Blind Source Separation Algorithms for the Analysis of Optical Imaging Experiments

vorgelegt von
Diplom Physiker
Ingo Schießl
aus Grafenau

Vom Fachbereich 13 - Informatik
Institut für Softwaretechnik und Theoretische Informatik der
Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften

- Dr. rer. nat.-

genehmigte Dissertation

Promotionsausschuß:

Vorsitzender: Prof. Günter Hommel
Berichter: Prof. Klaus Obermayer
Berichter: Prof. Jennifer Lund

Tag der wissenschaftlichen Aussprache: 17. Mai 2001

Berlin 2001

D83

Blind Source Separation Algorithms for the Analysis of Optical Imaging Experiments

–

Imaging and Analysis

Ingo Schießl

– PhD Thesis –

2001

Technische Universität Berlin
Fachbereich Informatik
Institut für Softwaretechnik und Theoretische Informatik
Fachgebiet Neuronale Informationsverarbeitung
Franklinstraße 28–29
D-10587 Berlin

Zusammenfassung

Im Rahmen dieser Arbeit wurde ein Algorithmus zur blinden Quellentrennung mit dem Namen “Extended Spatial Decorrelation (ESD)” entwickelt und mit bekannten “Independent Component Analysis (ICA)” Verfahren bezüglich der Robustheit gegen statistische Abhängigkeit der Quellen und Rauschen anhand von künstlichen Testdaten und “Optical Imaging (OI)” Daten verglichen. Optical Imaging von intrinsischen Signalen ist ein bildgebendes Verfahren bei dem die neuronale Aktivität in der Großhirnrinde nicht direkt, sondern anhand von mit der Stimulierung einhergehenden Veränderungen in den Streueigenschaften des Gewebes optisch gemessen wird. Diese intrinsischen Signale entstehen durch die Stoffwechselaktivität der Nervenzellen. Andere medizinische bildgebende Verfahren wie die Kernspintomographie basieren auf der Meßung derselben intrinsischen Signale und somit sind die in dieser Arbeit gewonnenen Erkenntnisse zum Teil direkt auf solche Verfahren transferierbar.

Das Problem bei der Auswertung der Daten besteht darin, daß die Meßung eine lineare Mischung aus der lokal gebundenen Stimulus spezifischen Antwort der Nervenzellen, dem “mapping signal”, und der gröberen metabolischen Aktivierung, dem “global signal”, sowie biologische Schwankungen und Rauschen enthält.

Zur Extrahierung des “mapping signals” bieten sich Verfahren zur blinden Quellentrennung an, da sie ohne Vorwissen über den Mischvorgang eine parameterfreie Lösung des Problems bieten. Vorraussetzung für die Anwendung von ICA ist die statistische Unabhängigkeit der Quellen. Diese Algorithmen verwenden die Optimierung von Eigenschaften der statistischen höheren Momente um die Komponenten wieder unabhängig zu machen und somit die Quellen zu schätzen.

Das hier vorgestellte ESD Verfahren benutzt zur Trennung nur Statistik zweiter Ordnung und stellt somit geringere Anforderungen an die Quellen. Die Annahme bei ESD besteht in der Autokorreliertheit der Quellen und der geringen Korrelation untereinander. Das selbe gilt für eine räumlich verschobenen Version.

Da man bei biologischen Signalen oft die Quellen nicht genau kennt, muß man einen quantitativen Vergleich der Entmischungen durch die berücksichtigten Verfahren an künstlichen Datensätzen durchführen. Bei der Untersuchung mit drei Testdatensätzen, mit verschiedenen statistischen Eigenschaften, wurde die Qualität der Entmischung bezüglich des Signal-Rausch-Verhältnisses getestet. Es zeigt sich, daß die verschiedenen Implementierungen des ESD Verfahrens bei der Trennung von künstlichen Daten mit den statistischen Eigenschaften des biologischen Signals den ICA Methoden überlegen sind.

Dieses Ergebniss bestätigt sich auch bei der Analyse der OI Daten. Dabei wurden Experimente untersucht, deren räumliche Antwort in der Sehrinde auf ein bestimmtes Stimulusregime bekannt ist. Zusätzlich wurden zur Beurteilung die aus Projektionen gewonnenen Zeitverläufe der geschätzten Quellen herangezogen.

Mittels der ESD Analyse können selbst die signalschwachen “single condition images” berechnet werden ohne Modellannahmen über den Mischprozeß oder die modulare Organisatzion der Sehrinde zu machen.

Acknowledgements

First of all I would like to thank Prof. Klaus Obermayer for giving me the opportunity to do the research for this thesis in his group. He provides a highly scientific environment and accelerates the development of everybody in his group by the standards he sets. The possibility to present the research on international level and the many international scientific contacts enriched this interdisciplinary work and will prove fruitful in the future.

Furthermore I want to thank Dr. Martin Stetter for the permanent scientific support during the last years and the inspiring discussions about the theoretical and biological aspects of this work.

Special thanks go to Prof. Jenny Lund and Prof. John Mayhew who patiently spent many hours of their precious time to introduce me to the biological aspects of this work and the experimental particularities. Their humane approach to the difficulties of the daily routine of such a interdisciplinary collaboration was encouraging for all of us.

Dr. Niall McLoughlin spent a lot of time to teach me about the details of the optical imaging experiment. I would like to thank him for the amicable working conditions at UCL London, the many inspiring discussions and the critical comments towards the publications.

In addition it was a pleasure to learn and work with the following people at UCL: Dr. Jonathan Levitt, Dr. Chris Tyler, Dr. Alessandra Angelucci, Dr. Achim Rumberger and Dr. Daniel Glaser.

Also I want to thank my colleagues at NI among which I like to particularly mention Dr. Peter Adorjan, Christian Piepenbrock, Dr. Michael Scholz, Hauke Bartsch and Roland Vollgraf. Special thanks to Roland Vollgraf for his comments to the theory sections of this thesis.

And finally I want to thank my parents Dr. Barbara Schießl and Dr. Anton Schießl for their never ending support that made all this possible for me. I want to dedicate this work to my father who died far to young in 1998.

List of Symbols

Symbol	Meaning
P_x	number of pixels in x direction
P_y	number of pixels in y direction
$P = P_x \times P_y$	number of pixels
t	time
t_1, t_2	start, end of stimulus
M_0	number of frames of rawdata
M_1	number of frames after binning
M	number of frames after binning and first frame analysis
$\mathbf{r}$	space vector in frame
$\mathbf{x}$	framestack and mixtures
$x_m(\mathbf{r})$	single frame in stack
$\mathbf{s} = s_j(\mathbf{r})$	sources
$\hat{\mathbf{s}}$	source estimates
$p(x)$	probability density function of x
μ	mean
σ^2	variance
σ	standart deviation
$\mathbf{C_x}$	covariance matrix
ρ_{xy}	correlation coefficient
Δ^2	Mahalanobis distance
$\mathbf{v}$	eigenvectors of $\mathbf{C_x}$
δ_{ij}	Cronecker delta
$\mathbf{B}, \mathbf{W}$	demixing matrix
$\mathbf{P}$	permutation matrix
$\mathbf{y}$	sphered dataset
$\mathbf{D}$	sphering matrix
H	Entropy
$kurt$	kurtosis
$\mathbf{C_s}$	cross correlation matrix
${\mathbf{C_s}}^{lm}$	element of cross correlation matrix
$\mathbf{A}$	mixing matrix
$\hat{\mathbf{A}}$	estimated mixing matrix
$\mathbf{I}$	identity matrix
$\mathbf{v}_D$	eigenvectors after sphering in ESD
λ_D	eigenvalues after sphering in ESD
$E(), <>$	exspectation value
E	cost function
$\delta\mathbf{r}_n$	noise robust sphering
Q	Number of common elements for a given lag in the cross-correlation function

Contents

Chapter 1

Introduction

" Alas, I have studied philosophy,
the law as well as medicine,
and to my sorrow, theology;
studied them well with ardent zeal,
yet here I am, a wretched fool,
no wiser than I was before."

Johann Wolfgang von Goethe, Faust (First Part)

The dilemma that a long time ago devoured Dr. Faust from Goethe (1808) seems to be one of the impasses we face in neuroscience today. The amount of knowledge acquired by scientists in the multiple disciplines of neuroscience in the last century is enormous and only the number of questions still open seems to be bigger. Nevertheless there is no reason to despair like Faust for the researcher as advances in technology accelerate in pace with the accrual of scientific knowledge and therefore the possibilities to process and evolve new veda.

The advent of computer aided medical imaging technologies brings with it the possibility of answering many open questions about the three dimensional functional architecture and the modular organisation of the human brain. Some of these topics revealed only a glimpse of their true nature through thousands of micro electrode recordings at various sites (saying this without curtailing the success and the ground breaking insights for the comprehension of the brain by electrode recordings!). In order to understand and interpret all the new data the classical scientific fields of chemistry, biology, physics and medicine left their parallel running paths long time ago and merged with new disciplines like computer science to form interdisciplinary research areas like computational neuroscience. Among those imaging techniques frequently used today are computer tomography (CT), single photon emission tomography (SPECT) and magnetic resonance imaging (MRI).

In the mid 1980's another new imaging technique was introduced to the neuroscience community by Blasdel and Salama (1986), the optical imaging of neural activity with voltage sensitive dyes in vivo across large areas of cortex (mm). Optical imaging is more invasive than the imaging techniques mentioned above, as it involves accessing the cortical surface. This approach revealed the two dimensional organisation of the visual cortex for the first time with a high spatial and temporal resolution. The principles of this method were then developed further by others (Grinvald et al., 1986; Ts'o et al., 1990; Bonhoeffer and Grinvald, 1991) to purely image intrinsic reflectance changes of the cortical tissue instead of dye signals with intrinsic characteristics. Optical imaging of intrinsic signals records the two dimensional neural activity patterns by detecting small activity related changes in the light reflectance of neural tissue under monochromatic illumination. Typical sources of intrinsic signals are the changes in blood volume, hemoglobin oxygenation, tissue scattering and cytochrome oxidase. Although the temporal resolution of intrinsic signals is poor compared to those of dyes, the non toxicity allows chronic experiments without tissue damage The optical imaging recordings contain a super-position of the individual intrinsic signals with various technical and biological noise sources. However we are only interested in a small portion of the overall mixture that is very localised to functional areas of the cortex and strongly correlated with a specific stimulus.

At nearly the same time that optical imaging was developed a group of algorithms was introduced into the field of statistical signal processing that addressed the problem of separating linear mixtures. The algorithms for blind source separation (BSS) and independent component analysis (ICA) make use of the statistical properties that the sources have before the mixing to extract the individual components from the mixture.

One critical point when working with separation algorithms on biological signals is how well the original sources fulfil the basic assumptions made. This problem becomes even more severe when we have no possibility to access the original signal sources separately, as it is the case of optical imaging. It is therefore difficult to give a quantitative measure for the success of the separation process and compare it to heuristic methods used in the analysis so far.

For this doctoral thesis we have developed BSS algorithms and investigated the applicability of these BSS- and standard ICA algorithms to the separation of intrinsic signals from optical imaging recordings. We compared the algorithms based on the Infomax principle introduce by Bell and Sejnowski (1995) and extended by Amari (1996) that makes use of statistics of all higher moments, the kurtosis optimisation algorithm of Hyvärinen and Oja (1997) that exploits only the fourth order moment, and the extended spatial decorrelation (ESD) algorithm (Schießl et al., 1998; Schießl et al., 1999) that was derived from an algorithm proposed by Molgedey and Schuster (1994). The ESD algorithm uses the second order statistic of the spatial structure of the recorded images to separate the sources.

To get a quantitative result for the success of this separation we have tested the algorithms first on artificial toy datasets with properties similar to the original data. We then applied the methods to the optical recordings of the visual cortex and interpreted the separation results by gaining the time course of the components by back projection on the original data and correlating them to the stimulus onset.

Chapter 2 presents an overview of the functional anatomy of the central visual pathway to aid the researchers understanding of the activation patterns one can expect to see in optical images of V1.

In the first part of *chapter 3* we describe the design of our optical imaging setup and some of its key components. In the second part we have a closer look at the sources and characteristics of the intrinsic signals and some wavelength specific properties that are important for the later analysis.

The first section of *chapter 4* explains the basic preprocessing of the raw data before we take a critical look at the heuristic analysis methods that are commonly used today. A introduction to the algorithms for BSS and principal component analysis (PCA) is given in the third part of *chapter 4*.

In the fourth part of *chapter 4* the motivation and the concepts behind ICA algorithms are explained.

The final section of *chapter 4* illustrates the derivation of our extended spatial decorrelation algorithm. A number of different instances of this method are described.

In the first section of *chapter 5* the different sets of artificial data and their statistical properties are introduced. Then we point out some of the difficulties with using the condition number of a matrix as a measure for how ill-conditioned the mixing process was.

In the third section of *chapter 5* the separation performance of the ICA algorithms are compared to the single shift ESD on the independent- and smooth sources. This is followed by a comparison between single shift ESD and multi shift ESD. Finally we examine the simulations using the regularized multi shift ESD.

Chapter 6 presents the results of optical imaging using standard analysis methods and filtering followed by a short presentation of the intermediate result we get from sphering. In the next three sections the ICA and ESD algorithms are applied to the optical imaging data sets from a ocular dominance experiment and from an orientation preference experiment.

Finally in *chapter 7* the results are discussed and the insights from the artificial data are compared to those of the real optical imaging data sets.

Chapter 2

The Central Visual Pathway

One of the best investigated and understood signal processing pathways of higher mammals is the central visual pathway. This is partially due to the fact, that it is relatively easy to access compared to other areas and can be stimulated in a straight forward fashion. In the following a short overview of a commonly used model for the visual pathway of macaque monkey is given as it is relatively similar to the human visual system and our main model for optical imaging.

The depth of the description is chosen to give the reader some understanding of the processing of the visual stimulus up to the striate cortex in primates, and does not claim to be comprehensive on this issue. Here the focus is on the description of the processing phenomena, that later lead to the patterns recorded with optical imaging of intrinsic signals. For a more detailed introduction to the anatomy and physiology of primates see Kandel et al. (1991).

In the description we briefly show how the visual world is mapped onto the retinas of both eyes and how the photoreceptors transform this information. The area that is mapped on either of the retinas is called the visual field, with the left and right half of the visual field called hemifield (for a sketch of the early visual pathway see fig.2.1).

The coded information is then transmitted by the two optic nerves into the central nervous system. On the way to the lateral geniculate nucleus (LGN) the optic nerve is separated at the chiasm in a fashion, such that binocular information from one hemifield is mapped to the contralateral LGN. After synapsing in the LGN another bundle of nerve fibers, the optic radiation, made up of the axons of LGN neurons, transmits the information to the primary visual cortex at the occipital pole of the neocortex.

2.1 The Retina

The retina is the first stage of visual information processing and acts like a analog-digital converter, as it converts electro magnetic signals (light) into coded electronic pulses. Further tasks include the adaptation to different light intensities, the

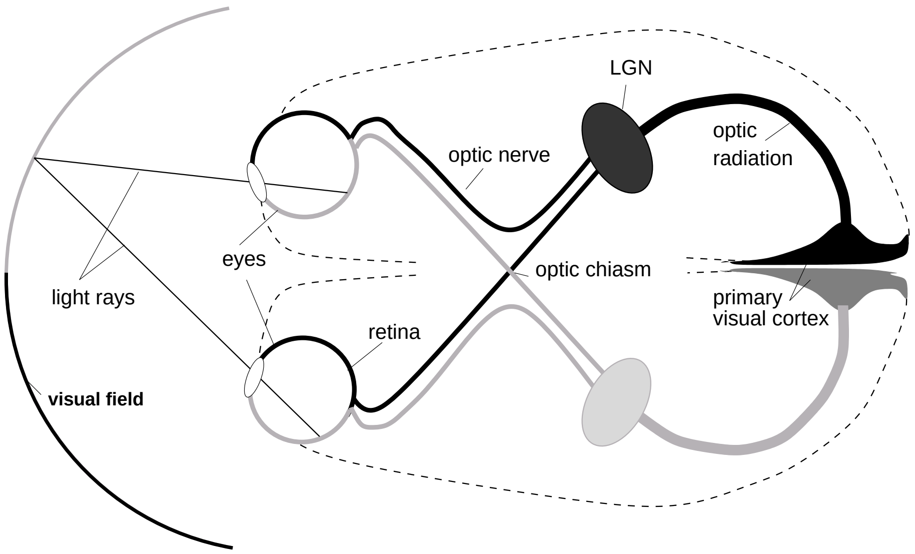


Figure 2.1: Schematic sketch of the early visual pathway of higher mammals. Once the light is focused by the lens on the retina, the information is coded into spike patterns and passed to the primary visual cortex across the optic nerve, the chiasm, the LGN and the optic radiation. Notice that the LGN and the visual cortex of each cerebral hemisphere process binocular information from the contralateral hemifield (adapted from Bauer (1999)).

detection of contrast and movement and the mediation of color perception.

The function of the retina is well understood and it is a good example for the understanding of how information is processed by complex neural circuits in the brain. This comes from the fact that it is derived during development from the neural ectoderm, which also gives rise to the brain .

2.1.1 The basic organization of the retina

The main task of the eye is to focus light entering the eye onto the retina with minimal distortion. This upside down projection of the outside world is then absorbed by the photoreceptor cells.

The retina is built up from three principal layers, that contain the photoreceptors, the different interneurons and the retinal ganglion cells (see fig.2.2). These three granular layers are separated by plexiform layers, that contain the majority of the synaptic connections. At the back of the retina lies the pigment epithelium, that contains melanin to absorb light and keep it from being reflected off the back of the eye. The other important task of the epithelial cells is to assist photoreceptors with important aspects of their metabolism. Therefore the photoreceptors have to contact the pigment epithelium, while other retinal cells are situated closer to the lens.

This leads to a "reverse" construction, with receptor cells pointing away from the

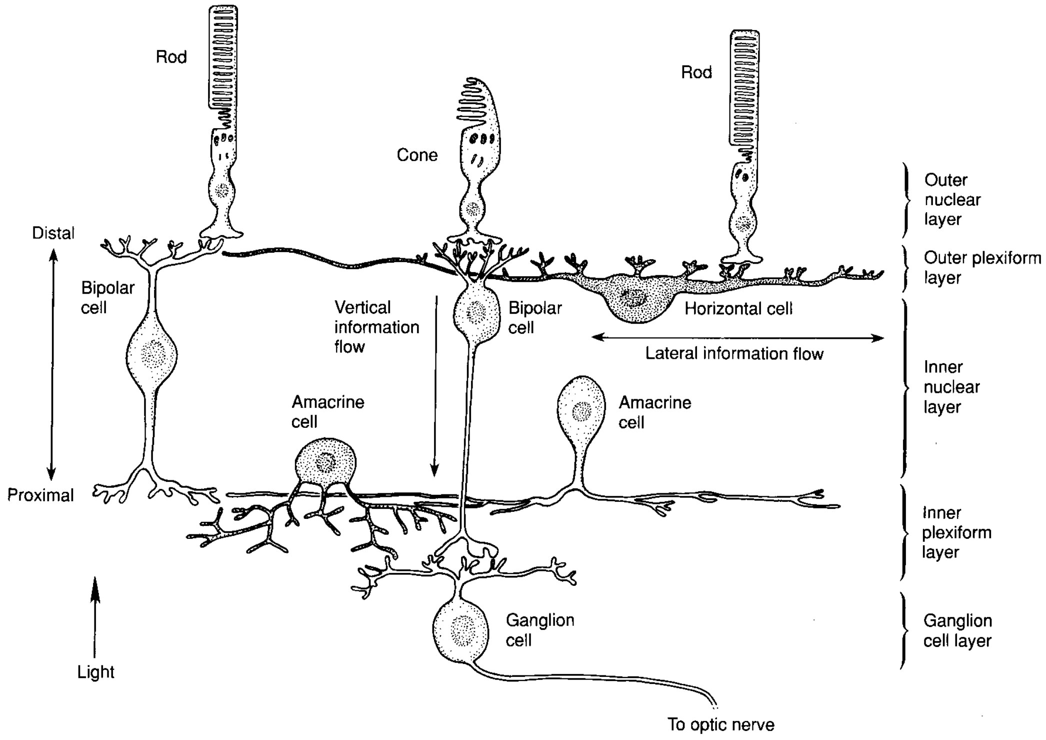


Figure 2.2: The retina is composed of three principal layers, that contain five classes of neurons: rods and cones, amacrine, horizontal and bipolar cells and the retinal ganglion cells. In the direct pathway the information is transmitted from the cones to the ganglion cells across the bipolar cells. The indirect pathway allows a horizontal flow of information mediated by horizontal- and amacrine cells (adapted from Kandel et al. (1991)).

light, that has to pass the other layers first. The only exceptions to this is the fovea where the cell bodies of the proximal retinal neurons are shifted to the side, and the optic disc where the optic nerve fibers leave the retina and no photoreceptors are present.

There are two basic types of photoreceptors, the rods and the cones, with the cones being subdivided into three distinguished types. The rods contain more photosensitive visual pigments than the cones and are therefore sensitive to dim light and responsible for night vision. The cone system has a lower sensitivity to light and mediates day vision. The brain obtains information about color by comparing the responses of the three cone types. The visual pigment of each type of cone is more sensitive to different wavelengths. The S- or B- cones respond best to wavelengths around 420 nm, sensed as blue light, the M- or G- cones are most sensitive to 513 nm green light and L- or R- cones have the maximum sensitivity at 558 nm, which is perceived as red light.

The intermediate layer of the retina is formed by three classes of inter neurons, the bipolar, horizontal and amacrine cells. The bipolar cells again can be distinguished into rod- or cone-bipolar cells, depending on where they get their input from. Both receive direct input from the photoreceptors in form of graded changes in membrane potential, and not action potentials.

The signal transduction from the photoreceptors to the ganglion cells can be by the direct (feed-forward) pathway or the indirect (lateral) pathway. In the first one the photoreceptors contact a bipolar cell and the bipolar cells contact a ganglion cell. The second pathway is characterized by the lateral information flow between bipolar cells and photoreceptors, which is mediated by horizontal and amacrine cells. Horizontal cells transfer information from distant cones to nearby bipolar cells. Some types of amacrine cells transfer information from distant bipolar cells to the ganglion cells (Kandel et al., 1991).

A single ganglion cell receives the combined input from several photoreceptors and transforms this information into spike patterns. This strong convergence concentrates the input of about 120 million photoreceptors per eye to about 1 million ganglion cells. The first stage of coding the visual scene is taking place.

2.1.2 Receptive Fields of Retinal Ganglion Cells

Ganglion cells are the output neurons of the retina. Most of their axons become myelinated after they leave the retina and form the optic nerve. This property allows relatively easy access to the response patterns of the ganglion cells to light stimulus by extracellular recordings of axons in the optic nerve.

In the early 50's it was found, that the specific connectivity and the physiological properties of the inter neurons are responsible for the contrast enhancing center surround organization (Kuffler, 1953). There are two principle types of center surround organizations in the retinal ganglion cells. ON cells increase their spike frequency when light stimulates the center region and reduce the spike rate if the inhibitory surround region is illuminated. OFF cells respond exactly the other way around.

The segregation of the visual information into off-center and on-center pathways is mediated by the bipolar cells and the horizontal cells. Apart from the distinction by the response properties, ganglion cells comprise several classes, of which three have been best explored: the magnocellular (M) ganglion cells, also called α- or parasol ganglion cells and the parvocellular (P) ganglion cells, or β- or midget ganglion cells, that project to the M-layers or the P-layers in the lateral geniculate nucleus (LGN) respectively after passing the optic chiasm. The third pathway is the X-pathway that projects into the inter-laminar zones of the LGN and terminates in the blob regions of V1.

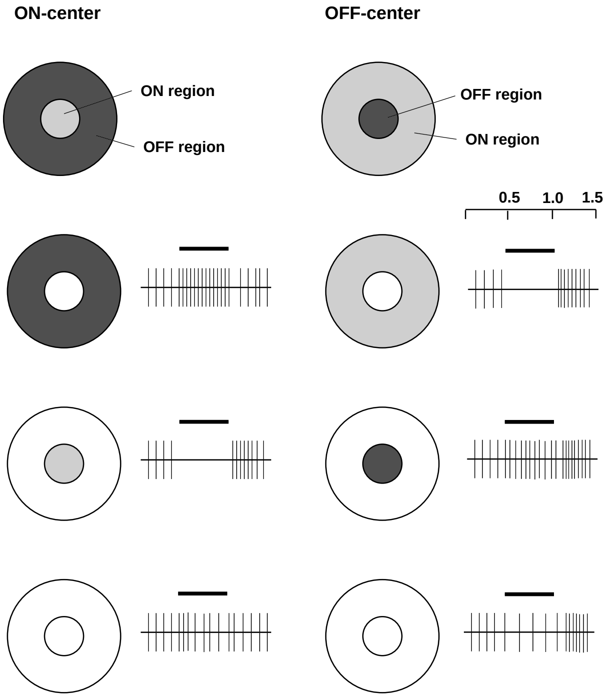


Figure 2.3: Receptive field center surround organization of retinal ganglion cells and their neural response patterns (after Kuffler (1953)). If the center of ON-center cells is stimulated (stimulus duration is marked by the bar above the spike pattern) it fires with increased rate. Illumination of the antagonistic surround region suppresses the cell response. Diffuse illumination of the whole receptive field has no effect (bottom row). The OFF-center cells (shown in the right column) respond the other way around.

2.2 The Lateral Geniculate Nucleus

The LGN is a part of the thalamus and the major thalamic relay stage in the central visual pathway of primates. It receives the afferent input from retinal-ganglion cells through the optic nerve. The LGN is formed by six layers of cell bodies, that are separated by inter-laminar zones rich in axons and dendrites but also contain cells that receive X inputs. The cells in the inter-laminar zones are called I-cells. The layers are innervated in an alternating fashion by ganglion cells from either the temporal hemiretina of the ipsilateral eye or the nasal hemiretina of the contralateral eye (see fig.2.4). The most ventral layers 1 and 2 contain large cell bodies and are inervated by M retinal ganglion cells; the dorsal layers which are characterized by small cell bodies are inervated by retinal P ganglion cells. There is evidence that retinal ganglion cells project in a one to one manner between retinal and geniculate P and M cells. Despite this fact the majority of connections are from other subcortical regions or the primary visual cortex. It is believed that the LGN acts as a relay station of the information flow to the

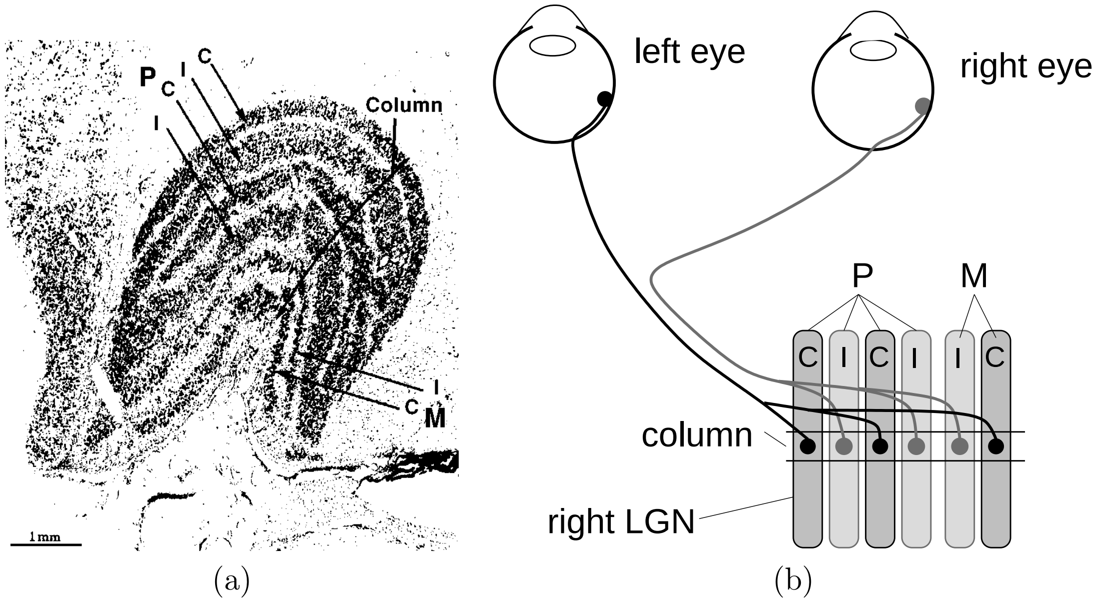


Figure 2.4: **(a)** Vertical slice through the LGN of a macaque monkey (adapted from Hubel and Wiesel (1977)), and **(b)** the retino-geniculate wiring pattern. “P” and “M” mark parvo- and magnocellular layers, “I” and “C” refer to layers driven by the ipsi- and contralateral eye respectively. All signals from a given small patch of the visual field arrive at a single column of the LGN (adapted from Stetter (2000)).

visual cortex and the modulatory input to the LGN controls the gain of the signal transmission (Coenen and Vendrik, 1972).

The termination pattern of the retinal ganglion cells in each layer is organized in a topographic way (Connolly and Essen, 1984). Each layer in the LGN contains a representation of the contralateral visual hemifield. The layers are stacked on top of each other, so that the topographic maps form a precise vertical register (Hubel and Wiesel, 1977). In the LGN the surface of the retina is not represented in an isometric way. The fovea has a much larger representation than the periphery of the retina (see fig.2.5). This arises from the fact, that there is a high density of retinal ganglion cells for the representation of the fovea. Due to the direct projection pattern half of the neural mass of the LGN represents the fovea and the areas perifoveal around it (Kandel et al., 1991).

In 1966 Wiesel and Hubel found that the receptive fields of LGN neurons are similar to those in the retina. They have circular center surround regions and also show the ON and OFF classification.

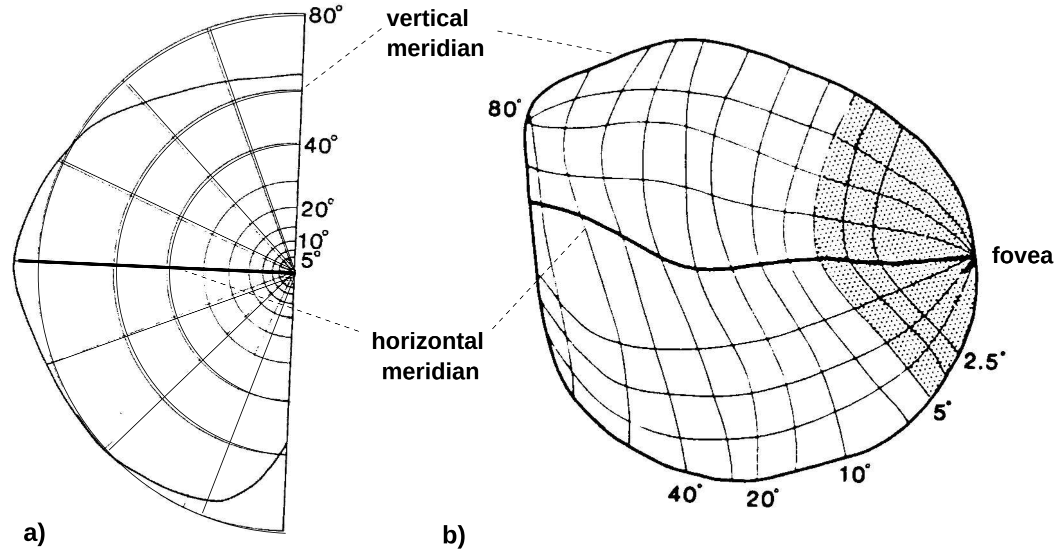


Figure 2.5: Anisometric representation of the visual field in the LGN. **(a)** Polar grid on the left visual hemifield that defines isoeccentricity (circles) and isopolar (rays) lines. The perimeter of the visual field is given by the dark line. **(b)** Mapping of the same grid on the layer 6 of the LGN. The central 5 % of the visual field are marked gray in both graphs (adapted from Connolly and Essen (1984))

2.3 The Primary Visual Cortex

The primary visual cortex is the first stage where receptive field properties significantly change. It is known that many features of its anatomy and physiology are very similar between macaque monkey and man (Levitt et al., 1996). Because of historical reasons and different qualities of the visual cortex in miscellaneous disciplines, it is called by many names. For long time people thought that all output from the LGN would be received by this area, so it was called primary visual cortex or V1, whereas its characteristic architecture of the layers that visibly appear as stripes in a vertical slice gave it the name striate cortex. As most of V1 is buried in the calacrine sulcus in man, a common clinical term is calacrine cortex. In the studies of Korbinian Brodmann the term area 17 was used and is widely commonly used today.

The projection from the LGN to the primary visual cortex is in an orderly point to point manner. Therefore the cortex of each hemisphere receives signals from the contralateral half of the field of view and the adjacent retinal points are mapped at adjacent points in the cortex, just like in the LGN (Daniel and Whitteridge, 1961). This representation of visual space on the cortex is called a topographical map.

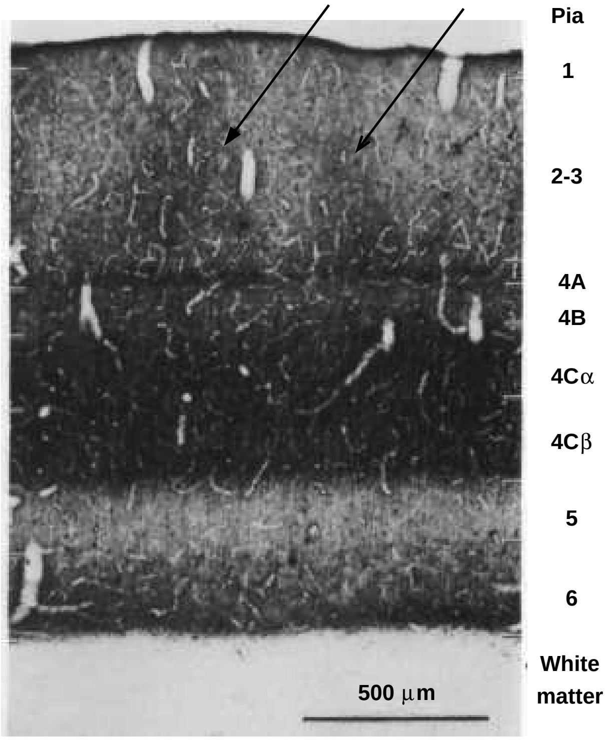


Figure 2.6: Microscope section of V1 from a macaque monkey with cytochrome oxidase staining to visualize the individual layers. Cytochrome oxidase is known to concentrate at zones of thalamic input and therefore the laminae 4C, 4A, the bottom part of layer 6 and the blobs (see arrows) in layers 2 and 3 are stained darker (adapted from Blasdel and Lund (1983)).

2.3.1 Anatomical Organization of V1

The total area covered by the primary visual cortex of macaques is about 1300 mm^2 (Hubel and Wiesel, 1977) and the input of 2 million LGN fibers is processed by 260 million cells in V1. These numbers give a rough idea about the rise of complexity in the information processing from the LGN to striate cortex.

The most obvious feature of V1, that also gave rise to the name striate cortex, is the organization into six principal layers 1 to 6 between the pial surface and the underlying white matter with the prominent stripe of white matter in upper layer 4, called “the stripe of Gennari” (Kandel et al., 1991). This visible property emerges from the presence of myelinated axons and the different cell densities between the layers. Layer 4 is further subdivided into three prominent layers 4A, 4B, 4C where 4C is further subdivided into 4Cα and 4Cβ.

All axons that project from the LGN to V1 terminate in the sublaminae 4Cα and 4Cβ (Hubel and Wiesel, 1972; Blasdel and Lund, 1983). Layer 4 therefore represents the input layer of the striate cortex. The majority of the P neurons project to the β division of layer 4C and a smaller population projects to layer 4A.

Single P axons provide terminals to the whole depth of the 4Cβ division, splitting the teritory laterally into alternating stripes of equal width for relays from each eye (Levitt et al., 1996). In layer 4A the P axons terminate in a cytochrome oxidase (CO) rich lattice, leaving small uninnervated lacunae.

Most M axons terminate to the lower two third of the α division of layer 4C with a small population of very large M axons also layered in the upper half of 4Cα. The M axons provide collaterals to layer 6 as they enter the cortex (Blasdel and Lund, 1983). Like the P axons, the M axons divide the 4Cα territory into alternating stripes of left and right eye inputs (Hubel and Wiesel, 1977).

The LGN I-cell population, located between and ventral to the LGN layers, target to the layer 2 and 3 CO blobs as well as layer 1. Layer 1, the so called molecular layer, is mainly comprised of the apical dendrites of lower pyramidal cells and horizontal cells running axons (see fig.2.6).

Pyramidal cells make up the major portion of cortical neurons with a presence of 80%. These cells are excitatory and form intrinsic local projections as well as long range projections to different areas. Their dendritic trees cover a volume with a diameter of 200 - 300 μm and can reach vertically over several layers. The local lateral axon projections of pyramidal cells are referred to as axon collaterals. In the layers 2 and 3 these projections form clusters of connections in a set of patches of approximately 250μm diameter, which can be located up to a few millimeters away from the cell body. Another type of excitatory neurons in layer 4, that have spherical instead of cylindrical dendritic trees are the spiny stellate cells.

2.3.2 Physiological Organization of V1

As stated before the receptive field properties of cells in the early visual pathway, before area V1, are organized in a circular center surround manner and do not change significantly between the retina and the LGN. Surprisingly in V1 the small circular stimuli trigger only a poor response in most of the cortical neurons. It was found that most of the cells outside the blob regions and layer 4C respond best to stimuli that show more elongated properties, like stripes or bars.

In 1968 Hubel and Wiesel described orientation and direction selectivity as the major response properties of cells in the striate cortex of primates. The orientation selective cells respond best when a bar with a specific orientation is chosen as stimulus. When the orientation of the bar is changed from the preferred within the same field of view, the strength of the response of the neuron declines gradually (Hubel and Wiesel, 1968). If the rate at which the cell fires under presentation of an optimal oriented bar, moving at 90° to the axis of orientation, changes significantly with a rotation of the stimulus of 180° (same orientation, opposite direction of motion), the cell is also direction selective. The findings from single cell recordings showed for some cells strong changes of the cell's response according to the positioning of the bar. This suggested a elongated receptive field with alternating on and off regions. These cells in V1 were classified as simple cells; a second major class , called complex cells, responded equally well wherever the bar

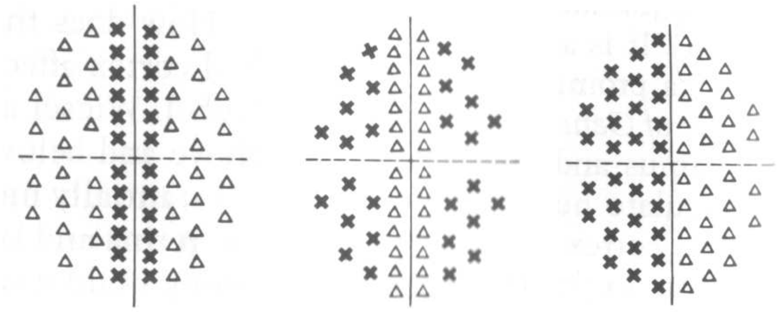


Figure 2.7: Examples of receptive fields of simple cells. There is a sharp separation between the ON and OFF zones. The receptive fields of simple cells are rectangular and oriented and the best stimulation is achieved when only the ON zone is stimulated with a bar in the right orientation (adapted from Kandel et al. (1991)).

is placed within its receptive field ((Hubel and Wiesel, 1968)).

Simple cells dominate in cells that are localized close to the inputs in layer 4C. They are assumed to receive the convergent input from three or more stellate cells of layer 4C, or from LGN, that have similar center surround organization and this convergence forms the rectangular orientation specific fields. The most effective stimulus coincides with the boundaries of the subdivision of the receptive fields ON and OFF zones in the same orientation and gives the maximum response when only the excitatory region is stimulated (see fig.2.7).

Complex cells are also pyramidal cells but found mainly in the layers 2, 3, 5 and 6 further away from 4C. They also have elongated receptive fields with a preferred orientation, but they lack clearly defined on-off regions like simple cells. Therefore the exact position of the stimulus within the receptive field is not that critical. Some of the complex cells are directly inervated from stellate cells in layer 4C, which can be either non-oriented or simple cells.

Most of the cells in the visual cortex and of layer 4C receive input from both eyes and are therefore called binocular cells. Their receptive field of the left and right eye cover the same regime of the visual field. Nevertheless these binocular cells are normally dominated by the input of one of the eyes and respond stronger, if the dominant eye is stimulated. Ocular dominance can be found in varying strengths up to cells only responding exclusively to one eye and therefore called monocular cells.

The strength of response of a neuron is often not only controlled by orientation or direction and ocularity, but also by the contrast (Albrecht and Hamilton, 1982) and other effects. DeAngelis et al. (1992) showed that a grating superimposed orthogonal to the optimally oriented grating can lead to so called cross-orientation suppression. Other studies have shown that even stimuli that lie outside the classical receptive field of a neuron can strongly modulate the response of a cell

(Blakemore and Tobin, 1972; Gilbert and Wiesel, 1990; Sillito et al., 1995; Levitt and Lund, 1997). This means that the response to a local feature depends on the visual context which is referred to as the contextual or surround effect.

2.3.3 Functional Organization of V1

So far we only looked at the individual response properties of neurons in the striate cortex, but not at the distribution of the cells with similar properties. As each area in the visual field, processed by the visual system, should be able to recognize all features the real world has to offer, a first thought could be, that the neural properties described above are randomly spread in depth and across the cortex. This is the so called salt and pepper model. Hubel and Wiesel were the first to investigate the receptive field properties across the cortex with electrode penetrations (Hubel and Wiesel, 1962; Hubel and Wiesel, 1968) and found a completely different strategy of cell organization.

The penetrations revealed a columnar organization of cells with similar response properties through the depth of all layers and a gradual periodic change across the cortex. Every orientation column has a width of about 30μm and within that distance a shift of about 10° in preferred orientation is encountered. They also found that the striate cortex is subdivided into regions of about 0.4mm width devoted either to the right or the left eye, derived from the stripe like monocular right and left eye relays from the LGN to layer 4C. These so called ocular dominance columns again repeat in an alternating fashion. Therefore the area of the striate cortex that contains a complete set of orientation columns (i.e. covers 180°) and a ocular dominance column of the left and the right eye is about 1mm^2 and is called a "hypercolumn" (see fig.2.8). The hypercolumns also form a regular and precise pattern across the primary visual cortex. The hypercolumn is the elementary processing module to analyze the features of a single spot on the retina.

So how are the orientation columns and the ocular dominance columns arranged with respect to each other across the cortex? To reveal a good presentation of this two dimensional organization with single electrode recordings is nearly impossible. A first visualization was possible with the staining of tissue with radioactively marked 2-deoxyglucose injected into the blood. As only metabolic active cells absorb it, one gets a regular pattern of active and inactive cells after the presentation of a oriented stripe pattern (Hubel et al., 1978).

More recently a technique was developed that uses a sensitive video- or CCD Camera to image the active and inactive regions on the cortical surface, the so called optical imaging technique. In the optical imaging of voltage sensitive dye signals (Blasdel and Salama, 1986; Blasdel, 1992a) the shift of the membrane potentials is imaged, whereas the optical imaging of intrinsic signals (Grinvald et al., 1986; Bonhoeffer and Grinvald, 1996) reveals the changes in light absorbency and reflection of active areas of the cortical tissue. The experimental setup and the analysis of optical imaging will be explained in much more de-

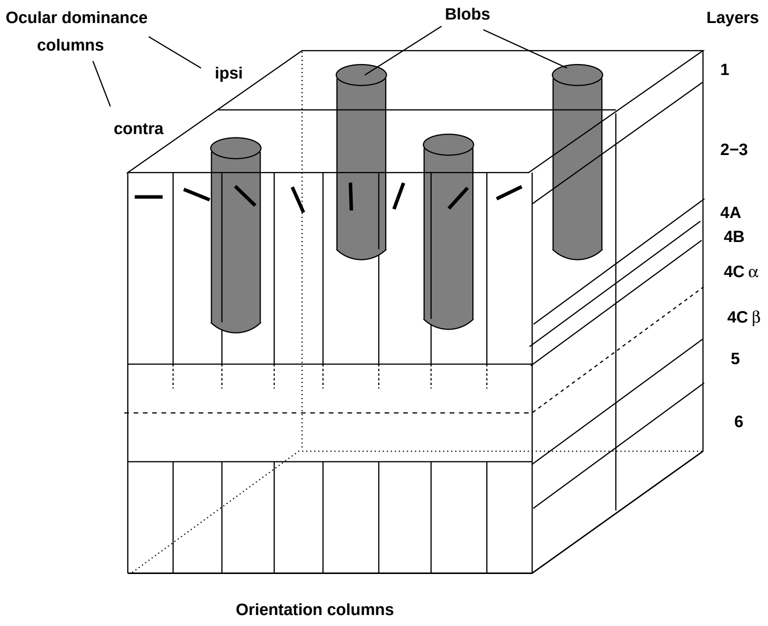
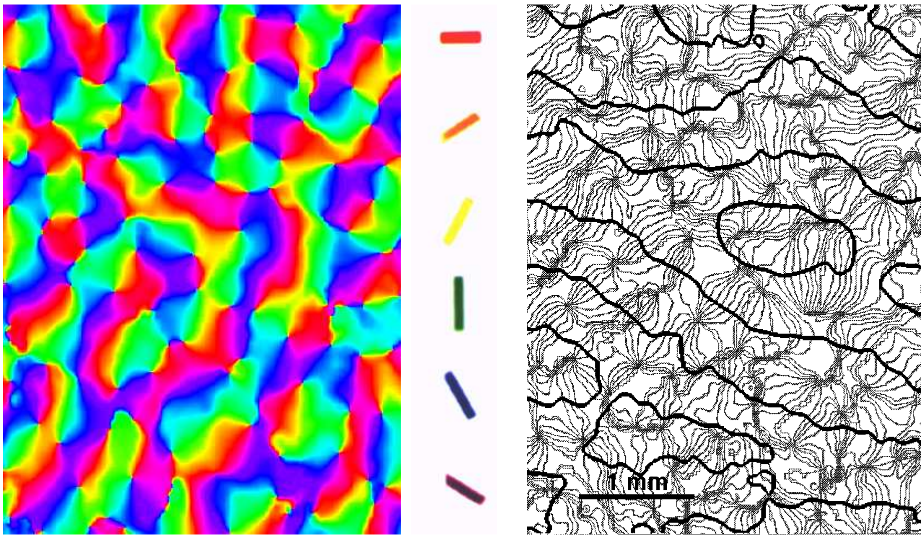


Figure 2.8: Schematic sketch of a hypercolumn. This basic module contains a complete set of orientations columns, that represent 180°, one right and one left ocular dominance column and several blobs. A hypercolumn can process a discrete region of the visual field and the complete presentation of the visual field on the cortex is represented by a regular pattern of hypercolumns.

tail in the following chapters. This method proved that the preferred orientation changes gradually across the cortex and form a regular orientation map (Blasdel and Salama, 1986; Obermayer and Blasdel, 1993). An example of such an orientation map from macaque V1 is shown in figure 2.9 where the different preferred orientations correspond to the colored bars on the right. The ocular dominance columns form a stripe like pattern that is strongly coupled with the orientation map. The singularities, that define areas of the orientation map where all degrees of orientation are present, tend to lie in the center of ocular dominance bands and the iso-orientation lines and the borders of the ocular dominance stripes intersect at nearly orthogonal angles (Obermayer and Blasdel, 1997; Müller et al., 2000).



(a) (b)

Figure 2.9: **(a)** Orientation preference map of V1 in macaque monkey. The bars on the side illustrate the color code for the preferred orientation of the cells imaged in the respective areas. **(b)** Superposition of an ocular dominance and an orientation map. The thick lines show the borders of ocular dominance stripes whereas the thin lines mark the iso-orientation contours within the orientation map (from Obermayer and Blasdel (1993)).

Chapter 3

Optical Imaging Setup and Data Acquisition

In order to understand the concepts of the modular organisation of the neocortex one has to obtain a two- or three dimensional dataset of activation of neurons, that is correlated with a specific stimulus presentation. Certainly the important work with single cell recordings now already revealed a good idea about what will be found once a two dimensional image is obtained.

One of the first visualisation of the cortical structure was obtained with the labelling of tissue with radioactively marked 2-deoxyglucose injected into the blood. For the labelling of the visual cortex a specific stimulus is presented to the eyes of the animal and the metabolically most active cells in the cortex concentrate most of the marker. The postmortem anatomical sections reveal a regular pattern of active and inactive cells depending on the stimulus paradigm (Hubel et al., 1978). This method delivered the first images of the two dimensional organisation of large areas of the cortex but the drawback of this method is clear. Only one stimulus regime could be investigated per animal. This leaves a lot of important question open. How would the response patterns change with a gradual change in the presented stimulus, like a change in contrast, bar width or orientation etc.? In the late 80's a imaging method was developed that has been proven a powerful technique to answer all these question. The optical imaging of neural activity (Blasdel and Salama, 1986; Blasdel, 1992a; Grinvald et al., 1986) is so far the only in vivo method to obtain detailed functional maps of the cortical architecture at sub-millimetre resolution. The latest results in functional magnetic resonance imaging (fMRI) (Kim et al., 2000a) also show promise of delivering a high spatial resolution but are still very controversial (Logothetis, 2000; Kim et al., 2000b).

The first optical images of wide areas of cerebral cortex were obtained by Blasdel and Salama (1986) using voltage sensitive dyes and a sensitive CCD camera to investigate patterns of neural activation. This method very directly displays the neural activity as the voltage-sensitive dyes are attached to the surface membranes of the neuron and alter their spectral properties by the Stark-effect when the electrical field in the membrane changes (Salzberg et al., 1973). Beside high spatial resolution, optical imaging with voltage-sensitive dyes also provides high temporal resolution. The temporal structure of the recordings show that the primary signal source is of intrinsic nature triggered by neural activity. The handicap of this method is that the voltage-sensitive dyes are photo toxic and destroy the cortical neurons over time. Therefore long or chronic experiments are not always possible.

A second form of optical imaging was introduced by Grinvald et al. (1986), the so called optical imaging of intrinsic signals. This method does not directly image the neural activity, but some of the metabolic responses in the area around the active neurons. Those imaged metabolic responses are the intrinsic signals. In the following sections we will extensively explain the experimental setup of optical imaging of intrinsic signals, the animal preparation and the biological origins of the intrinsic signals.

Optical imaging of neural activity has lead to a number on new insights of the functional organisation of the cortex in many species (Blasdel and Salama, 1986; Blasdel, 1992a; Blasdel, 1992b; Rao et al., 1997; Bosking et al., 1997; Bonhoeffer and Grinvald, 1991; Bonhoeffer and Grinvald, 1993; Bonhoeffer et al., 1995). This method also allowed a direct comparison of the anatomical wiring patterns and the functional activity in the cortex (Malach et al., 1993; Malach et al., 1994).

3.1 The Experimental Setup

The successful realization of an optical imaging experiment is very sensitive to many technical and physical parameters as well as an optimal animal preparation, that allow only a marginal variation from the ideal conditions. We now describe the overall experimental setup and then have a closer look at some of the individual components.

Figure 3.1 shows a rough sketch of the experimental setup for optical imaging. In the optical imaging of intrinsic signals the cortical area under investigation is illuminated with monochromatic light of a wavelength typically around 500-800 nm. To get the light to the cortical surface a craniatomy and duratomy are performed on the anaesthetised animal, that is fixed in a stereotaxic head frame. All the procedures carried out during the optical imaging experiment comply with the Home Office Animals (Scientific Procedures) Act 1986 regulations of the United Kingdom. As the signals we expect to record during the experiment are very small the cortex has to be stabilised again to reduce motion artifacts from respiration and heartbeat. This is done by mounting a steel chamber with dental cement to the skull. The top of the chamber is made of a cover glass, the so called cranial window. The chamber is filled with artificial cerebrospinal fluid (aCSF), saline or silicon oil. Through this cranial window a sensitive CCD camera is focused on the surface of the cortex. Some groups prefer to focus slightly below the surface but this leads to some problems described later in the text. Here we have one limitation of optical imaging as only cortical areas on the surface of the brain

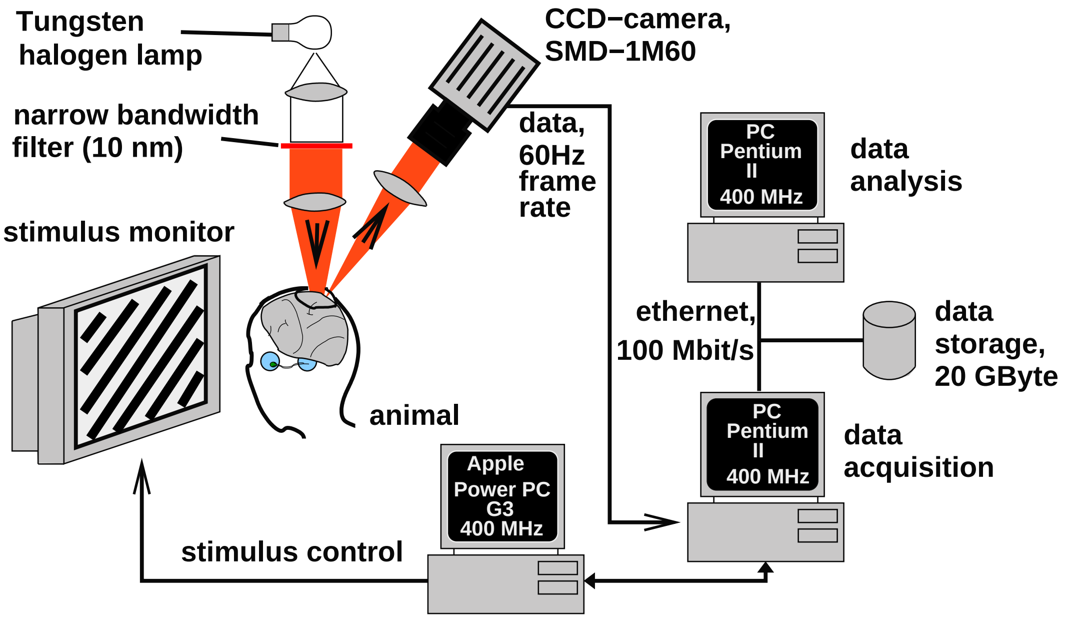


Figure 3.1: Sketch of the essential components for optical imaging. The cortical area of interest is illuminated with monochromatic light of a wavelength between $600 - 800nm$ from a highly stabilised light source. The sensitive low noise CCD camera focuses on the cortical surface and images the reflection patterns during stimulation. In the later analysis the differences between specific stimulus regimes can be visualised. The computer that controls the stimulus is connected via a "handshake" connection to the computer that controls the camera and the recording protocol. Due to the high camera speed the data is stored in a ring buffer first and then read out to the hard disk during the recovery periods.

that are not hidden in a sulcus can be investigated. During the imaging of the visual cortex the stimulus is presented to the animal by a computer monitor. The computer that controls the stimulus onset and regime is connected to the computer for the control of the data acquisition via a "handshake" connection to guarantee synchronisation with the recording. For preliminary analysis of the recorded data during an experiment the data are transfered to a second computer by a Ethernet connection. This online analysis delivers important feedback for the experiment. Apart from these optical imaging specific components the experimental setup also contains the apparatus for the anaesthesia and life support of the animal as well as monitors for expired CO_2, body temperature and tissue oxygenation. After this short outline we will now have a more detailed look at some of the key components and procedures of the optical imaging experiment.

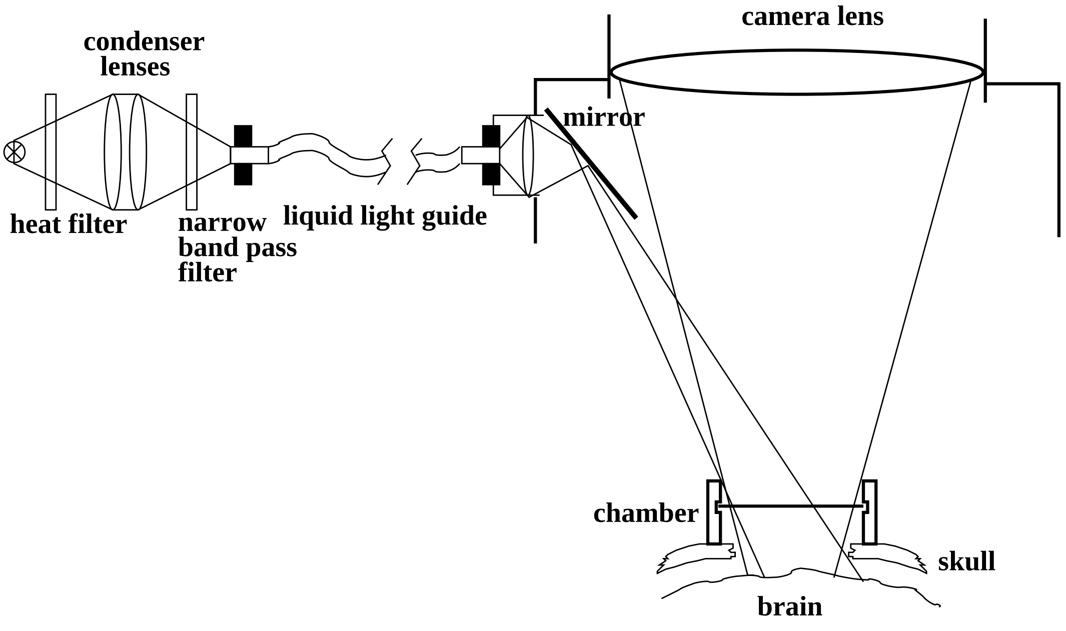


Figure 3.2: Typical light path for the illumination of the cortical surface with a $100W$ halogen light bulb. The wavelength of the light is determined by the heat filter and the narrow band width filter. The direct illumination on the cortex comes from a mirror inside a custom extension to the camera lens. This allows even illumination of the cortex because of a small illumination angle without specularities from the cover glass. The individual components are not drawn in proportion.

3.1.1 The Illumination

The amount of light back scattered from the cortical tissue containing information about the stimulus is less than 0.1% of the overall illumination (Bonhoeffer and Grinvald, 1996). This fact sets very high requirements in the stability of the light source to avoid artifacts from the illumination that are bigger than the signal. Most important is a power supply that guarantees that the ripple in the voltage from the AC to DC power conversion is smaller than 0.01%. Additionally a tungsten halogen light bulb should be used to achieve a highly stable light output. Arc bulbs are not an option, because the fluctuation in the light intensity is far too great.

There are several reasons why one has to get as high on intensity of light as possible. The photon emission, and therefore the number of photons remitted from the illuminated tissue, introduce a statistical error by photon shot noise. The number n of photons collected per unit time follow a Poisson distribution. If the expectation of the number of measured photons is N, the standard deviation of n around this value is $\sigma_n = \sqrt{N}$ (Papoulis, 1965). The relative statistical error for the measurement of the light reflectance is given by $\sigma_n/N = N^{-1/2}$. To achieve an accuracy of 0.01% in the measurement with this illumination, the CCD chip in the camera has to collect more than 10^8 photons per pixel.

Most of the light that is emitted from the bulb never even reaches the cortical surface due to numerous absorbers in the light path. Every surface of a lens or a mirror, that the light crosses, reduces the intensity. Figure 3.2 shows a typical light path for the optical imaging experiment. The major portion of the emitted light is lost in the narrow band width filter to obtain the monochromatic illumination we need. The bulb emits white light that is cut off at the low frequency side by a heat filter to avoid heat damage to the brain. From the remaining visible light only a $5 - 10nm$ wide portion around the chosen centre frequency passes through the narrow band width filter. The monochromatic light is then focused into a light guide of $1.5m$ length, where the intensity is reduced a further 40%. From a $100W$ halogen light source at full intensity and a $605nm$ filter with a with of $5nm$ we are left with $500\mu W$ at the end of the light guide (a $2 \cdot 10^5$ fold reduction in intensity). The camera lens system further reduces the light intensity.

3.1.2 The Imaging Camera

10^8 photons per pixel per frame of the camera have to be obtain to stay within the statistical noise limits of the signal response. This has two main consequences for the choice of the camera. The camera must have a high well capacity to collect the maximum number of photons before read out of the chip. Secondly the camera should have a high frame rate, because by later binning the effective well capacity can be raised as long as the CCD chip in each frame is nearly saturated. Furthermore the CCD chip in the camera should have a good quantum efficiency at the chosen wavelength.

The CCD camera SMD-1M60 by Silicon Mountain Design, Inc. delivers a 1024×1024 maximum resolution with square pixels and a frame rate of up to $60fps$. The true 12-bit colour depth provides up to 4096 distinct grey levels for capturing large inter scene light variations (dynamic range= $70dB$). The image is digitised in the camera head and transfered as a low noise digitised video signal to the computer, this allows capture of very low contrasts without noise in the image. In figure 3.3 the quantum efficiency for the camera is shown to have maximum values at the wavelengths used in optical imaging. The full well capacity is $250Ke^-$.

For focusing the image on the CCD chip a Nikon $50mm$ lens was used and two extension rings (pk12 and pk13). This gave a maximum resolution of $14.8\mu m \pm 0.5$ per pixel in the focal plane. A 256×256 image therefore covers an area of $3.7 \times 3.7mm^2$. The cell somas are typically $15 - 35\mu m$ and bigger (Kandel et al., 1991) but one would certainly need more than a few pixels to image the shape of a neuron. Furthermore in the following sections we will see that the scattering properties of the cortical tissue limit the resolution of the image and not the physical resolution of the lens and the camera.

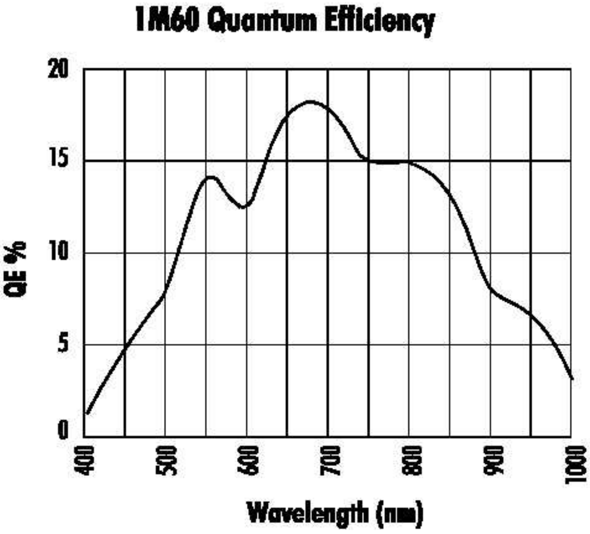


Figure 3.3: Plot of the quantum efficiency of the CCD camera SMD-1M60 from Silicon Mountain Design, Inc.. The maximum is reached at the wavelengths between $650 - 720nm$ that typically used for optical imaging of intrinsic signals.

3.1.3 The Head Chamber

A crucial component that determines the quality of the data obtained from optical imaging is the design of the head chamber. After the skull and the dura have been removed over the cortical area of interest the brain pulsates due to the animal's respiration and heart beat. These movements are far bigger than the resolution one can achieve with the imaging system and changes in reflection induced by this will mask the stimulus correlated signal. To stabilise the cortex a head chamber is mounted to the skull with dental cement.

In Figure 3.4 the chamber we developed for our optical imaging setup is shown. The basic setup is comparable to chambers used in most optical imaging systems (Blasdel and Salama, 1986; Haglund and Blasdel, 1992; Bonhoeffer and Grinvald, 1996) with the exception of a few details. The base ring (figure 3.4 d)) is mounted to the skull first with dental cement before the craniatomy. This way the dental cement on the base ring can have time to harden without the cortex being exposed and the evaporating bonding agent can not irritate the sensitive tissue. The gaskets (figure 3.4 c)) go below and above the cover glass (figure 3.4 e)) and are fixed with the closing ring (figure 3.4 b)) into the recess of the chamber body (figure 3.4 a)).

Now with the skull intact one can screw the assembled main body into the base ring and test the sealing of the chamber. A perfect sealing is essential for the long period of the experiment, as a slow leakage can lead to pulsations and a shift of the cortex that makes the direct comparison of trials impossible. Furthermore the slowly decreasing cranial pressure may cause oedema. Once the seal is perfect

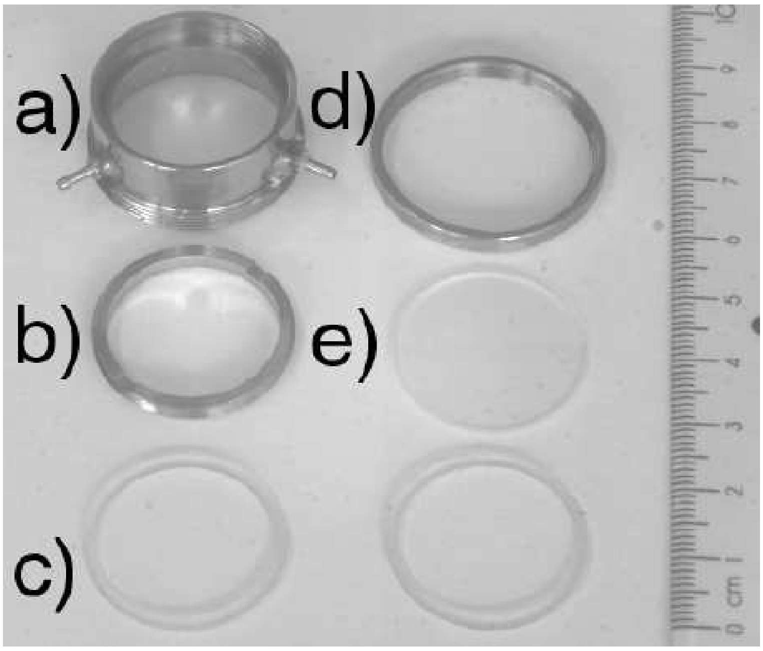


Figure 3.4: Individual parts of the head chamber used in our optical imaging experiments. **a)** Main body of the chamber with the two pipes for filling with aCSF. **b)** Closing ring to screw the cover glass (**e)**) into the main body with one of the gaskets (**c)**) on top and below.

the chamber body is removed and the craniatomy and duratomy can be performed with easy and unrestricted access to the surgical site.

A second important feature is the position of the filling pipes in the chamber body. They must be as close as possible to the bottom of the cover glass to enable the complete displacement of air in the chamber when filling it. Even small bubbles left will allow motion artifacts due to the compressibility of gases.

The material of the cover glass certainly should not have absorption bands at the chosen illumination wavelength.

3.1.4 The Stereotaxic Frame and Camera Holder

The stereotaxic frame is used to fix the animals head to ensure that motion artifacts from respiration are minimised. A standard stereotaxic frame with ear bars and one mouth and head bar is used. To the base plate of the frame the camera holder was attached, designed from camera tripod parts on a rail system. It has to be very rigid and should not allow a slow shift of the $1kg$ camera head over a long period of time. The whole setup must be vibration free and if necessary is mounted on an air table. To allow the imaging of different sized animals on both hemispheres a large degree of freedom for rotation and transversal travel of the camera is accommodated.

3.1.5 Animal Preparation

The careful preparation of the animal is key to the successful completion of an optical imaging experiment. The choice of anaesthetic and the given concentrations have a massive influence on the response properties of cortical neurons (Buchweitz and Weiss, 1986). The second critical procedure is the craniatomy and duratomy. Here one has to avoid damage to the cortex surface and bleeding from the dura into the chamber.

Here the preparation and surgical procedures for macaque monkey (*macacca mulatta*) are described as it was our main animal model, even if some of the data was collected from other species.

On the day prior to the experiment we start to starve the animal for 12 hrs to avoid vomiting during anaesthesia and intubation. The monkey gets an injection of Dexamethazone (intra muscular (i.m.) or intra venous (i.v.)) of $0.5mg/kg$ to reduce the risk of inflammation and oedema.

On the day of recording each animal was induced with a single injection (i.m.) of a mix of Ketamine ($10mg/kg$) and Xylazine ($0.5 - 1.0mg/kg$). Additionally Atropine ($0.05mg/kg$ i.m.) is administered to reduce salivation. After the animal had been intubated and a catheter for infusing electrolytes and drugs had been placed in either the saphenous vein, running from the back of the knee and then divides as it runs down the calf, or the cephalic vein, running along the anterior surface of the fore limb, it was ventilated with a $50 : 50$ mixture of O_2 and N_2O and 2% Isoflurane for surgery. For later maintenance the level is reduced to $1.5 - 0.75\%$. During the experiment the expired CO_2 and the tissue oxygenation are monitored additionally to the control of the body temperature, EEG, heart beat and respiration rate. The animal is kept on a heating blanket and fixed with two ear bars, one head bar, and a mouth bar into the stereotaxic frame.

After local sub-cutaneous injection of lignocaine the scalp is resected over the surgical area. Then the base ring of the chamber is mounted on the cranial surface using dental cement and tested for leakage. With a dental drill small corner holes are drilled slowly (to avoid heat damage) and then the bone is cut out between them. Carefully as large as possible a piece of dura is removed over the region of interest. Once all the bleeding has stopped the chamber body is screwed onto the base ring and filled with aCSF. When all air is displaced with fluid the chamber exhausts are sealed and the cranial window is ready for imaging.

Contact lenses are inserted into the animal's eyes and the light guide and camera are positioned. Following the completion of surgical procedures each animal was paralysed partially with vecuronium bromide to stabilise the eyes. Then the optical imaging experiment was started.

3.1.6 Optical Imaging Experiment and Data Collection

For the visual display of the stimulus a TV monitor was placed at the focal point of the animal. The focal point was determined roughly by back projection of the

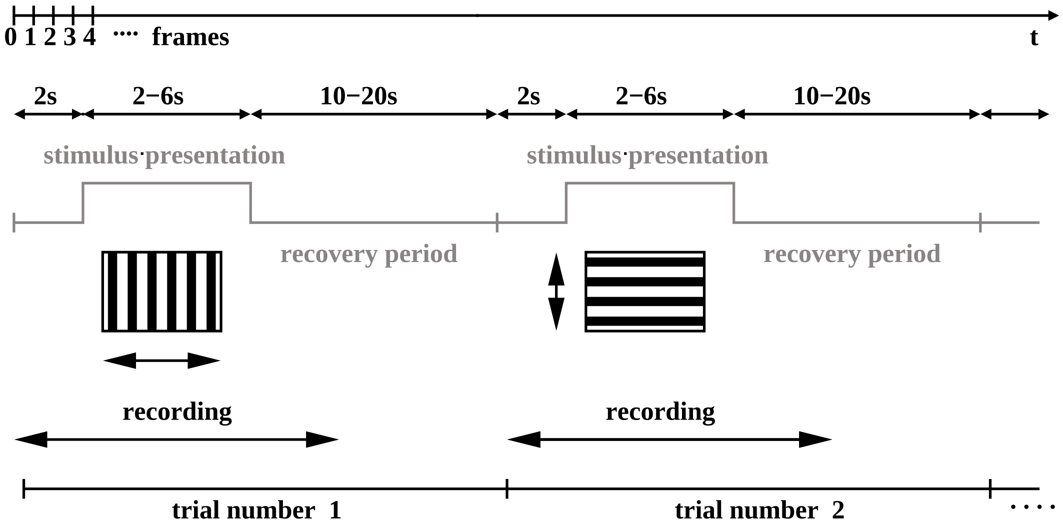


Figure 3.5: Standard scheme for the timing of an optical imaging experiment. The recording starts before the stimulus presentation and finishes after the end of the stimulus. The recovery period between the trials must be long enough to avoid activation artifacts from the previous stimulus. The presentation of different stimuli in the trials must not be alternating and can be displayed randomly.

foveas with an ophthalmoscope and verified with single cell recordings before the head chamber was closed.

Depending on the imaged cortical area, and the experimental goal, several different types of stimulus patterns were presented. Typical patterns were square wave gratings, sine wave gratings, checker board patterns or superimposed non harmonic square wave gratings. The patterns were moved at speeds between $1.0 - 4.0°/s$.

At the beginning of the optical imaging experiment a single frame under an illumination with high hemoglobin absorption (typically green light, 546 nm) of the whole exposed cortical area is recorded as a reference for the vessel pattern on the surface. This so called green image is used after the staining and sectioning of the cortical tissue to match the response patterns of the optical imaging with the anatomical structure of the imaged area, where the vessel pattern is crucial for spatial matching.

For the imaging of activity patterns a $605nm$ or $720nm$ filter was used. A single trial denotes the acquisition under one stimulus condition, whereas an experiment is the recording of several trials with alternating stimuli. $8 - 16$ trials per stimulus condition are recorded in one experiment. In a single trial the camera starts recording two seconds before the stimulus presentation to obtain data on the background activity and the DC level of reflectance for later analysis (see chapter 4). After two seconds the stimulus is presented for $2 - 8$ seconds while the recording continues (see figure 3.5). Once the stimulus stops the recording continues for two more seconds to obtain information about the signal decay. Before the next trial starts a recovery period of $8 - 16$ seconds elapses to avoid interference with the previous stimulus. In a typical optical imaging experiment the individual trials display so called orthogonal stimuli. Orthogonal stimuli are stimulus regimes that are expected to evoke responses in spatially disjunct cell populations in the cortex.

From all the trials with one stimulus condition the single condition image is calculated. The difference of two orthogonal single condition images yields a difference image. For a detailed description of the processing of the raw data see section 4.1.

To understand the design of a single experiment we examine two standard regimes: The recording of ocular dominance columns and orientation preference maps (for functional background section 2.3.3).

In an ocular dominance experiment the two orthogonal stimulus conditions are the alternating stimulation of the right and the left eye. So for half of the number of trials the right eye is occluded and for the other half the left eye is occluded. To avoid intensity changes in the mapped columns from the orientation preference of the cells within each half of the trials gratings of all orientations are presented, respectively $4 - 8$ distinct realizations (i.e. gratings at $0°, 45°, 90°, 135°$).

In an experiment for the mapping of orientation preference the stimuli can be presented binocularly or monocularly. The orthogonal stimuli are the alternating presentation of gratings with 90° difference in orientation (i.e. $0° - 90°$ or $45° - 135°$). In this regime each orientation delivers a single condition map and the difference image is obtained from the two orthogonal single condition maps. The final orientation preference map is obtained by vector summation of all the difference images.

Before we describe the analysis of the experiments in chapter 4 let us have a look at the actual origin of the recorded data, the intrinsic signals.

3.2 Sources and Characteristics of Intrinsic Signals

In the optical imaging of intrinsic signals it is not the activity of the neurons themselves that is monitored, but the changes in absorption and reflection of the cortical tissue that coincides with the stimulation of the neurons. These changes derive from metabolic activity and variations in the micro-circulation. Four main components have been described as sources of these so called intrinsic signals (Cohen et al., 1968; Cohen, 1973; Kreisman et al., 1995).

One component is due to the blood volume changes in the imaged area caused by capillary recruitment and increased blood supply by metabolic demand. A second component arises from the changes of the hemoglobin oxygen saturation of the blood in the active area. It is composed of the absorption properties of the oxy-hemoglobin (HbO_2) and the deoxy-hemoglobin (Hbr). Another main compo-

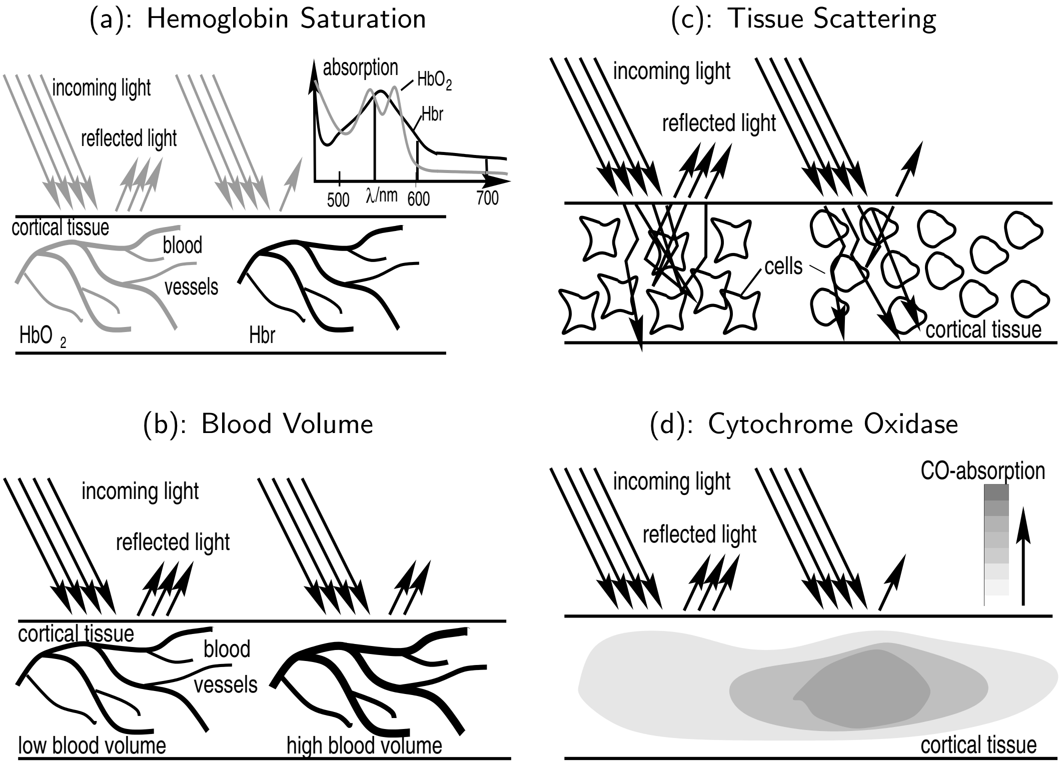


Figure 3.6: Biophysical origins of the intrinsic signals. **(a)** Due to their metabolic activity the neurons decrease the oxygen saturation of the hemoglobin. The ratio of the deoxy-hemoglobin component (Hbr) and the oxy-hemoglobin (HbO_2) is changed. The insertion shows the absorption book spectra of Hbr and HbO_2. The isobestic points are at the wavelengths where the spectra cross. **(b)** With the rising metabolic demand more blood is delivered to the tissue by increasing the blood volume. Mechanisms for the increase are capillary recruitment and vasodilation. **(c)** The excitation induces cell swelling in the neurons and therefore changes the scattering properties of the tissue. **(d)** Once the oxygenation state of the surrounding tissue changes the spectral behaviour of cytochrome oxidase is altered (adapted from Stetter (2000)).

nent of the intrinsic signal is caused by the change of the light scattering in the neural tissue due to cell swelling during activation. The final component we will consider here originates from the absorption characteristics of cytochrome oxidase, that change with the oxygenation state of the tissue. Compared to the temporal resolution in the millisecond regime one can obtain by imaging with voltage sensitive dyes (Shoham et al., 1999) all of the intrinsic signals are very slow ($0.5 - 2.0$ seconds onset and rise time). Many important questions about the columnar and modular organisation of the visual cortex do not depend on the very high temporal resolution but benefit from the excellent spatial resolution and the non invasiveness that optical imaging of intrinsic signals delivers.

In the measured data we have a superposition of all the intrinsic signals with biological noise artifacts such as respiration and heart beat, plus sensor noise sources due to photon shot noise and camera read out noise. Our goal is to separate from this mixture the intrinsic signal components that are closely localised in the areas of stimulus specific neural activity. This component is called the "mapping signal" because it will later form the functional maps we are looking for. Connected with this localised activity of the mapping signal is an overall increase of the background metabolic activity and blood volume changes on a coarser spatial scale to satisfy the oxygen demand. This so called global signal is also evoked by the stimulus but will not reflect the stimulus specific activation as it sometimes spreads over $3-5mm$ (Malonek and Grinvald, 1996). In chapter 4 we will introduce the methods to separate the mapping signal from the global signal and noise. All of the sources of intrinsic signals that will be described more closely now can contribute to both, the mapping and the global components (Frostig et al., 1990).

3.2.1 The Blood Oxygenation Component

The oxygen saturation level of hemoglobin changes either from metabolic demand or difference in the blood flow that alters the relative amounts of deoxy-hemoglobin (Hbr) and oxy-hemoglobin (HbO_2). All the blood related components have a large absorption of light at wavelengths in the range of $500-600nm$ (Malonek and Grinvald, 1996). In the resulting image a large absorption gives a lower intensity in the reflected light and therefore a darker area in the image. The insertion in figure 3.6 a) shows the wavelength dependency of the oxy and deoxy components. Important to mention are the points were the Hbr and HbO_2 spectra intersect as the intrinsic signals at these wavelengths are independent of the oxygenation level. These wavelengths are called isobestic wavelengths.

At wavelengths below $590nm$ the mapping component is only $5-10\%$ of the activity dependent reflection signal whereas at $605nm$ it makes up $39-50\%$ of the reflection signal (Bonhoeffer and Grinvald, 1996). This additional contribution to the mapping signal above $590nm$ most likely arises from the saturation level of the hemoglobin due to the increased oxygen consumption of the neurons. In the paper of Malonek and Grinvald (1996) the spectral and temporal characteristics of the Hbr and HbO_2 component in the mapping and the global signal were investigated.

The deoxy-hemoglobin component in the mapping signal starts to rise $200msec$ after the stimulus onset and continues to rise during a stimulus presentation up to eight seconds. Longer stimulus times beyond eight seconds lead to no more increase of the amplitude (same is true for HbO_2). Once the stimulus stops the Hbr component decays to baseline in $15-20s$. This slow decay must be considered in the recovery periods between the trials to avoid artifacts from the preceding stimulus. At $605nm$ the amplitude of the deoxy component is $\sim 30\%$ of the amplitude of the global component.

When nerve cells are active they metabolise oxygen and therefore oxygen diffuses from nearby capillaries to these cells. This leads to a decrease in oxygen saturation and the early deoxygenation component is therefore strongly localised to neural activity.

The oxy-hemoglobin component has a latency of $1-2s$. It rises continuously for even up to three seconds longer than the stimulus presentation. The decay to baseline is slower than in the Hbr component $(4-6s)$. The peak amplitude is about the same as in the HbO_2 but has a $1-3s$ delay.

3.2.2 The Blood Volume Component

The change in the blood volume during neural activity is due to local capillary recruitment or dilation of small veins or increased concentration of blood cells in the capillaries. These changes appear as an increase in hemoglobin absorption (Grinvald et al., 1991). The contribution of the blood volume component can best be investigated at isobestic wavelengths (i.e. $570nm$) where the oxygenation state of the blood does not influence the absorption. It was shown by Frostig et al. (1990) that functional maps can be obtained directly from blood volume changes. Capillaries and small veins seem to be recruited in a very localised fashion at the site of stimulus specific activation. Nevertheless the overall intrinsic signal from blood volume changes is much bigger than the mapping component and most of the activity dependent blood volume changes are regulated at a spatially spread domain that does not reflect the functional organisation.

3.2.3 The Tissue Scattering Component

The component of the intrinsic signal that causes changes in the scattering properties of the cortical tissue during activation originates from a swelling of the neurons. This change is induced by ion and water movement, expansion and neuro transmitter release (Cohen, 1973; Kreisman et al., 1995). The intrinsic signal by cell swelling is localised to regions of $100-300\mu m$ around the active neurons (MacVicar and Hochman, 1991). Experiments on hippocampal slices showed that the amplitude of the scattering component is only 0.01% in the beginning and builds up to more than 1.0% with repetitive stimulation for one second (Frostig et al., 1990). After the stimulus ends this component decays the fastest of all intrinsic signals (Malonek and Grinvald, 1996). The time course in a blood free slice preparation is independent of the wavelength and the amplitude only decreases slightly at long wavelengths.

3.2.4 The Cytochrome Oxidase Component

The chromophore cytochrome oxidase changes its absorption characteristics with the level of oxygenation of the surrounding tissue. This is maybe the most controversial component as the exact mechanisms are not fully understood. As cytochrome oxidase is present at higher concentrations in the cytochrome oxidase

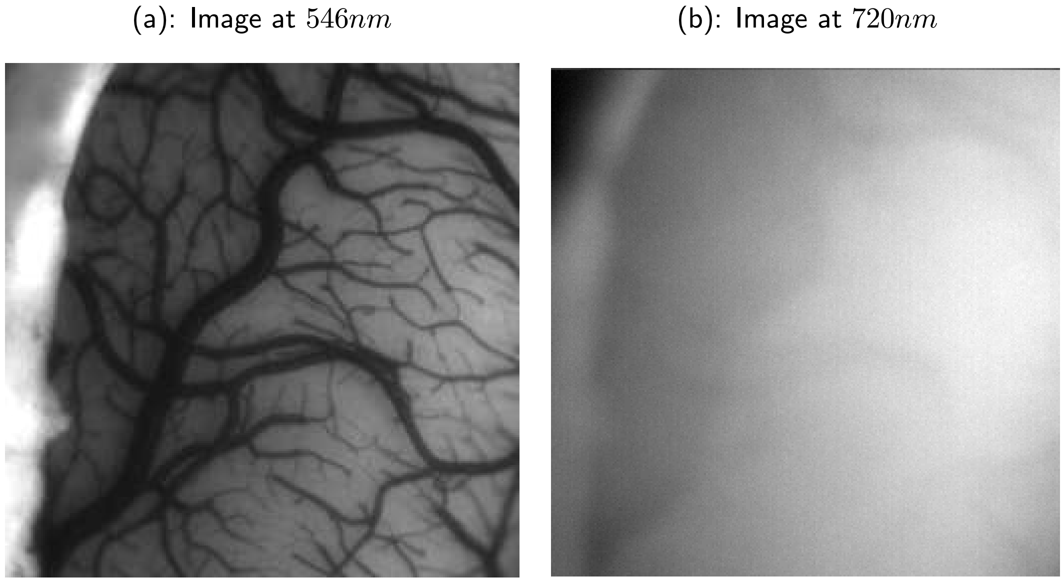


Figure 3.7: Difference in absorption and scattering of cortical tissue and superficial vasculature at $546nm$ (green light) and $720nm$ (near infra-red). The depth of focus is the same in both images. The blurring is caused by the bigger penetration of the tissue with light at higher wavelengths. The vessel pattern is represented less dominant because of the reduced absorption of hemoglobin at $720nm$. The images show a $3.7 \times 3.7mm^2$ area of visual cortex.

blobs in cortical layers $1-3$ of V1 and these blobs are supposed to be involved in colour vision, this signal could prove to be useful for the analysis of colour specific stimuli.

3.2.5 Wavelength Dependencies of Tissue Scattering and Artifacts

Comparing the overall absorption of the cortical tissue, including the capillaries and blood vessels, and therefore the change in brightness in the image at different wavelengths one can see that considerably more light is absorbed at wavelengths between $500-600nm$ than in wavelengths above (Malonek and Grinvald, 1996). At high wavelengths the image simply is brighter at the same illumination strength because of the reduced absorption. This change in absorption results from the decrease of absorption of hemoglobin in the capillaries, small arterioles and veins and is in agreement with the hemoglobin absorption book spectra (see insertion in figure 3.6 a)).

The second quantity that drastically changes with the wavelength is the scattering of light and therefore the blurring of the acquired image. The longer the wavelength is the deeper the light can penetrate the tissue. On its scattering path the single photon integrates the information from deeper layers, but less photons will backscatter and contribute to the image (Stetter and Obermayer, 1999).

An extensive study of a high resolution spatial comparison between optical maps from intrinsic signals obtained at different wavelengths and the overlaying vasculature through which these maps are recorded was done by McLoughlin and Blasdel (1998). In figure 3.7 the same piece of cortex is shown without changing the focus at different wavelengths. The vessel pattern that can be seen clearly at $546nm$ illumination is not just blurred due to increased scattering at $720nm$. Because of the lower absorption of hemoglobin at the higher wavelength the contrast and therefore the signal intensity of the vessel pattern is reduced. The size of the blurring due to light scattering has been estimated by Orbach et al. (1985) to be smaller than $200\mu m$. As a consequence in the data obtained at $720nm$ the proportion of the signal from vessel artifacts is reduced. McLoughlin and Blasdel (1998) found that the 600- and $720nm$ optical maps are strongly correlated in tissue compartments not underlying the surface vasculature, but while $720nm$ optical maps exhibit a flat distribution of signal strength across the various tissue compartments the regions associated with the surface vasculature exhibit much stronger signals in all the $600nm$ images. This effect can not be neglected as up to 40% of the cortex surface is covered with vasculature.

A commonly suggested method for getting “rid” of the vasculature artifacts at wavelengths like $605nm$ is to focus $300 - 400nm$ below the cortex using a lens of small focal depth (Grinvald et al., 1991). This is very misleading as the image might look like the blurred $720nm$ image now, but the signal content is different. The absorption signal of the vasculature certainly is still present and only smeared over a wider area therefore contaminating a even bigger region of the optical maps. In our setup we rather focus exactly on the cortex surface with a lens with a large focal depth and remove the vessel artifacts afterwards with sophisticated statistical methods.

One further advantage of long wavelengths is the reduction of the change in the absorption by the respiration and heart beat. The change in blood pressure during a respiratory cycle and the pulsation of the heart beat change the blood volume in the imaged area and strongly modulate the brightness in the image. As the influence of the blood volume component decreases with higher wavelengths this artifact can be reduced by changing from $530nm$ to $630nm$ by a factor of more than ten (Shoham et al., 1999) and another two fold by switching from $600nm$ to $720nm$ (McLoughlin and Blasdel, 1998).

Chapter 4

Statistical Signal Processing Algorithms

New techniques and advances are emerging in the field of statistical signal processing that deserve the attention of the biomedical and neuroscience community. Several algorithms have been proposed to separate multiple signal sources on the basis of their statistical properties, instead of the more common spectral features. These algorithms have the promise to lead to more accurate source modelling and more effective artifact rejection algorithms, two of the most challenging conditions faced in biomedical signal processing (Principe et al., 2000).

This so called Blind Source Separation (BSS) has been successfully applied to other biomedical data such as functional magnetic resonance imaging (fMRI) (McKeown et al., 1997; McKeown et al., 1998), electroencephalographical measurements (EEG) (Makeig et al., 1996), and cardiovascular signals (Vetter et al., 1999).

In chapter 3 we described how the optical imaging data for analysis are acquired. Some of the examples that will be shown are from species other than macaque monkey, but the basic recording technique is the same. The raw data from the optical imaging experiment contain the stacks from the recording of the individual trials. In those trials the different stimulus conditions were presented in random or alternating order. The task for the statistical analysis of these datasets is to reliably separate the signal that corresponds to the neural activity due to the stimulus presentation (mapping signal) from all other components that are described in chapter 3. We have no information about the mixing process itself nor do we know the exact spatial pattern of the sources that underly the recorded mixture. A further difficulty is the contamination of the data with high levels of noise.

In this chapter we will first explain how the basic preprocessing of the raw data is done and introduce a mathematical description of the dataset. Then established heuristic methods, like differential imaging are described and we have a look at some of the drawbacks with the use of these procedures on biological signals. The following section will show the differences between spatial and temporal analysis of the data proceeded by a description of BSS algorithms like principal component

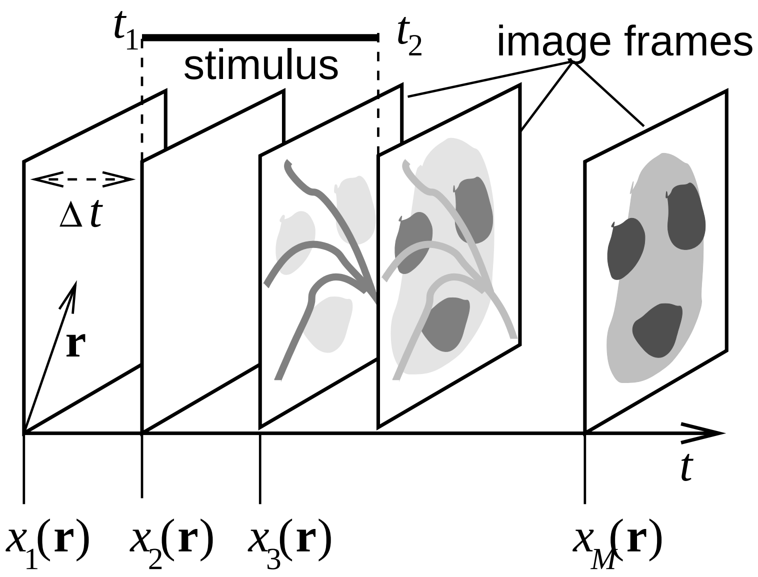


Figure 4.1: Outline of the data collection procedure for a single stimulus presentation: A stimulus is presented between times t_1 and t_2, while a camera takes a sequence of M_0 images $x_t(\mathbf{r}), t = 1, ..., M_0$ before, during, and after stimulation. The shaded patterns sketch the changes in reflectance over time, which are assumed to be made up of different spatial prototype patterns.

analysis (PCA) and independent component analysis (ICA).

In the last part we will have a look at a second order blind source separation algorithm called extended spatial decorrelation (ESD) and some extensions of it.

4.1 Basic Preprocessing of the Raw Data

A typical set of raw data, that was recorded as described in the chapter 3, contains the individual trials of the experiment plus some header files, that have information about the stimulus condition and camera parameters. The individual trials are numbered in ascending order and the format of the raw data is unsigned 16 bit integers. All trials contain the same number of M_0 frames $m_0 = 1, .., M_0$ with P_x x P_y pixels and the stimulus is always presented between the times t_1 and t_2 (see figure 4.1). As a first preprocessing step all the trials with the same stimulus condition are summed up frame by frame. This is justified by the assumption that all non metabolic random noise is reduced in the summation, whereas the prestimulus activity and the stimulus locked activity are enhanced to give a better signal to noise ratio. Examples of these random noise sources are the photon shot noise of the illumination and the read out noise of the CCD chip in the camera.

The result of this first step is one stack of M_0 frames with P_x x P_y pixels for each stimulus condition (single condition stack). With a frame rate of between

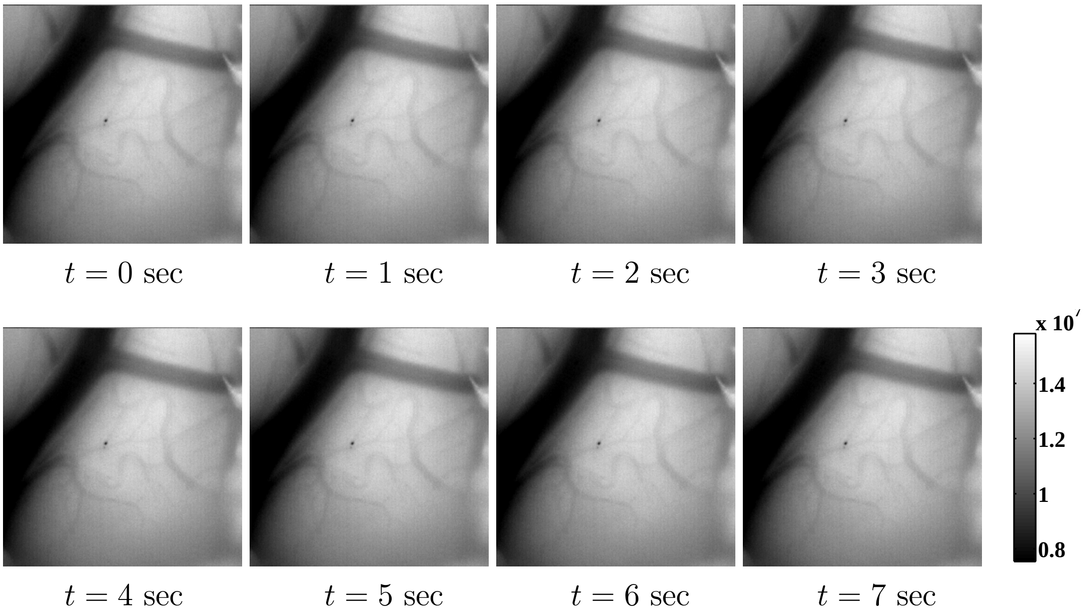
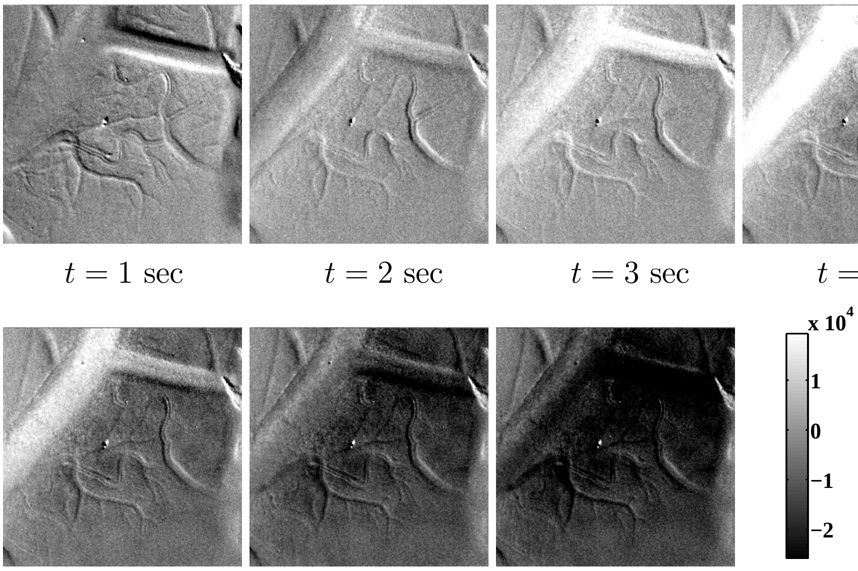


Figure 4.2: Single condition stack after the preprocessing steps of summation of the trials with same stimulus condition and binning of the frames. The $M_1 = 8$ frames now contain the information from one second of data collection. Due to the large contrast range within each frame, the small stimulus related mapping signal is still not visible.

$7.5fps$ and $30fps$ and a recording time of 5-15 seconds the number of frames in the stack M_0 is normally around 120.

For most of the analysis described later a much smaller number of frames is sufficient, saving computer memory and calculation time. The signal to noise ratio is enhanced even further by binning the M_0 frames in time as the slow changes of the signal components are not affected by the binning but the fast temporal changes of the unwanted noise cancel out. Normally the frames are binned by the factor of the recording frame rate, so each summed frame contains the recorded information of one second. This results in M_1 frames with P_x x P_y pixels each. For example when 32 trials were recorded in one experiment with two stimulus conditions and a frame rate of 15 fps and a recording time of 8 seconds we have 16 trials per condition each having $M_0 = 120$ frames. After the summation of the trials we have two stacks, one for each condition, with $M_0 = 120$ frames. The binning of the frames with a ratio of 15 ($\Rightarrow$ 15fps) results in two stacks, one for each condition, with $M_1 = 8$ frames (see figure 4.2). The size of the individual frame, i.e. P_x x P_y, has not changed during the whole procedure. All trials were recorded in the same manner, i.e. the stimulus always started at time $t_1 = 2$s after the beginning of the recording. Therefore it can be assumed that all frames of a stack before t_1 contain only baseline activity and the information about the unevenness of illumination, but no stimulus relevant signal (presuming that the



$t = 1$ sec $t = 2$ sec $t = 3$ sec $t = 4$ sec

$t = 5$ sec $t = 6$ sec $t = 7$sec

Figure 4.3: Same single condition stack as shown in figure 4.2, after first frame analysis. The former first frame contains all zeros and is neglected. So now the single condition stack contains $M = M_1 - 1 = 7$ frames. Without the pre- stimulus activity present the gray values of the image are spread over a much smaller range and make details visible, that could not be seen in figure 4.2.

recovery period was long enough). The stimulus correlated mapping signal is much smaller than the baseline activity and the variance of the illumination across each frame. To enhance the signal ratio of the stimulus related changes to the other components, the first frame of the summed and binned stack is subtracted from all subsequent frames. This method is called first frame analysis (Bonhoeffer and Grinvald, 1996; Blasdel, 1992a). As the first frame now contains all zeros it is neglected and the resulting single condition stacks have $M = M_1 - 1$ frames (see figure 4.3). If not explicitly stated all the data stacks considered from now on have undergone summation, binning and first frame analysis. Therefore the notation for a single condition stack with M frames m=1,..,M and P_x x P_y pixels for each frame is:

$$\mathbf{x}(\mathbf{r}) = x_m(\mathbf{r}) \tag{4.1}$$

with $\mathbf{r} = (r_x, r_y), 1 \leq r_x \leq P_x, 1 \leq r_y \leq P_y$.

4.2 Conventional Methods for Analysis of Optical Images

Since the beginning of optical imaging elaborate ways had to be found to extract the signal of interest from the recorded mixture. The difficulty in this analysis is that the mapping signal is only about 0.1% of the total intensity of the reflected light (Blasdel and Salama, 1986; Bonhoeffer and Grinvald, 1996). This means that all other components like biological noise and photon shot noise are bigger than the mapping signal. Before any analysis has taken place a lot of effort can be made to enhance the signal to noise ratio. If the images are always recorded at light intensities close to the camera saturation the photon shot noise is as low as possible. Furthermore the biological noise can be reduced, if the recording is synchronised with the respiration and the heart beat. Certainly also a high number of trials for summation enhances the signal quality.

4.2.1 Differential Imaging

One of the first methods that was introduced to reveal the stimulus dependent part of a optical imaging experiment was called differential imaging (Blasdel and Salama, 1986). In a first step the individual trials belonging to one stimulus condition are summed up and first frame analysis is applied. Depending on the recording method this is done directly in the computer memory during the experiment or in the later processing as in our case. The basic assumption for differential imaging is, that the only signal that changes with the presentation of different stimuli is the stimulus correlated mapping signal. Furthermore the chosen stimuli should be so called orthogonal stimuli. This means that the stimuli activate locally disjunct cell populations when presented to the animal. Typical orthogonal stimuli are the alternating stimulation of the left and right eye, that leads to the recording of ocular dominance bands or the presentation of gratings orthogonal to each other, that gives us orientation maps. To get a differential image, the two single condition stacks for the orthogonal stimuli are subtracted from each other frame by frame. This leads to a so called difference stack (see bottom row of figure 4.4 (a)). The sum of all frames from this difference stack is the difference image (see figure 4.4 (b)). Ideally in the difference image all signals that are not stimulus dependent are removed.

The first big drawback of this method is that not only the mapping signal in optical imaging data is correlated with the individual stimulus, but also parts of the global signal can be locked to only one stimulus or the DC level of the background activity might change between the trials.

The second disadvantage of this method is, that one needs a quite detailed knowledge of the cortical response to a stimulus in order to design a orthogonal pair of stimulus conditions. For classical stimuli like left eye - right eye stimulation for ocular dominance this might be true, but there are stimulus regimes like the

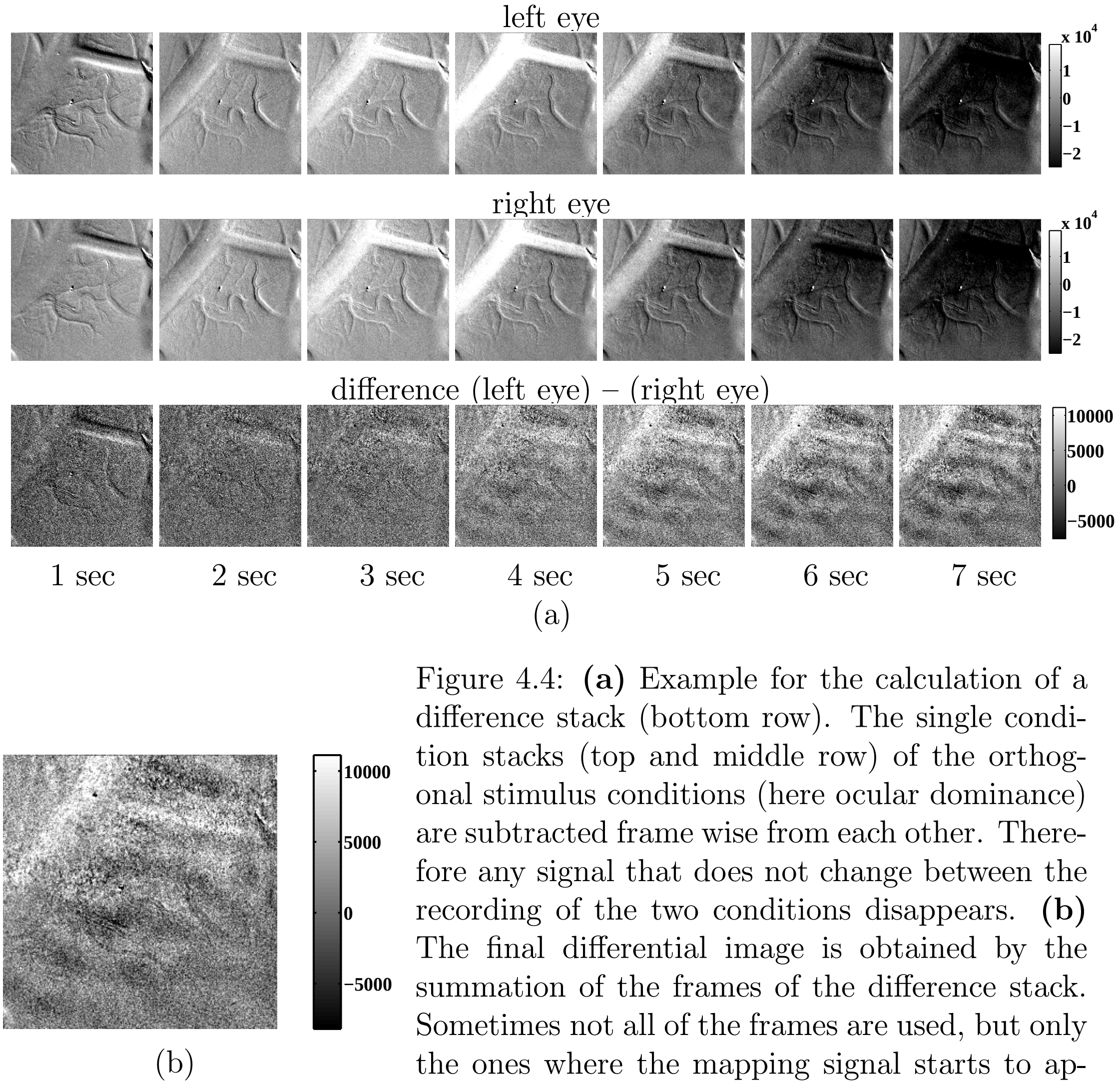


Figure 4.4: **(a)** Example for the calculation of a difference stack (bottom row). The single condition stacks (top and middle row) of the orthogonal stimulus conditions (here ocular dominance) are subtracted frame wise from each other. Therefore any signal that does not change between the recording of the two conditions disappears. **(b)** The final differential image is obtained by the summation of the frames of the difference stack. Sometimes not all of the frames are used, but only the ones where the mapping signal starts to appear.

activation of a cortical point spread function where an orthogonal stimulus can not be designed.

4.2.2 Cocktail Blank and Blank Images

The method of using cocktail blank or blank images for recovery of the mapping signal is described by Bonhoeffer and Grinvald (1996) and works as follows. In order to get rid of the baseline activity and any uneven illumination the image that was recorded with a stimulus is divided by a image of the unactivated cortex. This image of the unactivated cortex is called a blank image in this document. The other proposed method is to obtain a image of the cortex while it is activated in a uniform manner. This means in this image all areas of the cortex show the activity level they would have during their maximum stimulation. The image of the uniformly activated cortex is called cocktail blank. The recorded activity map is now divided by this cocktail blank.

Both methods have their disadvantages. Processing with the blank image often leaves strong activity related blood vessel artifacts in the maps, that can distort the image statistics. When the cocktail blank is used one introduces strong assumptions about the functional architecture of the cortical area under investigation (Bonhoeffer and Grinvald, 1996). It is difficult to ensure, if the stimulus regime chosen to obtain the cocktail blank activates the cortex uniformly and does not impose a pattern by itself.

This is especially true when calculating single condition maps using a cocktail blank image. Single condition maps are obtained by processing a data stack that results from the summation of trials of one stimulus condition. These maps should explicitly show the cortical response to a specific single stimulus. If in the processing of this map a cocktail blank is used one introduces the response information of other stimuli that were used for calculating the uniform activation of the cortex. Often orthogonal stimulus conditions are used to produce a cocktail blank. The single condition map, calculated with this cocktail blank, now carries the information of the orthogonal stimulus too and is more of a difference image than a single condition map.

A further problem with the correction of the uneven illumination by division is the introduction of a nonlinear relation. The method is justified with the small amplitudes of the mapping signal compared to the overall signal (Bonhoeffer and Grinvald, 1996). But the problem is not considered that the blank or cocktail blank image can have near zero values and therefore the division with this can result in arbitrary big numbers.

4.2.3 Bandpass Filtering

The last of the conventional methods regularly used to enhance the signal quality is the use of bandpass filters. The basic assumption here is that the individual signal components have different spatial frequencies and can be separated or cancelled out in the frequency domain after a discrete Fourier transform. In the frequency domain a rotation symmetric function is used to cut off the low frequency and the high frequency part of the transform. This is motivated by the assumption that all high frequency components are due to random noise and can not originate from the intrinsic signals because of the blurring of the cortical tissue by scattering (Stetter and Obermayer, 1999). The low frequency components of the image are known to derive from the global signal, that has a more coarse spatial distribution than the mapping signal, for example ocular dominance or orientation maps. Hence the mapping signal is believed to have spatial frequencies between the both.

Apart from the non trivial difficulty to choose a optimal filter function, that has smooth enough edges and does not introduce artifacts by its own back transform, one can never be sure that there is no overlap between the frequency bands for

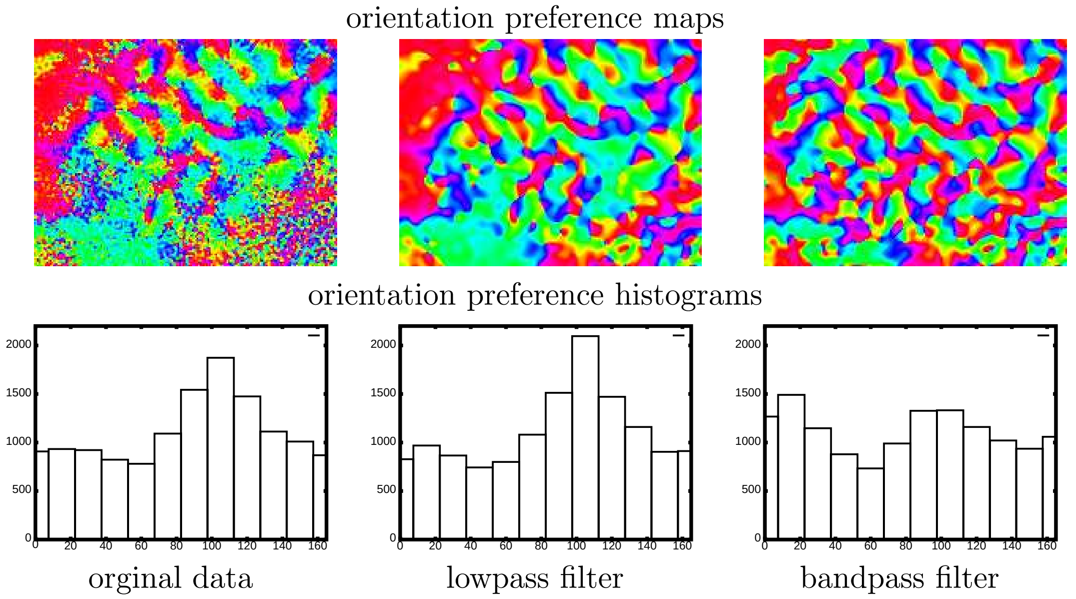


Figure 4.5: Influence of filtering on the statistics of optical imaging maps. The left column shows the orientation preference map (top) and the corresponding orientation preference histogram for unfiltered data. In the middle column a $(0/3.25)1/mm$ lowpass filter was applied. There is a slight change, but the overall distribution in the histogram stays about the same. This supports the assumption, that most of the high frequency components rise from noise. For the map of the right column a $(0.43/3.25)1/mm$ bandpass filter was used. Now we see a drastic change in the image statistics. The colours in the top row represent the preferred orientation of the cells (see also figure 2.9).

the individual signal components. For example the surface vessel pattern contains a wide spectral range due to their elongated nature and can not be separated in this way. There are several other reasons why bandpass filters seem to be critical. After the cut out of frequency components that are supposed to contain the mapping signal, the tails of the frequency distributions of the other sources are still contained in the signal and further processing has to be done. The signal to noise ratio has increased now, but a lot of interesting image parameters like the singularity location and the singularity density of orientation maps are very sensitive to such changes.

Figure 4.5 shows how these parameters can drastically change with the arbitrary selection of the highpass cut off frequency. The images in the left column show the orientation preference map (top) and the distribution of the individual orientation of the unprocessed data (bottom). After the application of a low-pass filter the high frequency speckles in the orientation preference map are lost (middle column) but more importantly the asymmetric distribution of the orientations has hardly changed. This supports the argument above, that due to tissue scattering

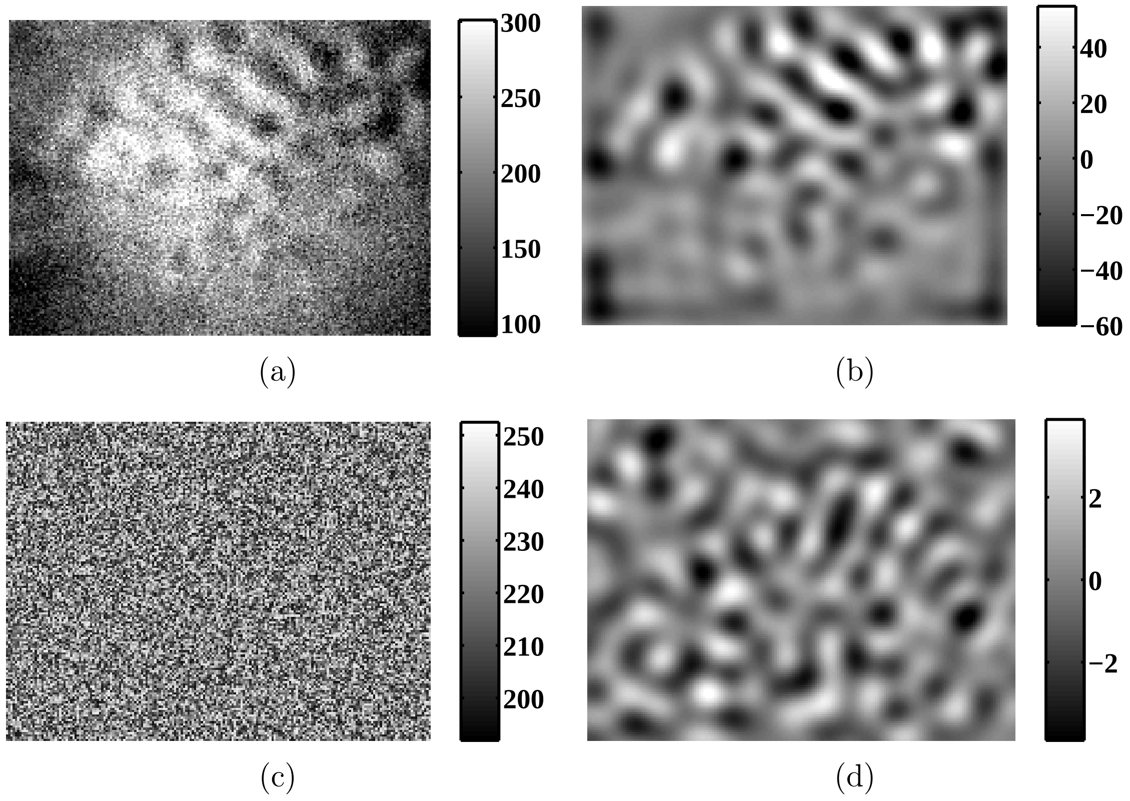


Figure 4.6: Bandpass filtering of white noise can result in image patterns that are not distinguishable from the real mapping signal. **(a)** 0-90 deg differential image from ferret primary visual cortex , and **(b)** the same image after bandpass filtering (highpass cutoff: 0.43 mm^{-1} ($\lambda = 2.3$ mm); lowpass cutoff: 1.63 mm^{-1} ($\lambda = 0.61$ mm)). **(c)** Random white noise image with same mean and variance than **(a)**, and **(d)** its bandpass-filtered version (same cutoff frequencies) (adapted from Stetter (2000)).

high frequency components rise from random noise. More critical is the introduction of the high-pass cut off (right column). The map now looks much better as the orientation patches are spread more evenly like one would expect them to be, but the distribution statistic has changed completely. The highpass cutoff value has to be chosen arbitrarily and this makes it difficult to believe in the resulting data.

The following example illustrates the difficulty in interpreting bandpass filtered results. As shown before the orientation maps from V1 in many species are arranged in a periodic pattern. The back transform of bandpass filtered noise can show very similar patterns and Rojer and Schwartz (1990) used bandpass filtered noise to describe the phenomenon of orientation maps. Figure 4.6 a) shows the differential image from a ferret V1 calculated after the response to a vertical and horizontal stripe pattern was recorded. In b) this image is bandpass filtered and the orientation pattern seems to be clearly visible. The image d) shows the band-pass filtered version of random white noise (image c)) that has the same mean and variance as a). For both b) and d) the same cut off frequencies were used. Even if the signal amplitude in the image generated from the noise is smaller than in the filtered optical image, the influence on the important near zero values of singularities can be crucial. This influence becomes even bigger when the recorded optical signal was not very strong.

All the methods introduced so far have been proven to be valuable tools for the analysis of optical imaging recordings, when great care is taken with respect to their limitations. A lot of technical effort is made during the experiment itself to make the basic assumptions for these methods more applicable.

The in vivo work does not always deliver good signal to noise ratios but the respect for the animal demands the use of the data and a careful analysis. Furthermore a lot of interesting stimulus designs for optical imaging do not provide an orthogonal stimulus condition and if completely new parameter regimes are explored or other cortical areas are investigated the pattern of critical responses might not be predictable anymore and hard to distinguish from artifacts.

All these reasons and difficulties have led us to the investigation of new analysis methods in the field of optical imaging. The algorithms we are looking for should be less heuristic and parameterised but make strong use of the statistical properties of the imaged mixture of sources and use these for separation. All these requirements are fulfilled by a group of algorithms that are gathered under the name blind source separation.

4.3 Introduction to Algorithms for Blind Source Separation

Blind source separation (BSS) is an emerging signal processing technique that aims at recovering unobserved signals or sources from a set of observed linear mixtures of theses sources. The mixtures are functions of space or time and are detected by a set of sensors. However no direct information about the original sources is available nor is the exact mixing process known, giving rise to the adjective “blind”.

The classical example for BSS is the so called cocktail party problem. Imagine two people speaking simultaneously. There are two microphones that are positioned at different distances from the speakers and therefore record a two different mixtures of the two voices. The two microphones give you two time signals, that we will call $x_1(t)$ and $x_2(t)$ with x_1 and x_2 being the amplitudes of the measurement and t the time index. Each of the recordings is the weighted sum of the speech signals emitted by the two speakers, who are the original sources and are denoted with $s_1(t)$ and $s_2(t)$. So now we can formulate this as a linear equation:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned} \tag{4.2}$$

where a_{11}, a_{12}, a_{21} and a_{22} are some parameters that depend on the distances of the microphones from the speakers. As we have only the recordings $x_1(t)$ and $x_2(t)$ and no information about the original sources $s_1(t)$ and $s_2(t)$ or the mixing process we need a procedure to estimate the sources from the recorded mixtures. With the BSS algorithms we can demixe the recordings so that we have two tracks with only the voice of one speaker on it.

So how does this apply to our optical imaging experiment? Let us have a look at figure 4.1 again (see page 38). We see a sketch of a data stack from an optical imaging experiment. As defined before in equation 4.1 the notation for a frame stack is $\mathbf{x}(\mathbf{r}) = x_m(\mathbf{r})$. When the stimulus is presented at time t_1 due to the metabolic activity in the cortex the absorption and reflection of the tissue changes. The changes each give rise to a characteristic pattern $s_j(\mathbf{r}, t), j = 1, ..., N$ in the data stack. For example, s_1 may describe reflectance changes due to the mapping signal, s_2 those due to the global signal and s_3 changes occurring within large blood vessels (Schießl et al., 2000c).

Under the assumption that the signal components are spatio/temporally separable, these patterns can be written as:

$$s_j(\mathbf{r}, t) = a_j(t)s_j(\mathbf{r}), \quad j = 1, ..., N \tag{4.3}$$

where $s_j(\mathbf{r})$ is the spatial pattern of the source s_j, that is the same at all times in each frame and $a_j(t)$ is the amplitude of the source s_j and makes it appear and disappear over time in the stack. Because of the small intensities we regard the overall recorded intrinsic signal as an instantaneous linear superposition of this set of spatial prototype patterns. The frames of the optical imaging dataset are the recordings of these superpositions and can be rewritten as:

$$x_m(\mathbf{r}) = \sum_{j=1}^{N} a_{mj}s_j(\mathbf{r}) \tag{4.4}$$

where a_{mj} is the amplitude in the m-th frame and gives the time-course of s_j by $a_{mj} = a_j(t_m)$.

This is now our noise free model of the optical imaging data. When we introduce a noise term now we must differentiate between the so called source noise and sensor noise. Source noise describes a form of noise that is introduced by an additional source s_j where the signal is Poisson- or white noise. This noise therefore appears to be the same in all mixtures just with different amplitudes and is easy to separate from the real signal with BSS. More of a problem is the sensor noise, that we encounter in optical imaging. Here the noise is added after the mixing process and is different for each recorded mixture. Origins of this form of noise are the photon shot noise of the illumination and the read out noise of the CCD chip in the camera. If we add the terms for the sensor noise, our model becomes:

$$x_m(\mathbf{r}) = \sum_{j=1}^{N} a_{mj} s_j(\mathbf{r}) + n_m(\mathbf{r}) \tag{4.5}$$

Instead of the sum notation we can combine the coefficients a_{mj} in the mixing matrix $\mathbf{A}$ and the spatial component of the sources in the source vector $\mathbf{s}(\mathbf{r}) = (s_1(\mathbf{r}), ..., s_N(\mathbf{r})$ and rewrite equation 4.5 in vector notation as

$$\mathbf{x}(\mathbf{r}) = \mathbf{A}\mathbf{s}(\mathbf{r}) + \mathbf{n}(\mathbf{r}) \tag{4.6}$$

For the introduction of the concepts of BSS algorithms we neglect the noise for a moment and use the vector form of the general mixing model in equation 4.4:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{4.7}$$

Sometimes we need the columns $\mathbf{a}_j$ of the matrix $\mathbf{A}$ and write equation 4.7 in the form

$$\mathbf{x} = \sum_{j=1}^{N} \mathbf{a}_j s_j \tag{4.8}$$

like in equation 4.4.

The task is to calculate the original sources $\mathbf{s}$ from the measurement $\mathbf{x}$ without knowing the linear mixing matrix $\mathbf{A}$ nor the sources $\mathbf{s}$. This means we have to find a demixing matrix

$$\mathbf{W} = \mathbf{A}^{-1} \tag{4.9}$$

that reverses the mixing introduced by $\mathbf{A}$. In most of the applications one can only find an estimate (marked by a hat) for the inverse of $\mathbf{A}$ and therefore only calculate an estimate of the original sources. Hence we formulate the demixing process as

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x} = \hat{\mathbf{A}}^{-1}\mathbf{x} \tag{4.10}$$

From the equation 4.7 we can easily see that there will be some ambiguities in the estimation of $\hat{\mathbf{s}}$.

The first ambiguity is that we can not determine the amplitude of the independent components . This is because both $\mathbf{s}$ and $\mathbf{A}$ are unknown and any scalar multiplier in one of the sources s_j should always be cancelled by dividing the corresponding column $\mathbf{a}_j$ of $\mathbf{A}$ by the same scalar. Another consequence of this is that we can not estimate the sign of the sources because we can multiply the independent component by -1 without affecting the model.

The second ambiguity is the order of the independent components. Again because of $\mathbf{s}$ and $\mathbf{A}$ being unknown we can freely change the order of the terms in the sum of equation 4.8. Formally a permutation matrix $\mathbf{P}$ and its inverse can be substituted in the model to give $\mathbf{x} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}$. The elements of $\mathbf{P}\mathbf{s}$ are the original independent variables s_j just in another order. The matrix $\mathbf{A}\mathbf{P}^{-1}$ is then a new unknown mixing matrix we have to estimate by the BSS algorithms (Hyvärinen and Oja, 1999). Luckily in most applications the order of the sources is of no importance.

4.3.1 Temporal vs. Spatial Analysis

In a data stack that contains a mixture of spatio-temporal separable signals, like in the optical imaging recordings, there are at least two ways to look at the data and analyse it. The first and more intuitive way is to look at the individual pixels of the frames and regard the values of a single pixel at the position $\mathbf{r}$ across all frames as the measurement. This is then a so called temporal analysis (see figure 4.7). Compared to our example of the cocktail party problem each pixel is a sensor, i.e. a microphone, and the measurement, i.e. the recording of the voices, are the M values of this pixel. In this case we have $P = P_x \times P_y$ sensors with M recorded data points each.

The other way to look at the dataset is by spatial analysis. Here we consider each individual frame to be a sensor and all the pixels in one frame are the measurement. In the words of the cocktail party example each frame now is a microphone and the $P = P_x \times P_y$ pixels are the measurement. In the analysis of optical imaging datasets with BSS algorithms we found certain advantages of a spatial analysis over temporal analysis. The individual sources in our linear mixture, the intrinsic signals, tend to have similar time courses after the presentation of the stimulus started. Therefore this is not a good feature for separation under information theoretic aspects, as the signal information is very similar and the size of the noise can be as big as the signal.

As shown in figure 4.7 we consider our individual sources s_j to have spatially fixed patterns, that rise and vanish over time. In this case each frame contains a mixture of prototype patterns that have well distinguishable statistical features.

A further advantage of spatial analysis is the dimensionality of the data space. After summation and first frame analysis we are typically left with about seven frames of the size $P_x \times P_y = 256 \times 256$ pixels. This means we have a seven dimensional data space with $P = 65536$ measurements. In the temporal analysis we end up with a 65536 dimensional space and the maximum number of measurements we can use is the number of frames M_0 in the raw data, typically around 120. This defines a very sparse problem and difficult to solve by BSS algorithms. There are certain methods of rearranging the raw stacks to reduce the sparseness of the data for temporal analysis (Stetter et al., 2000) but the default procedure in the following chapters will be spatial analysis.

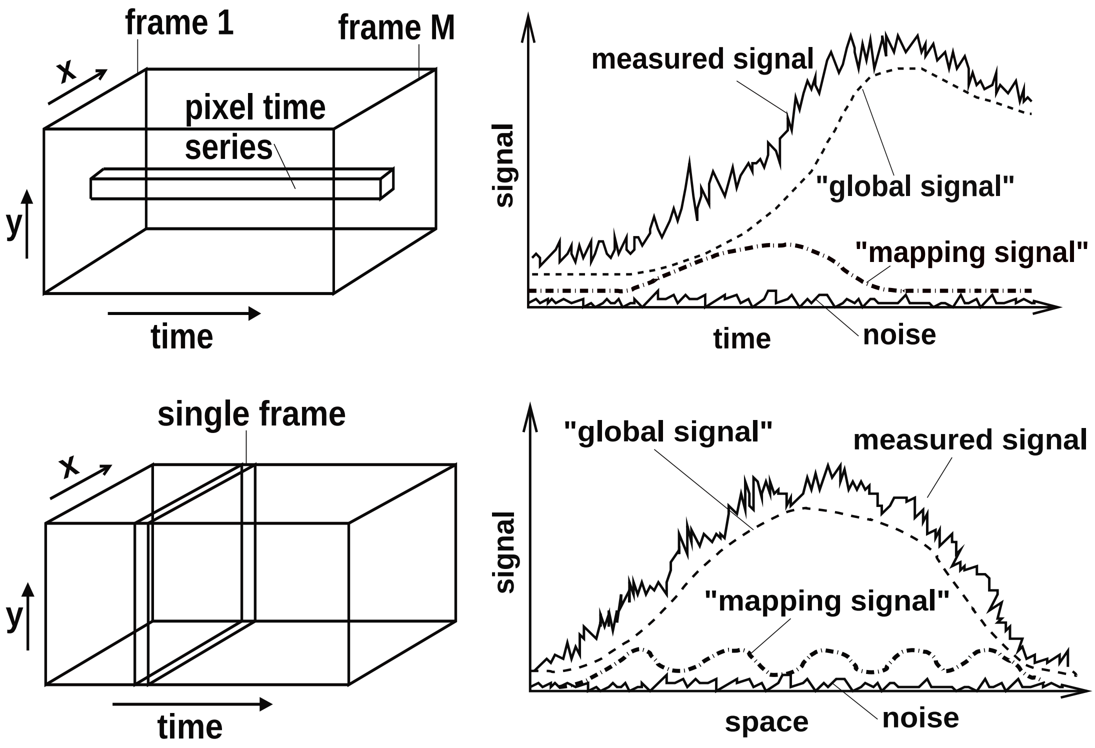


Figure 4.7: **(top row)**: In temporal analysis we consider each pixel to be the sensor and the values of this pixel in the single frames are the measurements. The right side shows the plot of the time course of the single components. **(bottom row)**: In the spatial analysis each frame is a sensor and the pixels in one frame are the measurement.

4.3.2 Basic Statistical Parameters

Before we derive the BSS algorithms in the following sections let us quickly repeat some basic statistical parameters with the example of the normal or Gaussian distribution.

The normal density function, for the case of a single variable, can be written in the form

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4.11}$$

where μ is called the "mean" or "average" and σ^2 the "variance". $\sigma = \sqrt{\sigma^2}$ is called the "standard deviation". The coefficient in front of the exponential in equation 4.11 ensures that $\int_{-\infty}^{\infty} p(x)dx = 1$. p(x) is called the "probability density function" (pdf). In the given example of the Gaussian distribution one has a high probability to draw a value x that is close to the mean μ (the maximum of the pdf

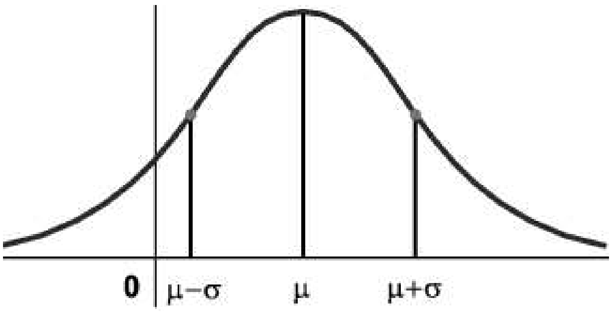


Figure 4.8: Plot of a Gaussian distribution around the mean μ with the standard deviation σ

$p(x)$ is at $x = \mu$, see figure 4.8). The definition for the mean μ is

$$\mu = E(\mathbf{x}) =< \mathbf{x} >= \int_{-\infty}^{\infty} xp(x)dx \tag{4.12}$$

$E(\cdot)$ or $< \cdot >$ are equivalent notations for the expectation value. Normally we do not know the pdf and only have a number of samples drawn from the distribution. In this case one can estimate the arithmetic mean $\hat{\mu}$:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{, where } x_i \text{ are realizations of } x \tag{4.13}$$

The variance σ^2 is defined by:

$$\sigma^2 = E((\mathbf{x}-\mu)^2) =< (\mathbf{x}-\mu)(\mathbf{x}-\mu) >= \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx \tag{4.14}$$

If $\mathbf{x}$ is a random sample of data from a normal distribution, the best unbiased estimate of its variance σ^2 is calculated with

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{4.15}$$

In the case of a two dimensional distribution of the two variables $(\mathbf{x}, \mathbf{y})$ we get the following statistical parameters:

$$\begin{aligned} \mu_x &= E(\mathbf{x}) =< \mathbf{x} >= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y)dxdy \\ \mu_y &= E(\mathbf{y}) =< \mathbf{y} >= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yp(x,y)dxdy \end{aligned} \tag{4.16}$$

are the expectation values for the variables $\mathbf{x}$ and $\mathbf{y}$.

$$\begin{aligned} \sigma_x^2 &= E((\mathbf{x}-\mu_x)^2) = E(\mathbf{x}^2) - (\mu_x)^2 \\ \sigma_y^2 &= E((\mathbf{y}-\mu_y)^2) = E(\mathbf{y}^2) - (\mu_y)^2 \end{aligned} \tag{4.17}$$

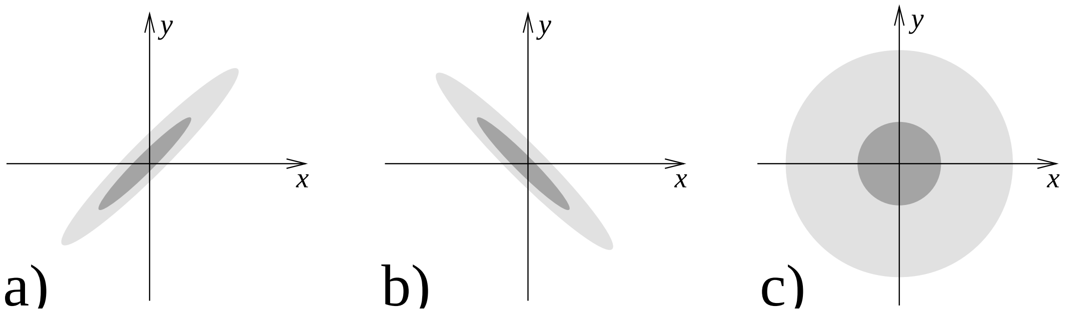


Figure 4.9: Sketch of the data distribution for a **a)** positive, **b)** negative and **c)** zero correlation between the samples **x** and **y**.

are the respective variances. The covariance of the samples **x** and **y** is defined by

$$\mathbf{C}_{xy} = E((\mathbf{x} - \mu_x)(\mathbf{y} - \mu_y)) = E(\mathbf{x}\mathbf{y}) - E(\mathbf{x})E(\mathbf{y}) \tag{4.18}$$

and the correlation coefficient is

$$\rho_{xy} = \frac{\mathbf{C}_{xy}}{\sigma_x \sigma_y} \tag{4.19}$$

The correlation coefficient is a measure for the dependency of the samples. In the figure 4.9 you can see an example of a a) positive correlation , b) a negative correlation and c) uncorrelated data. From the result $\rho_{yx} = 0$ one can only make the conclusion that **x** and **y** are statistically independent, if they have a two dimensional Gaussian distribution, with the following pdf

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\mathbf{C}_{xy}^2}} \exp\left(-\frac{1}{2(1-\mathbf{C}_{xy}^2)}\left(\frac{(\mathbf{x}-\mu_x)^2}{\sigma_x^2} - 2\frac{\mathbf{C}_{xy}(\mathbf{x}-\mu_x)(\mathbf{y}-\mu_y)}{\sigma_x\sigma_y} + \frac{(\mathbf{y}-\mu_y)^2}{\sigma_y^2}\right)\right)$$

In the final part of the statistics review we have a look at the d dimensional multivariate normal probability density that can be written in the following form (Bishop, 1995)

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\det(\mathbf{C}_\mathbf{x})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T {\mathbf{C}_\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \tag{4.20}$$

where the mean $\boldsymbol{\mu}$ is now a d-dimensional vector, $\mathbf{C}_\mathbf{x}$ is a $d \times d$ covariance matrix and $\det(\mathbf{C}_\mathbf{x})$ is the determinant of $\mathbf{C}_\mathbf{x}$. The pre-factor ensures again that $\int_{-\infty}^{\infty} p(\mathbf{x})d\mathbf{x} = 1$.

The pdf $p(\mathbf{x})$ is fully described by the mean $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{C}_\mathbf{x}$. The covariance matrix is a symmetric matrix and therefore has $d(d+1)/2$ independent components. There are d independent elements in $\boldsymbol{\mu}$ so the density function is completely specified once the $d(d+3)/2$ parameters have been determined. The quantity

$$\triangle^2 = (\mathbf{x} - \boldsymbol{\mu})^T {\mathbf{C}_\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{4.21}$$

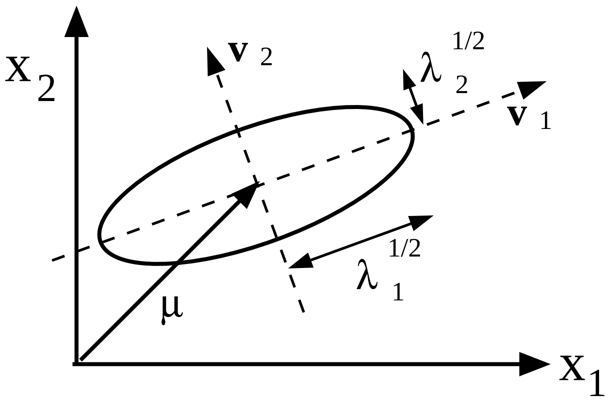


Figure 4.10: Plot of a two dimensional normal distribution. This distribution is completely determined by the mean vector $\boldsymbol{\mu}$ and a covariance matrix with eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$, and the corresponding eigenvalues λ_1 and λ_2.

that appears in the exponent of the equation 4.20 is called Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$.

It can be shown that the surfaces of a constant pdf for equation 4.20 are hyper-ellipsoids on which $\triangle^2$ is constant. Figure 4.10 shows an example for two dimensions. The principal axis of the hyper-ellipsoid are given by the eigenvectors $\mathbf{v}_i$ of $\mathbf{C}_\mathbf{x}$ which satisfy

$$\mathbf{C}_\mathbf{x}\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{4.22}$$

and the corresponding eigenvalues λ_i give the variances along the principal directions.

This Gaussian distribution can be simplified by considering a form in which the covariance matrix is diagonal

$$\mathbf{C}_{ij} = \delta_{ij}\sigma_j^2 \tag{4.23}$$

where δ_{ij} is the Kronecker delta. This is equivalent to decorrelating the x_i. Now the number of independent parameters is reduced to 2d because the principal directions are aligned with the coordinate axes (in figure 4.10 this would be a rotation around the ellipsoid centre, so that $\mathbf{v}_1$ and $\mathbf{v}_2$ point into the directions of the axes).

For a Gaussian distribution the components of $\mathbf{x}$ are now statistically independent as all second order statistics have vanished. The joint distribution can be written as the product of the distributions for each component separately in the form

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i) \tag{4.24}$$

This description of the factorising pdf's is the definition of statistical independence, and is also true for any non Gaussian pdf. All ICA algorithms are based on this definition.

We have now repeated all the basic statistical parameters we need to derive and understand the concepts of BSS algorithms.

4.3.3 Principal Component Analysis and Whitening

As the name implies principal component analysis (PCA) finds orthogonal directions, called principal components, within the probability density distribution of the data along which the variance has maximum values. In the example of the two dimensional Gaussian distribution the equation 4.22 is exactly doing this and the $\mathbf{v}_i$ in figure 4.10 illustrate the two principal components, as $\mathbf{v}_1$ shows the direction of the maximum variance and $\mathbf{v}_2$ the orthogonal direction with the next biggest variance. The vector notation for equation 4.22 is

$$\mathbf{C_x V} = \mathbf{\Lambda V} \tag{4.25}$$

where $\mathbf{C_x}$ is again the covariance matrix, $\mathbf{V}$ is a square orthogonal matrix (i.e.: $\mathbf{V}^{-1} = \mathbf{V}^T$), that has the eigenvectors $\mathbf{v}_i$ as columns and $\mathbf{\Lambda}$ is the diagonal matrix with the corresponding eigenvalues λ_i as diagonal elements, that are the variances along the corresponding eigenvectors. The elements of $\mathbf{C_x}$ are calculated as given in equation 4.18. Equation 4.25 holds for any pdf and not only for Gaussian distributions. Without loss of generality we can set a pdf to zero mean by subtracting the mean $\boldsymbol{\mu}$.

Let us apply this to our data model of the optical imaging stack with M frames and $P = P_x \times P_y$ pixels per frame. As the mean is subtracted ($\boldsymbol{\mu} = 0$) we can estimate the elements of $\mathbf{C_x}$

$$C_{x,mn} = E(x_m(\mathbf{r})x_n(\mathbf{r})) = \frac{1}{P}\sum_r x_m(\mathbf{r})x_n(\mathbf{r}) \tag{4.26}$$

where m, n stand for the rows and the columns of the matrix and $m, n \in 1...M$. In matrix notation this is

$$\mathbf{C_x} = \frac{1}{P}\mathbf{XX}^T \tag{4.27}$$

What are principal components in our data?

Figure 4.11 illustrates the principal components of a time series. In the left column the time series of several pixels along the same time-span are drawn. Now PCA tries to find the directions of maximal variance in the data, that are orthogonal. The largest variance is the change of the slow signal across the pixels (shown on the top right) and the second biggest variance is the change of the amplitude of the high frequency component (shown on the bottom right). There are two components temporal PCA can find in this example.

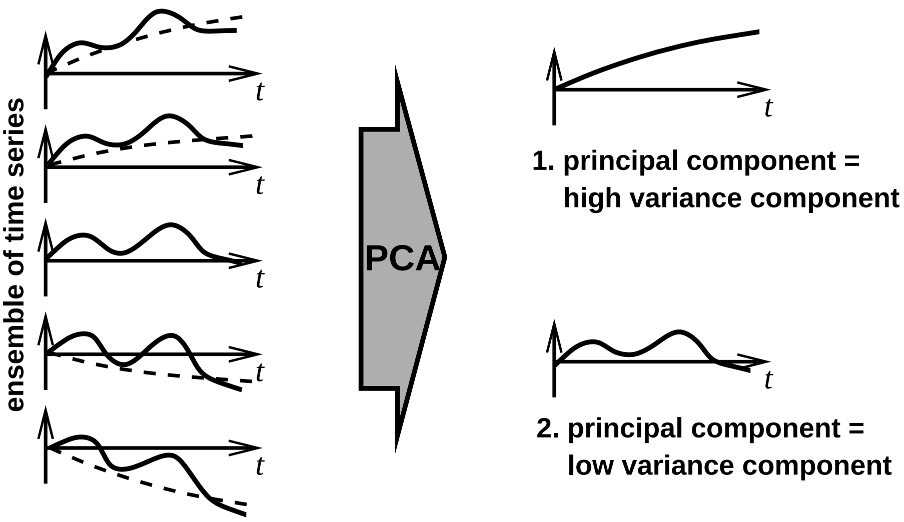


Figure 4.11: Illustration of PCA on a time series that contains two components. The left column shows the superposition of the signals in different pixels along the same time axis. The two components are separated by PCA in the order of the size of their variance.

PCA can be used for blind source separation by itself under certain statistical conditions of the data (Stetter et al., 2000) but we will mainly use it as a preprocessing step for ICA. As shown by Oja (1997) the problem of estimating the demixing matrix can be considerably simplified by whitening or sphering the datasets before the application of BSS algorithms. The goal of this procedure is to reduce the number of free parameters, that have to be estimated by the ICA algorithm. In our example of the Gaussian distribution in section 4.3.2 we have seen two effective ways of doing so already.

First of all we can subtract the mean vector $\boldsymbol{\mu}$ from our measured data and set it to zero mean ($E(\mathbf{x}) = 0$). This implies that $\mathbf{s}$ is zero mean as well, as can be seen by calculating the expectations of both sides of equation 4.7. In the case of a d-dimensional problem we have reduced the number of parameters to be estimated by d (because d parameters specify $\boldsymbol{\mu}$).

In the second step we reduce the number of parameters by aligning the directions of our principal components with the axis of the coordinate system. From equation 4.23 we know that this is true when the covariance matrix $\mathbf{C}_\mathbf{x}$ is diagonal. To achieve this rotation we simply multiply the data $\mathbf{X}$ with the transpose of the vectors $\mathbf{v}_i$, that are held in the matrix $\mathbf{V}$. Now each primary direction is specified by one variable instead of d variables, what reduces the number of variables for the estimation of the principal directions from d^2 to d.

In the third step of parameter reduction we use the ambiguity of ICA, that we can not determine the energies of the independent components (see page 48). As a consequence, we can fix the magnitudes of the independent components, as they

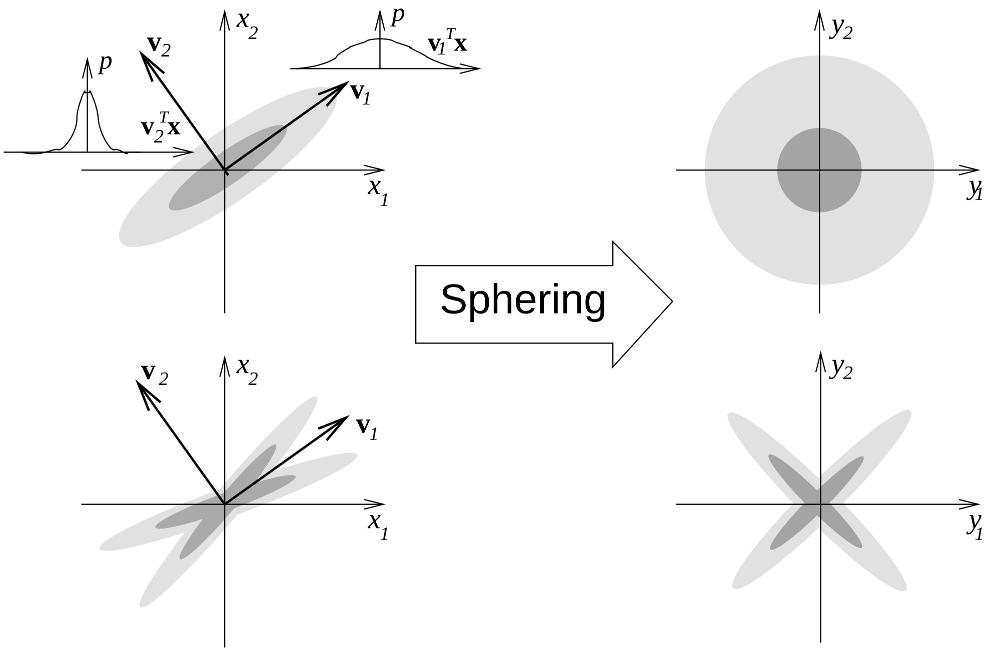


Figure 4.12: Illustration of the sphering procedure for a Gaussian pdf (top) and a non-Gaussian pdf (bottom). The sphering finds the principal axes and normalises the length of each equal to one (adapted from Stetter (2000)).

are arbitrary variables, to unit variance. This is done by dividing each vector $\mathbf{v}_i$ by its length $\lambda_i^{1/2}$ (see figure 4.10).

PCA solves the eigenvalue decomposition (equation 4.25) of the covariance matrix $\mathbf{C_x}$ and delivers the matrices $\mathbf{V}$ and $\mathbf{\Lambda}$ we need for the parameter reduction described above. The original measurement $\mathbf{X}$ is now transformed linearly by

$$\mathbf{Y} := \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^T\mathbf{X} =: \mathbf{DX} \tag{4.28}$$

what makes the components of $\mathbf{Y}$ uncorrelated and their variances equal to unity, in other words the covariance matrix of $\mathbf{Y}$ equals the identity matrix

$$E(\mathbf{YY}^T) = \mathbf{I} \tag{4.29}$$

The transformation of equation 4.28 is called sphering or whitening and $\mathbf{D}$ is the sphering matrix and $\mathbf{Y}$ are the sphered data. Figure 4.12 illustrates the process of whitening for two kinds of data distributions. In the top row we see the transform for a two dimensional Gaussian distribution. The sphering finds the principal directions and normalises the extend of each axis to one. Now the Gaussian distribution is circular symmetric and the probabilities factorise. For a joint Gaussian distribution all the statistical parameters are now determined, as the Gaussian pdf is fully specified by the second order statistics of the data.

The bottom row of figure 4.12 shows an example for a pdf, that is determined by features of higher than second order. The whitening procedure again makes

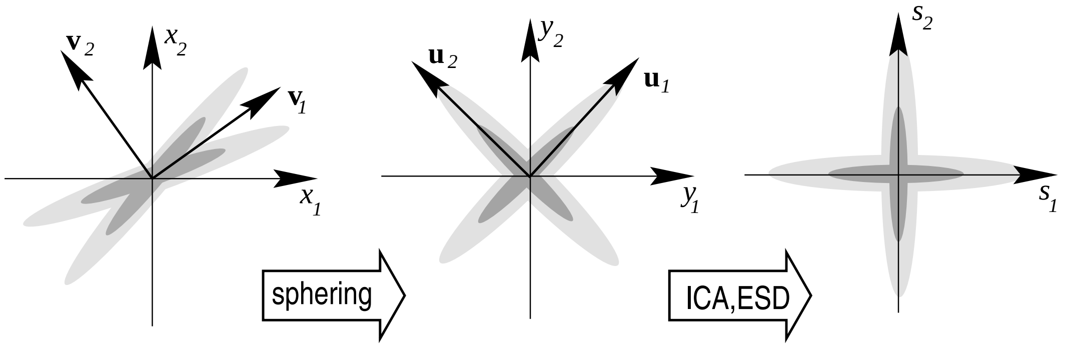


Figure 4.13: The demixing of the measurement is achieved in two steps. First the second order correlations are removed by the sphering procedure. In the second step an orthogonal matrix **U** is found by Independent Component Analysis or Extended Spatial Decorrelation as described in the following chapters (adapted from Stetter (2000)).

the second order moments vanish (zero mean, unit variance), but the directions of the independent components can only be revealed by methods, that consider the structure of higher order moments, i.e. ICA algorithms. For data that contain higher order structure we have reduced the problem of finding the independent components by applying PCA to the task of estimating a new matrix **U**, that is orthogonal and rotates the whitened data, so that the directions of the independent components are aligned with the coordinate axis, as is shown in figure 4.13.

Instead of having to estimate the d^2 parameters, that are the elements of the $d \times d$ demixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ we now only need to estimate the new orthogonal matrix **U** with $d(d-1)/2$ degrees of freedom. For large dimensions d this is only about half of the parameters.

A further advantage of the data preprocessing with PCA is the possibility of dimension reduction if there are less sources than mixtures. This done by looking at the eigenvalues of $\mathbf{C}_x$, that are contained as the variances in $\mathbf{\Lambda}$ and throw away the corresponding **v**'s, that have small λ's and therefore only contain noise.

In conclusion we have the mixing process $\mathbf{X} = \mathbf{AS}$ and need to find the demixing matrix $\mathbf{W} = \mathbf{A}^{-1}$, so that $\hat{\mathbf{S}} = \mathbf{WX}$. After this preprocessing we have the sphered dataset $\mathbf{Y} = \mathbf{DX}$, so that we now only have to estimate the orthogonal matrix **U** and end up with

$$\begin{aligned} \hat{\mathbf{S}} &= \mathbf{UY} = \mathbf{UDX} = \\ &= \mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^{\mathbf{T}}\mathbf{X} \end{aligned} \tag{4.30}$$

4.4 Independent Component Analysis

The term independent component analysis (ICA) describes a group of algorithms, that can estimate the original sources **s** from a recorded mixture by exploiting the

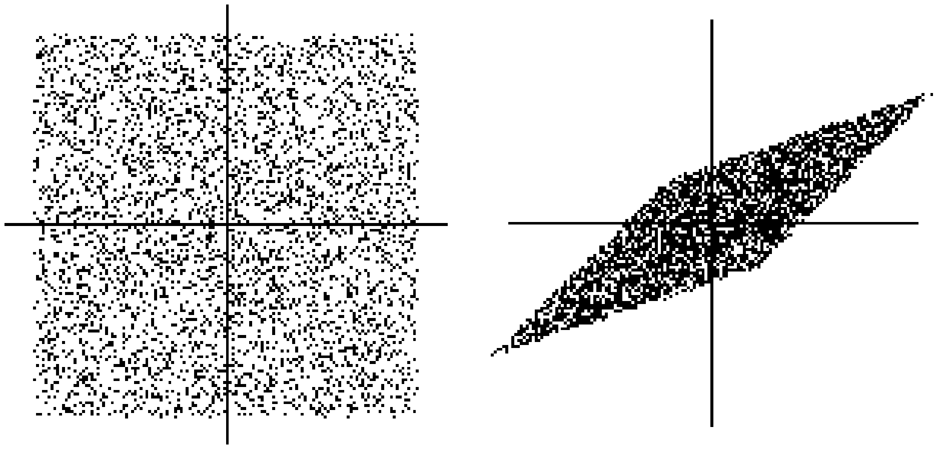


Figure 4.14: Intuitive understanding of statistical independence. **Left**: Plot of the probability density of the independent variables s_1 and s_2 with uniform distribution. **Right**: Joint distribution of the variables after the mixing with $\mathbf{A}$. The variables are not statistically independent anymore, as the maximum or minimum of one variable completely determines the value of the other variable. In the case of independence on the left we can make no prediction about the one variable from the other (adapted from Hyvärinen and Oja (1999)).

statistical properties of the higher order moments of the probability distribution. PCA is limited to second order. Some basic assumptions have to be fulfilled by the sources and the mixtures, so that ICA is able to work.

The most important assumption for ICA is, that there is a representation with independent components of the recorded mixtures or equivalently the original sources must have been statistical independent in the first place. We know the mathematical definition for statistical independence from equation 4.24. If the product of the individual pdf's of the source signals equals the pdf of the overall data distribution of the signals, then the signals are independent. Most of the time we do not know the sources $\mathbf{s}$ or the mixing matrix $\mathbf{A}$, therefore ICA transforms the measurements $\mathbf{x}$ in a way, that the resulting estimates $\hat{\mathbf{s}}$ fulfil equation 4.24 as good as possible. Once the densities satisfy the equation 4.24 we know that the estimates $\hat{\mathbf{s}}$ equal the original sources $\mathbf{s}$ up to a permutation and a scaling factor (remember the ambiguities of ICA on page 48). Let us stress at this point again, that we look at our measurements $\mathbf{x}$ like the measurement of a random variable, therefore the time structure is of no importance, because we only use the pdf's (we can randomly scramble the data points within the individual x_j and still get the same results).

There is also a more intuitive understanding of statistical independence, that is illustrated in figure 4.14. Let us assume two independent variables s_i, that have a uniform distribution between two arbitrary values. The joint density of s_1 and s_2 is then uniform on a square. The left plot of figure 4.14 shows data points randomly drawn from this distribution. Now we mix these two independent components with the mixing matrix $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -1 & 2 \end{pmatrix}$ and we get two mixed variables x_1 and x_2. The joint density of x_1 and x_2 is a uniform distribution on a parallelogram (right plot). We can see now that the random variables x_1 and x_2 are not independent anymore, because if we go to one of the minimum or maximum values of x_1, the value of x_2 is completely determined. We can predict the value of one distribution from the value of the other distribution. In the left plot for each value of s_1 all values of s_2 are possible, so they are independent. This gives us the marginal pdf of s_1 and s_2 by

$$p_1(s_1) = \int p(s_1, s_2) ds_2 \tag{4.31}$$

and similarly for s_2. Then we define that s_1 and s_2 are independent if and only if the joint pdf is factorisable in the following way (Hyvärinen and Oja, 1999):

$$p(s_1, s_2) = p(s_1)p(s_2) \tag{4.32}$$

This is true for any number of random independent variables s_j. From this and the independence assumption we can derive a important property of independent random variables. Given two functions h_1 and h_2 we always have

$$E(h_1(s_1)h_2(s_2)) = E(h_1(s_1))E(h_2(s_2)) \tag{4.33}$$

This can be proven as follows:

$$\begin{aligned} E(h_1(s_1)h_2(s_2)) &= \int\int h_1(s_1)h_2(s_2)p(s_1, s_2)ds_1ds_2 \\ &= \int\int h_1(s_1)p(s_1)h_2(s_2)p(s_2)ds_1ds_2 \\ &= \int h_1(s_1)p(s_1)ds_1 \int h_2(s_2)p(s_2)ds_2 \\ &= E(h_1(s_1))E(h_2(s_2)) \end{aligned} \tag{4.34}$$

One example for h_1 and h_2 is the potency, that will be important when we consider the higher order moments of the distributions.
For $h_1(s_1) = (s_1)^\alpha$ and $h_2(s_2) = (s_2)^\beta$ we have

$$E((s_1)^\alpha(s_2)^\beta) = E((s_1)^\alpha)E((s_2)^\beta) \quad \text{for any} \quad \alpha, \beta \tag{4.35}$$

From this we can see that uncorrelatedness is a weaker argument, than statistical independence. Two random variables s_1 and s_2 are uncorrelated if their covariance is zero

$$E(s_1s_2) - E(s_1)E(s_2) = 0 \tag{4.36}$$

If the variables are independent, they are uncorrelated, what follows directly from equation 4.33 with $h_1(s_1) = s_1$ and $h_2(s_2) = s_2$.

The other way around this is not true, because uncorrelatedness does not imply independence. For example assume that (s_1, s_2) are discrete valued and follow a distribution, that the pairs are drawn with a probability $p = 1/4$ equal to any of the following values: $(0, 1), (0, -1), (1, 0), (-1, 0)$ (Hyvärinen and Oja, 1999). Then s_1 and s_2 are uncorrelated, but

$$E(s_1^2 s_2^2) = 0 \neq 1 = E(s_1^2)E(s_2^2)$$

so the equation 4.33 is violated and the variables are not statistically independent.

The second fundamental restriction to ICA is that the independent components must be non Gaussian for ICA algorithms to work. In the top row of figure 4.12 we see the result of a two dimensional Gaussian distribution after sphering. The density is completely symmetric and has no information on the directions of the columns of the mixing matrix $\mathbf{A}$. Therefore we can only estimate the $\hat{s}_j$ up to a orthogonal transform. If only one independent component is Gaussian then ICA still works.

4.4.1 Estimation of Extremal Kurtosis

The advantage of ICA over PCA is, that it uses the information of the higher order statistics in the data to find the estimates $\hat{\mathbf{s}}$ of the original independent sources $\mathbf{s}$ and is therefore not bound to find only orthogonal components. The higher order statistic is expressed in the higher order moments, or cumulants respectively. The first cumulant is the mean μ of a distribution. The second cumulant represents the variance σ^2. For a symmetric distribution the third order cumulant, that is referred to as skewness, vanishes. The fourth order cumulant is called the kurtosis and is defined for a random variable $\mathbf{x}$ by:

$$kurt(x) = E(x^4) - 3(E(x^2))^2 \tag{4.37}$$

As we know from equation 4.17 the variance equals one as we have zero mean unit variance data. This simplifies the right hand side of equation 4.37 to

$$kurt(x) = E(x^4) - 3 \tag{4.38}$$

The kurtosis is the lowest cumulant that holds information about the statistics of the pdf after decorrelation by whitening. As figure 4.15 shows, the kurtosis is a measure for the gaussianity of a distribution. For a Gaussian distribution the kurtosis is zero, for a supergaussian distribution the kurtosis is positive and negative for subgaussian distributions.

So how can we use this information of the kurtosis to find the independent components from the mixture? The solution was introduced by Hyvärinen and

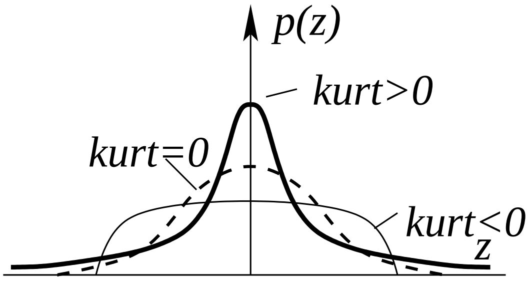


Figure 4.15: Kurtosis *kurt* as a measure of nongaussianity of a pdf. A Gaussian distribution has a kurtosis equal to zero as it has no higher than second order statistic (dashed line), supergaussian densities a positive (thick line) and subgaussian densities a negative kurtosis (thin line). The supergaussian distributions have heavier tails than a Gaussian pdf, which makes the calculation of the kurtosis sensitive to outliers in the data.

Oja (1997). The key is the central limit theorem of probability theory, that shows that the distribution of a sum of independent nongaussian variables tends towards a Gaussian distribution. Thus the sum of two such variables has a distribution that is closer to a Gaussian than any of the two original random variables, i.e. the nongaussianity gets more and more lost. The following example illustrates the demixing procedure for a two dimensional model $\mathbf{x} = \mathbf{A}\mathbf{s}$. Let us assume the independent components s_1, s_2 with zero mean, unit variance and kurtosis values $kurt(s_1), kurt(s_2) \neq 0$. For the kurtosis of independent random variables s_1, s_2 the following holds:

$$kurt(s_1 + s_2) = kurt(s_1) + kurt(s_2) \tag{4.39}$$

and

$$kurt(\alpha s_1) = \alpha^4 kurt(s_1) \tag{4.40}$$

To estimate one of the independent components $\hat{s}_j$ consider a linear combination of the x_j with $\hat{s} = \mathbf{w}^T\mathbf{x} = \sum_j w_j x_j$. The vector $\mathbf{w}^T$ has to be determined and if $\mathbf{w}^T$ equals a row of $\mathbf{W} = \mathbf{A}^{-1}$ then $\hat{s}_j$ is one of the sources. We make the transformation

$$\mathbf{z} = \mathbf{A}^T\mathbf{w} \tag{4.41}$$

and get

$$\hat{\mathbf{s}} = \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{A}\mathbf{s} = \mathbf{z}^T\mathbf{s} = z_1 s_1 + z_2 s_2 \tag{4.42}$$

From the equations 4.39 and 4.40 we have $kurt(\hat{\mathbf{s}}) = z_1^4 kurt(s_1) + z_2^4 kurt(s_2)$. On the other hand we made the constraint that the variance of $\hat{\mathbf{s}}$ equals 1. This implies a constraint on $\mathbf{z}$: $E(\hat{\mathbf{s}}^2) = z_1^2 + z_2^2 = 1$. Geometrically this means that the vector $\mathbf{z}$ is constrained to the unit circle. The optimisation problem is now : what are the maxima of the function $|kurt(\hat{\mathbf{s}})| = |z_1^4 kurt(s_1) + z_2^4 kurt(s_2)|$ on the unit circle?

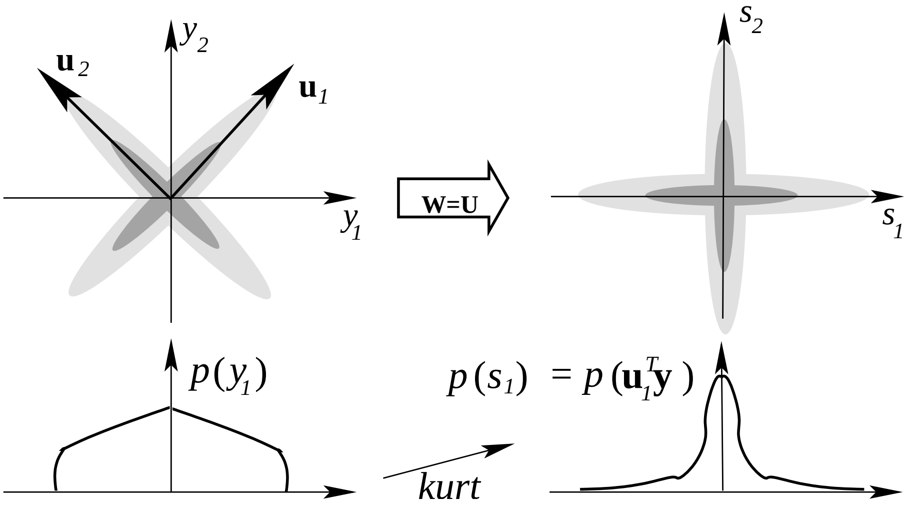


Figure 4.16: Independent Component Analysis by estimation of extremal kurtosis. After sphering the FastICA algorithm finds the direction in the joint densities, along which the absolute kurtosis of the projected data becomes maximal. The example shown consists of two supergaussian sources. On the left hand side we see the result after sphering. The bottom plot shows the projection of the mixture on the axis. After the separation (right hand side) the projection of the densities is more supergaussian than in the mixture (adapted from Stetter (2000)).

In the article by Delfosse and Loubaton (1995) it is shown that the maxima are at the points where exactly one of the z_j is zero. Because of the unit circle the other z_j is then equal to 1 or -1 and we get from equation 4.42 that $\hat{s}_j$ is one of the original sources s_j and therefore the problem is solved.

For many dimensions the same principle is used by finding one row $\mathbf{w}^T$ of $\mathbf{W}$ and then repeating the procedure on the orthogonal subspace. The fast fixed point algorithm for ICA, that was introduced by Hyvärinen and Oja (1997), works on this principle and yields the rows of the unmixing matrix $\mathbf{W} = \hat{\mathbf{U}}$ one at a time. The first row $\mathbf{w}_1^T$ of $\mathbf{W}$ is obtained by the fixed point iteration from the sphered data $\mathbf{y}$ (remember $\mathbf{y} = \mathbf{Dx}$)

$$\mathbf{w}_1(iter+1) = E(\mathbf{y}(\mathbf{w}_1^T(iter)\mathbf{y})^3) - 3\mathbf{w}_1(iter) \tag{4.43}$$

while for the remaining rows after each iteration step the resulting vector $\mathbf{w}_j$ is projected into the subspace orthogonal to $w_1, ..., w_{j-1}$.

In Figure 4.16 the concept of estimation of extremal kurtosis is illustrated. After sphering of the data the algorithm finds an orthogonal transform $\mathbf{U}$ so that the projection of the pdf of the data onto the new axis has its maximum absolute value of the kurtosis.

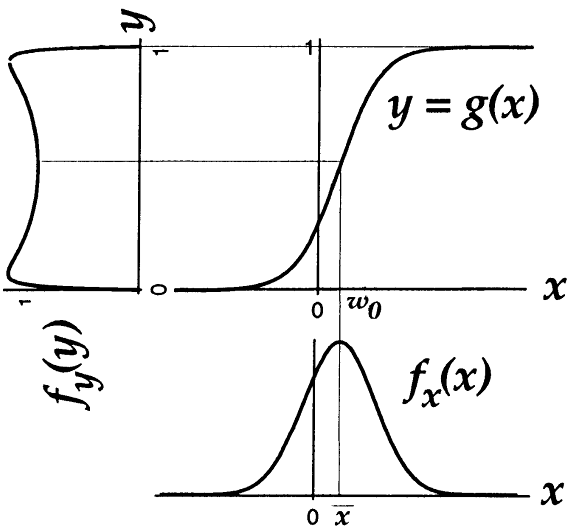


Figure 4.17: The optimal information transfer of the input $\mathbf{x}$ with the pdf $f_{\mathbf{x}}(\mathbf{x})$ through a sigmoidal neuron with the non-linear transfer function $g(\mathbf{x})$ is achieved when the resulting density $f_{\mathbf{y}}(\mathbf{y})$ is as uniform as possible (maximum entropy!). This is achieved when the mean and the variance of $\mathbf{x}$ is matched with the threshold ω_0 and the slope ω of $g(\mathbf{x})$. This property will be used in the learning step of the Infomax algorithm (from Bell and Sejnowski (1995)).

4.4.2 The Infomax Principle

A second class of ICA algorithms makes use of all higher moments of any order to explore the probability density of the mixture and find the independent components. To do this we use a generalisation of the Linsker (1989) Infomax principle that was introduced by Bell and Sejnowski (1995). The Infomax principle was described by Laughlin (1981) in the following way: when inputs are to be passed through a sigmoid function, maximum information transmission can be achieved when the sloping part of the sigmoid is optimally lined up with the high density parts of the inputs (see figure 4.17). To understand what that means for our ICA problem we will introduce now a measure for the information content of a stochastic variable and then we will see that maximising the information transfer is equal to reducing the redundancy between sources, which is equal to making the sources independent.

The measure we are looking for was originally developed in the context of thermodynamics and introduced to information theory by Shannon (1948). It is called Entropy H and can be interpreted as the degree of disorder of a system or the information content a variable carries. For probability distributions p which are functions of variables $\mathbf{x}$ we define the entropy to be (Bishop, 1995):

$$H(\mathbf{x}) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \tag{4.44}$$

The entropy H has its maximum value for a bounded variable $\mathbf{x}$ if it has a flat probability distribution, i.e. the probability density $p_i(\mathbf{x})$ to draw a value x is equal for all $x \in (a, b)$, where a and b are the boundaries. In this state the system has its maximum disorder. On the other hand H has the smallest value close to zero, when all the probability mass is concentrated in one value, i.e. the $p(\mathbf{x}) = \delta(x-\mu)$. This very sharply peaked distribution has the highest possible order.

In the information theoretic interpretation of the entropy, H tells us the amount of information, or equivalently the 'degree of surprise', which is obtained when we learn that a particular value was drawn. When one probability p_i equals 1 (so the others must be 0) then the event is certain to occur, and there is no surprise when the event is found to occur (and no information is received). Again when all p_i are equally distributed, we have the maximum surprise (or information transfer) when a value x_i is drawn.

From the definition of independence in equation 4.24 and the entropy H in equation 4.44 we see that for independent $\mathbf{s}$ the following is true

$$0 = (\sum_j H(s_j)) - H(\mathbf{s}) \tag{4.45}$$

So any non zero value on the left side of equation 4.45 would derive from a mixing of the independent variables $\mathbf{s}$ and is a measure for redundancy of information in the components of the mixture $\mathbf{x} = \mathbf{As}$. This gives the definition of the mutual information (MI)

$$MI(\mathbf{x}) = (\sum_j H(x_j)) - H(\mathbf{x}) \tag{4.46}$$

Note that this expression is identical to the Kullbach-Leibler distance (KLD) of the joint density and the factorisation

$$\begin{aligned} KLD(p(\mathbf{x}), \prod_j p(x_j)) &= \int p(\mathbf{x}) \ln \left(\frac{p(\mathbf{x})}{\prod_j p(x_j)} \right) \\ &= (\sum_j H(x_j)) - H(\mathbf{x}) \end{aligned} \tag{4.47}$$

The KLD is a distance measure between two distributions. This shows us again that the mutual information will be minimal, if the variables are statistically independent. So we solve the ICA problem if we make the mutual information vanish.

One solution to this was introduced by Bell and Sejnowski (1995) from a neural network point of view. The authors proved that maximising the information transfer reduces the redundancy between the units in the output layer. Starting from the Infomax principle (see page 63) they show that for continuous deterministic mappings the mutual information between the inputs and outputs can be maximised by maximising the entropy of the outputs alone. Do not confuse this with the mutual information between the outputs themselves. Here the mutual information between the inputs and the outputs of a network is a measure for the information transfer through a network. Now Shannon (1948) has shown that the uniform distribution of a random variable with finite support has the largest entropy H of all possible distributions. So once the distribution of the elements of the outputs have a uniform distribution they are independent.

The last thing we have to prove now is that the inputs of the non-linear invertible transform $\mathbf{\Phi}$ are independent as soon as the outputs are independent. The source estimates $\hat{\mathbf{s}}$ are transformed component wise by a non linear invertible function $\mathbf{\Phi}$

$$z_l = \phi_l(\hat{s}_l), \quad \mathbf{z} = \mathbf{\Phi}(\mathbf{W}\mathbf{x}), \quad (\mathbf{\Phi})_l = \phi_l \quad \text{(here} \quad \phi_l = \tanh) \tag{4.48}$$

in a way that the data vector is restricted to a M-dimensional hypercube $-1 \leq z_l \leq 1, \quad l = 1, ..., M$. Because the transform $\mathbf{\Phi}$ acts component wise it does not change any dependencies between the components. Therefore if the z_l are independent, the $\hat{s}_l$ are independent too. This can be shown with the transformation rules for probability densities (Papoulis, 1965)

$$p(\mathbf{s}) \quad = \quad q(\mathbf{z}) det \left(\frac{d\mathbf{z}}{d\mathbf{s}} \right) = q(z) det \mathbf{J}. \tag{4.49}$$

with

$$\mathbf{J}_{kl} \quad = \quad \frac{\partial z_k}{\partial s_l} = \delta_{kl} \varphi'_l(s_l) \tag{4.50}$$

where the prime denotes the first derivative of a function and δ_{kl} denotes the Kronecker delta, which is 1 if $k = l$ and 0 otherwise. Once the pdf of the transformed vector $\mathbf{z}$ factorises, we obtain the following term for the source density:

$$\begin{aligned} p(\mathbf{s}) &= \left(\prod_l q_l(z_l) \right) \left(\prod_l \varphi'_l(s_l) \right) \\ &= \prod_l q_l(\varphi_l(s_l)) \varphi'_l(s_l) \\ &\equiv \prod_l p_l(s_l). \end{aligned} \tag{4.51}$$

The equations (4.51) say that as soon as $q(\mathbf{z})$ factorises, the source density $p(\mathbf{s})$ factorises as well and we have found the independent components.

The confusing bit of the ICA with the Infomax principle is the two step algorithm that is iteratively used. The first step is the calculation of estimates $\hat{\mathbf{s}}$ with

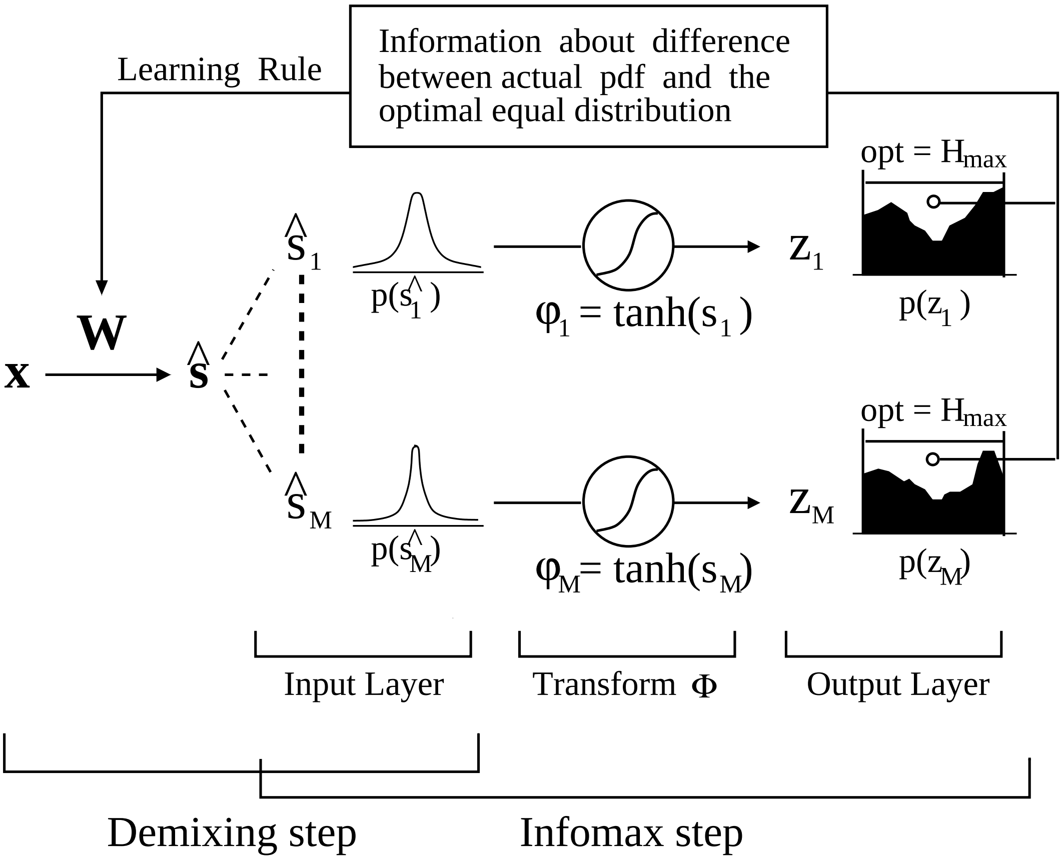


Figure 4.18: Sketch of the two step iteration algorithm of ICA with the information-maximisation approach. *Logical concept*:

- *If the mutual information between source estimates $\hat{s}_i$ is 0 they are independent.*
- *Maximise information transfer between input and output layer ⇔ reduce redundancy between units in the output layer.*
- *Mutual information, i.e. information transfer, between the input layer and the output layer can be maximised in a network with invertible, continuous, deterministic mapping, by maximising the entropy between the outputs alone ⇒ maximum entropy of outputs ≡ independent outputs.*
- *The maximum entropy for a bounded random variable is the uniform distribution.*
- *Because the transform Φ acts component wise the inputs $\hat{s}_i$ are statistically independent when the outputs z_i are independent ⇒ ICA problem is solved.*

a demixing matrix $\mathbf{W}$ from our measurement $\mathbf{x}$, so $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$. The first $\mathbf{W}$ is chosen randomly. This step is the standard demixing part and contains no Infomax concept so far.

The second step is the Infomax step, where we take the $\hat{\mathbf{s}}$ and get a measure how independent they are with the help of a nonlinear transfer function $\mathbf{\Phi}$. If the $\mathbf{z} = \mathbf{\Phi}\hat{\mathbf{s}}$ have a uniform distribution they are independent and therefore our estimates $\hat{\mathbf{s}}$ are also independent. We found the $\mathbf{W} = \mathbf{A}^{-1}$. If the $\mathbf{z}$ are not uniformly distributed we use a learning rule to change $\mathbf{W}$ and calculate new estimates $\hat{\mathbf{s}}$ and start a new Infomax step. With the right learning rule the distribution of the $\mathbf{z}$ will converge to the uniform distribution in a hypercube.

To end with a perfect uniform distribution the marginal source densities s_l would have to match the derivatives of the nonlinearity. As the source densities are unknown, this is normally not given, but a approximately correct nonlinearity still gives good separation results and makes the z_l converge to a flat distribution.

The learning rule for the iterative change of $\mathbf{W}$ with the Infomax principle is given in Bell and Sejnowski (1995) and a extensive proof of it is derived in the appendix of this paper. This learning rule performs a gradient ascent in the information that the outputs transmit about the inputs by noting that the information gradient is the same as the entropy gradient for invertible deterministic mappings. This involves the calculation of the inverse of $\mathbf{W}$, which is computationally expensive and should be avoided. Instead of taking the actual gradient one can take its product with a positive definite matrix $\mathbf{W}^T\mathbf{W}$. The resulting so called natural gradient, first introduced by Amari (1996), has a positive inner product with the original gradient and points therefore into the same overall direction (Parra, 1998).

This gives us the following learning rule for the matrix elements of $\mathbf{W}$

$$\hat{\mathbf{s}} = \mathbf{W}(k)\mathbf{y} \tag{4.52}$$

$$\mathbf{W}(k+1) = \mathbf{W}(k) - \mathbf{\Phi}(\hat{\mathbf{s}})\hat{\mathbf{s}}^T\mathbf{W}(k) \tag{4.53}$$

In the beginning we stated that this method uses the statistics of all higher moments. So where in the procedure did this happen? The higher order statistics are accessed through the use of the static non-linear function. The Taylor series expansion of the non linearity yields the higher order terms. By passing the information through this non-linearity the algorithm enables the network to find higher order forms of redundancy inherent in the inputs.

4.5 Extended Spatial Decorrelation

The two ICA algorithms we have considered so far make the basic assumption that the measurement $\mathbf{x}$ is a linear mixture of statistically independent sources, i.e. their probability densities factorise. To generate this separation the higher order moments in their pdf's are used. These algorithms do not use the time or space structure of the measurement. Therefore any random scrambling (i.e.

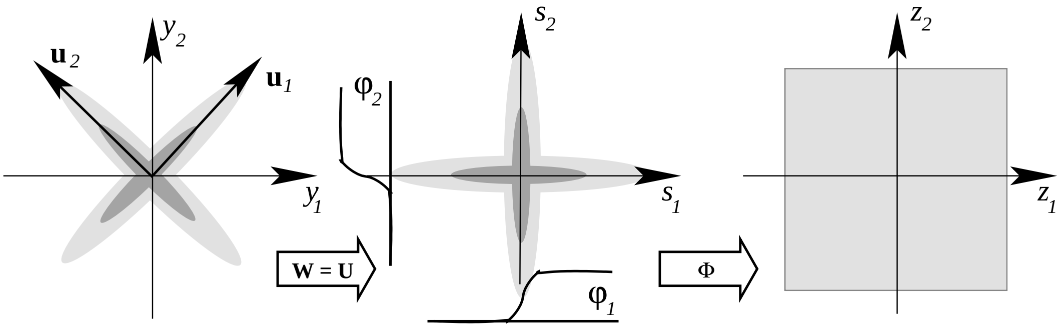


Figure 4.19: Independent Component Analysis with information maximisation. After sphering (left) the orthogonal transform **U** is applied (middle). To control if the sources are independent after the transformation with **U** another transform **Φ** is applied. If the distribution of **z** is uniform (right), independence is achieved. If not, **U** is modified and the next iteration step starts. The three graphs show the final perfectly separated result (adapted from Stetter (2000)).

permutation) of the data in a measurement x_i would still deliver the same separation result. One main drawback of the separation with higher order moments in practice is that we have to make a estimation from a measured sample of finite length. Cumulants like the kurtosis are very sensitive to outliers and therefore need a high number of data in the measurements to make a reliable estimation. This is also one of the reasons for their high sensitivity to noise.

We now introduce a class of algorithms, that only relies on the second order statistics of the data to find the unmixing matrix **W** and by doing this has a more robust estimation from a realistic sized data set. The following algorithms explicitly make use of the structure in the original sources **s** and do not rely on the statistical independence assumption. This makes them less sensitive to small dependencies in the sources, that can be present in real world data. So strictly seen they are not ICA algorithms (they do not make use of the equation 4.24) but BSS algorithms as the sources **s** and the matrix **A** are not known.

A second order BSS algorithm for the separation of a mixture of one dimensional time series was introduced by Molgedey and Schuster (1994). We extended the algorithm for the analysis of two dimensional spatial structures (Otto et al., 1998; Schießl et al., 1998; Schießl et al., 1999; Schießl et al., 2000c; Schießl et al., 2000b). This algorithm is called extended spatial decorrelation (ESD) and makes use of the simultaneous decorrelation of the sources and shifted versions of them. We then extended the algorithm for use with multiple shifts (Schöner et al., 1999; Schöner et al., 1999b; Schöner et al., 2000) and added a regularization term (Schießl et al., 2000a).

In the following sections we introduce cross-correlation function and the concept of how it can be used for blind source separation with ESD. Then the different versions of ESD, i.e. single shift ESD, multi shift ESD and regularized ESD are derived.

Let us derive now how we justify the basic assumptions we make for ESD later on. We know from the definition of covariance in equation 4.18 and equation 4.36 that two source estimates $\hat{s}_l$ and $\hat{s}_m$ are uncorrelated if their covariance is zero. As we only consider the expectation values E$(\cdot)$ equation 4.36 is also true for shifts $\Delta\mathbf{r}$ within the zero mean sources and can be reformulated as

$$E(\hat{s}_l(\mathbf{r})\hat{s}_m(\mathbf{r}+\Delta\mathbf{r})) = E(\hat{s}_l(\mathbf{r}))E(\hat{s}_m(\mathbf{r}+\Delta\mathbf{r})) = 0 \quad \forall \quad l \neq m, \Delta\mathbf{r} \tag{4.54}$$

With the mixing matrix $\mathbf{A}$ and the demixing matrix $\mathbf{W}$ we have $\hat{\mathbf{s}} = \mathbf{WAs}$ which leads us to

$$\begin{aligned} E(\hat{s}_l(\mathbf{r})\hat{s}_m(\mathbf{r}+\Delta\mathbf{r})) &= E\left(\sum_i (\mathbf{WA})_{li} s_i(\mathbf{r}) \sum_j (\mathbf{WA})_{mj} s_j(\mathbf{r}+\Delta\mathbf{r})\right) \\ &= \sum_i \sum_j (\mathbf{WA})_{li} (\mathbf{WA})_{mj} E\left(s_i(\mathbf{r}) s_j(\mathbf{r}+\Delta\mathbf{r})\right) \\ &= \sum_i (\mathbf{WA})_{li} (\mathbf{WA})_{mj} E\left(s_i(\mathbf{r}) s_i(\mathbf{r}+\Delta\mathbf{r})\right) \\ &= \sum_i (\mathbf{WA})_{li} (\mathbf{WA})_{mi} (\mathbf{C}_\mathbf{s}(\mathbf{r}))_{ii} \end{aligned} \tag{4.55}$$

This condition for a non singular $(\mathbf{WA})$ is only fulfilled if the matrix $(\mathbf{WA})$ has only one element in each column and row that is not equal to zero, which is the definition of the permutation matrix $\mathbf{P}$, that we know already from the second ambiguity of ICA on page 49. From this we see that we solved the BSS problem with considering only second order statistics, if equation 4.55 is true for any $\Delta\mathbf{r}$. One exception is the extreme case where all s_i have the same auto-correlation function, because then equation 4.55 would be fulfilled by any orthogonal matrix $(\mathbf{WA})$. Therefore one of the assumptions for ESD, that we formulate later in the text will be , that this exception is excluded, i.e. auto-correlations are different for all s_i.

This demonstrates how ESD makes explicit use of the spatial structure in the patterns s_i. Equation 4.54 demands, that for all $\Delta\mathbf{r}$ the correlations vanish. Later we will see that we can only consider a limited number of shifts, for computational reasons. But let us first introduce the cross-correlation function and have a look at some of its properties.

The definition of the cross-correlation function between two source patterns $s_l(\mathbf{r})$ and $s_m(\mathbf{r})$ is

$$C_\mathbf{s}^{lm}(\Delta\mathbf{r}) = E\left(s_l(\mathbf{r}) s_m(\mathbf{r}+\Delta\mathbf{r})\right) = \frac{1}{Q}\sum_\mathbf{r} s_l(\mathbf{r}) s_m(\mathbf{r}+\Delta\mathbf{r}) \tag{4.56}$$

The shift $\Delta\mathbf{r}$ is often called the lag of the correlation function and the right hand side of equation 4.56 is the deterministic cross-correlation sequence, that we have to use as we only know the measured values (i.e. pixels) but not the continuous function. Each lag $\Delta\mathbf{r}$ delivers one value $C_{\mathbf{s}}^{lm}(\Delta\mathbf{r})$ and as the number of overlapping pixels between s_l and s_m decreases with the size of $\Delta\mathbf{r}$ we normalise with the value Q, that is the number of pixels the two sources still have in common. If $l = m$ we get the auto-correlation $C_{\mathbf{s}}^{ll}(\Delta\mathbf{r})$ of a source s_l. To avoid confusion between the terms covariance and auto-correlation recall that we assume zero mean and unit variance data and therefore both terms are mathematically the same.

To get a feeling for the meaning of a cross-(auto-)correlation function let us have a look at two zero mean one dimensional time series. The first time series is a simple sine function $s_1(t) = \sin(\omega\pi t)$ and the second time series is white noise $s_2(t) = randn(t)$. In figure 4.20 on the left hand side we see the plot of a sine wave (**a)**) and its auto-correlation sequence (**c)**) without correction by the value Q. For the zero lag we always get a maximum as the variables in the sum of equation 4.56 are the same and we therefore never have negative values in the sum. With increasing lag the correlation coefficient becomes smaller until the shift π (0.5 on the x axis) where we always add up pairs of $s_1(t)s_1(t+\pi)$ with negative value. Now the correlation coefficient rises again up to the shift 2π. As we did not divide by Q this maximum is smaller as there is less overlap and therefore less elements in the sum.

The right hand side plot of figure 4.20 demonstrates why a slow decay of the auto-correlation function is a measure for the smoothness of the source s_l. In the case of the white noise (**b)**)we get again the maximum for the zero lag. But in contrast to the sinusoidal signal a small shift has a drastic effect on the size of the $C_{s_2}^{ll}$, because the values are random and therefore the sum of equation 4.56 cancels out over the average (**d)**). So a smooth decaying $C_{\mathbf{s}}^{ll}$ tells us that the neighbouring values $s_l(t)$ and $s_l(t+1)$ are similar.

Extended spatial decorrelation makes the following assumptions:

i): The original sources s are smooth in space, which is reflected in a smooth auto-correlation function.

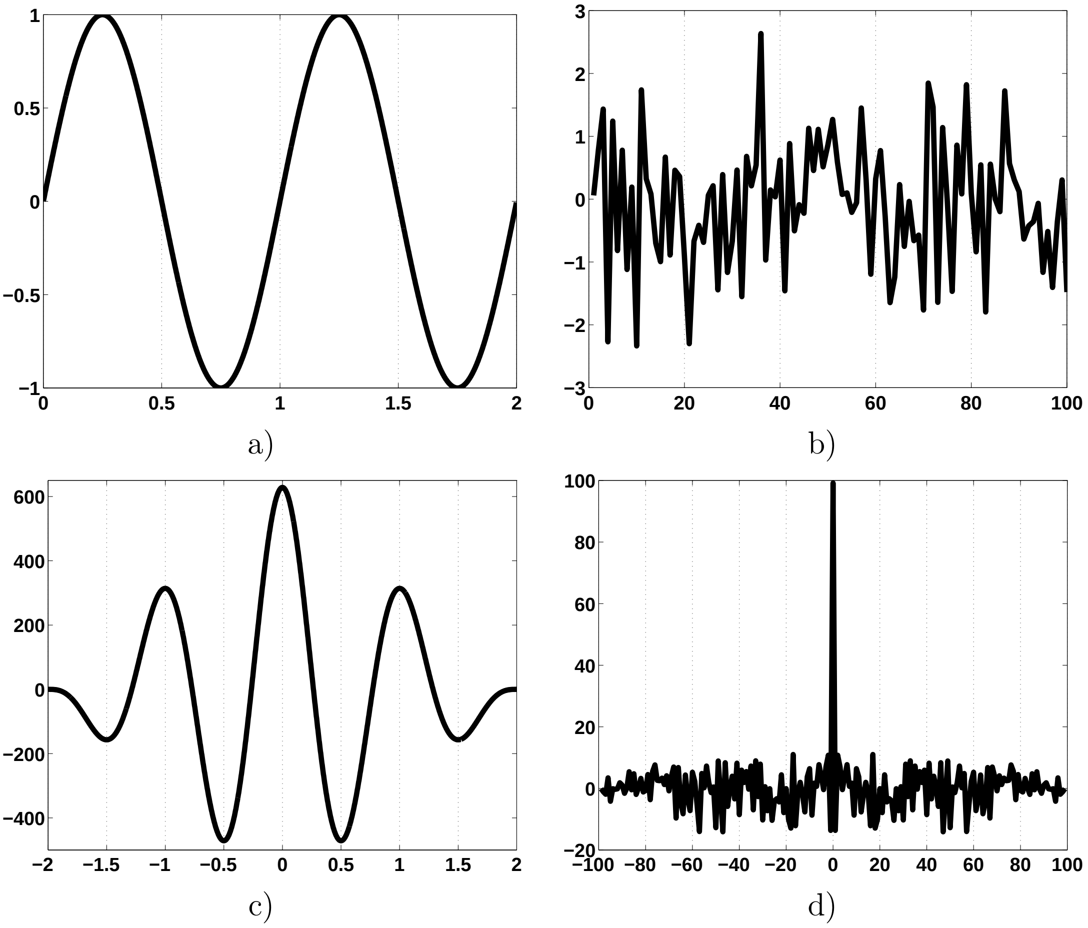
ii): Different sources s_l, s_m are uncorrelated among themselves. This is expressed in a small (ideally vanishing) cross-correlation value $C_{\mathbf{s}}^{ml}(0)$.

iii): Sources s_l are uncorrelated with shifted versions of s_m, i.e $C_{\mathbf{s}}^{lm}(\Delta\mathbf{r})$ vanishes for all $\Delta\mathbf{r}$.

iv): The auto-correlation functions must be different between all sources.

For M given sources s we can write all auto- and cross-correlations in form of a matrix $\mathbf{C}_{\mathbf{s}}(\Delta\mathbf{r})$ with $(\mathbf{C}_{\mathbf{s}}(\Delta\mathbf{r}))_{lm} := C_{\mathbf{s}}^{lm}$. The diagonal elements of $\mathbf{C}_{\mathbf{s}}(\Delta\mathbf{r})$ are the auto-correlations and the off diagonal elements are the cross-correlations. Therefore the assumptions i), ii) and iii) just request that $\mathbf{C}_{\mathbf{s}}(\Delta\mathbf{r})$ is a diagonal matrix for all $\Delta\mathbf{r}$, which gives the following rule for the matrix elements $C_s^{lm}(\Delta\mathbf{r})$

$$C_s^{lm}(\Delta\mathbf{r}) = \delta_{l,m} C_s^{lm}(\Delta\mathbf{r}) = C_s^{ll}(\Delta\mathbf{r}) = \mathbf{\Lambda}(\Delta\mathbf{r}) \quad \forall \quad \Delta\mathbf{r} \tag{4.57}$$



a)

b)

c)

d)

Figure 4.20: Examples for auto-correlation functions of a sine wave and random white noise. In **a)** we see a plot of a sinus over a period of 4π. The label of the x axis is given in multiples of 2π, i.e. a full period. In **c)** the plot of the auto-correlation function of this sine wave is shown. Now the label on the x axis denotes the shift of the function, i.e. the lag, in the calculation of the correlation coefficient in multiples of 2π. The right column shows an example of random white noise (**b)**) and its auto-correlation below (**d)**). Here the x axis simply denotes the number of the measurement and the lag respectively. The auto-correlation coefficient shows a smooth transition for the smooth sine wave, whereas the correlation value nearly disappears for a shift bigger than zero for the white noise. Therefore we say that the sinus has a good auto-correlation and the white noise has no auto-correlation.

where $\mathbf{\Lambda}(\Delta\mathbf{r})$ denotes the diagonal matrix of the auto-correlation coefficients.

So let us summarise what we have so far: From the equation 4.55 we derived our assumptions i)-iv) for ESD and got the mathematical formulation in equation 4.57. This is equivalent to a simultaneous decorrelation of all s for all $\Delta\mathbf{r}$.

As we do not know the original source $\mathbf{s}$ we have to answer the question now, how we can minimise the correlations between the estimates $\hat{s}_l$ for all $\Delta\mathbf{r}$. Equation 4.54 is also true for the $\hat{\mathbf{s}}$ and in the case of uncorrelated $\hat{\mathbf{s}}$ the covariance matrix $\mathbf{C}_{\hat{\mathbf{s}}}(\Delta\mathbf{r})$ is diagonal:

$$(\mathbf{C}_{\hat{s}}(\Delta\mathbf{r}))_{lm} = E(\hat{s}_l(\mathbf{r})\hat{s}_m(\mathbf{r}+\Delta\mathbf{r})) = 0 \quad \forall \quad l \neq m, \Delta\mathbf{r} \tag{4.58}$$

With the relation $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$ we get

$$\mathbf{C}_{\hat{s}}(\Delta\mathbf{r}) = \mathbf{W}\mathbf{C}_{\mathbf{x}}(\Delta\mathbf{r})\mathbf{W}^T \tag{4.59}$$

and analogously from $\mathbf{x} = \mathbf{A}\mathbf{s}$

$$\mathbf{C}_x(\Delta\mathbf{r}) = \mathbf{A}\mathbf{C}_{\mathbf{s}}(\Delta\mathbf{r})\mathbf{A}^T \tag{4.60}$$

Due to the mixing process the $\mathbf{C}_x(\Delta\mathbf{r})$ are not diagonal anymore, but still symmetric.

The task of the ESD algorithm is to obtain the separated estimates $\hat{\mathbf{s}}$ by calculating the covariance matrices $\mathbf{C}_x(\Delta\mathbf{r})$ for a (good) set of $\Delta\mathbf{r}$ and optimise the matrix $\mathbf{W}$ in a way that the resulting $\mathbf{C}_{\hat{s}}(\Delta\mathbf{r})$ are diagonal simultaneously.

4.5.1 Single shift ESD

The simplest application of the ESD concept is the use of the zero shift decorrelation matrix $\mathbf{C}_{\mathbf{x}}(0)$ and a second decorrelation matrix $\mathbf{C}_{\mathbf{x}}(\Delta\mathbf{r})$. This delivers $M(M+1)$ equations

$$\begin{aligned} \mathbf{C}_{\mathbf{x}}(0) &= \sum_l \mathbf{A}_{il}\mathbf{A}_{jl}\mathbf{C}_{\mathbf{s}}^{ll}(0) \\ \mathbf{C}_{\mathbf{x}}(\Delta\mathbf{r}) &= \sum_l \mathbf{A}_{il}\mathbf{A}_{jl}\mathbf{C}_{\mathbf{s}}^{ll}(\Delta\mathbf{r}) \end{aligned}$$

for the $M(M+1)$ unknowns $\mathbf{A}_{i\neq j}, \quad \mathbf{C}_{\mathbf{s}}^{ll}(0), \quad \mathbf{C}_{\mathbf{s}}^{ll}(\Delta\mathbf{r})$. The mixing matrix $\mathbf{A}$ only has $M(M-1)$ unknown variables instead of M^2 as one might expect, due to the fact that the scaling of the columns has no effect on the mixing and therefore a fixed diagonal can be assumed without loss of generality. In the matrix notation and $\mathbf{W} = \mathbf{A}^{-1}$ we get

$$\mathbf{W}\mathbf{C}_{\mathbf{x}}(0)\mathbf{W}^T = \mathbf{C}_{\mathbf{s}}(0) = \mathbf{\Lambda}_{\mathbf{s}}(0) \tag{4.61}$$

$$\mathbf{W}\mathbf{C}_{\mathbf{x}}(\Delta\mathbf{r})\mathbf{W}^T = \mathbf{C}_{\mathbf{s}}(\Delta\mathbf{r}) = \mathbf{\Lambda}_{\mathbf{s}}(\Delta\mathbf{r}) \tag{4.62}$$

The $\mathbf{C}_{\mathbf{s}}$ equal the $\mathbf{\Lambda}_{\mathbf{s}}$, i.e they are diagonal matrices with the auto-correlations as values, because the sources where uncorrelated. For now we assume perfect separation by the ESD method so the $\hat{\mathbf{s}}$ and the $\mathbf{s}$ are the same. We see that $\mathbf{W}$ diagonalises $\mathbf{C}_{\mathbf{x}}(0)$ and $\mathbf{C}_{\mathbf{x}}(\Delta\mathbf{r})$ simultaneously and if $\Delta\mathbf{r}$ is chosen so that

$${\mathbf{C}_{\mathbf{s}}}^{ll}(0){\mathbf{C}_{\mathbf{s}}}^{mm}(\Delta\mathbf{r}) \neq {\mathbf{C}_{\mathbf{s}}}^{ll}(\Delta\mathbf{r}){\mathbf{C}_{\mathbf{s}}}^{mm}(0) \quad \forall \quad l \neq m \tag{4.63}$$

then the problem is solvable up to $M!$ permutations (Molgedey and Schuster, 1994).

We now can solve our BSS problem by solving equations 4.61 and 4.62 in two steps. Once we have the preliminary matrix $\mathbf{D}$, that solves the first equation we only apply orthogonal transforms $\mathbf{U}$ to solve the second equation and therefore do not destroy the condition for equation 4.61. This way we find the demixing matrix $\mathbf{W} = \mathbf{A}^{-1} = \mathbf{UD}$. The first step diagonalises the zero shift correlation matrix $\mathbf{C_s}(0)$ and as we assume the sources $\mathbf{s}$ to be zero mean and unit variance, $\mathbf{C_s}(0)$ equals the identity matrix $\mathbf{I}$. Therefore equation 4.61 is nothing else than the sphering we know from section 4.3.3 and we get

$$\mathbf{DC_xD}^T = \mathbf{I} \tag{4.64}$$

and as $\mathbf{C_x}(0)$ is symmetric we can derive the eigenvalue equation

$$\mathbf{C_x}(0)\mathbf{V}_D = \mathbf{V}_D\mathbf{\Lambda}_D \tag{4.65}$$

with

$$\mathbf{D} = \mathbf{\Lambda}_D^{-\frac{1}{2}}\mathbf{V}_D^T \tag{4.66}$$

After the sphering the data $\mathbf{y} = \mathbf{Dx}$ have zero mean and unit variance and no cross correlations, i.e. $\mathbf{C_y}(0) = \mathbf{I}$.

Now we look for the orthogonal matrix $\mathbf{U}$ that diagonalises $\mathbf{C_y}(\Delta\mathbf{r})$. We see from

$$\mathbf{C_y}(\Delta\mathbf{r}) = \mathbf{DC_xD}^T = \mathbf{DAC_s}(\Delta\mathbf{r})\mathbf{A}^T\mathbf{D}^T \tag{4.67}$$

that $\mathbf{C_y}(\Delta\mathbf{r})$ is symmetric, because $\mathbf{C_s}(\Delta\mathbf{r})$ is diagonal (assumption iii) for ESD). So the eigenvalue equation

$$\mathbf{C_y}(\Delta\mathbf{r})\mathbf{V}_U = \mathbf{V}_U\mathbf{\Lambda}_U \tag{4.68}$$

delivers

$$\mathbf{U} = \mathbf{V}_U^T \tag{4.69}$$

Because orthogonal transforms preserve the vector lengths and angles between the vectors, applying $\mathbf{U}$ to $\mathbf{C_y}(0)$ does not destroy the results of the first step and we get the overall demixing matrix $\mathbf{W}$ from $\mathbf{W} = \mathbf{UD}$, that diagonalises simultaneously

$$\begin{aligned}
\mathbf{WC_x}(0)\mathbf{W}^T &= \mathbf{UDC_x}(0)\mathbf{D}^T\mathbf{U}^T = \mathbf{I} && (4.70)\\
\mathbf{WC_x}(\Delta\mathbf{r})\mathbf{W}^T &= \mathbf{UDC_x}(\Delta\mathbf{r})\mathbf{D}^T\mathbf{U}^T = \mathbf{\Lambda}(\Delta\mathbf{r}) && (4.71)
\end{aligned}$$

and solves the ESD problem.

Both the sphering in step one and the calculation of the eigenvalue equation in step two can be solved analytically without the use of a learning network and are much faster than the ICA algorithms in section 4.4 and the multi shift ESD algorithms in the following section.

In real world datasets we often have the problem that $\mathbf{C_s}(\Delta\mathbf{r})$ is not perfectly diagonal and therefore $\mathbf{C_y}$ is not perfectly symmetric. In this case we use the artificially symmetrised matrix $\frac{1}{2}(\mathbf{C_y}(\Delta\mathbf{r}) + \mathbf{C_s}(-\Delta\mathbf{r}))$ in the second step and still get a satisfying separation.

The advantage of single shift ESD, apart from the computational speed, is that equation 4.57 only has to be fulfilled for a single shift $\Delta\mathbf{r}$. The ESD assumptions about the sources are not that restrictive anymore, but the shift is arbitrary and the separation result depends a lot on a good choice of $\Delta\mathbf{r}$. This good choice of $\Delta\mathbf{r}$ for single shift ESD will always be a guess as we again do not know the original sources $\mathbf{s}$ in equation 4.57 to make this judgement.

By our experience a small shift of only a few pixels (i.e. 5 pixels) delivers good results, as the auto-correlation of the smooth sources does not change to much, but the auto-correlation of the noise will drop drastically. This brings us to a problem we generously neglected so far, the presence of noise in the mixtures $\mathbf{x}$.

4.5.2 Multi shift ESD

In the presence of sensor noise we got the notation for our mixing model in equation 4.6. The introduction of the noise term has the effect, that the sphering in the first step of the single shift ESD tries to compensate not only for the effects of the mixing matrix $\mathbf{A}$ but also for the additional noise

$$\mathbf{C_x}(0) = \mathbf{A}\mathbf{C_s}(0)\mathbf{A}^T + E\left(\mathbf{n}(\mathbf{r})\mathbf{n}(\mathbf{r})^T\right) \tag{4.72}$$

Even if the sources $\mathbf{s}$ fulfilled equation 4.57 our source estimates $\hat{\mathbf{s}}$, that inherently carry the noise compensation from sphering might not do so anymore. The effect of this is that the $\mathbf{C_{\hat{s}}}(0)$ and $\mathbf{C_{\hat{s}}}(\Delta\mathbf{r})$ are not diagonal and the basis of ESD, that we derived from equations 4.61 and 4.62 is destroyed. In order to still solve the ESD problem we have to find a non-orthogonal matrix $\mathbf{U}$ in the second step to achieve simultaneous diagonalisation of the correlation matrices.

In the ESD assumption we said that all $\mathbf{C_s}(\Delta\mathbf{r})$ should be diagonal for any $(\Delta\mathbf{r})$. If we now use multiple $\mathbf{C_x}(\Delta\mathbf{r})$ for simultaneous diagonalisation we can increase the robustness of the method against the influences from the noise compensation. As we can not rely on the diagonal form of our $\mathbf{C_{\hat{s}}}(\Delta\mathbf{r})$ anymore we have to derive a optimisation procedure that diagonalises the correlation matrices for many shifts as good as possible.

The neural network algorithm to solve this task is called multi shift ESD and was introduced in (Schöner et al., 1999; Schöner et al., 1999b; Schöner et al., 2000). We use a gradient descent algorithm to find the absolute minimum of a cost function, that diagonalises the equation $\mathbf{W}\mathbf{C_x}(\Delta\mathbf{r})\mathbf{W}^T = \mathbf{\Lambda}(\Delta\mathbf{r}) \quad \forall(\Delta\mathbf{r})$.

One possible cost function E to solve this problem after sphering is

$$\begin{aligned} E(\mathbf{U}) &= \sum_{\Delta\mathbf{r}} \sum_{l \neq m} \langle \hat{\mathbf{s}}_m(\mathbf{r}) \hat{\mathbf{s}}_l(\mathbf{r} + \Delta\mathbf{r}) \rangle_{\mathbf{r}} \\ &= \sum_{\Delta\mathbf{r}} \sum_{l \neq m} \left(\left(\mathbf{U}\mathbf{C}_{\mathbf{y}}(\Delta\mathbf{r})\mathbf{U}^T \right)_{lm} \right)^2 \end{aligned} \tag{4.73}$$

The gradient descent on this cost function enables us to a find non-orthogonal matrix $\mathbf{U}$ as claimed before. All this cost function does is to score the ability to remove the cross-correlation functions between the source estimates (i.e. make the sum of all off-diagonal elements of $\mathbf{U}\mathbf{C}_{\mathbf{y}}(\Delta\mathbf{r})\mathbf{U}^T$ as small as possible) while keeping the auto correlation functions finite for all $\Delta\mathbf{r}$. To avoid the trivial solution as the minimum of the cost function the diagonal elements of the estimated mixing matrix $\hat{\mathbf{A}}$ are fixed to 1.

$$(\hat{\mathbf{A}})_{ll} \equiv 1 \tag{4.74}$$

In practice we do not use all possible shifts, but a limited number. Normally we have chosen about 12 shifts that are arranged in a star like pattern around the zero shift. One first applies the sphering and then calculates the correlation matrices for the given shifts. Then the gradient descent is applied to the calculated $\mathbf{C}_{\mathbf{y}}(\Delta\mathbf{r})$. The gradient of $E(\mathbf{U})$ is calculated numerically. To accelerate the gradient descent we used the conjugate gradient method described by Press et al. (1988) with dynamic step-width adaptation published by Rüger (1996). "In order to achieve dynamic parameter adaptation, it is necessary to modify the learning algorithm under consideration: evaluate the performance of the parameters in use from time to time, compare them with the performance of nearby values, and (if necessary) change the parameter setting on the fly. This requires that there exists a measure of the quality of a parameter setting, called performance, with the following properties: the performance depends continuously on the parameter set under consideration, and it is possible to evaluate the performance locally, i.e. at a certain point within a inner loop of the algorithm (Rüger, 1996)".

Instead of applying sphering first and then calculate $\mathbf{U}$ one can also minimise a cost function that optimises the full separating matrix $\mathbf{W}$

$$E(\mathbf{W}) = \sum_{\Delta\mathbf{r}} \sum_{l \neq m} \left(\left(\mathbf{W}\mathbf{C}_{\mathbf{x}}(\Delta\mathbf{r})\mathbf{W}^T \right)_{lm} \right)^2 \tag{4.75}$$

It implicitly calculates the sphering matrix $\mathbf{D}$ and the demixing matrix $\mathbf{U}$, but is less stable and more likely not to converge.

Noise-Robust Sphering

One possibility to make the gradient descent more stable is the so called noise-robust sphering. The non-orthogonality of $\mathbf{U}$ was forced by the influence of the noise on the result of the sphering (Müller et al., 1999). This influence can be weakened by using a symmetrised shifted correlation matrix with a small shift vector $\Delta\mathbf{r}_n$ (i.e. one pixel) instead of the zero shift covariance matrix (Müller et al., 1999)

$$\begin{aligned}\hat{\mathbf{C}}_{\mathbf{x}} &= \mathbf{A}\mathbf{C}_{\mathbf{s}}(\Delta\mathbf{r}_n)\mathbf{A}^T + \langle \mathbf{n}(\mathbf{r})\mathbf{n}^T(\mathbf{r}+\Delta\mathbf{r}_n)\rangle_{\mathbf{r}} \\ &\approx \mathbf{A}\mathbf{C}_{\mathbf{s}}(\Delta\mathbf{r}_n)\mathbf{A}^T \approx \mathbf{A}\mathbf{C}_{\mathbf{s}}(0)\mathbf{A}^T \end{aligned} \tag{4.76}$$

As we know from figure 4.20 **d)** this destroys the correlation for white noise but has hardly any influence on the correlations of the smooth sources.

4.5.3 Regularized multi shift ESD

So far we only considered BSS algorithms that do not need any information about the mixing matrix **A** and the sources **s**. Only the assumption of factorising pdf's for ICA and the assumptions about the cross- and auto-correlation in case of ESD must be fulfilled by the **s**. Let us have a closer look at the properties of the cost function in equation 4.73 now.

In the paper by Molgedey and Schuster (1994), the constraint to fix the diagonal auto-correlation values to 1 was suggested to avoid trivial solutions of **W**. However under certain conditions the gradient descent with the constraint on the diagonal elements of $\hat{\mathbf{A}}$ does not prevent **W** from becoming arbitrarily small and decreasing the cost function without diagonalising $\mathbf{C}_{\hat{\mathbf{s}}}(\Delta\mathbf{r})$. We will only give a short example here to illustrate this weakness. An extensive discussion on cost functions for ESD and their limitations can be found in (Vollgraf, 2000).

Consider two covariance matrices

$$\mathbf{C}_{\mathbf{s}}^1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \tag{4.77}$$

and

$$\mathbf{C}_{\mathbf{s}}^2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.78}$$

for two given source signals, which are mixed using the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}. \tag{4.79}$$

Figure 4.21 shows a contour plot of the cost function with respect to the two off-diagonal elements of the estimated mixing matrix $\hat{\mathbf{A}}$. The hyperbolic ridge corresponds to values at which $\hat{\mathbf{A}}$ becomes singular and causes an infinite value of the cost function. Solid lines mark gradient descent trajectories for five different initialisations. It can be seen, that depending on the initialisation, the gradient descent procedure may succeed to find the true mixing matrix (trajectories 1,3), or may diverge, leading to arbitrary small **W** and hence trivial minima (trajectories

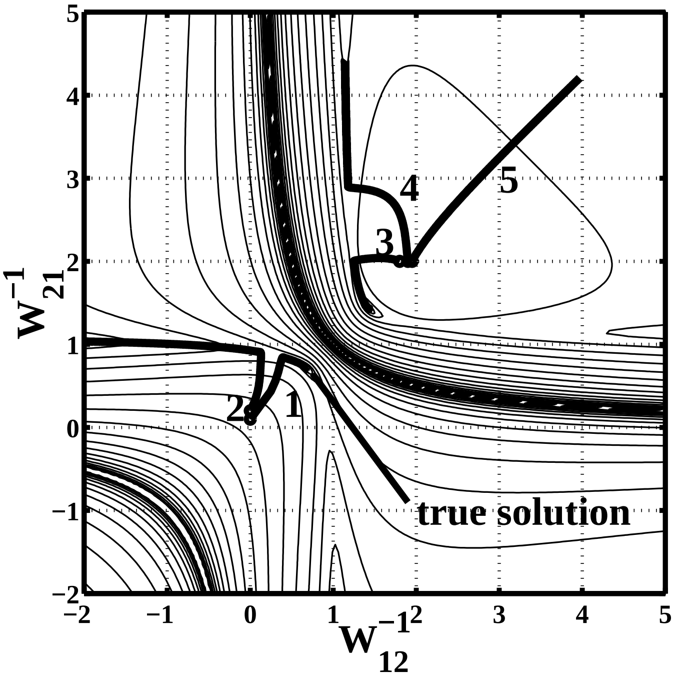


Figure 4.21: Surface of the cost function equation (4.74) with the constraint $(\hat{\mathbf{A}})_{ll} = 1$ for two sources and the mixing matrix provided in the text. The two hyperbolic ridges are regions with singular $\mathbf{W}^{-1}$ and have infinite values of the cost function (adapted from Vollgraf et al. (2000)).

2,4,5). Due to these properties, the constraint (4.74) does not seem to be suitable for gradient based joint diagonalisation.

In order to prevent the gradient descent from running into the wrong minima in the publication by Schießl et al. (2000a) an additional so called regularization term was introduced. These regularization terms specify prior knowledge we have about the time course of the signal of interest. One possibility of such prior knowledge about the time course of a source in optical imaging is the metabolic response to the stimulus onset. Because $\mathbf{W}^{-1}$ should be close to $\mathbf{A}$ after the decorrelation process, we introduce a regularization term, which punishes derivation of $\mathbf{W}^{-1}$ from an estimated $\hat{\mathbf{A}}$. As column j of $\mathbf{A}$ represents the time course of signal j, we weight this derivation by the confidence we have in our prior knowledge α_j:

$$E_{\mathbf{W}}(\mathbf{W}) = \sum_j \alpha_j \cdot \sum_i \left((\mathbf{W}^{-1})_{ij} - A_{ij}\right)^2 \tag{4.80}$$

The farther the distance (sum of squared differences) is from the assumed time series, the higher the regularization costs are. Altogether we get a cost function

$$E(\mathbf{W}) = E_{\mathbf{S}}(\mathbf{W}) + E_{\mathbf{W}}(\mathbf{W}), \tag{4.81}$$

which is minimised using the gradient descent procedure mentioned above.

Strictly speaking the regularized multi shift ESD is not a BSS algorithm anymore as we make explicit assumptions about the mixing matrix **A** and therefore the mixing process. This algorithm is rather a hybrid method that uses spatial information in the data as well as the temporal structure.

Chapter 5

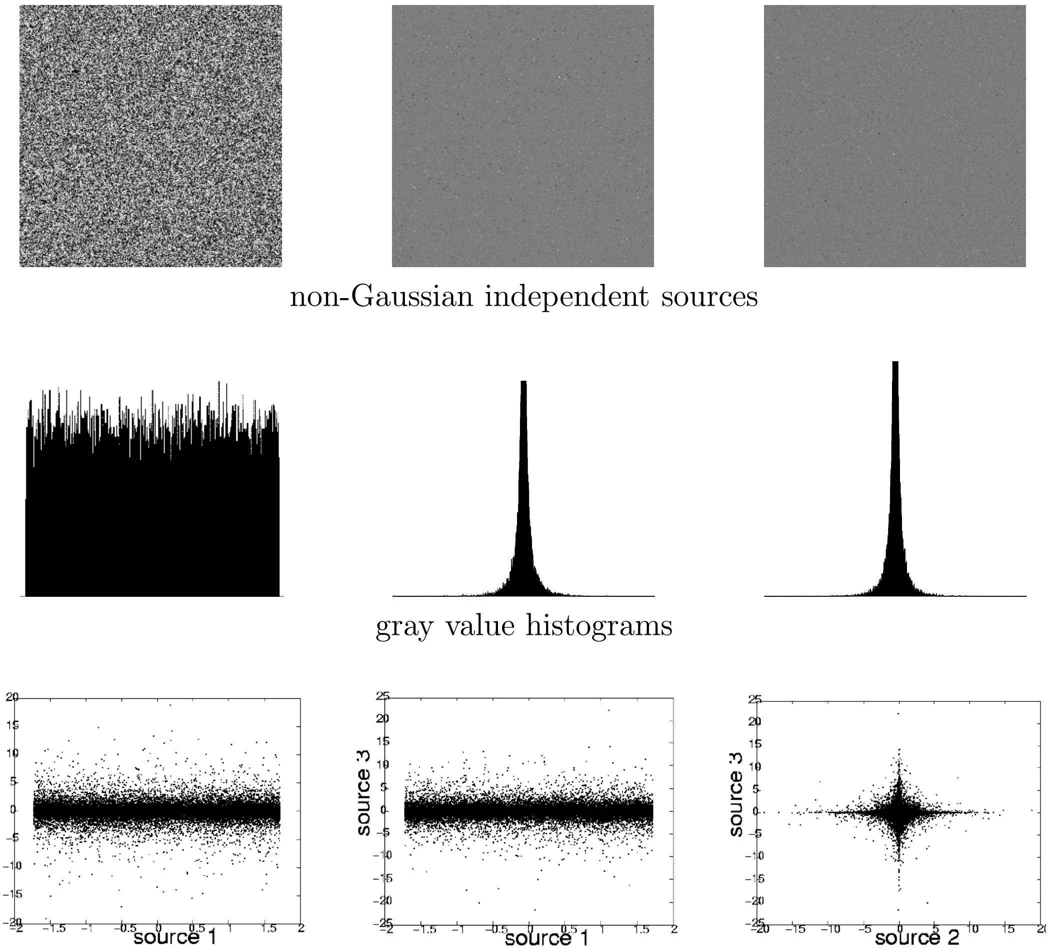
Performance Test on Artificial Data

In order to make a quantitative study of the performance of the individual separation algorithms we must compare the original sources with the separation results. We will introduce a measure that will tell us on a continuous scale from zero to one how good these estimates are. The problem we have with our datasets form the imaging of the intrinsic signals in vivo is that we do not know the spatial distribution of the underlying original sources, as they differ slightly from individual to individual. There are possibilities to get a better idea by simultaneous electrode recordings and imaging or the staining of the tissue, but still the biological and metabolic activities we can control in this manner are different in time and space than the intrinsic signals.

To overcome this weakness we created different sets of artificial data that on the one hand should have similar statistical properties to the optical images and on the other hand allow us to explore dependencies of the separation quality of the algorithms on special features of the data. We have seen in chapter 4 that the individual algorithms are based on different assumptions. This will certainly be reflected in the separation performance depending on how well the assumptions are fulfilled. So keep in mind that a poor separation quality does not indicate that the BSS or ICA algorithm is less elaborate but that it is not suitable for the given problem.

5.1 The Artificial Datasets

We have used three different artificial datasets for testing the algorithms. In the later results only the meaningful combinations of datasets and algorithms that underline the tendencies will be presented due to space limitations.



non-Gaussian independent sources

gray value histograms

joint probability distribution

Figure 5.1: Illustration of the statistical parameters for the non-Gaussian independent sources. The **top row** shows the individual sources. As they were designed to only meet the assumptions made by the ICA algorithms about the pdf's they contain no structure. In the **middle row** we can see the histograms of the gray values that illustrate the source distribution. The flat histogram on the left shows that source one is sub-Gaussian whereas the other two are super-Gaussian. From the joint distributions in the **bottom row** we can detect that the sources are independent.

5.1.1 Independent Sources

This dataset was designed to fulfill the basic assumptions of the classical ICA algorithms like kurtosis optimization and Infomax. That is the sources should be as non-Gaussian as possible (i.e. have higher order moments) and should be statistical independent. For all the tested datasets we calculated three images. For this set the first image was a dataset with a sub-Gaussian gray value distribution.

It was calculated by taking a random equal distribution and has a kurtosis of $kurt \approx -1.2$. The other two sources are super-Gaussian distributions that were calculated by taking the power of three of a random normal distribution. These super-Gaussian sources had a kurtosis around $kurt \approx 40$. For each mixture the three sources were calculated again and with the random values (hence the kurtosis changed a bit (therefore the $\approx$)).

Figure 5.1 shows three generated images in the top row and reminds us again that ICA algorithms purely consider the pdf's but no structure.

In the second row we see the histograms of the gray value distributions. For the two supergaussian images we have clipped the y-axis to be able to display the tails of the distribution.

In the third row we see the joint distributions and as we know from figure 4.14 the distributions are independent because we can make no prediction from one value of the first about the a value of the other distribution.

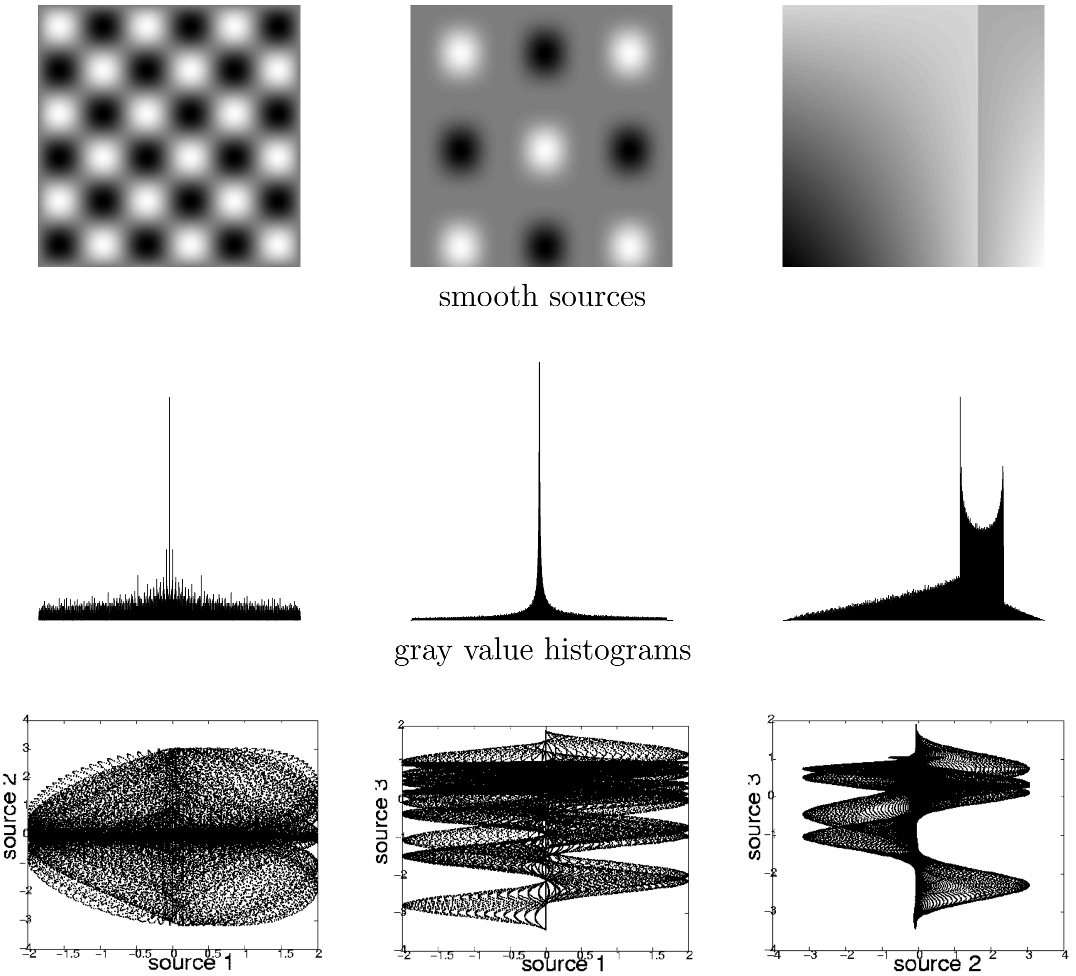
The three independent sources are generated randomly and therefore there is no structure in the images. The auto-correlations of the three images are similar to that of white noise and have a strong decay for even small shifts (see example in figure 4.20 d)). The two dimensional auto- and cross-correlations are not shown for this test dataset as they again only show a peak in the middle.

5.1.2 Smooth Sources

The sources we refer to as smooth sources were generated so that they yield properties of our real optical images. All the images are calculated from superimposed two dimensional sine functions with different frequencies. The second image is additionally multiplied with itself two times. The third image is rescaled on top of that by the pixel maximum times a sign function. Therefore we have spatial smooth structures in the images and they are subgaussian (smooth source 1, $kurt \approx -0.75$) or supergaussian (smooth source 2, $kurt \approx 2.1$; smooth source 3, $kurt \approx 0.6$).

In figure 5.2 the top row displays the three smooth sources. We now have the spatial smooth structure like in the optical images. The first two images should represent intrinsic patterns with different spatial frequencies whereas the third smooth source illustrates the gradual change in intensity across the image due to illumination or global response. The second row displays the gray value histograms again and as stated above the first image has a negative kurtosis and the other two a positive kurtosis.

The bottom row plots the joint pdf's of the smooth sources and we can see that the sources we generated from sine functions have very elaborate dependencies and are not independent anymore. But there is also no simple correlation structure and the density is well spread in the probability space, so that the independence assumption is not fulfilled but good enough to let the ICA algorithms work. The violation of the independence assumption is also expected to be true for the biological sources, as they are all triggered by the same event. But again as we do



smooth sources

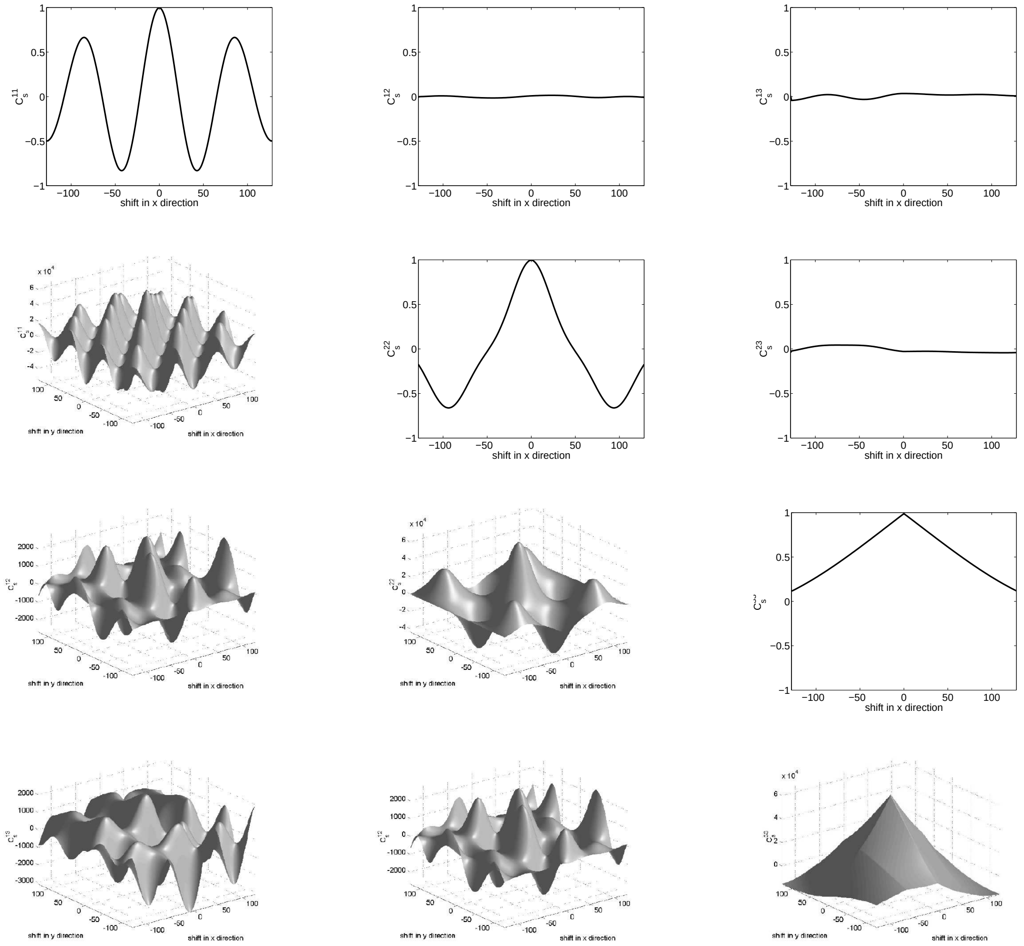
gray value histograms

joint probability distribution

Figure 5.2: Illustration of the statistical properties of the smooth sources. The **top row** displays the three smooth sources and the **middle row** shows the gray value histograms. From the histogram of the first source one might think it is supergaussian but the tails are to heavy compared to the thin peak in the middle, so that the kurtosis is negative. The other two sources are supergaussian again. The joint probabilities in the **bottom row** show complex statistical dependencies as the sources were calculated from different sine functions.

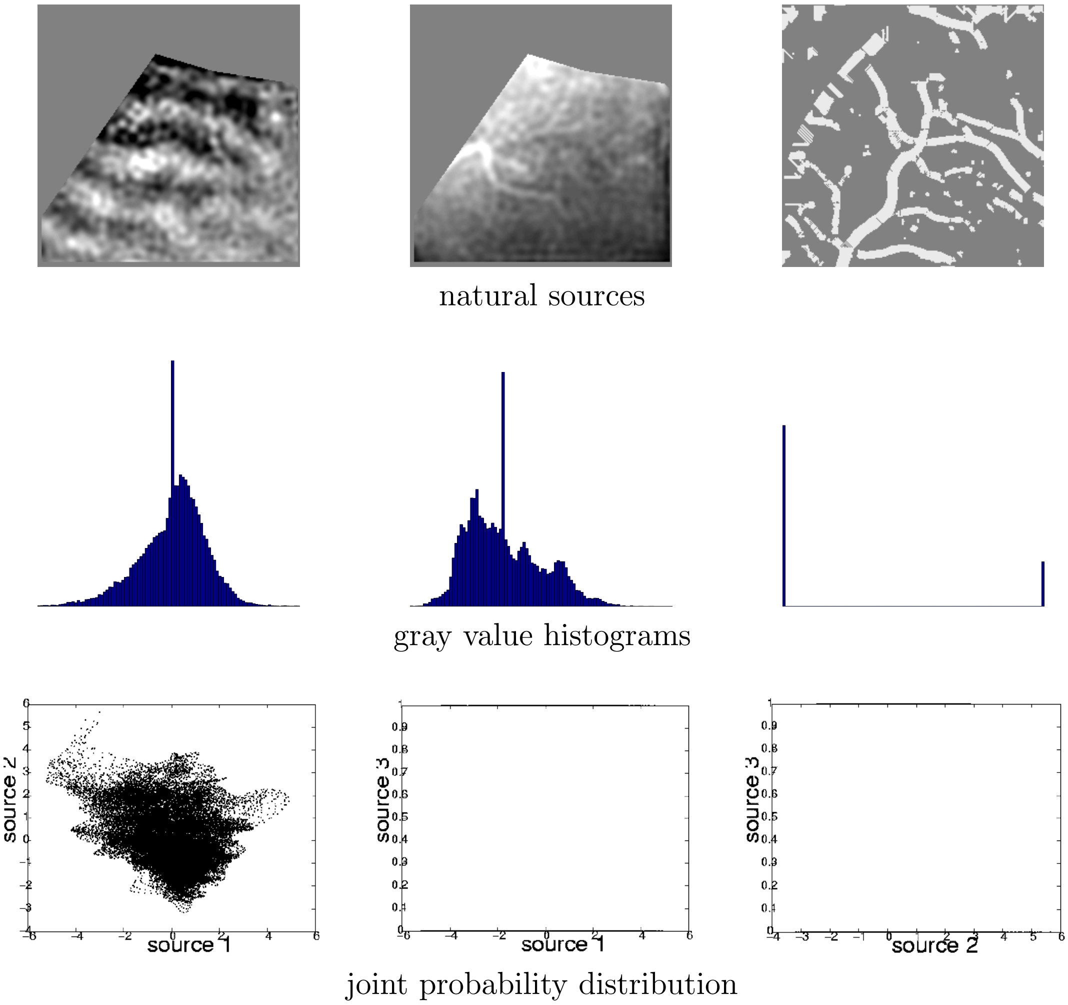
not know the biological sources this is hard to prove.

Figure 5.3 shows the auto- and cross-correlations for the three smooth sources. The lower triangle displays the two dimensional surface of the correlation functions of the images. The surfaces are not normalized to display the smoothness across the whole range. The upper triangle shows the auto- and cross-correlations along the horizontal mid-line of the correlation surface and is normalized to one. As we can see the assumption for ESD is fulfilled, the auto-correlations (diagonal



auto- and cross correlations

Figure 5.3: The auto- and cross-correlations for the smooth sources. The **upper triangle** shows a cut through the two dimensional correlation functions of the smooth sources displayed in the **lower triangle**. The graphs in the upper triangle are normalized so we can see that the cross-correlations (off diagonal elements) are much weaker than the auto-correlations (diagonal elements) as required for the ESD algorithm. The surface displays are scaled to their individual maximum to display the smoothness of the correlation functions.



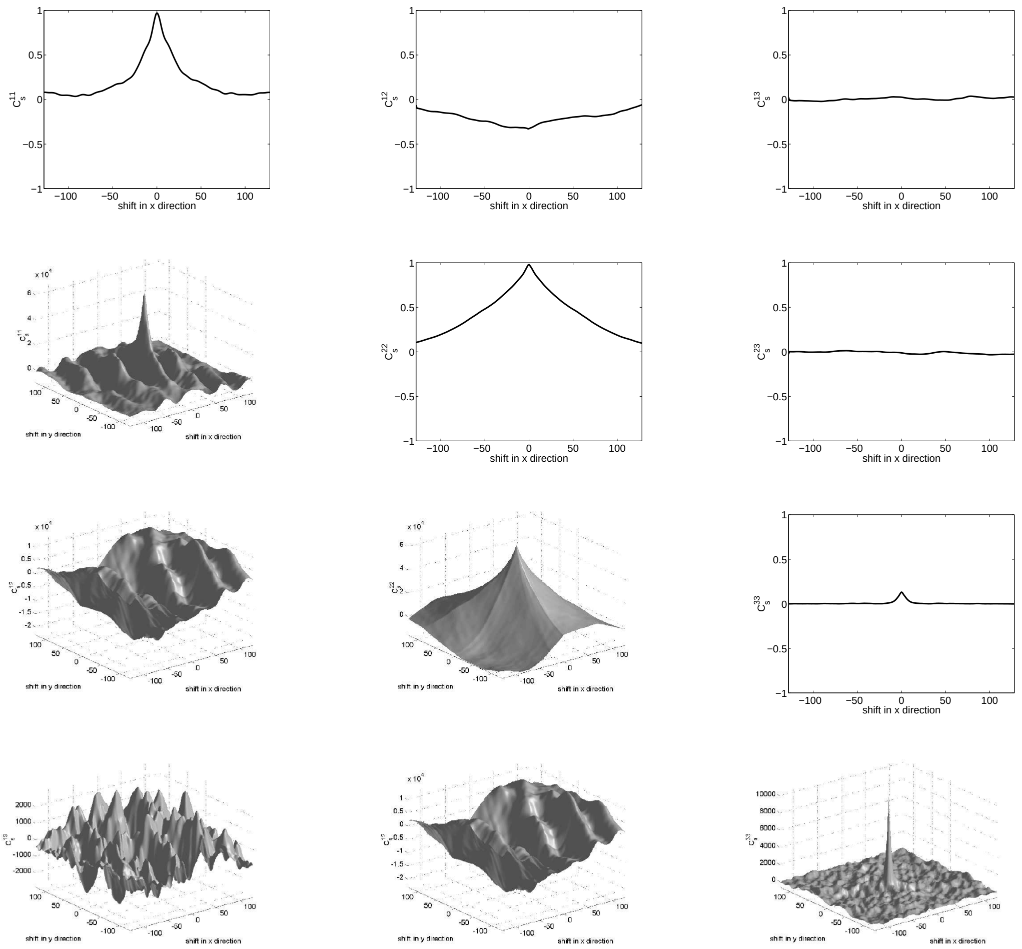
natural sources

gray value histograms

joint probability distribution

Figure 5.4: Illustration of the statistical properties of the natural sources. The **top row** displays the three natural sources and the **middle row** shows the gray value histograms. The histogram of the vessel pattern is restricted to the values 0 and 1 as this is the mask we received from a wavelet analysis. All three sources are supergaussian. The joint probabilities in the **bottom row** display no dependencies for the first two sources. The pdf's with the third source are split to the two values of the third source but still we can make no prediction from one pdf about the other.

elements) are strong compared to the cross-correlations (off diagonal elements). The correlations for the shifted versions are not shown, but from the smoothness of the correlation graphs one can see that it is nearly the same for small shifts.



auto- and cross correlations

Figure 5.5: The auto- and cross-correlations for the natural sources. The **upper triangle** shows a cut through the two dimensional correlation functions of the natural sources displayed in the **lower triangle**. The assumptions for ESD are not perfectly fulfilled anymore as the auto-correlation of the third natural source is also quite weak compared to the other auto-correlations. Also the smoothness is lost due to the thinned structure of the vessels.

5.1.3 Natural Sources

In the final test data set with the name natural sources we used results from optical imaging after the analysis, where we assumed a successful separation. The top row in figure 5.4 shows the selected images. The first image is an ocular dominance map with a masked region. The second image displays a global illumination gradient and the third image a vessel pattern. All there sources are supergaussian with a kurtosis of $kurt = 3.18; 1.32; 0.3$ respectively. The histograms of the first two sources in the middle row show the peak for the zero gray value due to the masking of the images. The histogram of the third source contains only the values zero and one. This awkward pdf also has an impact on the shape of the joint probability distributions. The joint pdf of the first two sources is shown on the left in the bottom row. It shows some structure and the two sources are not perfectly independent but independent enough. The joint pdf that contain the third source are split between the two values 0 and 1 on the y-axis. But still we can make no prediction from one source about the value of the other.

Figure 5.5 displays the correlations between the sources themselves and among each other. The two dimensional auto-correlation surface of the ocular dominance pattern nicely reflects the periodic structure the image contains. Again the third source makes an exception as the auto-correlation function is quite small (upper triangle last plot) compared to the other auto-correlations and not smooth as displayed by its correlation surface. For all cross-correlations it is true that they nearly vanish as assumed by the ESD algorithms.

5.2 The Reconstruction Error and Condition Number

For testing the separation performance of the introduced algorithms on the artificial data sets with additive sensor noise we need an error measure for the quality of the reconstruction. In the case of a perfect blind separation our source estimates $\hat{\mathbf{s}}$ would look like the original sources. Due to the ambiguities of ICA they could have a different sign, i.e. look like a negative of the original, and the order of the sources from one to three can be different, i.e. a permutation (see figure 5.10. Therefore the quality of reconstruction was scored by calculation the covariance matrix $G_{i,j} = \langle \hat{\mathbf{s}}_\mathbf{i} \mathbf{s}_\mathbf{j} \rangle$ between the true sources and the estimates. By definition a permutation matrix only contains one non-zero element with the value 1 in each row and column. Unfortunately we will not always get perfect separations especially at high noise levels.

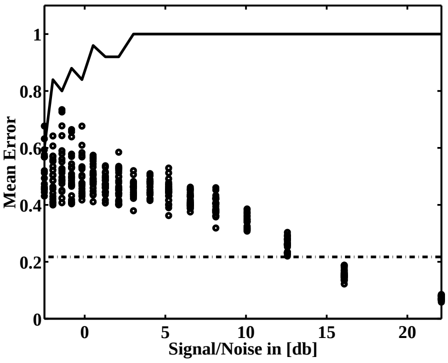
In figure 5.6 a) we see a typical result from the multiplication of $\hat{\mathbf{s}}$ and $\mathbf{s}$. So far it does not look like a permutation matrix. We now determine the maximum absolute value of each row (boxed value) and normalize the elements of this row by division with the absolute value of this maximum. The result of this is shown in b). At this point we make the judgment, if we count the separation as a successful trial or not.

$$\begin{pmatrix} -2.1611 & 3.8979 & \boxed{-4.6167} \\ \boxed{3.6116} & -0.2532 & -1.7981 \\ 2.2964 & \boxed{4.9549} & 3.3318 \end{pmatrix}$$

a)

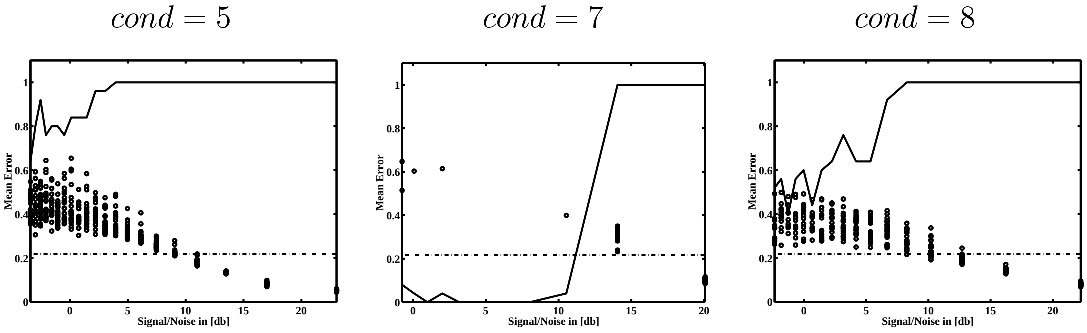
$$\begin{pmatrix} 0.4681 & 0.8443 & \boxed{1.0000} \\ \boxed{1.0000} & 0.0701 & 0.4979 \\ 0.4635 & \boxed{1.0000} & 0.6724 \end{pmatrix}$$

b)



c)

Figure 5.6: Calculation of the success rate and the mean reconstruction error RE. **a)** shows the result from $(\mathbf{s}\hat{\mathbf{s}}^T)$. Due to the normalization this is equal to the correlation matrix. In **b)** each row is normalized to the maximum absolute value (boxed numbers). If one column contains two maxima the matrix is no permutation matrix and the separation failed. The plot in **c)** shows the final result of a complete test. For detailed explanation see text.


$cond = 5$ $cond = 7$ $cond = 8$

a) b) c)

Figure 5.7: The condition number of a matrix is an unreliable measure of how ill-conditioned a matrix is. In general the separation quality decreases with the rising condition number **a)-c)** as the percentage of successful separation goes down and the mean reconstruction error and the variance for each noise level rises. But one has to take care of outliers like in **b)** that result in an arbitrary breakdown of the separation.

If a single column contains two 1's we do not have a permutation and therefore the sources were not separated. The separation failed. If we have a permutation matrix the non-maximum values tell us how good or clean the separation was. In the ideal case they should all be zero. As a quantitative measure we use the mean reconstruction error (RE) as suggested by Koehler and Orglmeister (1999):

$$RE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N-1}(\sum_{j=1}^{N}\frac{|G_{ij}|}{max_k|G_{ik}|} - 1) \tag{5.1}$$

where $G_{i,j}$ are the matrix elements and k denotes a row of the matrix.
For a complete test result like in shown in figure 5.6 c) we mix the chosen toy dataset with a fixed mixing matrix **A** and add random white noise to each of the resulting mixtures. The noise is re-calculated for each of the three mixtures at a given noise level and therefore sensor noise. Now we apply the separation algorithm and calculate RE from the results. This procedure is repeated several times (normally 15-25 times) for each noise level.

The noise level is gradually increased until a signal to noise ratio of at least 1 is reached, that is equal to $0dB$. The definition of the signal to noise ratio (SNR) in decibel (dB) is

$$SNR = 10 \times \log_{10}\left(\frac{\sigma^2(signal)}{\sigma^2(noise)}\right)[dB] \tag{5.2}$$

The maximum noise level is set by taking the maximum variance σ^2 of the mixtures and rise the variance of the noise step wise until we have reached this value. Therefore sometimes even higher noise levels than $0dB$ are shown in the plots. The solid line in the plot of figure 5.6 **c)** displays the percentage of successful separations (where 0 = 0% and 1 = 100%). The dashed line shows the percentage of permutation matrices one generates by random calculation of 3×3 matrices. If the solid line goes below the dashed line the result is worse than by chance. Each circle in the plot shows the reconstruction error for a successful separation. This means if twenty trials per noise level are run and 100% were successful (solid line at 1) we have 20 circles at this noise level. If only 50% were successful (solid line at 0.5) we are left with 10 circles. In case of very stable results the circles are drawn on top of each other and may appear as one.

Apart from the sources, the noise level and the separation algorithm the attributes of the mixing matrix **A** are critical for the separation result. The more ill-conditioned the mixing is, the more difficult is the separation. Ill-conditioned for the source separation problem means: Can we calculate the inverse of **A**, the demixing matrix **W** and does **W** contain arbitrary big or small elements? An indication for the accuracy of the results from matrix inversion is the condition number of a matrix. The definition of the condition number is

$$cond(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-}\| \tag{5.3}$$

to a given vector norm.
So the condition number is the ratio of the largest singular value of the matrix to the smallest, because the norm $\|\mathbf{A}\|$ is equal to the largest singular value of **A**. The singular value decomposition of a matrix **A** results in a diagonal matrix **S** of the same dimension as **A** and the s_{ii} are the singular values. The calculation is done with two unitary matrices so that $\mathbf{A} * \mathbf{S} * \mathbf{V}^T$. This is the computationally most efficient way to calculate the condition number and used by the standard

C or matlab routines. A matrix with a condition number of $cond = 1$ is ideally conditioned and the higher *cond* the more difficult is the separation.

In figure 5.7 we can see that this is generally true that with rising condition number the separation is more difficult and therefore the reconstruction error becomes bigger. But under arbitrary conditions there can be outliers in the trend as shown in figure 5.7 b). So when we make a judgment about the performance of the separation for a given condition number we have to consider the results with several matrices with the same condition number to eliminate such influences.

5.3 Choosing the right shift for single shift ESD

In order to find out which single shift applied in the single shift ESD separation gives the best and most stable results we ran a simulation on different toydata sets. In this simulation we used the mixing matrix**A** with a condition number of 11 (see equation 5.5). The representative example depicted (figure 5.8) is the result for the mixing of the smooth sources (see figure 5.2).

The range of shifts along the x- and y-axis was -100 to 100 pixels in a 256x256 pixel image. For each shift the reconstruction error (equation 5.1) was calculated. The simulation was repeated ten times for each shift and the mean value of the successful separations was stored at the shift location. In a second matrix the number of successful trials was stored to have a measure for the stability of the separation for a given shift. If all ten trials to separate the mixture were unsuccessful according to the criteria in the calculation of the reconstruction error, the error was set to 1. So for a perfect separation the reconstruction error is 0, represented by a black pixel in figure 5.8 **a**), **b**) and **c**), and 1 for complete failure, represented by a white pixel. The complete simulations for all shifts were calculated at different noise levels. For the simulations in **a**) no noise was added, in **b**) the signal to noise ratio was around $5db$, while in **c**) the signal to noise ratio was $0db$. In figure 5.8 **d**) the success rate for the $0db$ noise level is shown.

The first three images show that the black areas representing perfect separations become much smaller with rising noise levels. Also the white areas where no separation was possible grow with higher noise levels. The symmetry of the underlying original sources is reflected in the pattern of separation quality as we are using a correlation criterion for separation. This property though is of no help for selecting the shift in real world data, as we do not know the patterns of the original sources.

In the no noise case **a**) for very small shifts up to 3x3 pixels the separation fails as the unshifted and shifted correlation matrix are too similar for the type of smooth patterns we are looking at and therefore no additional information is gained. The shifts from 5x5 to 9x9 pixels deliver good separation results. In the case with the added noise all small range shifts deliver good separations in the case of the smooth sources. There are areas with better separation qualities (more

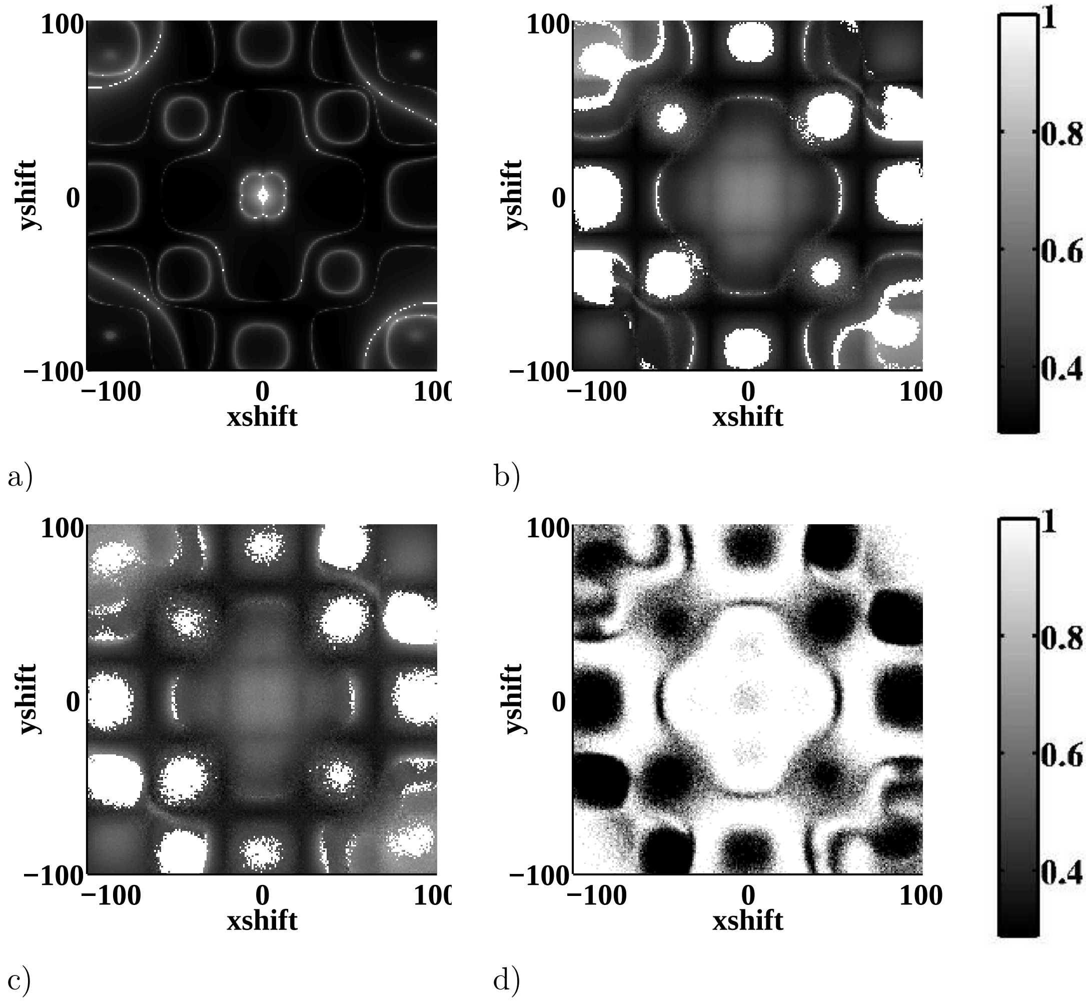


Figure 5.8: Simulation of the separation results of single shift ESD at different signal to noise levels (**a)** for no noise, **b)** for 5 db and **c)** at around 0 db) for shifts from −100 to 100 along both axes of the smooth sources. No filtering was applied. The separation success was scored with the reconstruction error (see equation 5.1). The black areas show no reconstruction error whereas for shifts where the reconstruction failed completely the pixels are white. Image **d)** shows the rate of success of separation at the high noise level (**c)**)in this simulation. If all trials were separated the value equals one (white) and zero (black) for no successful trial. For details see text.

black) but they could not be predicted in the real datasets.

Another important factor for choosing the right shift is how reproducible the separation result is. In **d**) we see the distribution of the number of successful separations of the high noise case in **c**). If ten of ten trials were separated the pixel value is 1 (white). These areas correlate with the areas of low reconstruction error but do not exactly match the patterns. Thus it is possible to have a low reconstruction error but not be very reliable in separation success.

So how does one choose the best shift for the single shift ESD? From the results presented here and the experience from many other separations with different sources one has to select a shift that gives the highest probability for a low reconstruction error and a good success rate. Without knowing the original sources in our experience shifts of 5x5 to 9x9 work the best for the reasons given above. For the rest of the thesis in the case of single shift ESD a shift of (5x5) was chosen.

5.4 Comparison of ICA with single shift ESD

To begin with we have to investigate how the two classical ICA algorithms perform compared to our single shift ESD algorithm on the different test datasets at various noise levels. For this comparison we will consider the independent sources and the smooth sources as they have clearly designed properties. The natural sources will be used in later comparisons once we have experience with the separation behavior under the defined conditions.

Remember that the independent sources favor the ICA algorithms while the smooth sources have properties similar to the optical imaging datasets.

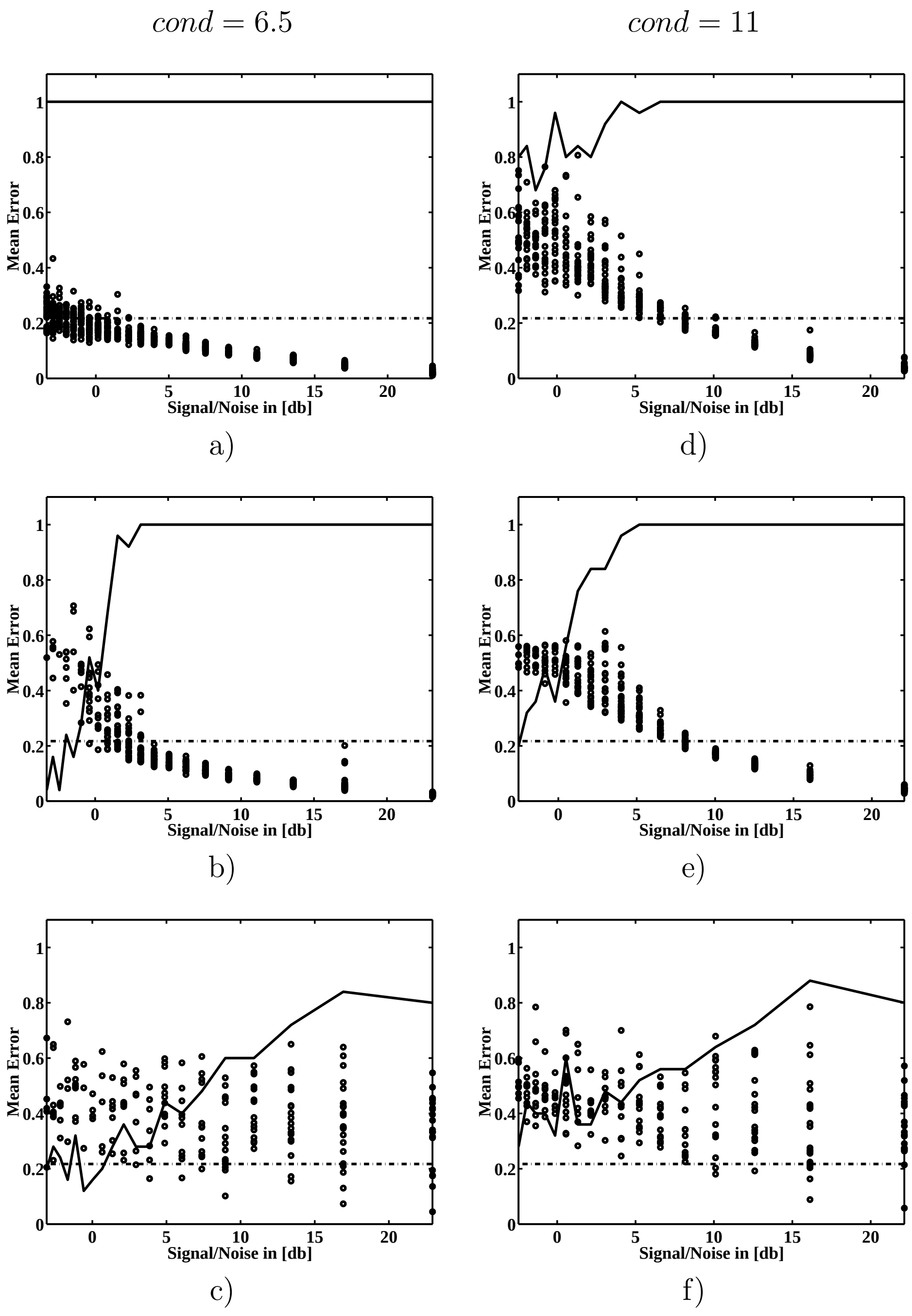
5.4.1 Test Results with the Independent Sources

This test series was run with one mixing matrix $\mathbf{A}$ having a condition number of 6.5 and one with 11:

$$\mathbf{A}_{cond=6.5} = \begin{pmatrix} 0.39 & -0.56 & 0.78 \\ 0.08 & 0.44 & 0.57 \\ -0.64 & -0.95 & -0.82 \end{pmatrix} \tag{5.4}$$

$$\mathbf{A}_{cond=11} = \begin{pmatrix} 0.74 & 0.41 & 0.93 \\ 0.41 & 0.97 & 0.73 \\ 0.52 & 0.72 & 0.45 \end{pmatrix} \tag{5.5}$$

In the left column of figure 5.9 we see the performance for the matrix with the lower condition number and on the right for the matrix with the higher condition number. The top row shows the results for the kurtosis optimization, the middle column for the Infomax algorithm and in the bottom row for the single shift ESD. For the condition number of 6.5 the kurtosis optimization *a*) sustains a 100% successful separation rate even for signal to noise ratios worse than $0dB$. Also the mean reconstruction error is low ($RE < 0.2$) for all noise levels up to $0dB$ with a small variance in the quality of the results in each single step (small spread of the circles). The Infomax algorithm *b*) performs nearly as well up to a noise level of $3dB$. From there on with a decreasing signal to noise ratio the percentage of successful separations drops drastically and the mean error rises. Also the



a) d) b) e) c) f)

Figure 5.9: Test results for the simulations with the **independent sources**. The left column shows the results for the matrix with the condition number 6.5 and the right column for the condition number 11. The kurtosis optimization algorithm **a)**, **d)** performs well up to high noise levels. The Infomax algorithm shows a decreasing performance from a signal to noise ratio of $\approx 3dB$ downwards as the percentage of the successful separations declines and the variance of the separation quality increases. The single shift ESD algorithm **c)**, **f)** shows the worst performance on the dataset with the independent sources, as they do not fulfill the needed assumptions about the auto- and cross-correlations. Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices by random generation of 3x3 matrices.

reliability is reduced as we get a reconstruction error up to $RE = 0.75$ in single instances.

With this dataset single shift ESD *c*) has the worst performance of all three algorithms because the assumptions about the smoothness of the sources are not fulfilled. For all noise levels we get a huge variance in the separation quality ($\Delta RE \approx 0.5$) and the number of successful separations starts to decline already around $16dB$. This result for all three algorithms is also reflected in the trials using the matrix with a condition number $cond = 11$. Because this matrix is more ill-conditioned and therefore the separation task more difficult, the performance of the three algorithms is slightly worse, expressed by a higher mean error and a lower number of successful trials. Only the Infomax algorithm seems to perform a little better below $3dB$, but as we said before this changes slightly with the individual run.

We can summarize that ICA algorithms perform well on the separation problem with the independent sources up to high noise levels and difficult mixing conditions. The single shift ESD algorithm on the other hand is clearly not designed for this kind of data and performs poorly.

5.4.2 Test Results with the Smooth Sources

Now we run the same test on the dataset with the smooth sources that were developed to meet the properties of the optical imaging datasets.

In figure 5.10 we see the smooth sources and their mixtures for the individual steps of the test with the single shift ESD algorithm and the mixing with the lower condition number matrix. After the whitening with PCA *c*) the separation is still incomplete but after the application of single shift ESD the sources are separated again *d*). Here we can see how the ambiguity of ICA affects the images. The order of the estimated sources in *d*) is permuted compared to *a*) and the source estimate in the middle of column *d*) has the opposite sign of the original.

In figure 5.11 we see the actual separation results for all three algorithms. The basic arrangement is the same as in figure 5.9. Now the performance is reversed compared to the test before. On the smooth dataset the single shift ESD algorithm delivers the best separation performance. The percentage of successful separations in *c*) is 100% even below $0dB$ for the mixing matrix with $cond = 6.5$. The quality of the reconstruction is slightly increasing with increasing noise but still very good with $RE = 0.3$ at $0dB$. The ICA algorithms $a), b)$ do not completely fail like ESD on the other dataset, nevertheless the performance is poor compared to the result shown in *c*). Especially the drop in the number of successful separations at higher noise levels is striking. This is relevant for us because in the real data sets we have signal to noise ratios of $0dB$ due to the small intensity of the mapping signal. Again all the trends are verified in the test with the high condition number matrix just that the overall performance of the three algorithms is slightly worse due to the difficult mixing matrix $\mathbf{A}$.

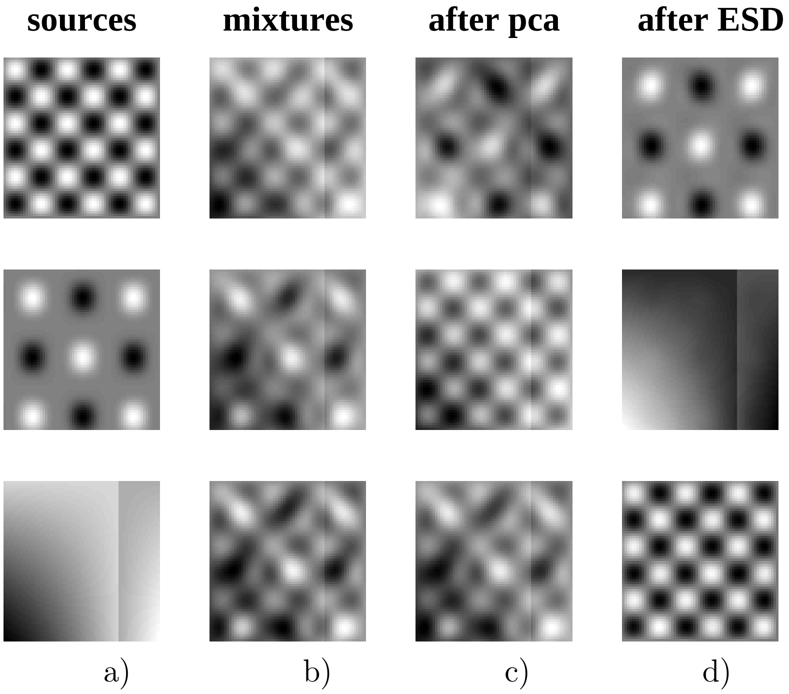


Figure 5.10: Example for the separation tests with the smooth sources **a)** and single shift ESD. In column **b)** we see the mixture that is produced with the low condition number matrix and additive sensor noise ($\sigma^2 = 0.1$). After the sphering with PCA in **c)** the source separation is still incomplete. Not until the application of single shift ESD with a shift vector of $\mathbf{\Delta r} = (5, 5)$ the sources are successfully separated. Column **d)** also illustrates the consequences of the ambiguities of ICA as the sources are permuted and the middle estimate has the opposite sign.

In section 4.2.3 we have learned that lowpass filtering can be applied to the data as the tissue scattering does not allow high frequencies in the intrinsic signals and they therefore descent from noise sources. In the following test we investigated how this lowpass filtering would affect the separation quality for our smooth sources. After the mixing of the sources and the application of the noise the mixtures were lowpass filtered with a cut off frequency of $25 cycles/256 pixels$.

In figure 5.12 we can see that the lowpass filtering increased the performance of all algorithms as the spectral power of white noise is constant across all frequencies and therefore reduced by the filter. In c) and f) the result for single shift ESD on the filtered data is plotted. More than the other algorithms ESD profits from the lowpass as the images become smoother and with them the underlying mixed sources. This smoothness is one of the assumptions of the algorithm. In c) the mean reconstruction error stays below $RE = 0.2$ even at the high noise level of $0dB$.

This demonstrates that lowpass filtering will enhance the ability of ESD to separate the sources from our optical imaging data.

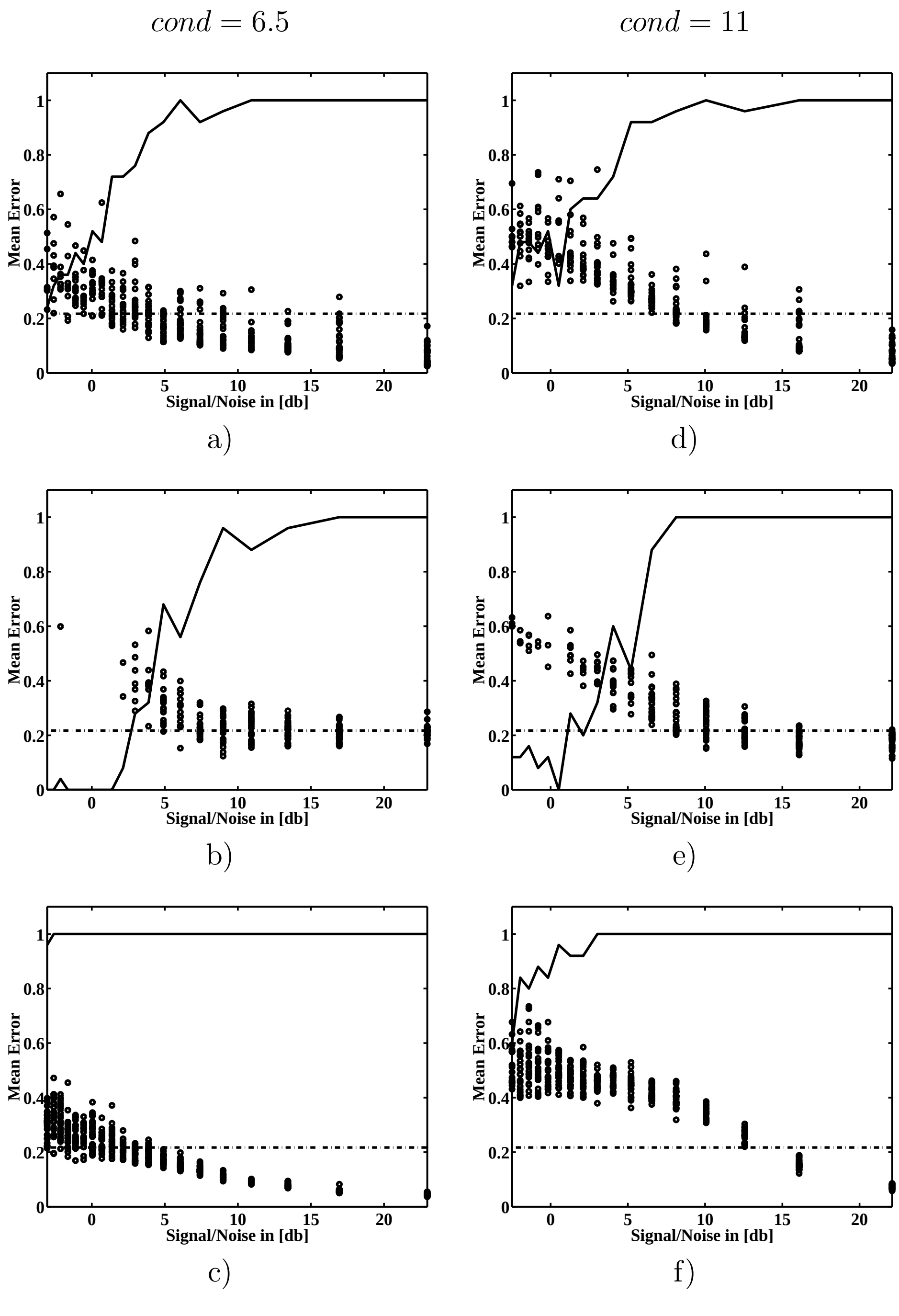
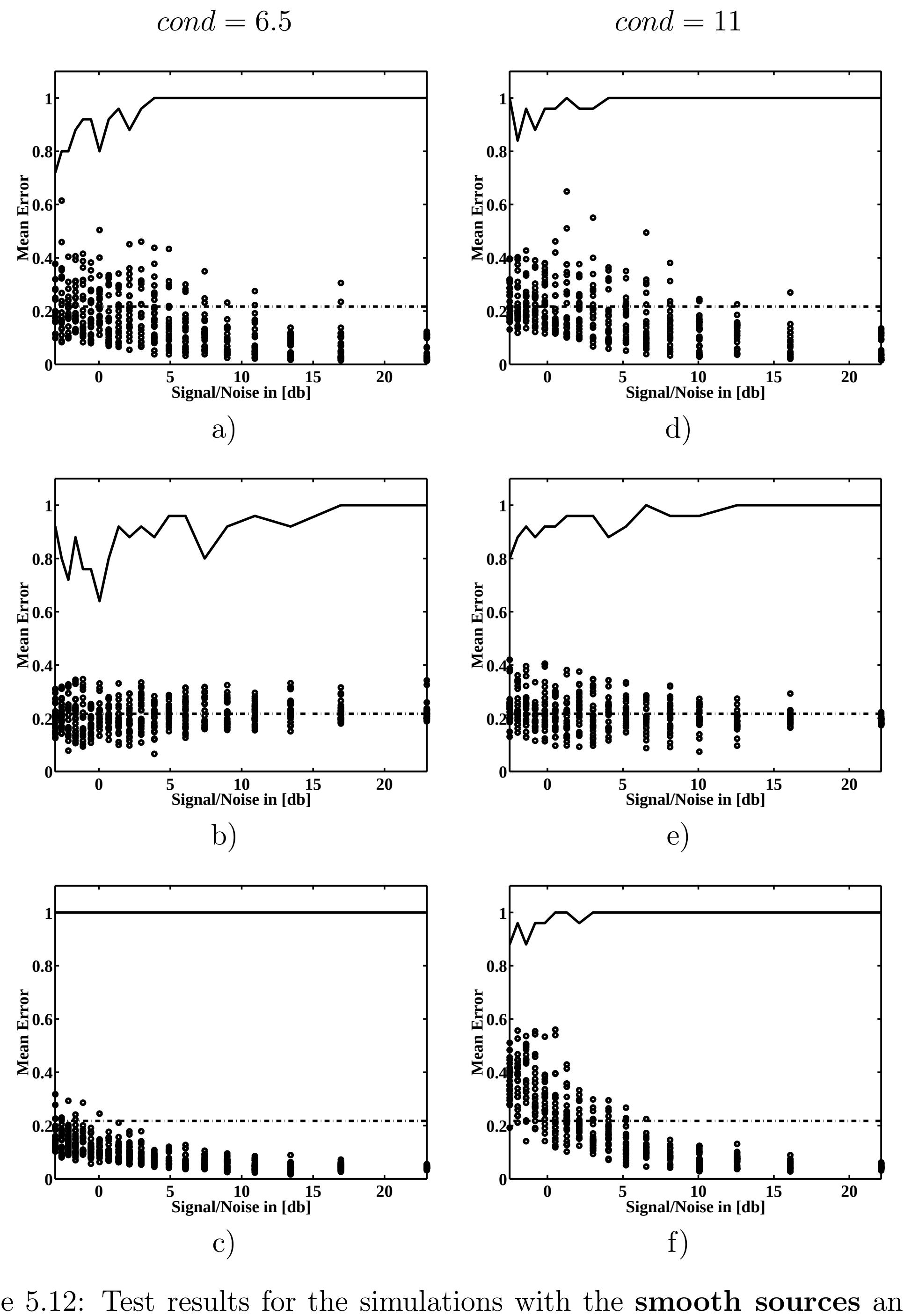


Figure 5.11: Test results for the simulations with the **smooth sources**. Unlike the ESD algorithm in the tests with the independent sources (see figure 5.9) the ICA algorithms (**first row:** kurtosis optimization, **second row:** Infomax) do not break down completely on the separation task with the smooth sources, because the sources are nongaussian and therefore partially fulfill the ICA assumptions. We can also see in **c)** and **f)** that the single shift ESD algorithm has the best performance on the artificial dataset with the statistical properties of optical images. Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices by random generation of 3x3 matrices.



$cond = 6.5$

$cond = 11$

a)

d)

b)

e)

c)

f)

Figure 5.12: Test results for the simulations with the **smooth sources** and the application of a lowpass filter with a cut off frequency of $25 cycles/256 pixels$ on the mixtures. All the algorithms (**first row:** kurtosis optimization, **second row:** Infomax, **third row:** ESD) show a better performance as the spectral power of the white noise is reduced. The ESD algorithm in **c), d)** benefits the most from this procedure, because the source become smoother after the filtering and therefore even more supportive of the ESD assumptions. Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices by random generation of 3x3 matrices.

In conclusion, the tests with the smooth sources tell us that the single shift ESD algorithm has the most stable and the best separation performance of the tested algorithms. This is due to the fact that the properties of optical imaging data, that are reflected by the smooth sources, support the assumptions of the ESD algorithm. As soon as the independence assumption of the ICA methods is not fulfilled by the data the performance of the kurtosis optimization and the Infomax algorithm decreases significantly. Consequentially we will concentrate on the further development and improvement of our ESD algorithm for the application on the real data.

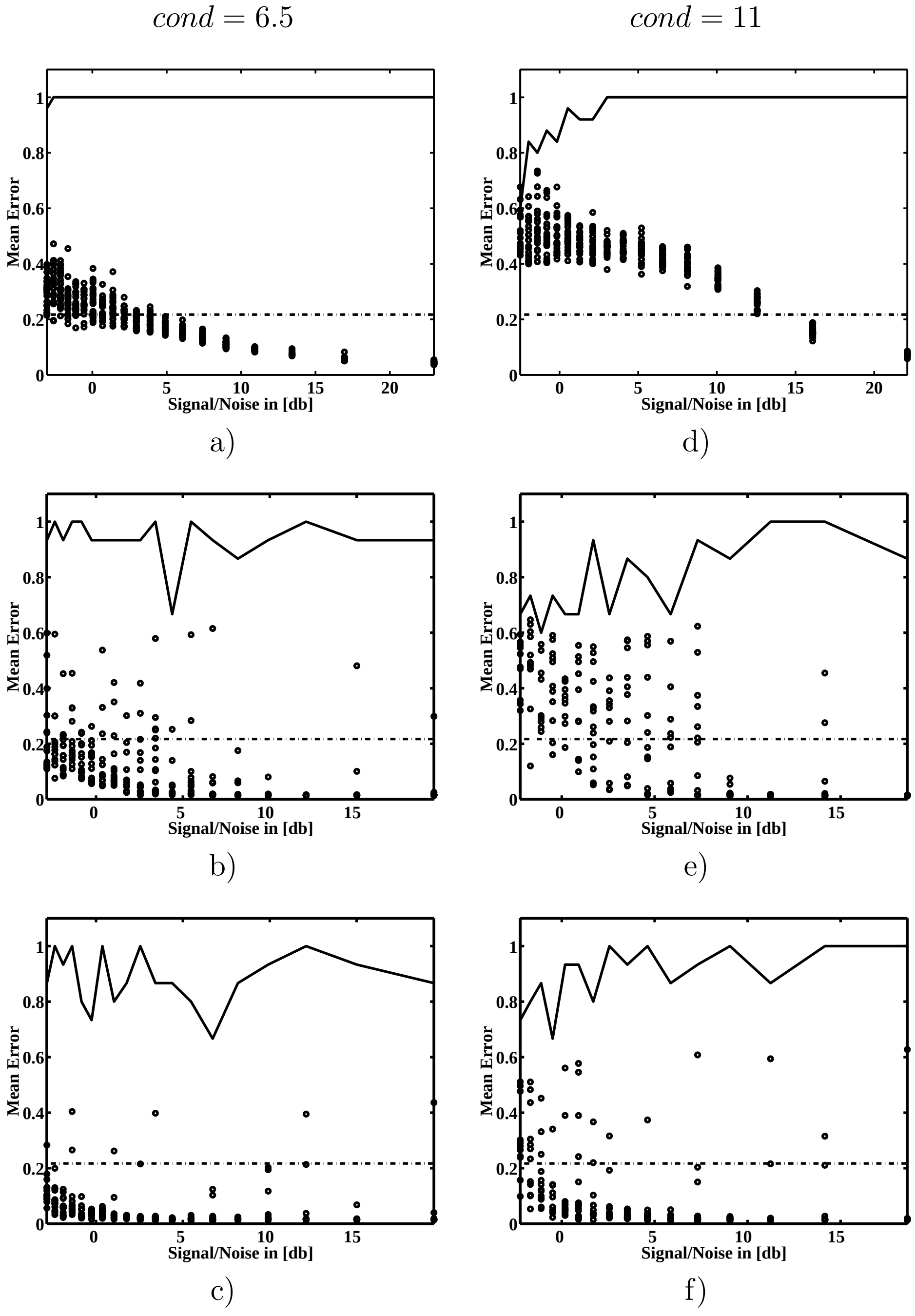
5.5 Comparison of single shift ESD with multi shift ESD

From the test results so far we have seen that the level of the sensor noise is the most disrupting influence on the separation quality of single shift ESD. In section 4.5.2 we introduced the multi shift ESD algorithm. The multiple shifts introduce redundancy by approximate simultaneous diagonalisation of the corresponding cross-correlation matrices. This should make the method less sensitive to sensor noise (Schöner et al., 2000). The cost function for both methods is the same, but single shift ESD performs an analytic calculation and multi shift ESD an optimization procedure by gradient descent.

In figure 5.13 the test results with the smooth sources are displayed. In the left column the mixing matrix with *cond* = 6.5 was used and in the right column the one with *cond* = 11. In the top row the results for the single shift ESD from figure 5.11 are shown again for direct comparison. As we want to investigate the reduced sensitivity of multi shift ESD to sensor noise no filtering was applied to the data

The simulations with the multi shift ESD (figure 5.13 b), e)) revealed a quite ambiguous result. Individual trials showed an excellent separation for both condition numbers, whereas other trials performed even worse than single shift ESD. The big variance in the separation quality and therefore the reduced reliability do not speak for the multi shift algorithm. Also the success rate is reduced compared to the analytic solution. On the other hand the great quality of some of the trials is tempting. Nevertheless for the application on the real data we need blind confidence in the method, as we do not have the possibility to make a quantitative judgment about the result.

In section 4.5.2 we introduced the concept of noise robust sphering from Müller et al. (1999). The bottom row in 5.13 shows the results after separation with using noise robust sphering. The multi shift ESD algorithm is much more stable now and the majority of trials show a high quality in the separation, i.e. very low reconstruction error. In figure 5.14 a direct comparison between single shift ESD



$cond = 6.5$

$cond = 11$

a)

d)

b)

e)

c)

f)

Figure 5.13: Comparison of the separation quality between single shift ESD and multi shift ESD with the **smooth sources**. The **top row** shows the results for single shift ESD without filtering. In the **middle row** the performance of multi shift ESD is displayed. Individual trials have an excellent separation quality whereas others perform badly. The variance in separation quality is big for both matrices. With the application of noise robust sphering the multi shift ESD is stabilized and the majority of trials have a small reconstruction error RE (**bottom row**). Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices by random generation of 3x3 matrices.

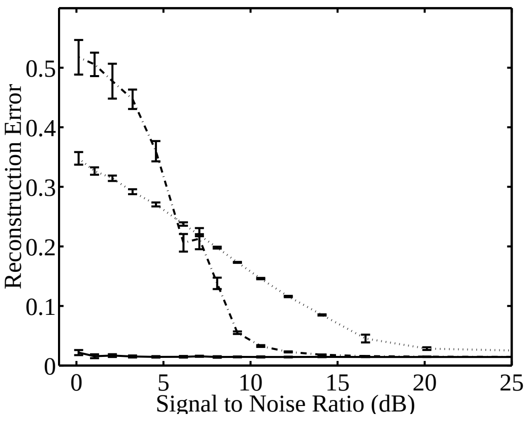


Figure 5.14: Direct comparison of the single shift ESD (dotted line), multi shift ESD (dashed-dotted line) and multi shift ESD with noise robust sphering (solid line) for the separation with the high condition number matrix. The lines display the mean reconstruction error at the individual noise levels and the error bars the variance. The multi shift ESD algorithm with noise robust sphering has a superior separation quality.

(dotted line), multi shift ESD (dashed-dotted line) and multi shift ESD with noise robust sphering (solid line) for the separation with the high condition number matrix is plotted. The lines show the mean of the reconstruction error and the error bars the variance. One can see that the multi shift ESD with noise robust sphering exhibits a superior performance at all noise levels. Therefore from now on when we use multi shift ESD noise robust sphering is applied. In fairness it has to be pointed out, that this type of plot facilitates the performance of the multi shift algorithms as it neglects the percentage of successful separations.

5.6 Comparison of multi shift ESD with regularized ESD

The regularized ESD algorithm uses additional information we have about the mixing process to stabilize the convergence of the gradient descent. We want to point out again that we now deliberately leave the field of parameter free estimations and introduce knowledge about the original sources. If the assumptions are incorrect we force a wrong result.

In order to test if the regularization term improves the separation we created a new3 $\times$ 3 matrix $\mathbf{A}$. We know that in our mixing model the column a_j of $\mathbf{A}$ represents the time course of the source s_j. In figure 5.15 we see the time course we implemented in the matrix for the $\mathbf{A}$ for the three sources.

The first source has a rising and then descending time course (dashed-dotted line), the second source has a continuously rising time course (solid line) and the third source stays constant (dotted line). The matrix has a condition number of $cond = 4.8$.

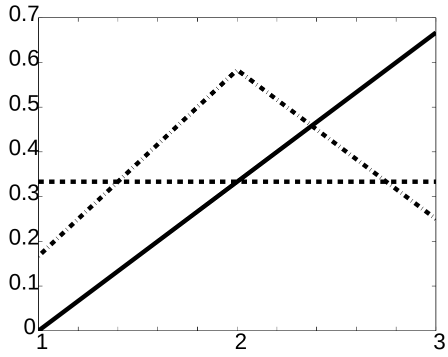


Figure 5.15: The 3×3 mixing matrix **A** for the testing of the regularization terms. The dashed-dotted line is the time course for the first source, the solid line for the second source and the dotted line for the third source.

5.6.1 Regularization of all Sources

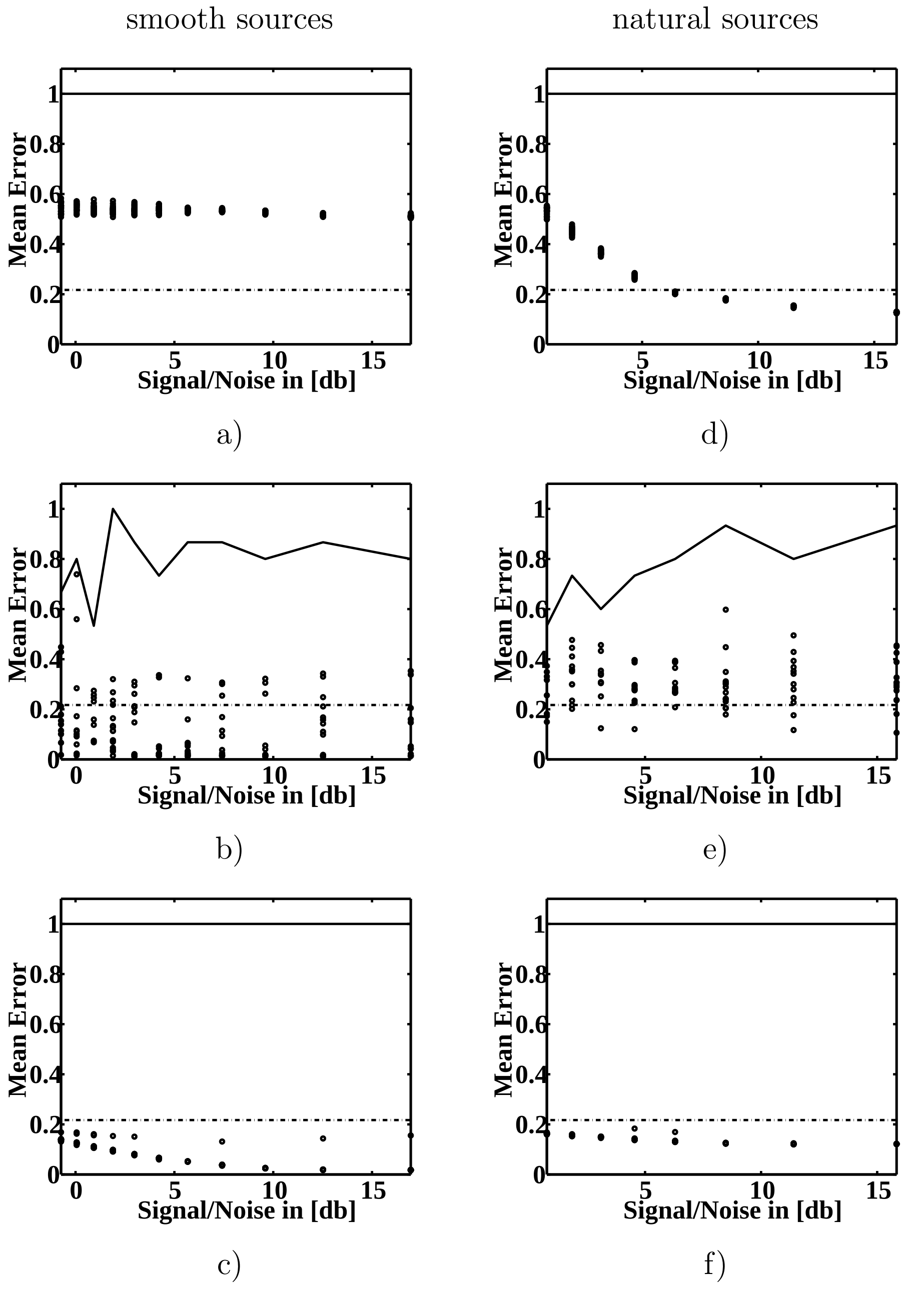
For the first set of tests we are using a regularization term that introduces knowledge for all three sources. In figure 5.16 we see the results of the simulations with the smooth sources in the left column and the natural sources in the right column. Even if we do not directly consider the single shift ESD in this section, we have calculated the separation results with the new matrix for the demonstration of the increased difficulty of the separation task (figure 5.16 a), d)). The middle row b), e) shows the separation results for the multi shift ESD algorithm. As in section 5.5 the quality of individual separations can be far better ($RE \approx 0.0$) than with the single shift ESD, but due to the convergence of the gradient descent into the wrong minima the reliability is poor (big variance of results on individual noise levels).

The separation results of multi shift ESD with regularization of all three sources eliminates this short coming (c), f)). As well for the smooth sources as the natural sources the rate of successful separation is 100% across the whole noise image. Also the mean reconstruction error stays below $RE = 0.2$ for all trials and has a small variance. This method is therefore more reliable and successful than the single shift ESD and the multi shift ESD on our artificial datasets.

5.6.2 Regularization of one Source

In a real optical imaging experiment it is very likely that we have only knowledge about the time course of a part of the original source, i.e. those sources that are correlated with the stimulus onset. To simulate this situation we calculated and scored the regularization term for only the fist column of the assumed **A** during the optimization. In figure 5.17 a), c) we can see that the number of successful separations is reduced again and the mean error and the variance is bigger than in the case with prior knowledge on all the sources.

Part of this seeming decrease in the separation performance is that the source with the prior knowledge in the regularization term is well separated, whereas in some trials the other sources are still mixed. In this case the calculation of


smooth sources

natural sources

a)

d)

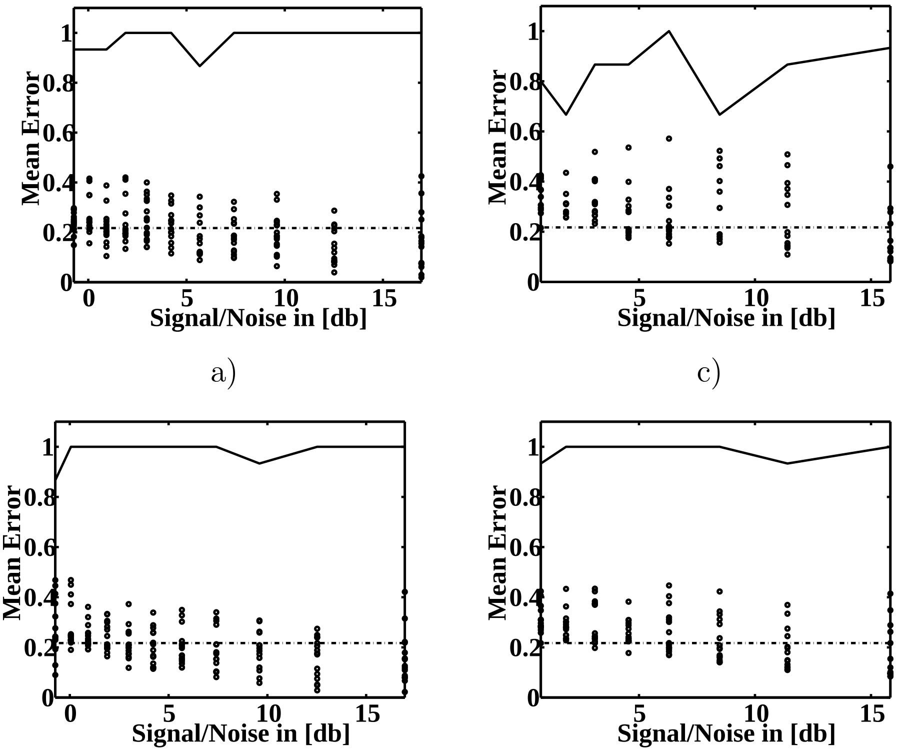
b)

e)

c)

f)

Figure 5.16: Mean reconstruction error as a function of the signal to noise ratio in *dB* (15 trials per noise level). The **left column** shows the results of the separation for the **smooth sources** and the **right column** for the **natural sources** after mixing with the 3x3 matrix **A** . The **first row** shows the results for ESD with only two shifts (here [0,0;5,5]). The **second row** shows the results the multi shift ESD. In the **third row** the results with multi shift ESD and the regularization term on the time course of all sources is displayed. Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices by random generation of 3x3 matrices.



a)

c)

b)

d)

Figure 5.17: Mean reconstruction error as a function of the signal to noise ratio in *dB* (15 trials per noise level). The **left column** shows the results of the separation for the **smooth sources** and the **right column** for the **natural sources**. In the **first row** the results with multi shift ESD and the regularization term on the time course of only the first sources is displayed. A better and more stable convergence of the gradient descent can be achieve by initializing the estimate of $\mathbf{W}$ with the inverse of a matrix, that has the assumed time course in the first column and random noise in the others (bottom row). Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices by random generation of 3x3 matrices.

the mean reconstruction error as introduced in equation 5.1 is not appropriate anymore. Figure 5.18 shows such a typical result a 0 *dB*. We found that we can stabilize the separation performance by initializing $\mathbf{W}$ with the inverse of a matrix, that has the assumed time course in the first column and random noise in the others (see figure 5.17 b,d).

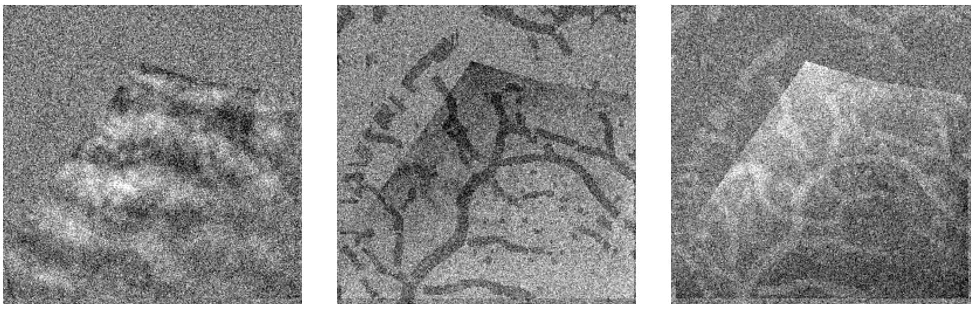


Figure 5.18: The three separated natural sources after application of multi shift ESD and the regularization term on only the first time course of the first source at the high signal to noise ratio of 0 dB. The source of interest is well separated, whereas the other two sources are still mixed.

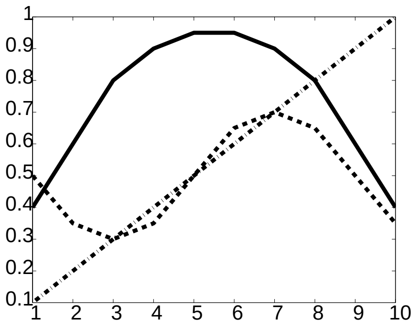
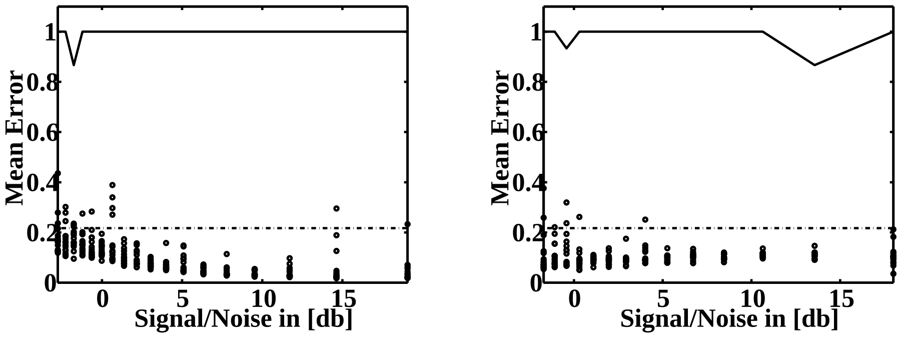


Figure 5.19: The 3 × 10 mixing matrix **A** for the testing of the regularization terms. The dashed-dotted line is the time course for the first source, the solid line for the second source and the dotted line for the third source.

5.6.3 Results using a 3×10 mixing matrix A

In the optical imaging experiment one normally measures a higher number of video frames and therefore mixtures than there are underlying sources. To simulate this property we now use a 3x10 mixing matrix **A** (see figure 5.19). The dashed-dotted line is the time course for the first source, the solid line for the second source and the dotted line for the third source.

The resulting ten frames after the mixing contain the mixtures of the original sources at ten succeeding points in time. We now have to estimate a 10x10 demixing matrix **W**. Applying the regularization term to the first three of the ten estimated sources we can force the underlying original sources into the first three frames. This way we can still use the calculation of the mean reconstruction error as given in equation 5.1. Figure 5.20 shows the separation result for the smooth sources (a) and the natural sources (b) up to signal to noise ratios even lower than 0 dB. It shows that the minimization of the cost function in equation 4.73 converges well independent of the number of mixtures. The percentage of successful separations is nearly 100% over the whole noise range. Also the variance in the mean reconstruction error is low at all noise levels.



a)

b)

Figure 5.20: Mean reconstruction error as a function of the signal to noise ratio in dB (15 trials per noise level). For both plots the original sources where mixed with a 3x10 matrix. For the demixing process the multi shift ESD algorithm with the regularization term for the first three of the ten mixtures was used. **a)** shows the result for the smooth sources and **b)** the result for the natural sources. Circles: individual trials. Solid line: percentage of successful trials. Dashed line: percentage of permutation matrices by random generation of 3x3 matrices.

We have to remember that with this spatio-temporal hybrid method knowledge about the time course of the original sources is introduced. When we have non symmetric matrices **A** and we apply a prior for only one source the other sources can be anywhere in the resulting stack and therefore are misjudged by the error measure RE.

Chapter 6

Separation of Optical Imaging Data

After the successful testing of the individual algorithms on the artificial datasets we will now study the performance on the datasets we achieved from optical imaging of intrinsic signals. For most of the comparisons we have a detailed look at two selected experiments that have a stimulus regime where the associated mapping signal is well known. The first contains images from an ocular dominance (od) experiment of V1 from macaque monkey. The second set contains an orientation preference (op) experiment of V1 from cat. The second dataset was given to us for the collaboration in Schießl et al. (1998) and Stetter et al. (2000).

Before we can start with the source separation we have to apply the basic preprocessing procedures describe in the section 4.1. For all the later results first frame analysis and binning of the frames in time was used. In figure 4.4 the frames after the preprocessing for the ocular dominance experiment are shown. The two single condition stacks and the difference stack of the orientation preference experiment for the orientations $0° - 90°$ are shown in figure 6.1. The stacks for the $45° - 135°$ orientation look similar. Like in the ocular dominance stack we can not see the mapping signal in the single condition stacks emerge over time, but in the difference stack the mapping signal becomes visible after the stimulus onset.

6.1 Differential Imaging and Filtering

In the section 4.2 we described several methods for the analysis of the data with the concept of differential imaging. For the standard differential imaging the frames of the difference stack are summed up. Due to the different times courses of the individual intrinsic signals it is sometimes advantageous to only sum up the early frames after the stimulus onset as the vessel and blood flow artifacts are slow compared to the cell swelling. In figure 6.2 the result of this selective summation is shown in a). Another method for differential imaging suggested the division with a blank image, that was recorded before the stimulus onset. The result of

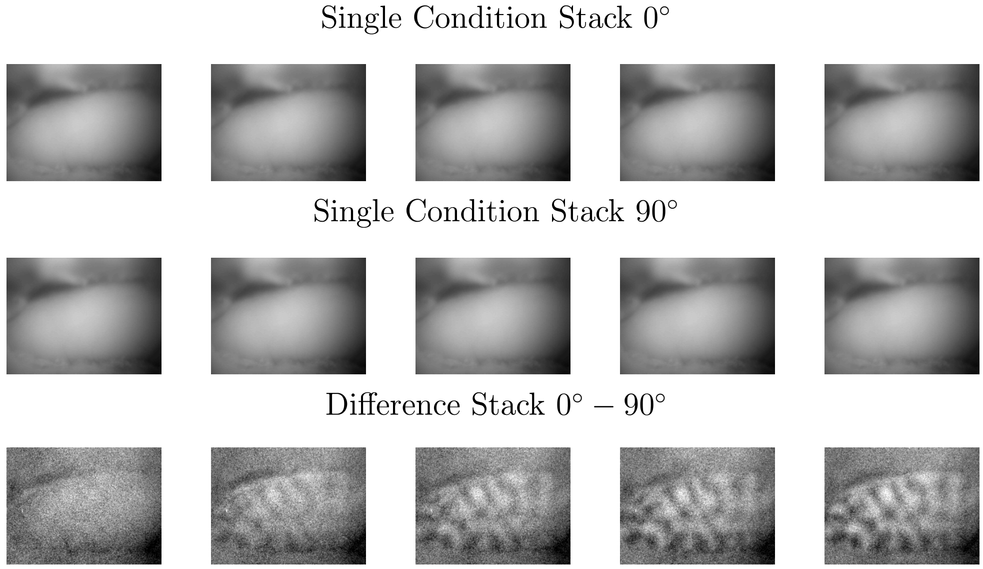

Single Condition Stack 0°
Single Condition Stack 90°
Difference Stack 0° − 90°

Figure 6.1: Single condition stacks and difference stack for the orthogonal stimulus condition 0° − 90° of a orientation preference experiment. The corresponding 45° − 135° datasets look similar.

this procedure is shown in c) for the ocular dominance data. In both images the patterns of the mapping signal is visible but the quality has not enhanced much compared to the single frames of the difference stack. The signal still contains strong artifacts. The large vessel in the top third of image c) might always be a problem, as it is big enough to absorb the underlying mapping signal, but there are still small vessel artifacts left in the area of the mapping signal.

To enhance the signal quality lowpass filtering can be applied if necessary. The right column in figure 6.2 shows the images after a lowpass filter was applied. The contrast is enhanced as the noise is filtered out but the artifacts are still present. The filtering of data is often used when orientation preference maps are calculated. In figure 6.3 we see the two difference images for a) 0° − 90°, and b) 45° − 135° from which the orientation preference map in c) is calculated. The patchy appearance of the orientation preference map is somewhat messed by the high frequency noise. After the lowpass filtering d) the organisation of the orientation preference map becomes visible. But as we pointed out in section 4.2.3 filtering must be handled with great caution.

For instant feedback of the stimulus response during an experiment we can display the a set of single condition and difference images that are recorded online during the trials. Figure 6.4 shows an example of the display. In the top row the first frame analysed images are shown and below the mean gray value of the region of interest during the recording of all trials is plotted. The big dips in the mean time course represent the change in absorption due to the stimulus presentation. Therefore we know that the cortex responds to the stimulation. The differential

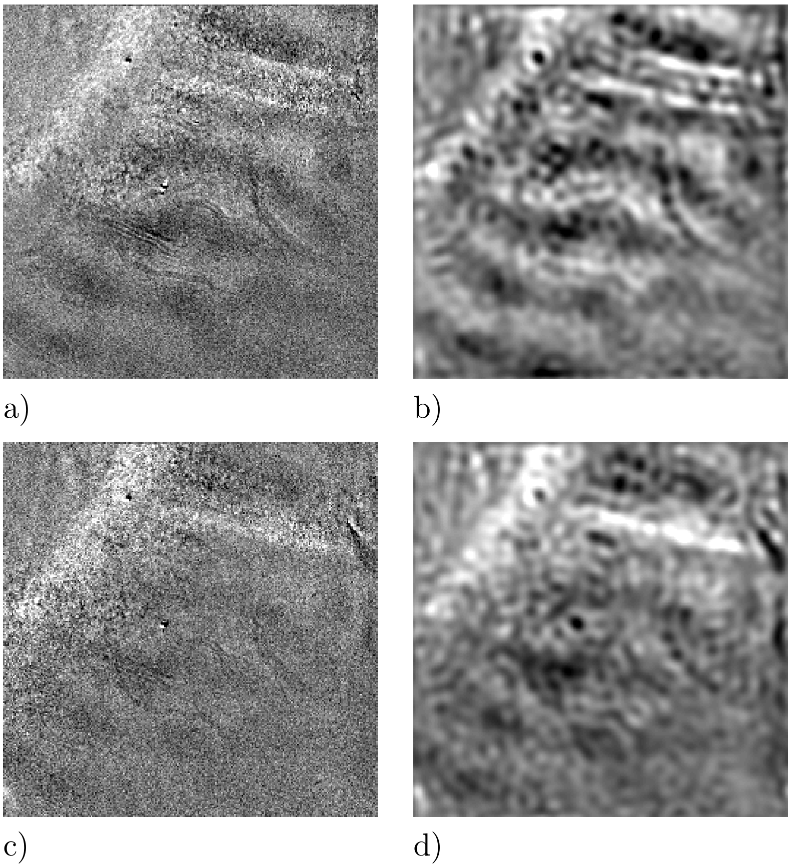


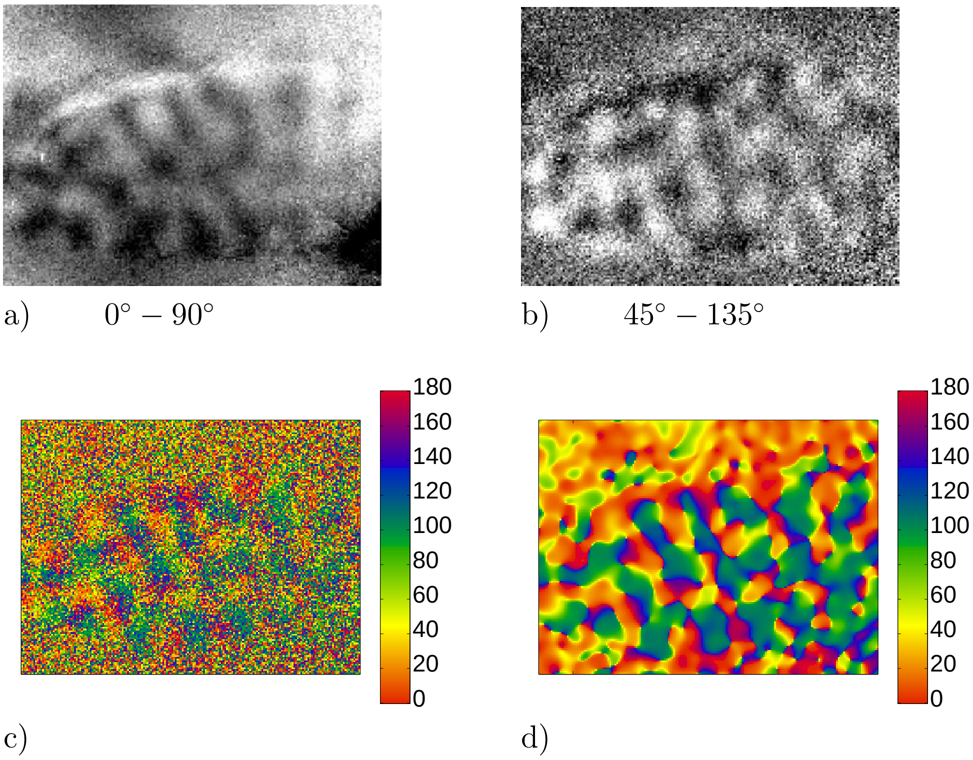
Figure 6.2: Analysis results from the standard analysis of the ocular dominance experiment. In **a)** only the frames in the first second of the stimulus presentation are summed up to reduce the influence of vessel artifacts. In **c)** the result after the summation of all frames from the difference stack and division of the blank image is show. **b), d)** show the results after lowpass filtering of the difference stack $(25cyl/256pixels)$.

image in the top right only sums up the frames during the first second after the stimulus onset to keep the influence from the slower vascular response small.

6.2 Principal Component Analysis

We normally use PCA in the context of blind source separation only for whitening of the data and if necessary to perform dimension reduction. Under certain condition PCA can deliver the aspired separation of the sources after the frames of the trials have been specially rearranged (Stetter et al., 2000).

Here we have a look at one example of the ocular dominance data stack after sphering. In figure 6.5 we see the first ten principal components (first and third column) with the largest eigenvalues from the data stack with 120 frames. To the



a) 0° – 90°

b) 45° – 135°

c)

d)

Figure 6.3: Orientation preference maps from the differential image of the presentation of a 0° and 90° square wave grating and the presentation of a 45° and 135° square wave grating. In **c**) we see the result without lowpass filtering. The patchy structure of the map is not clearly visible due to noise. After the application of the lowpass filter **d**) the orientation preference map is visible.

right of each image we can see the time course of the principal component. This time course is calculated by back projection of the component onto the difference stack by matrix multiplication. This procedure is the same for the BSS and ICA algorithms. This time course is very important for the interpretation of the results as we do not know exactly what the spatial distribution of the mapping signal should look like, but we know the timing of the stimulus presentation.

Principal component 120 shows mainly the big vessel artifact and the corresponding time course clearly correlates with the heart beat of the animal (18 dips in 120 frames at a frame rate of $15fps \Rightarrow 135$ beats per minute). The principal component 118 has the best representation of the mapping signal after PCA. In the time course thought we can see that the source is not perfectly separated yet, as the response to the stimulus (gradual rise of the DC level) is strongly contaminated by a respiration artifact (6 dips in 120 frames at $15fps \Rightarrow 45$ breaths per minute). Component 112 for example shows noise in the image and the time course and could be neglected in the process of dimension reduction. The interpretations from the time course become even more important if we use a stimulus regime

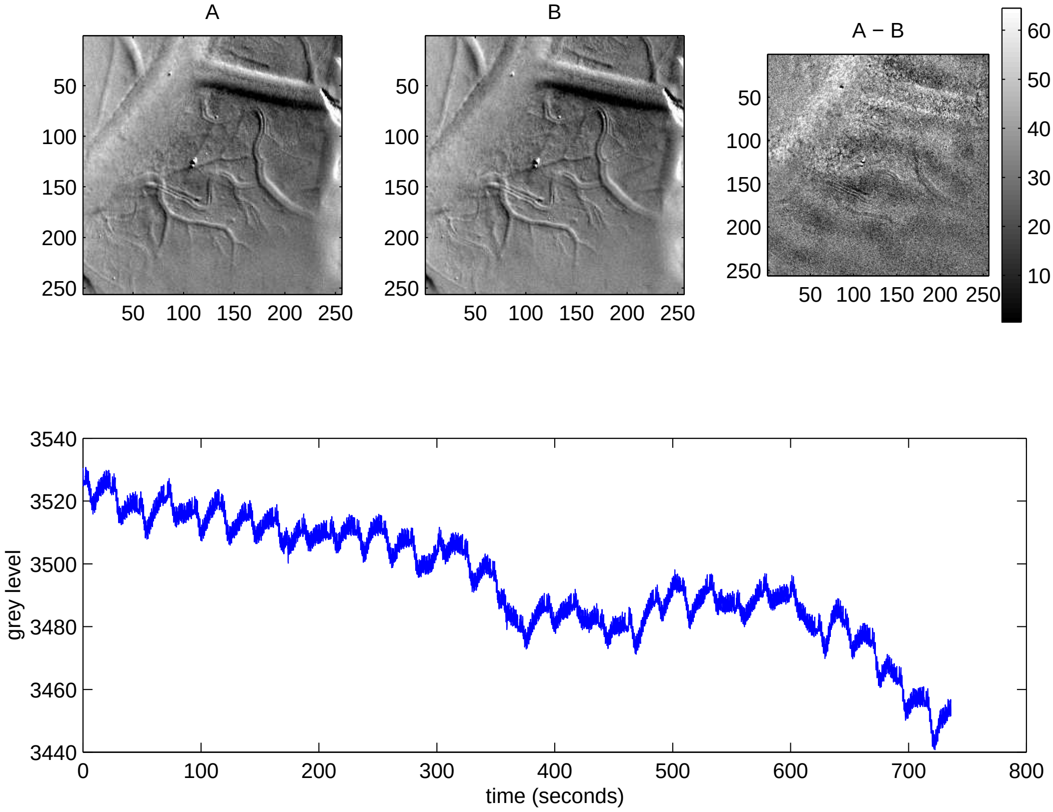


Figure 6.4: Representation of the online analysis during the recording of the ocular dominance experiment. The images **A** and **B** are calculated by subtraction of frames from the stimulus onset until one second after from image frames before the stimulus. This is done for all trials and all stimulus conditions. The results from the single trials are summed up then. Image **A** shows the single condition for the left eye and **B** for the right eye. **A** − **B** is the difference image calculated from **A** and **B**. The plot shows the mean gray value for the duration of the whole experiment (32 trials). The big dips represent the response to the stimulus presentation. The slow drift in the time series is caused by a slow change in biological parameters over the 12 minutes of the trial like a change in the oxygenation state or paralysis.

where we do not know how the spatial distribution of the cortical response looks like.

Figure 6.5 demonstrates that PCA by itself was not able to separate the sources for this experiment. Therefore we will now have a look at methods that take into account higher order statistics.

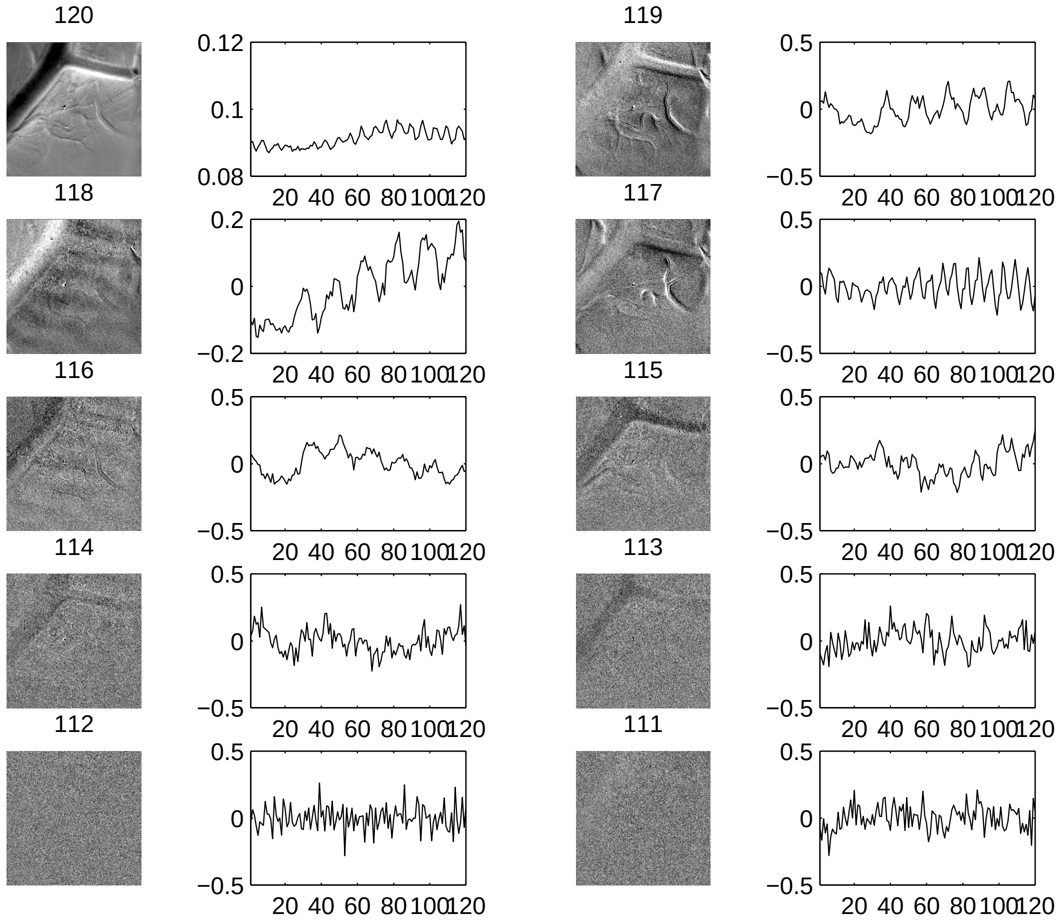


Figure 6.5: The first ten principal components and their time courses. The principal components are sorted in descending order according to the largeness of their eigenvalue. The time courses are calculated by back projection of the individual component onto the data stack they were calculated from. The run of the curve gives important information about the signal content of the image. The same procedure of back projection is applied to the independent components in the later analysis.

6.3 Performance of the ICA Algorithms

After basic preprocessing and PCA the data stacks were processed with the Infomax algorithm and the kurtosis optimisation algorithm. The data stacks before ICA contain the principal components sorted by their eigenvalues. The result of the separation with ICA have the individual independent components as the single frames. Due to the ambiguity of ICA there is no specific order in the components. The individual frames from the results of ICA were then back projected on the original data stack to get the time course of the independent components. The greater the number of frames available in the original data for back projection the more time points we can display. In order to get any reasonable results from the optical images after the analysis with ICA masks were applied to the stacks after PCA to cover the regions in the image that are outside the exposed cortex or showed a very large vessel. This reduces the content of noise in the image. We expect to have the best signal to noise ratio after we subtracted the orthogonal stimulus conditions, because the biological background activity common to both single condition stacks is removed. Therefore we start with the analysis of the difference stacks.

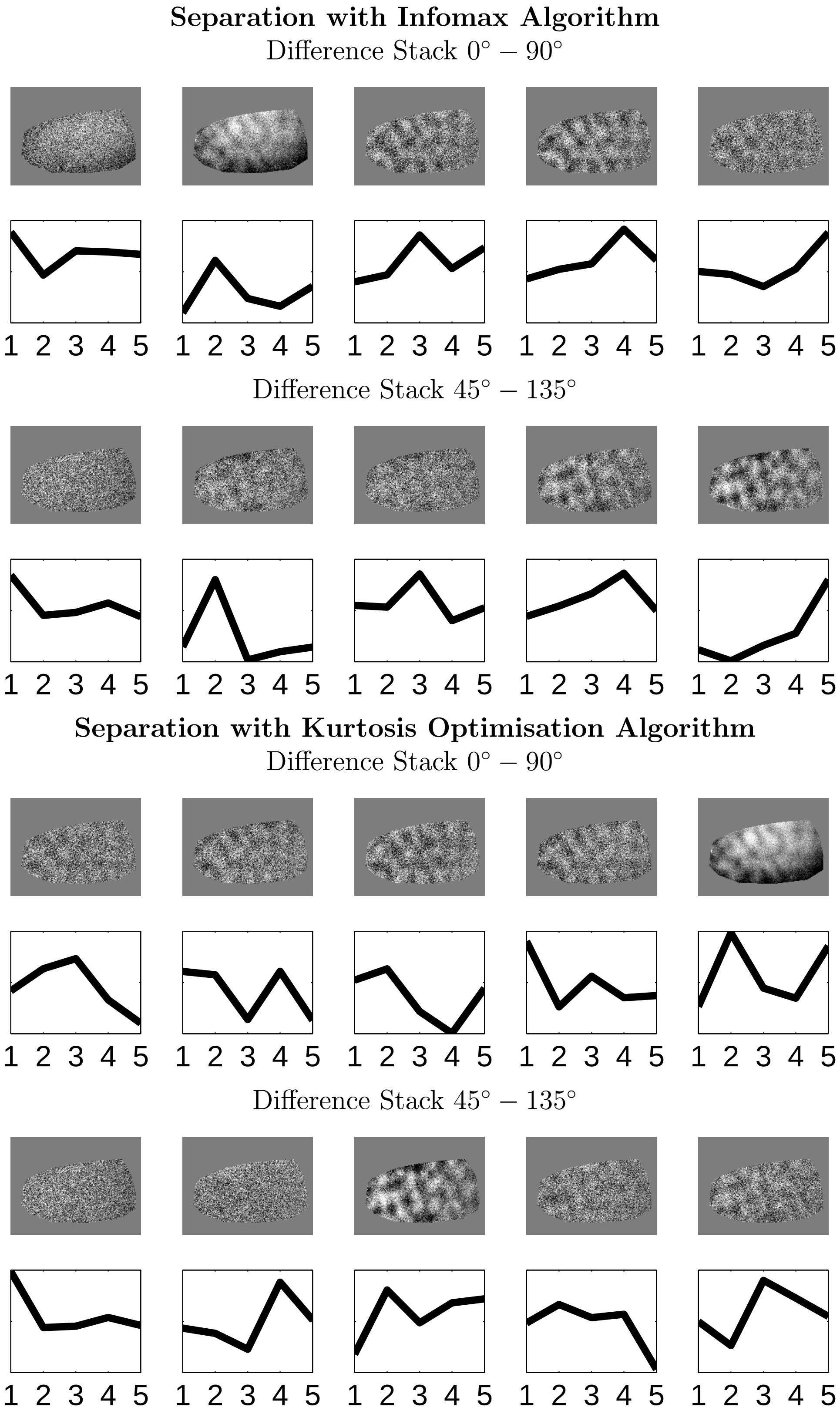
6.3.1 Separation of Difference Stacks

In the dataset from the orientation preference experiment we have two difference stacks. The first is from the subtraction of the 0° stimulus minus the 90° stimulus and the second from the 45° stimulus minus the 135° stimulus. In the stacks we have only five frames from the recording and therefore only a rough plot of the time course. The interval between frames is $600ms$ and the stimulus was presented during the whole time of the trial.

In figure 6.6 we see the separation results of the two ICA algorithms on the two difference stacks of the orientation preference experiment and their back projections. For a rough estimate how the patterns we expect to see look like have a view at the images a) and b) in figure 6.3 again. Neither the Infomax- nor the kurtosis optimisation method show a clear separation of the mapping signal from the biological- and sensor noise. In only a few frames like the last frame of the kurtosis optimisation of the $0° - 90°$ difference stack one can vaguely anticipate the expected pattern. The corresponding time courses of the independent components do not reflect the temporal gradient of the mapping signal that should start to rise or decline from beginning to the end as the stimulus was presented from the first frame to the last.

So let us have a look at the separation result with the difference stack from the ocular dominance experiment. The origin data stack before binning contains 120 frames ($8s$ at $15fps$) and therefore the time course calculated from the stack is finer now. The bottom bar in each plot represents the duration of the stimulus. In figure 6.7 the independent components from the Infomax algorithm are printed in the first column and the time courses in the second. The stripe like pattern we know from figure 6.2 a)-d) is not projected into one component and therefore not separated. The best result is the last frame of column one, but again the time course does not reflect the stimulus presentation.

From the results on the artificial datasets we know that lowpass filtering can enhance the performance of the ICA algorithms. Therefore we applied a lowpass filter ($25cyl/256pixels$) to the data stack before ICA. In the right half of figure 6.7 the separations of the Infomax algorithm on the filtered data is displayed. The quality of the separation has improved and the fourth and sixth image from the top show the ocular dominance pattern. The time course of the fourth image



Separation with Infomax Algorithm

Difference Stack 0° – 90°

Difference Stack 45° – 135°

Separation with Kurtosis Optimisation Algorithm

Difference Stack 0° – 90°

Difference Stack 45° – 135°

Figure 6.6: Separation results of the tested ICA algorithms and their back projections on the two difference stacks from the orientation preference experiment. The stimulus was presented during the whole trial ($3s$).

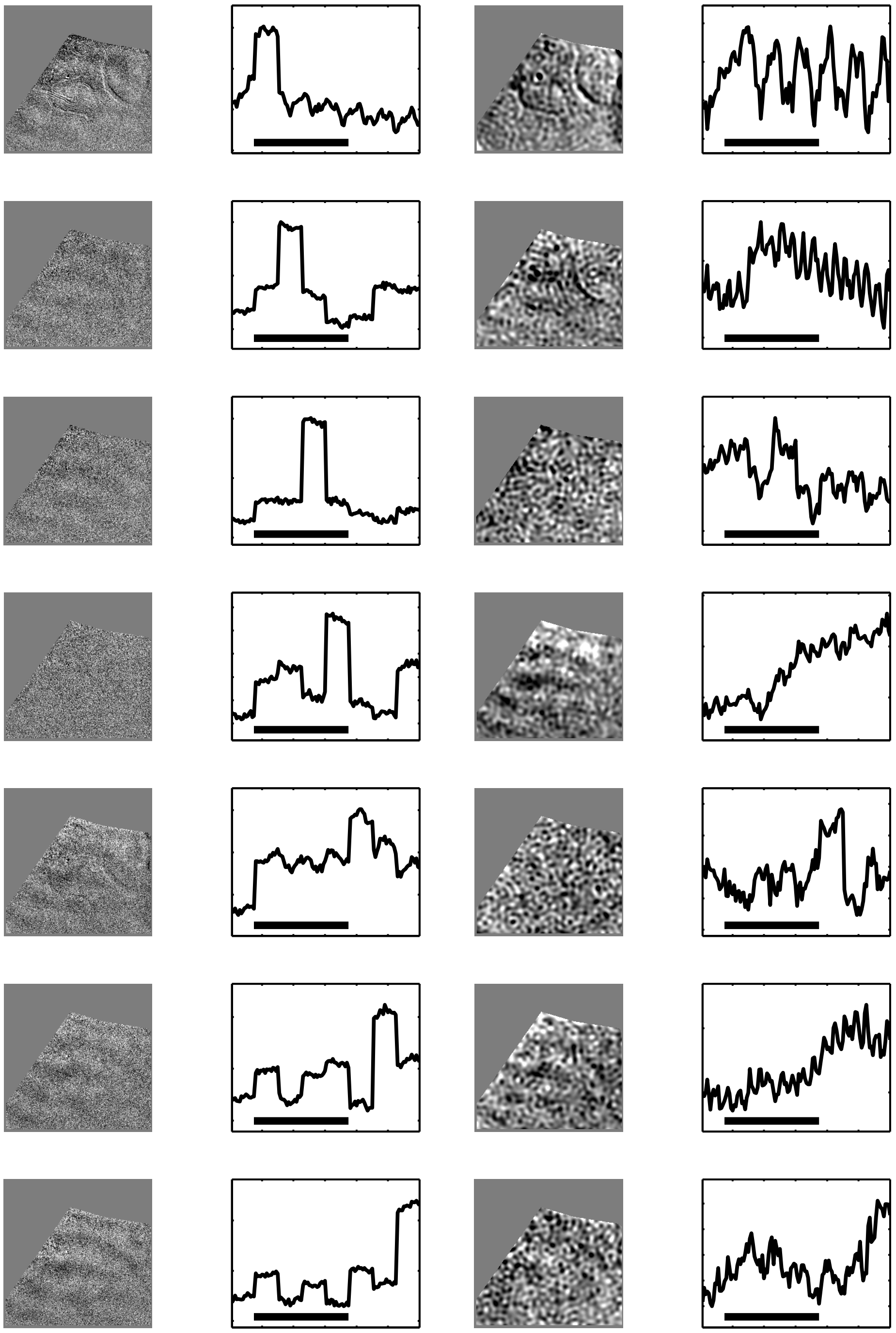


Figure 6.7: Separation results of the **Infomax algorithm** and the back projections on the difference stack from the ocular dominance experiment (**left half**). In the **right half** the result after lowpass filtering is shown ($25cyl/256pixels$). (Bottom bar: Stimulus duration $4s$).

shows a positive trend after the stimulus onset. But again the separation is not perfect as the mapping signal is still present in several independent components. The time course of the component in the first image from the top shows the respiration artifact again.

The same analysis was done now with the kurtosis optimisation algorithm. The arrangement for the results from this in figure 6.8 is the same as in figure 6.7. The independent components from the unfiltered data in the left column again do not represent the stripe like ocular dominance pattern clearly. In the third to the fifth source it is slightly visible but the time courses are arbitrary. After the filtering kurtosis optimisation is capable of separating the mapping signal well enough to make it visible. Again it is present in two sources (second and third). The correlated time courses show a clear stimulus related change but are still strongly contaminated by artifacts. The top source contain the change in absorption from respiration and the lower sources contain mainly noise.

6.3.2 Separation of Single Condition Stacks

The more difficult task is the analysis of the single condition stacks because the signal to noise ratio is worse than in the difference stacks. The single condition stacks are not processed with a cocktail blank as this would introduce information of the orthogonal stimulus condition.

In figure 6.9 we see the separation results of the two ICA algorithms on the 0° stimulation single condition stack of the orientation preference experiment as one example of the results on unfiltered data. For both algorithms (Infomax top rows, kurtosis optimisation bottom rows) the independent components do not show the mapping signal. If anywhere one can anticipate the pattern in the second component of the top row and the fourth component of the third row. In both separations we can see that the algorithms are able to separate a global artifact (first row, last image; third row , third image) that is triggered by the stimulus onset (see time course) and the reaches a constant level from the second time frame on. This is a good example for a global signal that appears with the stimulus presentation but is not stimulus specific, i.e. it is also present in the separations of the single condition stacks of the other orientations (not shown). So this artifact would be removed by the subtraction used in the calculation of the difference stack.

In figure 6.10 a example of the separation with ICA on the filtered single condition stack of the right eye stimulation of the ocular dominance experiment is shown. The independent components and their time courses from the Infomax algorithm are shown in the left half and the results from kurtosis optimisation in the right half. Here we have a better separation due to the lowpass filter with some interesting features.

Both algorithms separate the vessel artifacts into two independent components with different time courses (left: images 1 and 2; right: images 2 and 3).Both vessel components show a strong response to the stimulus onset. The stripe like

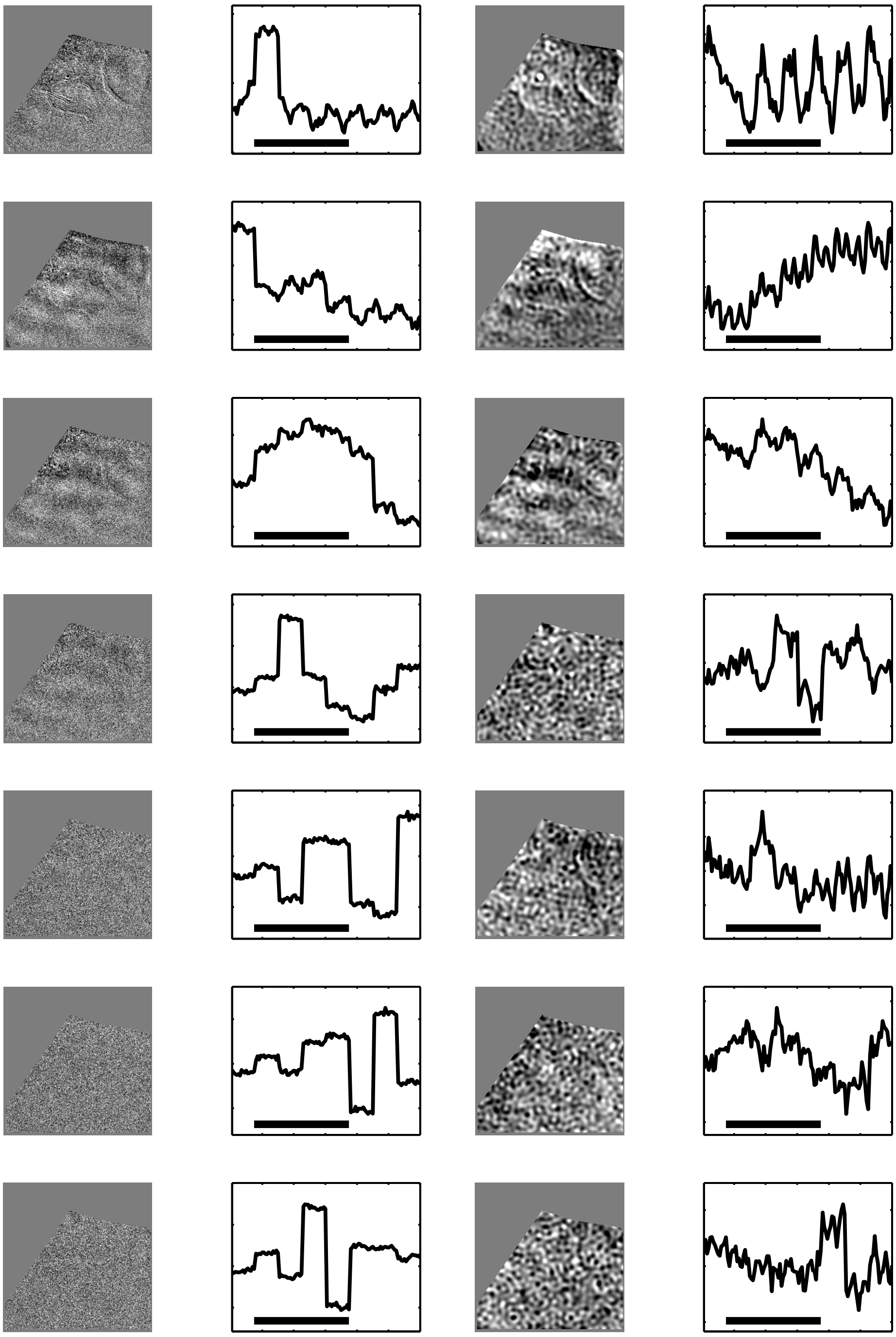
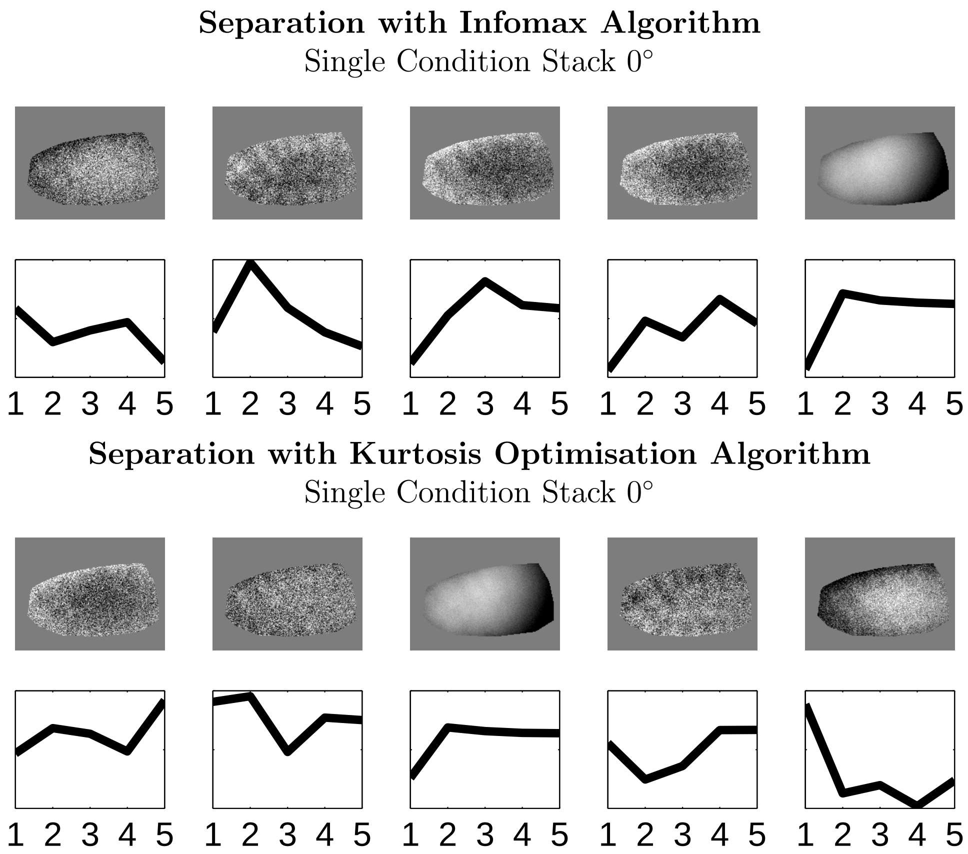


Figure 6.8: Separation results of the **kurtosis optimisation algorithm** and the back projections on the difference stack from the ocular dominance experiment (**left half**). In the **right half** the result after lowpass filtering is shown($25cyl/256pixels$). (Bottom bar: Stimulus duration $4s$).



Separation with Infomax Algorithm
Single Condition Stack 0°

Separation with Kurtosis Optimisation Algorithm
Single Condition Stack 0°

Figure 6.9: Separation results of the tested ICA algorithms and their back projections on the single condition stack with the 0° orientation from the orientation preference experiment. The results of the single condition stacks with the other three orientations show similar results. The stimulus was presented during the whole trial ($3s$).

patterns of the ocular dominance stripes is slightly visible in the third frame on the left and the first frame on the right. The time course though is not exactly what we would expect, as it should continue to rise after the end of the stimulus presentation. Interestingly all the components show some response to the stimulus and none reflects just arbitrary noise what means that the signal is not separated from the noise.

6.4 Performance of single shift ESD

The single shift ESD algorithm makes no assumptions about the independence of the sources, only about their correlation. Nevertheless it is not restricted to orthogonal solutions like PCA. From the results of the simulations with the smooth artificial data we expect this algorithm to handle the optical images better than

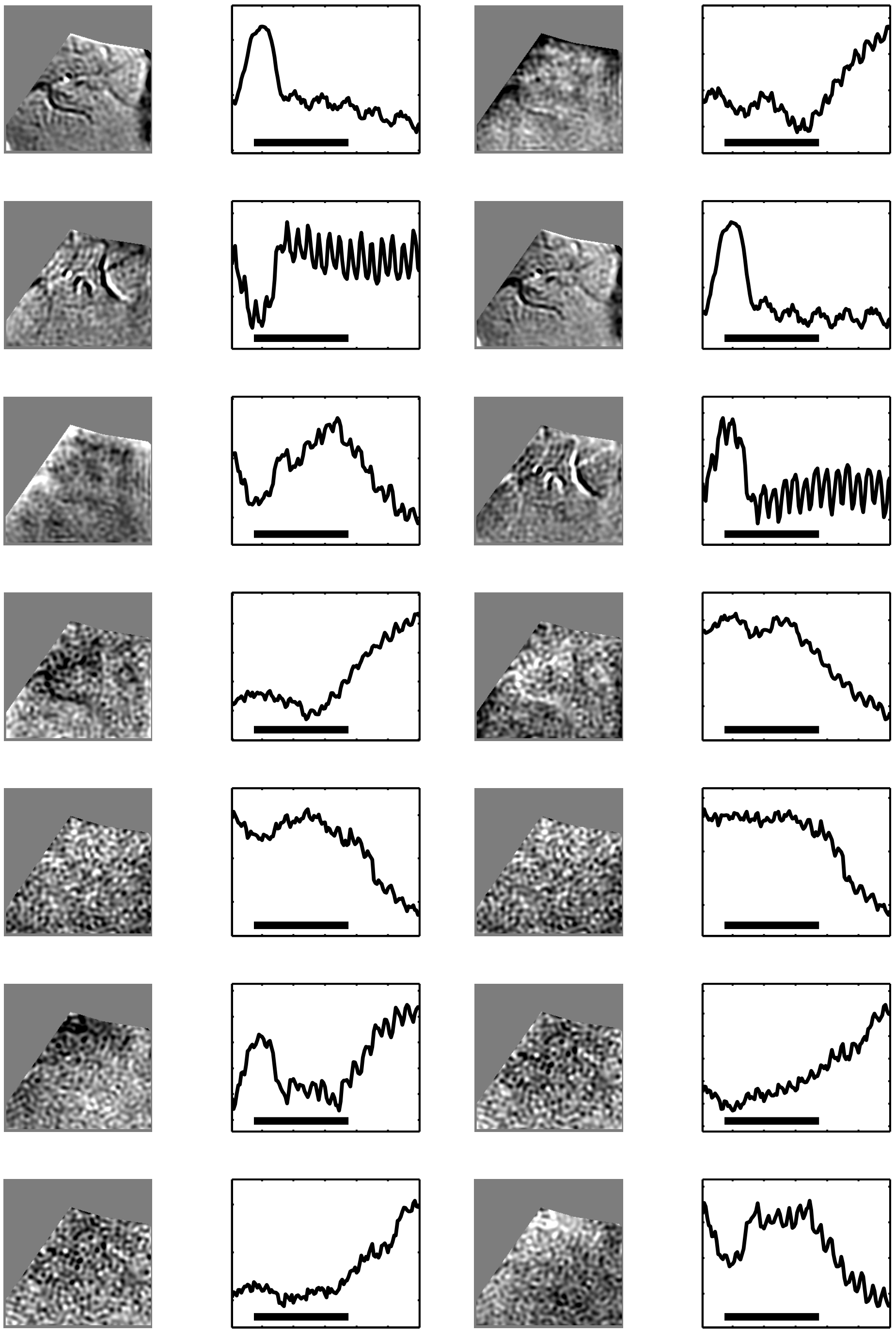


Figure 6.10: Separation results of the Infomax algorithm (**left half**) and the kurtosis optimisation algorithm (**right half**) and the back projections on the single condition stack with the stimulation of the right eye from the ocular dominance experiment after lowpass filtering ($25 cyl/256 pixels$). The results for the left eye are similar. (Bottom bar: Stimulus duration $4s$).

the ICA algorithms. We will start with the easier task of separating the difference stacks and then look at the separation of single condition stacks.

6.4.1 Separation of Difference Stacks

In order to be able to compare the performance of single shift ESD with the ICA algorithms the separation on the orientation preference experiment with the mask was calculated first. Not all the results will be shown but the examples are representative.

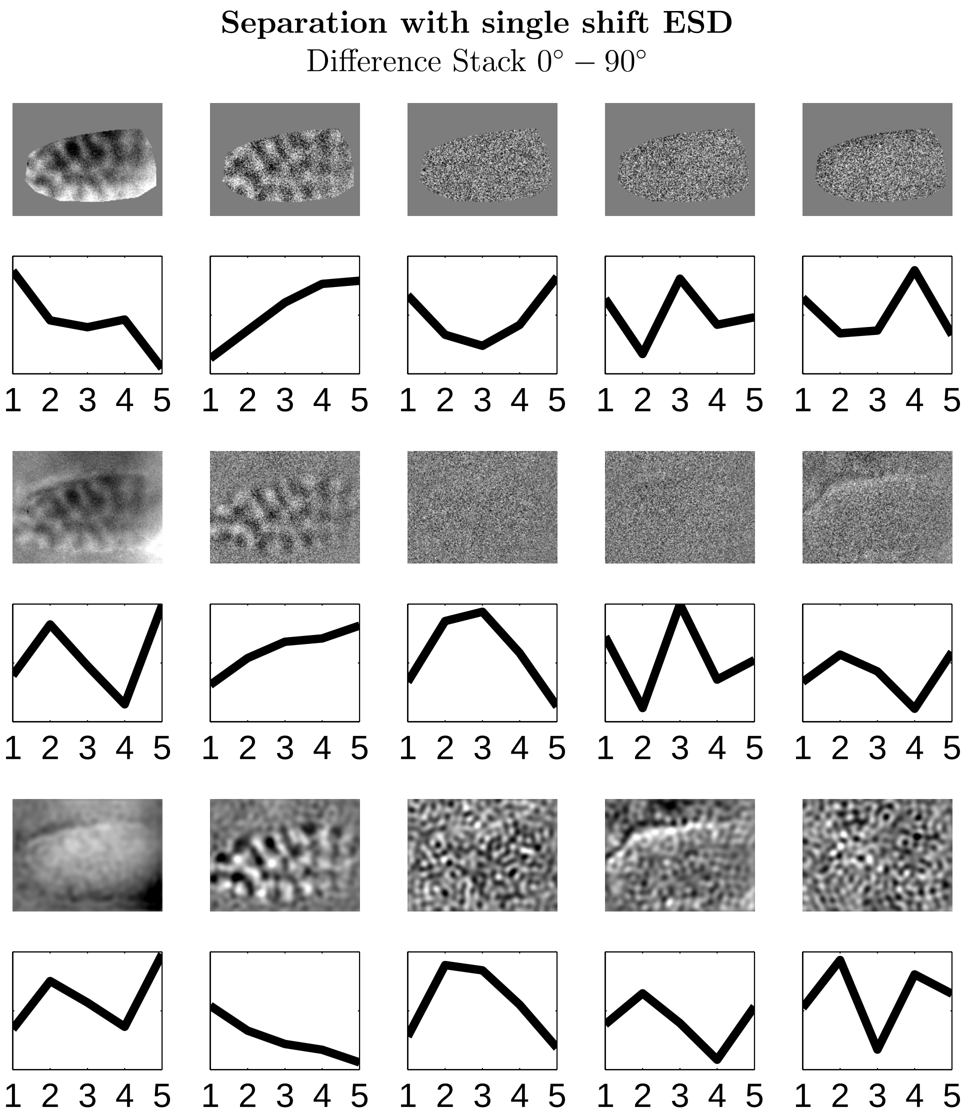
In the first two rows of figure 6.11 the estimates from the $0° - 90°$ difference stack of the orientation preference experiment with the mask are shown. The mapping signal is now clearly visible with better contrast than in the results from ICA. Nevertheless it is present in the first and second source. The other sources contain noise. The time course of the second component is close-by the one we would expect from the experiment (see section 3.2). Therefore the separation has a better quality than with ICA but is not yet optimal.

To see more of the underlying sources the mask was omitted now and the single shift ESD was applied again. The middle two rows show that the calculated components contain the mapping signal. We can see that a ridge from the back folded dura causes a motion artifact that is partially separated (last component) but also present in the first component together with some global signal and parts of the mapping signal. In order to enhance the separation quality the lowpass filter ($25cyl/256pixels$) was applied to the difference stack and the separation was repeated. In the bottom two rows the mapping signal (second image) is separated from the motion artifact of the dura (fourth image) and the global background variation (first image). The time course of the mapping signal shows the expected decline. The global signal in the first image is underlying in the first images of the upper separations just with opposite sign.

For the separation of the ocular dominance data we started with the unmasked and unfiltered difference stack. The results in the first two columns of figure 6.12 show that the ocular dominance stripes become visible in the first image together with the big vessel. The time course shows the predicted delayed response to the stimulus onset with small modulations from the vessels artifact. The lower four source estimates contain noise. It seems that the vessel is big enough to bury the mapping signal, so we applied the mask again.

The result in the middle two columns has a better contrast now as the gray values are only distributed across the ocular dominance pattern. The influence of the vessel artifact on the time course is reduced. The artifacts from the small vessels are projected into the components two to four.

We can further enhance the separation quality by applying the lowpass filter. Now the first component in the right two columns of figure 6.12 shows the ocular dominance stripes with good contrast and the appropriate time course. The temporal trend of the second image reflects the respiration artifact. The lower images



Separation with single shift ESD
Difference Stack 0° − 90°

Figure 6.11: Separation results of the single shift ESD algorithm and their back projections on the 0° − 90° difference stack from the orientation preference experiment. The **top two rows** show the components from the masked stack. The **middle two rows** display the unmasked components and the **bottom two rows** were calculated from the lowpass filtered stack ($25cyl/256frames$). The stimulus was presented during the whole trial ($3s$).

mainly contain noise.

6.4.2 Separation of Single Condition Stacks

The success of the separation with the single shift ESD on the single condition stacks followed the trend we had in the separation of the difference stack. In the

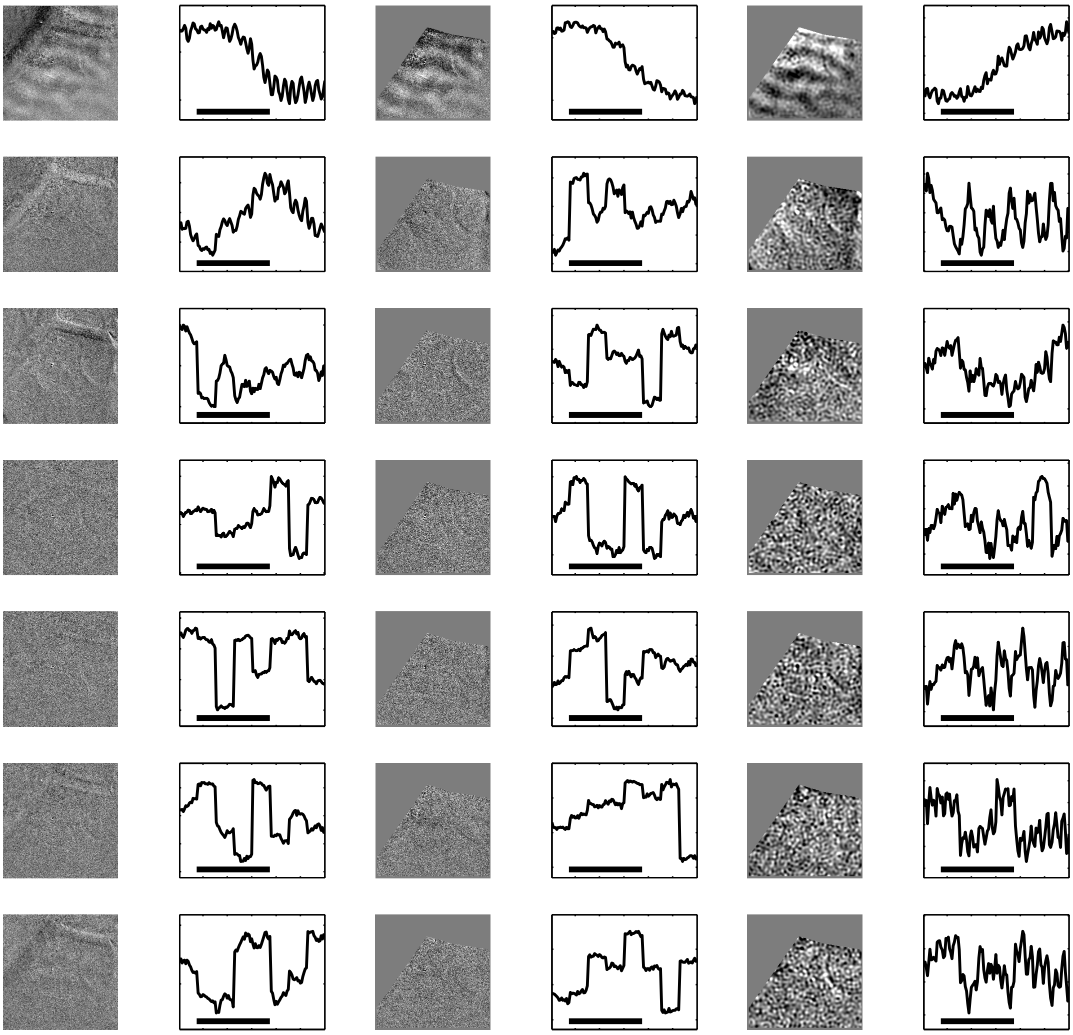


Figure 6.12: Separation results of the single shift ESD algorithm and the back projections on the difference stack from the ocular dominance experiment. On the (**left**) the result without mask and no filtering is shown. In the **middle** the large vessel was masked. On the **right** the result with the mask and lowpass filtering is shown ($25cyl/256pixels$). (Bottom bar: Stimulus duration $4s$).

case of the 0° single condition stack from the orientation preference data we can see in the first two rows of figure 6.13 that the separation with the region of interest fails. The source that is separated contains the global artifact. The mapping signal is not visible. The same is true after we applied no mask to the data before analysis. In the middle two rows the resulting source estimates show a separation from the noise (first two images) but the mapping signal seems to be hidden in the third component under the noise. This problem is solved again by the application of the lowpass filter ($25cyl/256pixel$). In the bottom two rows of figure 6.13 the mapping signal is separated into the second estimate and the global artifacts into the first and third component. The time course of the stimulus specific pattern show a steady rise, even if it is very coarse due to the limited number of frames.

Therefore we look at the ocular dominance data now. In figure 6.14 the separation of the unfiltered data (left half) was only able to find the global activation in the mixture (top image). After the lowpass filtering the mapping signal is found in the mixture and separated from the artifacts (right half second image from top). The time course shows a response to the stimulus onset with a slow slope. Ideally it should continue to rise a little longer after the end of the stimulus but the slope is clearly different from the responses of the background- and vessel artifacts in the other components. In contrast to the separation results with the ICA algorithms on the same filtered single condition stack (see figure 6.10) the stimulus specific component is extracted from the recorded mixture.

6.5 Performance of multi shift ESD

We know now that the concept of ESD is able to find the mapping signal under conditions were the ICA methods fail. Due to the redundant information we have by applying the multi shift algorithm the separation should be more stable and less sensitive to noise. The final goal would be to complete the separation without filtering. To test if the multi shift ESD manages to solve this task we have a look at the separation results from the ocular dominance experiment.

6.5.1 Separation of Difference Stacks

In the case of the separation of the sources from the ocular dominance experiment the single shift ESD showed a good separation without the application of filtering. The lowpass filter was used to enhance the contrast of the result.

In figure 6.15 the left two columns show the separation result from the separation of the difference stack with the multi shift ESD. A weak lowpass filter ($50cyl/256pixels$) was applied. The mapping signal is well separated in the third source from below. The corresponding time course shows a steady increase of the signal strength with a short delay after the stimulus onset. The strength of the mapping signal continues to rise after the stimulus was turned off. The other sources mainly contain respiration artifacts or random noise.

Separation with single shift ESD

Single Condition Stack 0°

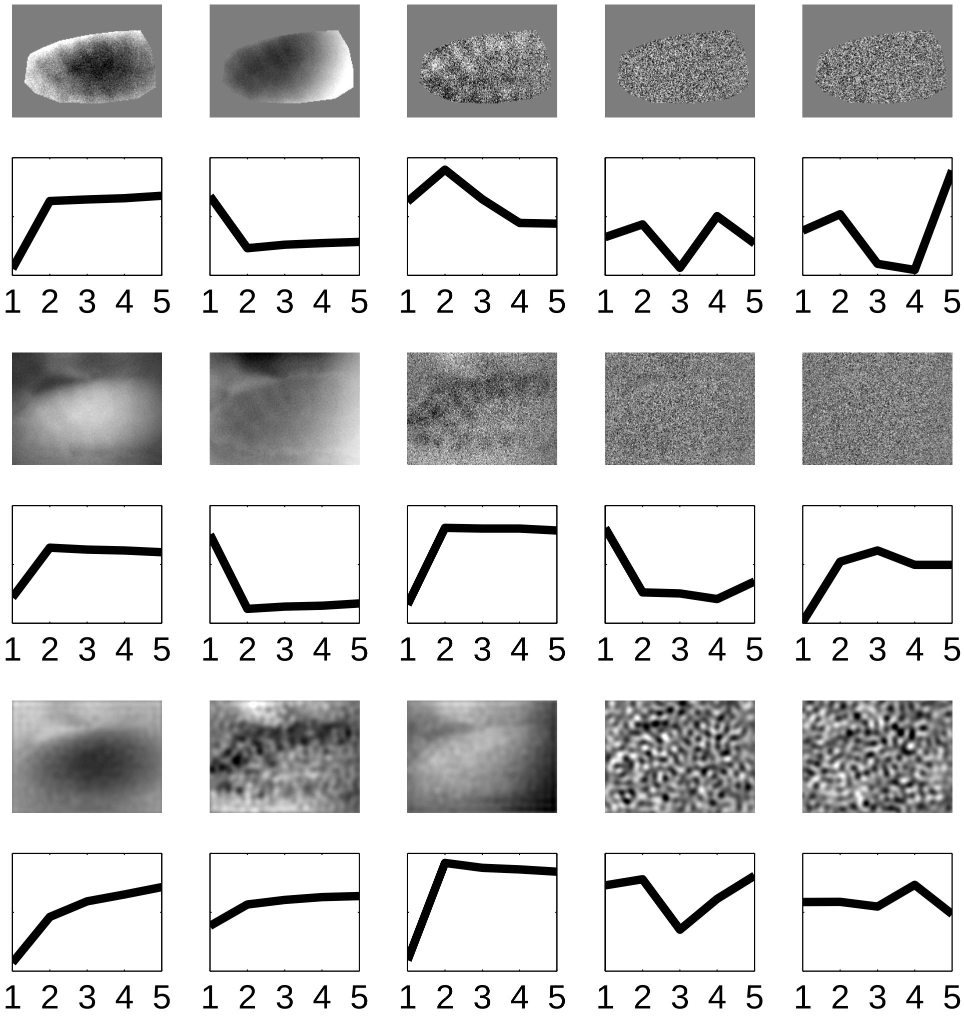


Figure 6.13: Separation results of the single shift ESD algorithm and their back projections on the 0° single condition stack from the orientation preference experiment. The **top two rows** show the components from the masked stack. The **middle two rows** display the unmasked components and the **bottom two rows** were calculated from the lowpass filtered stack ($25cyl/256frames$). The stimulus was presented during the whole trial ($3s$).

6.5.2 Separation of Single Condition Stacks

To demonstrate the separation quality of the multi shift ESD on a single condition stack concentrate on the analysis of the data from the stimulation of the left eye with multi frequency gratings at various orientations. In figure 6.15 we see the results from this part of the ocular dominance experiment after the application of

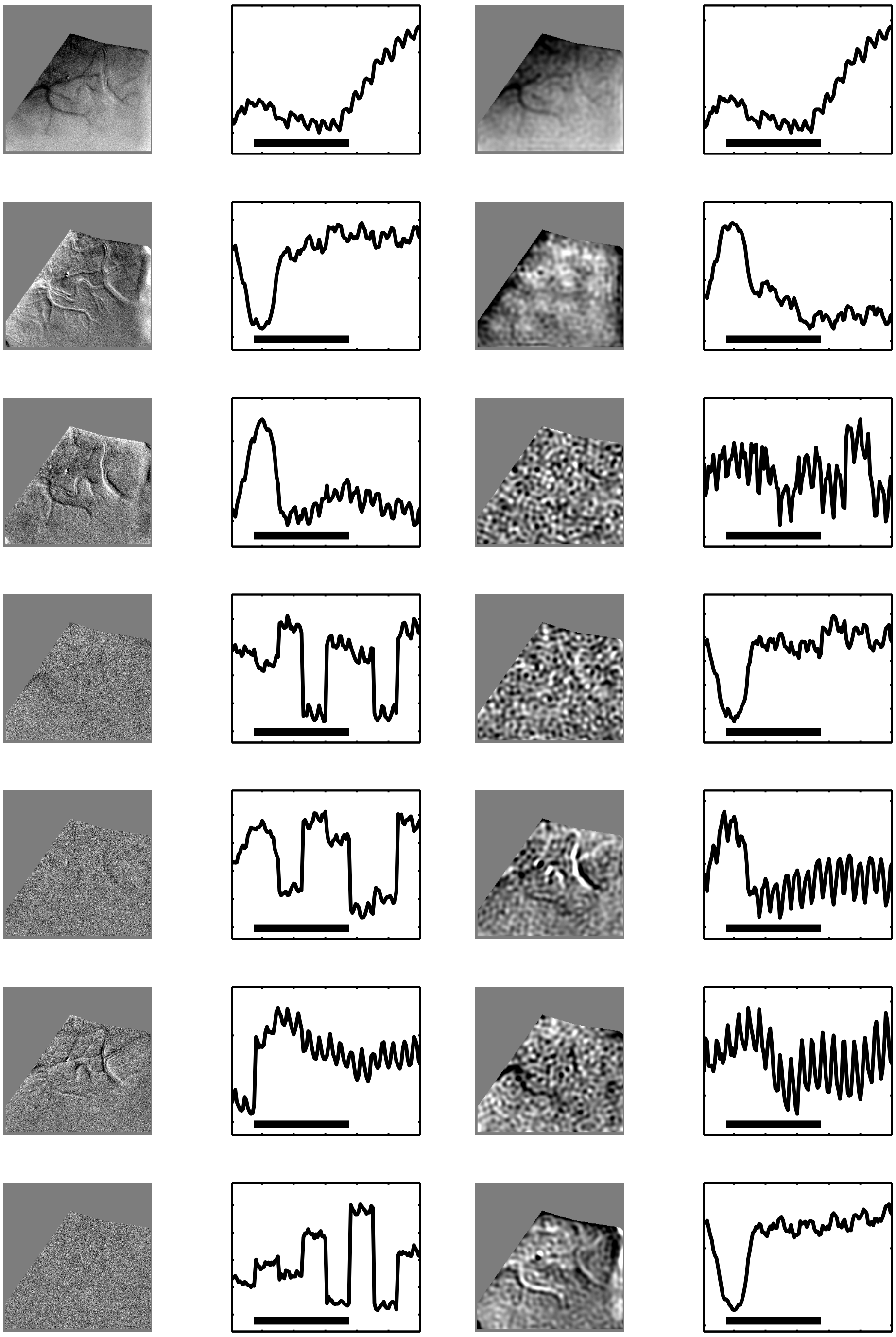


Figure 6.14: Separation results of the single shift ESD algorithm and the back projections on the single condition stack from the ocular dominance experiment (right eye stimulation). On the **left** the result without filtering is shown. The large vessel was masked. On the **right** the result with the mask and lowpass filtering is shown ($25cyl/256pixels$). (Bottom bar: Stimulus duration $4s$).

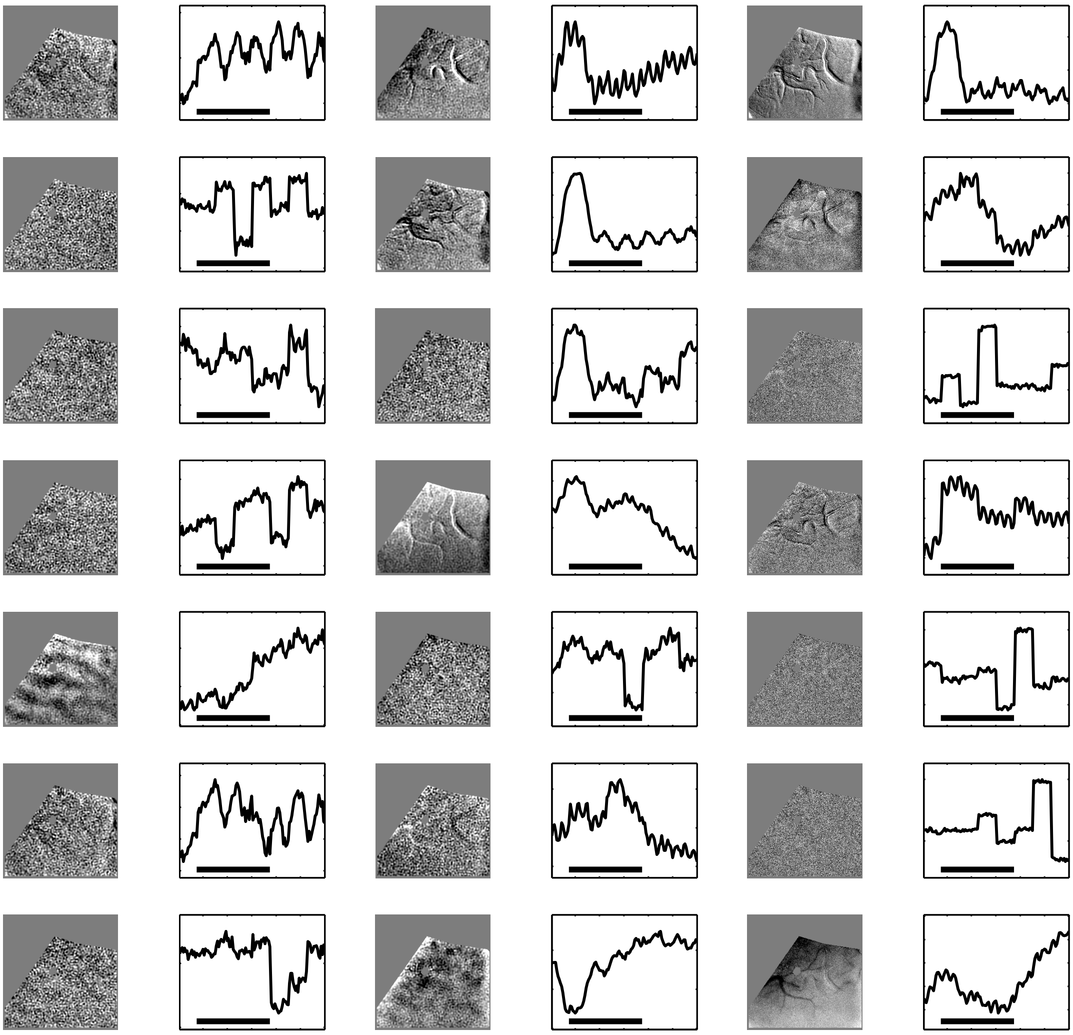


Figure 6.15: Separation results of the multi shift ESD algorithm and the back projections on the difference stack (**left two columns**) and the single condition stack (left eye) (**middle and right columns**) from the ocular dominance experiment. Lowpass filtering is applied to the difference stack and the single condition stack in the middle. In the case of multi shift ESD much less filtering is necessary to receive a result from the single condition stack ($50cyl/256pixels$). Without the application of any lowpass filter the separation is incomplete (**right two columns**). (Bottom bar: Stimulus duration $4s$).

the weak lowpass filter ($50cyl/256pixels$) in the middle two columns and without any filtering in the right two columns.

In the separation with multi shift ESD and the weak lowpass filter the mapping signal is visible in the last component of the middle columns. The time course also agrees with the result we got from the single shift ESD and the stronger filter (see figure 6.14 right columns, second image). Therefore with multi shift ESD we can reduce the strength of the influence of the lowpass filter compared to single shift ESD to get the separation of the mapping signal. The application of a lowpass filter is well justified (see section 4.2.3) and should have no influence on the mapping signal. Nevertheless the weaker the influence of the lowpass filter the less dependent we are on that assumption.

When we applied the multi shift ESD to the single condition stack without any filtering the separation was incomplete and the mapping signal could not be extracted (figure 6.15 right columns).

So the application of the multi shift ESD algorithm again moved the limits of the successful separation of single condition stacks towards a completely parameter free separation. With the data at hand the preprocessing in form of a weak lowpass filter could not be completely eliminated. Compared to the massive influence on the data with the use of the conventional analysis methods the application of the ESD algorithms is a big improvement.

Chapter 7

Discussion

In this work we developed and introduced a blind source separation algorithm called extended spatial decorrelation for the analysis of data sets from the optical imaging of intrinsic signals. Since the early days of optical imaging the separation of the signal components that are exclusively correlated with a specific stimulus condition (mapping signal) from the overall recorded mixture has been a challenging task. Due to the nature of the recording we face a highly noise corrupted mixture with the mapping signal only contributing 0.1% of the overall signal amplitude. A very critical look has to be taken at the standard methods used in the analysis of the optical imaging recordings of intrinsic signals.

Very often strong data pre-processing is already performed during the recording of the data due to intentional blurring by focusing below the cortical surface. This blurring is equivalent to strong lowpass filtering of the images. Misleadingly it is often stated after the recording with focus below the surface, that no further pre-processing like filtering was applied. This is true in a certain sense as the filtering was done by hardware during the experiment. The difference between this type of lowpass filtering and the lowpass filter we apply in some cases of the data pre-processing is the size of the filter kernel. For this study we choose the blur caused by the filter small enough to only cause the high frequency noise component in the recordings to be cancelled but leave the smooth image features like global signal, mapping signals, vessel patterns etc. nearly unchanged. The aim of focusing below the cortical surface is to get rid of the vessel artefacts by smearing them out over a wider area due to the shallow depth of field of the imaging lens. Certainly this does not get rid off the vessel artefacts but distributes them over the whole image. This blurring is not even uniform because of the curvature of the cortex and the difficulty in positioning the camera perpendicular to the imaged surface. Now one has the situation that the signal artefact from the vessels, one wanted to get rid of in the first place, is spread uncontrollably over the image instead of having it at the well defined locations of the arteries and veins. A more elaborate approach should try to extract the vessel artefacts into a separate component that can be removed from the mixture.

Another problematic but often used method of pre-processing is the bandpass filtering of the data. The wrong choice of the cut off frequencies can introduce artefacts that are very similar to the signal components one is looking for. This is nicely demonstrated in figure 4.6.

The most common form of extracting the mapping signal from the recording is the calculation of a difference image as described in chapter 4.2.1. The assumption that all signal components except the mapping signal remain the same for two orthogonal stimulus conditions and will therefore be removed by the subtraction in the calculation of a difference image is not fulfilled. Because other more global features are linked with each of the orthogonal conditions individually the subtraction can not get rid off them though it enhances the signal quality a lot. Further enhancement of the signal to noise ration can be gained by first frame analysis.

A even stronger assumption about the organisation of the visual cortex is introduced with the use of a cocktail blank image. The cocktail blank image is supposed to represent the reflectance pattern of the uniformly activated cortex. By division or subtraction of this image from a difference of single condition image the non stimulus related more global artefacts should be removed. In chapter 4.2.2 it is shown why this assumption has to be handled carefully especially when calculating a single condition image. From a single condition experiment one wants to get information about the activity related to only the chosen stimulus condition and can not introduce information of orthogonal or other stimulus conditions via the cocktail blank image.

To avoid these parametric assumptions and to be neutral in regard to pattern extraction we concentrated on the exploration of parameter free separations methods: the blind source separation (BSS) and independent component analysis (ICA) algorithms. These algorithms use statistical features like second and/or fourth order moments for the source separation rather than heuristic or parametric assumptions like the methods mentioned before.

Among the ICA methods two concepts have been proven to work well for the classical cocktail party problem. One is the Infomax algorithm of Bell and Sejnowski (1995) that uses the statistics of all higher order moments in the probability distributions of the sources and the other is the kurtosis optimisation method of Hyvärinen and Oja (1997) that exploits the properties of the fourth order cumulant. Both algorithms are based on the assumption that the original sources are statistically independent.

This independence assumption is often not fulfilled properly in the recording of a mixture of biological signals that are triggered by the same event, i.e. the presentation of a visual stimulus. Therefore we derived the expanded spatial decorrelation algorithm from a BSS method that was published by Molgedey and Schuster (1994). The assumption behind the ESD algorithm is the smoothness of the underlying sources and the vanishing cross correlations, all of which are properties of second order statistics. From the scattering properties of the cortical tissue we know that the smoothness assumption is well fulfilled (Stetter and Obermayer, 1999).

A general problem in evaluating and comparing the performance of BSS algorithms is that the separation result has to be compared to the underlying original sources. This is not possible with in vivo data as the exact spatial distribution of sources for an individual animal is not know. Therefore we designed three different sets of artificial data. The first dataset contains independent sources and fulfils the assumptions made in ICA. The second data set was developed to meet the statistical properties of the optical images and the third is composed of separation results from real data. The separation quality is measured by the mean reconstruction error (RE) (see equation 5.1).

The simulation with all three artificial datasets were performed at different noise levels ranging from no noise to a noise level of $0db$. The classical ICA methods delivered very good separation results on the first artificial dataset that fulfilled the independence assumption but does not have smooth sources. In this case the performance of the ESD algorithms was poor as expected due to the lack off auto- and cross correlations. But the simulations with the artificial datasets two and three (smooth and natural sources) clearly demonstrated that the different instances of the ESD algorithm perform superior on the artificial data that reflect the statistics of the optical images (Schießl et al., 1998; Schießl et al., 1999). This is especially true at the high noise levels we find in vivo. The single shift ESD shows a higher quality in separation than the ICA algorithms (Schießl et al., 2000c). This result can be enhanced even further by using the multi shift ESD algorithm with noise robust sphering. This approach is less susceptible to noise because it has redundant information about the sources (Schöner et al., 1999; Schöner et al., 2000). Another explanation for the bad performance of the ICA methods might be explained by the fact, that it is more difficult to estimate higher order moments in the presence of strong sensor noise from a limited number of data points, than the estimation of second order moments.

With the introduction of the regularization term to the cost function of ESD one deliberately leaves the path of parameter free estimation in order to introduce additional knowledge about the time course of the sources (Schießl et al., 2000a). If the assumptions are wrong the result will necessarily also be wrong. But if the assumptions are right the tests with the artificial data show that this is the most promising method. Unfortunately up to now the regularization term implies some restrictions on the size and appearance of the demixing matrix **W** so that the method could not yet be applied to the real data sets.

The insights of the performance differences between ICA and ESD we have gained from the tests with the artificial data were reflected in the results with real data. This shows, apart from the fact that the artificially created data were well designed, that these biological signals with the high noise levels are not suited for the classical ICA methods. The ESD assumption about the smoothness of the sources is supported by tissue scattering properties. From the first and second source of the natural images in the test data we know that the auto-correlations of a ocular dominance pattern and a vessel pattern is about five times bigger than the cross-correlations.

The shift vector for single shift ESD should be chosen in a way that the auto-correlations of the sources as well as the cross-correlations of the mixtures should be big. Furthermore it showed that small shifts like (5×5) pixels have the highest probability for good and reliable separation results. Though ESD only uses second order statistics it is not bound to orthogonal solutions like PCA. On the tested optical imaging data sets the ICA algorithms were not able to clearly separate the mapping signal from the artefacts in the difference stacks even after lowpass filtering. The single shift ESD algorithm on the other hand could clearly separate the mapping signal in difference stacks after filtering. For the analysis of the single condition stacks the ICA methods proved to be inadequate and the stimulus specific signal could not be extracted. Both the single shift ESD and the multi shift ESD separated the mapping signal in the single condition stack and the later method often needed less or no filtering. Apart from that fact the two instances of ESD did not show a that big performance difference on the real data sets.

Once a source is separated from the mixture one has to make a judgement about its metabolic origin. This proves especially difficult in the analysis of new stimulus regimes where the pattern of cortical activation is unknown. An important tool to aid this judgement is the corresponding time course one can calculate by back projection on the data stack. This tool is delivered by our non parametric statistical methods and is not available in this form from the standard analysis methods for optical images.

We have introduced a method that allows us to analyse data with much lower signal to noise ratios compared to what has been possible so far. This not only extends the stimulus regimes that can be investigated with optical imaging, but also helps to analyse data from sub-optimal recording conditions. As the statistical properties of the estimates are untouched we can also have more belief in the statistical results like pinwheel densities etc. we get from the data. The metabolic activities reflected in the intrinsic signals are also used for imaging with other techniques like the blood-oxygenation level-dependent (BOLD) functional magnetic resonance imaging (fMRI) for localising brain function of humans in vivo. Because of the statistical similarities of the data from these different medical imaging techniques the results and findings from this work show the same analysis methods are highly applicable to another. Both the experimental technique and the field of blind source separation are quite recent and so far the properties of BSS have not been exploited for the separation of two dimensional medical imaging data.

One of the goals for the application of the non invasive optical imaging technique is the use during brain surgery on humans. Due to size of the craniatomy the stable conditions we have in the experiment with the steel chamber will not be reproduced. This will introduce many artefacts that can not be removed with the standard methods as they are not stimulus locked. We believe that it is here that the robustness of the ESD algorithms will prove themselves especially valuable.

Bibliography

Albrecht, D. G. and Hamilton, D. B. (1982). Striate cortex of monkey and cat: contrast response function., *J. Neurophysiol.* **48**: 217–237.

Amari, S. (1996). Neural learning in structured parameter spaces - natural riemannian gradient., *in* M. C. Mozer, M. I. Jordan and T. Petsche (eds), *Advances in Neural Information Processing Systems*, Vol. 9.

Bauer, U. (1999). *Computational models of neural circuitry in the macaque monkey primary visual cortex.*, Cuvillier Göttingen.

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

Bishop, C. M. (1995). Neural networks for pattern recognition, Clarendon Press, Oxford.

Blakemore, C. and Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex., *Exp. Brain Res.* **15**: 439–440.

Blasdel, G. G. (1992a). Differential imaging of ocular dominance and orientation selectivity in monkey striate cortex., *J. Neurosci.* **12**: 3115–3138.

Blasdel, G. G. (1992b). Orientation selectiviy, preference, and continuity in monkey striate cortex., *J. Neurosci.* **12**: 3139–3161.

Blasdel, G. G. and Lund, J. S. (1983). Termination of afferent axons in macaque striate cortex, *J. Neurosci.* **3**: 1389–1413.

Blasdel, G. G. and Salama, G. (1986). Voltage-sensitive dyes reveal a modular organization in monkey striate cortex., *Nature* **321**: 579–585.

Bonhoeffer, T. and Grinvald, A. (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns, *Nature* **353**: 429–431.

Bonhoeffer, T. and Grinvald, A. (1993). The layout of iso-orientation domains in area 18 of cat visual cortex: optical imaging reveals a pinwheel-like organization., *J. Neurosci.* **13**: 4157–4180.

Bonhoeffer, T. and Grinvald, A. (1996). Optical imaging based on intrinsic signals: The methodology, *in* A. Toga and J. C. Maziotta (eds), *Brain mapping: The methods*, Academic Press, Inc., San Diego, CA, pp. 55–97.

Bonhoeffer, T., Kim, D. S., Malonek, D., Shoham, D. and Grinvald, A. (1995). Optical imaging of the layout of functional domains in area 17 and across the area 17/18 border in cat visual cortex., *Eur. J. Neurosci.* **7**: 1973–1988.

Bosking, W. H., Zhang, Y., Schofield, B. and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex., *J. Neurosci.* **17**: 2112–2127.

Buchweitz, E. and Weiss, H. R. (1986). Alterations in perfused capillary morphometry in awake vs. anesthetized brain, *Brain Res* **2**(377(1)): 105–111.

Coenen, A. M. L. and Vendrik, A. J. H. (1972). Determination of the transfer ratio of cat's geniculate neurons through quasi-intracellular recordings and the relation with the level of alertness., *Exp. Brain Res.* **14**: 227–242.

Cohen, L. B. (1973). Changes in neuron structure during action potential propagation and synaptic transmission., *Physiol. Review* **53**: 373–418.

Cohen, L. B., Keynes, R. D. and Hille, B. (1968). Light scattering and birefringence changes during nerve activity., *Nature* **218**: 438–441.

Connolly, M. and Essen, D. V. (1984). The representation of the visual field in parvocellular and magnocellular layers of the lateral geniculate nucleus in the macaque monkey., *J. Comp. Neurol.* **226**: 544–564.

Daniel, P. M. and Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys, *J. Physiol.(Lond.)* **159**: 203–221.

DeAngelis, G. C., Robson, J. G., Ohzawa, I. and Freeman, R. D. (1992). Organization of suppression in receptive fields of neurons in cat visual cortex., *J. Neurophysiol.* **68**: 144–163.

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

Frostig, R. D., Lieke, E. E., Ts'o, D. Y. and Grinvald, A. (1990). Cortical functional architecture and local coupling between neuronal activity and the microcirculation revealed by in vivo high-resolution optical imaging of intrinsic signals., *Proc. Natl. Acad. Sci USA* **87**: 6082–6086.

Gilbert, C. D. and Wiesel, T. N. (1990). The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat., *Vision Res.* **30**: 1689–1701.

Goethe, J. W. (1808). Faust (part 1), *in* translation by P. Salm (ed.), *Bantam Classics*, 1985 edn, Bantam Books.

Grinvald, A., Bonhoeffer, T., Malonek, D., Shoham, D., Bartfeld, E., Arieli, A., Hildesheim, R. and Ratzlaff, E. (1991). Optical imaging of architecture and function of the living brain, *in* L. R. Shire, N. M. Weinberger, G. Lynch and J. C. McGaugh (eds), *Memory: Organization and Locus of Change*, Oxford University Press, New York, Oxford, pp. 49–85.

Grinvald, A., Lieke, E., Frostig, R. D., Gilbert, C. D. and Wiesel, T. N. (1986). Functional architecture of cortex revealed by optical imaging of intrinsic signals., *Nature* **324**: 361–364.

Haglund, M. M. and Blasdel, G. G. (1992). Video imaging of neuronal activity., *in* Stanford (ed.), *Monitoring Neural Activity*, Oxford Press, chapter 4, pp. 85–111.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex., *J. Physiol.* **160**: 106–154.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex., *J. Physiol.* **195**: 215–243.

Hubel, D. H. and Wiesel, T. N. (1972). Laminar and columnar distribution of geniculo-cortical fibers in macaque monkey., *J. Comp. Neurol.* **146**: 421–450.

Hubel, D. H. and Wiesel, T. N. (1977). Functional architecture of macaque monkey visual cortex., *Proc. R. Soc. Lond. B* **198**: 1–59.

Hubel, D. H., Wiesel, T. N. and Stryker, M. P. (1978). Anatomical demonstration of orientation columns in macaque monkey., *J. Comp. Neur.* **177**: 361–379.

Hyvärinen, A. and Oja, E. (1997). A fast fixed point algorithm for independent component analysis., *Neural Comput.* **9**: 1483–1492.

Hyvärinen, A. and Oja, E. (1999). Independent component analysis: A tutorial, http://www.cis.hut.fi/projects/ica/, Helsinki University of Technology.

Kandel, E. R., Schwartz, J. H. and Jessel, T. M. (1991). *Principles of Neural Sciences*, Prentice Hall International London.

Kim, D.-S., Duong, T. Q. and Kim, S.-G. (2000a). High-resolution mapping of iso-orientation columns by fMRI, *Nature Neurosci.* **3**: 164–169.

Kim, D.-S., Duong, T. Q. and Kim, S.-G. (2000b). Reply to "can current fMRI techniques reveal the micro-architecture of the cortex?", *Nature Neurosci.* **3**(5): 414.

Koehler, B.-U. and Orglmeister, R. (1999). Independent component analysis using autoregressive models., *in* J.-F. Cardoso, C. Jutten and P. Loubaton (eds), *Proceedings of the ICA99 workshop*, Vol. 1, pp. 359–363.

Kreisman, N. R., LaManna, J. C., Liao, S.-C., Teh, E. R. and Alcala, J. R. (1995). Light transmittance as an index of cell volume in hippocampus slices: optical differences of interfaced and submerged positions., *Brain Res.* **693**: 179–186.

Kuffler, S. W. (1953). Discharge patterns of and functional organization of mammalian retina., *J. Neurophysiol.* **16**: 37–68.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity, *Z. Naturforsch.* **36**: 910–912.

Levitt, J. B. and Lund, J. S. (1997). Contrast dependence of contextual effects in primate visual cortex., *Nature* **387**: 73–76.

Levitt, J. B., Lund, J. S. and Yoshioka, T. (1996). Anatomical substrates for early stages in cortical processing of visual information in the macaque monkey, *Behav. Brain Res.* **76**: 5–19.

Linsker, R. (1989). How to generate ordered maps by maximizing the mutual information between input and output signals., *Neural Comput.* **1**: 402–411.

Logothetis, N. (2000). Can current fMRI techniques reveal the micro-architecture of cortex?, *Nature Neurosci.* **3**(5): 413.

MacVicar, B. A. and Hochman, D. (1991). Imaging of synaptically evoked intrinsic optical signals in hippocampal slices., *J. Neurosci.* **11**: 1458–1469.

Makeig, S., Bell, A. J., Jung, T.-P. and Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data., *in* D. Touretzky, M. Mozer and M. Hasselmo (eds), *Advances in Neural Information Processing Systems*, MIT Press, Cambridge MA, pp. 145–151.

Malach, R., Amir, Y., Harel, M. and Grinvald, A. (1993). Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex., *Proc. Natl. Acad. Sci. USA* **90**: 10469–10473.

Malach, R., Tootell, R. B. and Malonek, D. (1994). Relationship between orientation domains, cytochrome oxidase stripes, and intrinsic horntal conections in squirrel monkey area v2.nizo, *Cereb. Cortex* **4**: 151–165.

Malonek, D. and Grinvald, A. (1996). Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping., *Science* **272**: 551–554.

McKeown, M. J., Jung, T.-P., Makeig, S., Brown, G., Kindermann, S. S., Lee, T.-W. and Sejnowski, T. J. (1998). Spatially independent activity patterns in functional MRI data during the Stroop color-naming task., *Proc. Natl. Acad. Sci. USA* **95**: 803–810.

McKeown, M., Martin, J., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A. and Sejnowski, T. (1997). Blind separation of functional magnetic resonance imaging (fMRI) data, *Advances in Neural Information Processing.*

McLoughlin, N. P. and Blasdel, G. G. (1998). Wavelenght-dependent differences between optically determined functional maps from macaque striate cortex, *Neuroimage* **7**: 326–336.

Molgedey, L. and Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations, *Phys. Rev. Lett.* **72**: 3634–3637.

Müller, K.-R., P, P. and Ziehe, A. (1999). Jadetd: Combining higher-order statistics and temporal information for Blind Source Separation (with noise)., *in* J.-F. Cardoso, C. Jutten and P. Loubaton (eds), *Proceedings of the 1. ICA99 Workshop, Aussois*, Vol. 1, pp. 87–92.

Müller, T., Stetter, M., Hübener, M., Gödecke, I., Chapman, B., Löwel, S., Sengpiel, F., Bonhoeffer, T. and Obermayer, K. (2000). An analysis of orientation and ocular dominance patterns in the visual cortex of cats and ferrets., *Neural Comput.* p. in press.

Obermayer, K. and Blasdel, G. G. (1993). Geometry of orientation and ocular dominance columns in monkey striate cortex., *J. Neurosci.* **13**: 4114–4129.

Obermayer, K. and Blasdel, G. G. (1997). Singularities in primate orientation maps., *Neural Comput.* **9**: 555–567.

Oja, E. (1997). The nonlinear pca learning rule in independent component analysis., *Neurocomputing* **17**: 25–45.

Orbach, H. S., Cohen, L. B. and Grinvald, A. (1985). Optical mapping of electrical activity in rat somatosensory and visual cortex., *J. Neurosci.* **5**: 1886–1895.

Otto, T., Stetter, M., Müller, T., Sengpiel, F., Hübener, M., Bonhoeffer, T. and Obermayer, K. (1998). Source separation of intrinsic signals from image sequences of cat area 17., *26. Göttingen Neurobiology Conference.*

Papoulis, A. (1965). *Probability, random variables, and stochastic processes.*, McGraw-Hill Book Company.

Parra, L. C. (1998). An introduction to independent component analysis and blind source separation, Sarnoff Corporation, CN-5300, Princton, NJ 08543.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1988). *Numerical Recipes in C*, Cambridge University Press, Cambridge, U.K.

Principe, J. C., Cerruti, S. and Amari, S. (2000). Special topic section on advances in statistical signal processing for medicine, *IEEE Trans. Biomed. Engin.* **47**: p.565.

Rao, S. C., Toth, L. J. and Sur, M. (1997). Optically imaged maps of orientation preference in primary visual cortex of cats and ferrets., *J. Comp. Neurol.* **387**: 358–370.

Rojer, A. S. and Schwartz, E. L. (1990). Cat and monkey cortical columnar pattern modeled by bandpass-filtered 2D white noise., *Biol. Cybern.* **62**: 381–391.

Rüger, S. M. (1996). Stable dynamic parameter adaptation., *in* D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds), *Advances in Neural Information Processing Systems.*, Vol. 8, MIT Press Cambridge, MA, pp. 225–231.

Salzberg, B. M., Davila, H. V. and Cohen, L. B. (1973). Optical recording of impulses in individual neurones of an invertebrate central nervous system., *Nature* **246**: 508–509.

Schießl, I., Schöner, H., Stetter, M., Dima, A. and Obermayer, K. (2000a). Regularized second order source separation, *in* P. Pajunen and J. Karhunen (eds), *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland*, Vol. 2, pp. 111–116.

Schießl, I., Stetter, M., Mayhew, J. E. W., Askew, S., McLoughlin, N., Levitt, J. B., Lund, J. S. and Obermayer, K. (1999). Blind separation of spatial signal patterns from optical imaging records., *in* J.-F. Cardoso, C. Jutten and P. Loubaton (eds), *Proceedings of the 1. ICA99 Workshop, Aussois*, Vol. 1, pp. 179–184.

Schießl, I., Stetter, M., Mayhew, J. E. W., McLoughlin, N., Lund, J. S. and Obermayer, K. (2000b). Blind signal separation from optical imaging data, *in* G. Sommer, N. Krüger and C. Perwass (eds), *Mustererkennung 2000, DAGM-Symposium Kiel, 13-14. September 2000*, Vol. 2, pp. 91–98.

Schießl, I., Stetter, M., Mayhew, J. E. W., McLoughlin, N., Lund, J. S. and Obermayer, K. (2000c). Blind signal separation from optical imaging recordings with extended spatial decorrelation, *IEEE Trans. Biomed. Engin.* **47**: p.573.

Schießl, I., Stetter, M., Otto, T., Sengpiel, F., Hübener, M., Bonhoeffer, T. and Obermayer, K. (1998). Signal extraction from optical imaging data from cat area 17 by blind separation of sources., *Soc. Neurosci. Abstr.* **24**: 9.

Schöner, H., Stetter, M., Schießl, I., Mayhew, J. E. W., Lund, J. S., McLoughlin, N. and Obermayer, K. (1999). Blind separation of noisy mixtures by iterative decorrelation.., Learning workshop, Snowbird, Utah, April 6-9.

Schöner, H., Stetter, M., Schießl, I., Mayhew, J. E. W., Lund, J. S., McLoughlin, N. and Obermayer, K. (1999b). Noise-robust blind separation of sources for optiocal imaging of intrinsic signals., *Soc. Neurosci. Abstr.* **25**: 783.

Schöner, H., Stetter, M., Schießl, I., Mayhew, J., Lund, J., McLoughlin, N. and Obermayer, K. (2000). Application of blind separation of sources to optical recording of brain activity, *in* S. Solla, T. Leen and K.-R. Müller (eds), *Advances in Neural Information Processing Systems NIPS 12*, MIT Press, pp. 949–955.

Shannon, C. E. (1948). A mathematical theory of communication., *Bell Syst. Tech.* **27**: 379.

Shoham, D., Glaser, D. E., Arieli, A., Kenet, T., Wijnbergen, C., Y.Toledo, Hildesheim, R. and Grinvald, A. (1999). Imaging cortical dynamics at high spatial resolution with novel blue voltage-sensitive dyes, *Neuron* **24**: 791–802.

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J. and Davis, J. (1995). Visual cortical mechanisms detecting focal discontinuities., *Nature* **378**: 492–496.

Stetter, M. (2000). *Exploration of Cortical Function*, Habil Thesis, TU Berlin.

Stetter, M. and Obermayer, K. (1999). Simulation of scanning laser techniques for optical imaging of blood-related intrinsic signals., *J. Opt. Soc. Am. A* **16**: in press.

Stetter, M., Schießl, I., Otto, T., Sengpiel, F., Hübener, M., Bonhoeffer, T. and Obermayer, K. (2000). Principal component analysis and blind separation of sources for optical imaging of intrinsic signals., *NeuroImage* **11**: 482–490.

Ts'o, D. Y., Frostig, R. D., Lieke, E. E. and Grinvald, A. (1990). Functional organization of primate visual cortex revealed by high resolution optical imaging., *Science* **249**: 417–420.

Vetter, R., Vesin, J., Celka, P. and Scherrer, U. (1999). Observer of the autonomic cardiac outflow in humans using non-causal blind source separation., *in* J.-F. Cardoso, C. Jutten and P. Loubaton (eds), *Proceedings of the 1. ICA99-Workshop, Aussois*, pp. 161–166.

Vollgraf, R. (2000). *Vergleich convolutiver Dekorrelationsverfahren zur blinden Quellentrennung*, Diplomarbeit, Fachbereich Informatik, Technische Universitaet Berlin.

Vollgraf, R., Stetter, M. and Obermayer, K. (2000). Convolutive decorrelation procedures for blind source separation, *Proceedings of the ICA2000 workshop.*