rabi, J.R. Zacharias and P. Kusch. A new method of measuring nuclear magnetic moment. *Phys. Rev.*, 53:318, 1938. 1

[63] A. Souloumiac J. F. Cardoso. Blind beamforming for non gaussian signals. *IEE-Proceedings-F*, 140(6):362–370, 1993. 48, 53

[64] Jay Yellen Jonathan Gross. *Graph Theory and its Applications*. CRC Press, 1998. 67

[65] S. Y. Kung K. I. Diamantaras. *Principal Component Neural Networks*. John Wiley & Sons, Inc., 1996. 33

[66] D. S. Kim, D. Q. Duong, and S. G. Kim. High resolution of iso-orientation columns by fmri. *Nature Neuroscience*, 3:164–169, 2000. 8

[67] M. A. Koch, D. G. Norris, and M. Hund-Georgiadis. An investigation of functional and anatomical connectivity using magnetic resonance imaging. *NeuroImage*, 16(1):241–250, 2002. 62

[68] T. Kohonen. *Self-organizing maps*. Springer, 1995. 11

[69] G. Kruger, A. Kleinschmidt, and J. Frahm. Dynamic mri sensitized to cerebral blood oxygenation and flow during sustained activation of human visual cortex. *Magn. Reson. Med.*, 35:797–800, 1996. 8

[70] K.K. Kwong, J.W. Belliveau, D.A. Chesler, I.E. Goldberg, R.M. Weisskoff, B.P. Poncelet, D.N. Kennedy, B.E. Hoppel, M.S. Cohen, R. Turner, H. M. Chang, Brady T. J., and B. R. Rosen. Dynamic magneticresonance imaging of human brain activity during primary sensory stimulation. *Proc Natl Acad Sci U S A*, 89(12):5675–5679, 1992. 1

[71] P.G. Lauterbur. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*, 242:190–191, 1973. 1

[72] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000. 48

[73] Bart De Moor Lieven De Lathauwer and Joos Vandewalle. Independent component analysis and (simultaneous) third-order tensor diagonalization. *IEEE Transactions on Signal Processing*, 49(10):2262–2271, 2001. 48, 53

[74] Ralph Linsker. An application of the principle of maximum information preservation to linear systems. In *Advances in Neural Information Processing Systems*, volume 1, pages 186–194, 1989. 52

[75] Ralph Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4:691–702, 1992. 52

[76] T. T. Liu, W. M. Luh, E.C. Wong, P. A. Bandettini, T. Obata, L. R. Frank, and R. B. Buxton. On the nonlinear relation between bold and cbf. *Proc. Intl. Soc. Mag. Reson. Med.*, 8:948, 2000. 8

[77] N. K. Logothetis, H. Guggenberger, S. Peled, and J. Pauls. Functional imaging of the monkey brain. *Nature Neuroscience*, 2:555–562, 1999. 8

[78] Nikos K. Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412:150–157, 2001. 7, 21

[79] G. Lohmann and S. Bohn. Using replicator dynamics for analyzing fmri data of the human brain. *IEEE Transactions on Medical Imaging*, 21(5):485–492, 2002. 62

[80] M. J. Lowe, M. Dzemidzic, J. T. Lurito, V. P. Mathews, and M. D. Phillips. Correlations in low-frequency bold fluctuations reflect cortico-cortical connections. *NeuroImage*, 12(5):582–587, 2000. 62

[81] P. Mansfield. Multi-planar image formation using nmr spin-echos. *J. Phys. C: Solid State Physics*, 10:L55–L58, 1977. 1

[82] P. McCullagh. *Tensor Methods in Statistics*. Chapman and Hall, New York, 1987. 52, 53

[83] A. R. McIntosh. Mapping cognition to the brain through neural interactions. *Memory*, 7(5-6):523–548, 1999. 62

[84] A. R. McIntosh and F. Gonzalez-Lima. Structural equation modeling and its application to network analysis in functional brain imaging. *Brain Map*, 2:2–20, 1994. 62

[85] A. R. McIntosh, L. Nyberg, F. L. Bookstein, and E. Tulving. Differential functional connectivity of prefrontal and medial temporal cortices during episodic memory retrieval. *Human Brain Mapping*, 5(4):323–327, 1997. 62

[86] M. J. McKeown, T. P. Jung, S. Makeig, G. G. Brown, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski. Analysis of fmri data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188, 1998. 36

[87] M. L. Mehta. *Random Matrices.* Academic Press, London, 2 edition, 1991. 41, 42

[88] V. H. Vu N. Alon, M. Krivelevich. On the concentration of eigenvalues of random symmetric matrices. *Israel Journal of Mathematics*, 131:259–268, 2002. 42, 43

[89] J.-P. Nadal and N. Parga. Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer. *Network*, 5:565–581, 1994. 52

[90] S. Ogawa, TM. Lee, A. R. Kay, and Tank D. W. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci USA*, 87:9868–9872, 1990. 1

[91] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 1984. 21, 32, 109

[92] A. J. Peltier and D. C. Noll. T2* dependence of low frequency functional connectivity. *NeuroImage*, 16(4):985–992, 2002. 61, 62

[93] Karl Magnus Petersson Thomas E. Nichols Jean-Baptiste Poline and Andrew P. Holmes. Statistical limitations in functional neuroimaging. ii. signal detection and statistical inference. *Phil. Trans. R. Soc. Lond. B*, 354:1261–1281, 1999. 29

[94] N. U. Prabhu. *Stochastic processes.* Collier-Macmillan, 1965. 64

[95] W. H. Press, editor. *Numerical recipes in C.* Cambridge, Univ. Press, 1996. 33, 64, 65

[96] Friston KJ Holmes AP Worsley KJ Poline JP Frith CD Frackowiak RSJ. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995. 19, 26, 29

[97] Ogawa S, Tank DW, Menon R, Ellermann JM, Kim SG, Merkle H, and Ugurbil K. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc Natl Acad Sci USA*, 89:5951–5955, 1992. 1

[98] A. Rucinski S. Janson, T. Luczak. *Random Graphs.* John Wiley & Sons, Inc., 2000. 71

[99] C. P. Slichter. *Principles of Magnetic Resonance.* Springer Series in Solid-State Sciences. Springer, 1978. 1

[100] D. M. Y. Sommerville. *An Introduction to the Geometry of N Dimensions.* Dover, New York, 1958. 127

[101] A. Soshnikov. Universality of edge of the spectrum in wigner random matrices. *Commun. Math. Phys.*, 207:697–733, 1999. 41

[102] O. Sporns, G. Tononi, and G. M. Edelman. Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex*, 10:127–141, 2000. 62, 107

[103] T. Stein, C. Moritz, M. Quigley, D. Cordes, V. Haughton, and E. Meyerand. Functional connectivity in the thalamus and hippocampus studied with functional mr imaging. *American Journal of Neuroradiology*, 21(8):1397–1401, 2000. 62

[104] J. A. Thomas T. M. Cover. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991. 50, 51

[105] McKeown MJ. Sejnowski TJ. Independent component analysis of fmri data - examining the assumptions. *Human Brain Mapping*, 6(5-6):368–372, 1998. 47

[106] Irene Tracey. Brief introduction to fmri - physiology. http://www.fmrib.ox.ac.uk/fmri_intro/physiology.html (Oct. 2002). URL: `http://www.fmrib.ox.ac.uk/fmri_intro/physiology.html`. 7

[107] R. Turner, D. Le Bihan, CTW. Moonen, D. DesPres, and J. Frank. Echo-planar time course mri of cat brain oxygenation changes. *Magn Reson Med*, 22:159–166, 1991. 1

[108] R.C. Weast, editor. *Handbook of Chemistry and Physics.* Chemical Rubber Company, Cleveland, Ohio, 1972. 2

[109] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math.*, 62:548–564, 1955. 41

[110] E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Ann. Math.*, 67:325–328, 1958. 41

[111] K. J. Worsley, J. Cao, T. Paus, M. Petrides, and A. C. Evans. Applications of random field theory to functional connectivity. *Human Brain Mapping*, 6(5-6):364–367, 1998. 62

[112] K.J. Worsley. Local maxima and the expected euler characteristic of excursion sets of chi^2, f and t fields. *Advances in Applied Probability*, 26:13–42, 1994. 29

[113] K.J. Worsley. Estimating the number of peaks in a random field using the hadwiger characteristic of excursion sets, with applications to medical images. *Annals of Statistics*, 23:640–669, 1995. 29

[114] A. M. Yaglom, editor. *An Introduction to the Theory of Stationary Random Functions.* Prentice-Hall, Inc., 1962. 64

[115] X. Zhao, R.W. Cox, W.-M. Luh, and S.-J. Li. Clique analysis to study functional connectivity. In *Proc. Soc. Magn. Reson.*, volume 3, page 1479, 1998. 62

Acknowledgement

First of all I wish to thank Prof. Theo Geisel for the opportunity to join his group and the extraordinary environment of open mindedness and freedom along with the pleasant surrounding he created. Further, I am very much indebted to Michael Herrmann for the support he gave me particularly during the final stage of my thesis, for endless discussions, nightly proofreading, and great sushi.

I thank Prof. Jens Frahm and his group, particularly Jürgen Baudewig, Peter Dechent, Jürgen Finsterbusch, and Peter Fransson for providing us with fMRI data and for helpful discussions. Jürgen Finsterbusch I thank for proofreading, helpful suggestions and providing me with the figures for the chapter about fMRI.

I thank Hans-Ulrich Bauer for having introduced me to the institute and supported my work in the first months of being here. Ragnar Fleischmann during the years was a cheerful friend, sharing the preference for good cooking (he makes the best Tiramisu in the world), and was always ready to help with any problems that occured. By telling me the secret of formatting figure captions Udo Ernst has saved much of my time he spent on the other hand in the most pleasant way by continuously showing up with cakes from Cron&Lanz knowing that I couldn't resist to.

I am grateful to Fred Wolf for reading part of my manuscript and Markus Diesmann for helping with C++ related problems. Wolf-Dieter Brandt, Denny Fliegner and Yorck-Fabian Temme did a great job as system administrators in keeping the computers running. I enjoyed the discussions with Hans Ekkehard Pleßer, Matthias Kaschube, Marc-Felix Otto, Michael Schnabel, and Mathias Puhlmann, with whom I shared the office. Marc Timme was very helpful in finishing his thesis about the same time and providing me with all the necessary information as to the administrative procedure. Björn Naundorf I thank for his articles in the journal of applied bakery. The secretaries Corinna Trautsch, Rita Bartels, Regina Wunderlich, Alexandra Behling, and Agnes Bleile were always very friendly and supportive in administrative struggles. It has been a pleasure to have as colleagues Matthias Bethge, Dmitri Bibitchkov, Dirk Brockmann, Michael Buschermöhle, Michael Denker, Moritz Hiller, Sven Gödeke, Lars Hufnagel, Roland Ketzmerick, Tsampikos Kottos, Karsten Kruse, Stefan Liehr, Norbert Mayer, Abigail Morrison, Alexander Ossipov, Klaus Pawelzik, Felix Petri, Frédéric Piéchon, Holger Schanz, Dennis Springsguth, Frank Steinbach, Tom Tetzlaff, Mathias Weiß, and Alexander Zumdieck.

Finally I would like to thank my family for their support in many ways during the years.

Curriculum Vitae

Name	Silke Dodel
Nationality	German
28.11.1969	born in München

1975-1977	Grund- und Hauptschule Büsnau
1977-1978	École Jean-Moulin, Massy
1978-1979	Grund- und Hauptschule Büsnau
1979-1988	Fanny-Leicht-Gymnasium, Stuttgart-Vaihingen
1988	Abitur

1988-1993	Studies in Mathematics, Russian Philology and Theology at the Universities of Tübingen and Stuttgart
1993-1997	Studies in Mathematics and Physics at the Universities of Tübingen, Cologne and Tampere (Finland)
1995-1996	Master's Thesis at the Tampere University of Technology in Finland (Solution of the Inverse Problem in EEG Measurements using Statistical Methods)
1997	Staatsexamen in Mathematics and Physics at the University of Tübingen
2000	Staatsexamen in Russian Philology at the University of Göttingen
1997-2002	Max-Planck-Institut für Strömungsforschung in Göttingen, Department of Nonlinear Dynamics headed by Prof. Theo Geisel