Lehrstuhl für Steuerungs- und Regelungstechnik
Technische Universität München

FEATURE EXTRACTION IN NON-INVASIVE BRAIN-COMPUTER INTERFACES

Moritz Grosse-Wentrup

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr.-Ing. Klaus Diepold
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr.-Ing./Univ. Tokio Martin Buss
	2. Priv.-Doz. Micah M. Murray, PhD, Universität Lausanne/Schweiz

Die Dissertation wurde am 21.04.2008 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 24.06.2008 angenommen.

Abstract

By inferring the intention of human subjects from signals generated by the central nervous system (CNS), Brain-Computer Interfaces (BCIs) provide an alternative means of communication for subjects with damages to the peripheral nervous system, e.g., caused by neuro-degenerative diseases such as amyotrophic lateral sclerosis or brain stem stroke. While state-of-the-art BCIs based on non-invasive recording modalities enable elementary communication, more complex tasks, such as the control of a robotic endeffector, remain beyond the feasibility of current systems.

In this thesis, it is argued that the primary cause for this limitation is the inadequacy of present algorithms for feature extraction, i.e., of algorithms that aim to extract those characteristics of the data recorded from the CNS providing most information on the BCI-user's intention. The main contribution of this thesis in addressing this problem is threefold. In terms of supervised feature extraction, the framework of information theoretic feature extraction is employed to derive an algorithm that is, under some assumptions, optimal in terms of maximizing mutual information of the BCI-user's intention and extracted features. In terms of unsupervised feature extraction, an algorithm based on beamforming methods is designed that optimally extracts signals originating in certain regions of interest within the brain. Due to its unsupervised nature, this algorithm is very robust and requires substantially less training data than supervised approaches. Both algorithms are validated experimentally and shown to outperform state-of-the-art approaches for feature extraction in non-invasive BCIs. Finally, a theoretically founded and experimentally validated explanation for the success of Independent Component Analysis (ICA) in the analysis of EEG/MEG recordings in general, and as tool for feature extraction in BCIs in particular, is provided that resolves the apparent contradiction between the requirement of ICA of at least as many sensors and sources and the physiological implausibility of this assumption.

In summary, it is argued that the main limitation for feature extraction in non-invasive BCIs is insufficient knowledge on how cognitive states are encoded in signals generated by the CNS. The thesis concludes with a discussion why future research on feature extraction in non-invasive BCIs should take into account the nature of the brain as a complex network with time-varying connectivity patterns.

Contents

Chapter 1

Introduction

1.1 Motivation

Any act of human communication depends on volitional muscle control. When we speak to another person, use our fingers to type a text on a keyboard, or engage in any other type of communication, we rely on our ability to produce goal directed muscle activations. While such actions are initiated within the central nervous system (CNS), no muscle activation, and thus no communication, is possible without the peripheral nervous system. The peripheral nervous system comprises all nerves and neurons outside the CNS, i.e., outside the brain and the spinal cord, and provides the connection between the brain and the rest of the body. As such, it is much more exposed and less protected than the CNS. What happens if the peripheral nervous system is injured and the connection between the CNS and the rest of the body is affected? Depending on the severity of the damage the resulting effects may range from mild impairment up to a so called locked-in state - a state in which a person becomes imprisoned in her/his body without being able to communicate with the outside world. One disease that inevitably leads to a locked-in state is amyotrophic lateral sclerosis (ALS), a degenerative disease that affects motor neurons. During the progress of the disease patients gradually loose control over their motor system, until all voluntary and involuntary motor control is lost. Prominent patients with ALS include the physicist Stephen Hawking and the recently deceased painter Jörg Immendorf. Diseases such as ALS, however, are not the only cause of damage to the peripheral nervous system. Accidents and strokes are other frequent causes for a loss of voluntary motor control. While these impairments rarely lead to a locked-in state, they also constitute a significant decrease in the affected patients' life quality. Scenarios such as these provide the motivation for research on Brain-Computer Interfaces (BCIs). BCIs are devices that enable communication without using the peripheral nervous system. They solely rely on signals generated by the CNS, which are used to infer the BCI-user's intention. BCIs thereby provide a new output channel for the brain that can be used to replace or assist a damaged peripheral nervous system. While BCIs can be realized by a variety of means (discussed in Section

1.2), the work of this thesis only concerns non-invasive BCIs, i.e., BCIs that solely utilize signals generated by the CNS that can be recorded without penetrating the skull. Subsequently, if not stated differently, the term BCI refers to non-invasive BCIs.

Nowadays BCIs hardly enable more than basic communication to those with severe damages to the peripheral nervous system. In the future, however, the use of BCIs will not be restricted to basic communication. The control of a wheelchair by a BCI already is an active field of research [LNF+06], and the extension to more powerful robotic systems only a matter of time. For example, a locked-in patient might be equipped with a BCI that enables the control of a humanoid robot. This robot would be used as a replacement for the patient's body, enabling her, at least to some extent, to participate in every day life.

The primary goal of research on BCIs is the construction of a neuro-prothesis that can replace or assist the peripheral nervous system, but this is by far not its only purpose. In fact, the main task in constructing a BCI is the development of powerful tools to analyze and interpret signals generated by the brain. As such, the advances in research on BCIs provide new tools that are of large value for understanding the way information is processed by our brains. However, besides improving the life quality of disabled patients and enabling advances in neuroscientific research, there are also less noble areas of application. BCIs will probably find the largest proliferation as input devices for video games. Taking into consideration the wide success of alternative input devices, such as movement sensors for video games, and the simple fact that it is fun to control a video game just by thought, a BCI that can be sold at a reasonable price for private use is likely to become a large financial success.

Before these visions turn into reality, several major obstacles have to be overcome. One of these obstacles is the currently very low information transfer rate (ITR) of BCIs [1]. The amount of information that can be send through a BCI roughly determines the complexity of the device that can be controlled with it. While the ITR of current BCIs suffices to write short sentences [BGH+99] or, after intensive subject training, control a computer cursor [WM04], the reliable control of more complex systems, such as a humanoid robot or just a robotic arm, requires a significant increase in the amount of bits that can be send per second.

There exist two principal approaches to increase the ITR of a BCI. First, new experimental paradigms can be designed that allow for higher ITRs. The experimental paradigm of a BCI consists of a set of rules that determine which thoughts should be executed by a subject to express a certain intention. These thoughts lead to pattern changes in the signals recorded from the CNS, which can be detected and used to infer the BCI-user's intention. The number of intentions that can be expressed by an experimental paradigm determines an upper bound on the amount of information that can be transmitted. Research in this field aims at discovering paradigms that

[1] The concept of ITR does not apply to BCIs in a straightforward manner, as discussed in Section 2.3. For now, however, it suffices to accept ITR as a measure of the performance of a BCI.

a) allow expressing a multitude of intentions, b) lead to strong and distinct pattern changes in the data recorded from the CNS for each expressed intention, and c) can be used by disabled subjects without extensive subject training. The second approach to increase the ITR is to focus on the available data and improve the inference of the user's intention from the recorded signals. This approach can be roughly subdivided into feature extraction and statistical inference, although there is some overlap between these two concepts. In the framework of BCIs, the goal of statistical inference is to design algorithms that learn to optimally infer the BCI-user's intention from the recorded signals. Feature extraction, in contrast, is not concerned with actual inference, but with extracting those components/characteristics of the recorded data that are optimal for inferring the user's intention. Feature extraction can thus be seen as a pre-processing of the recorded data, with the aim to facilitate a subsequent inference. While machine learning algorithms for statistical inference are highly developed and can be applied in a straight-forward manner to BCIs, feature extraction for BCIs is still largely in its infancy. In BCIs, the data recorded from the CNS is usually high-dimensional, non-stationary, and has a low signal-to-noise (SNR) ratio, i.e., the components of the data providing information on the BCI-user's intention are deeply buried in ongoing background activity of the brain. This combination poses problems that are seldomly encountered in other areas of signal processing or machine learning. Consequently, few algorithms exist that are suitable for feature extraction in the context of BCIs.

The motivation for the work presented in this thesis is the conviction that the lack of advanced methods for feature extraction constitutes the main bottleneck for a significant increase in ITR of BCIs. Consequently, the main topic of this thesis is the development of algorithms for BCIs that extract those characteristics of signals recorded from the CNS that are optimal for inferring the BCI-user's intention.

1.2 State-of-the-Art of Brain-Computer Interfaces

A multitude of approaches to realizing a BCI exist. The most influential of these, from a historical and state-of-the-art perspective, are briefly presented in this section. A more detailed discussion of the components employed in state-of-the-art non-invasive BCIs is carried out when appropriate in Chapters 3 - 6.

In general, BCIs can be realized by invasive- and by non-invasive means. Invasive BCIs infer the user's intention from signals recorded directly inside the CNS, e.g., from local field potentials or single cell activity. This offers the advantage of providing direct access to information processing within the brain, but poses a significant medical risk and raises ethical concerns. Non-invasive BCIs, on the other hand, only utilize signals that can be recorded without penetrating the skull. These include signals such as the electric or magnetic field of the brain, measured by electroencephalography (EEG) and magnetoencephalography (MEG), or the hemodynamic response modulated by neuronal activity and measured by functional magnetic res-

onance imaging (fMRI) or near infrared spectroscopy (NIRS). Recordings of these signals can be performed without medical risks for the subject, but have the disadvantage of only providing measures of neural mass activity, i.e., only superpositions of the signals generated by hundred-thousands of single neurons can be measured. As a consequence, BCIs based on non-invasive methods currently only achieve a fraction of the ITR achieved by invasive BCIs.

1.2.1 Invasive Approaches

Research on invasive approaches is mostly carried out within the USA. While measurements of local field potentials are increasingly seen as an alternative to single-cell recordings in the context of invasive BCIs (cf. [HWCM06], [MRV+03]), most groups still focus on decoding movement intentions from firing patterns of single neurons located in motor areas of the cortex. Due to the medical risks and ethical concerns associated with brain implants most experiments, with one notable exception, are carried out with non-human primates. One of the main problems, that all groups employing invasive methods face, is the insufficient stability of recordings obtained from chronically implanted electrodes. Subsequently, an overview of research groups developing invasive BCIs is given. While care has been taken to include the most significant work, this overview is necessarily biased.

MotorLab, University of Pittsburgh

The group of A. Schwartz, now at the University of Pittsburgh, was the first one to realize online control of a neuroprosthetic device in three dimensions [THS02]. Direction tuning properties of single-cells, recorded from motor and pre-motor areas, were used to enable two Rhesus macaques to move a three-dimensional cursor to one of eight locations on a three-dimensional grid. Interestingly, Schwartz et al. could show that the tuning properties of the recorded cells adapted to the neuroprosthesis. This lead to improved movement accuracy with training and, more importantly, decreased the number of cortical units necessary for movement prediction.

Laboratory of Miguel A. L. Nicolelis, Department of Neurobiology Duke University Medical Center

The group of M. Nicolelis at Duke University uses single-cell recordings from large neuronal ensembles in non-human primates to predict several motor parameters such as hand position, velocity, and gripping force. These parameters are then used to control a neuroprosthetic device or enable reaching and grasping movements in virtual environments (cf. [CLC+03] and the references therein). Interestingly, recordings are not confined to a single area of cortex, but simultaneous recordings from multiple sites are obtained. The recordings from all sites are then shown to

contribute, to varying degrees, to the estimation of the desired motor parameters. This is in contrast to most other groups, which aim to record from brain regions specialized for certain motor tasks.

Neural Prosthetic Systems Laboratory, Stanford University

The Neural Prosthetic Systems Laboratory at Stanford University, headed by K. Shenoy, employs single-cell recordings from dorsal pre-motor cortex for motor inference. Contrary to the groups of Schwartz and Nicolelis they do not aim to translate neural activity into continuous movement commands. Instead, they predict the intended target location of reaching movements from single-cell activity. Using this approach they obtained a maximum ITR of 6.5 bits/s [SRY+06], which is the highest ITR reported for BCIs so far.

Cyberkinetics / Donoghue Lab, Brown University

The ”BrainGate”, developed by Cyberkinetics, a company founded by J. Donoghue of Brown University, is the first invasive BCI actually tested on a human subject [HSF+06]. Electrodes were implanted in primary motor cortex of a human subject with tetraplegia, and single-cell activity was used to enable control of a computer cursor in two dimensions. While this was an important study, in terms of proving that results obtained by invasive BCIs on non-human primates transfer to human subject, the limited functionality of the BCI and questionable benefit to the human subject raises serious ethical concerns.

1.2.2 Non-invasive Approaches

Contrary to invasive BCIs, non-invasive approaches can not record single-cell activity but measure neural mass action of many hundred-thousands of neurons. This aggravates the direct decoding of motor plans, since tuning characteristics of single neurons can not be utilized for inference. While there is some evidence that the electrical field of the brain does provide detailed information on kinematic parameters [SKM+07], all currently employed non-invasive BCIs are based on experimental paradigms: specific thoughts are carried out by subjects to express certain intentions. Non-invasive BCIs can thus be characterized by the experimental paradigm that is employed. Typically, one research group focuses on only one type of paradigm, although there are exceptions to this rule. Subsequently, the work of some of the most influential groups working on non-invasive BCIs is presented. A more comprehensive review of work on non-invasive BCIs is given in [WBM+02].

Laboratory of E. Donchin, University of South Florida

The name of E. Donchin is associated with the P300, a positive deflection in the EEG measured over parietal areas that occurs approximately 300 ms after an infre-

quent stimulus. Building upon the P300, Donchin et al. were the first to realize a non-invasive BCI in 1988 [FD88]. They arranged the letters of the alphabet (and some additional symbols) in a 6x6 matrix and consecutively flashed random rows or columns of this matrix. By concentrating on a certain letter subjects could spell words, since only flashing of those rows and columns including the letter the subject concentrated on would elicit a P300. This basic principle still serves as the experimental paradigm of many recent BCIs, with most research directed at improving detection of a P300 (cf. [RGMA05] and [SYTI05]). Non-invasive BCIs based on the P300 do not require any subject training and are especially suited for spelling devices in which one out of many symbols has to be selected. However, they are only of limited use for control of an end-effector such as a computer cursor or a robotic device.

Institute of Medical Psychology and Behavioral Neurobiology, Eberhard Karls Universität Tübingen

The group of N. Birbaumer, head of the Institute of Medical Psychology and Behavioral Neurobiology at the Eberhard Karls Universität Tübingen, is another pioneering group of research on non-invasive BCIs. Their so called ”Thought-Translation-Device” is based on slow cortical potentials (SCPs), i.e., the DC electric potential on the scalp. SCPs can be intentionally modulated by subjects, which can be used to answer simple yes/no questions or write short sentences [BGH$^+$99]. The significance of the work of Birbaumer et al. is the fact that their BCI was the first to be operated by subjects with amyotrophic lateral sclerosis (ALS), thereby providing the first proof of principle that BCIs are indeed suited for paralyzed subjects. A drawback of using SCPs for communication is the extensive training time of several months necessary to master this paradigm. As a consequence, modulating SCPs has been widely discarded as a suitable paradigm for non-invasive BCIs.

Wadsworth Center, New York State Department of Health

The group of J. Wolpaw at the Wadsworth Center, New York State Department of Health, proposed a BCI similar in principle to the one of Birbaumer et al. in 1991 [WMNF91]. Instead of using SCPs, they trained their subjects to modulate the strength of the EEG mu-rhythm, i.e., the variance of the EEG signal in the 8-12 Hz frequency range. Over the period of several weeks of training healthy subjects thereby gained control over a cursor in one dimension. In 2004, Wolpaw et al. published results on two-dimensional cursor control, also using modulations of EEG rhythms [WM04]. The significance of this work is that it was the first study to show that subjects could achieve independent volitional control over different EEG rhythms. By independently modulating the variance of the mu- (8-12 Hz) and beta-rhythm (approximately 18-25 Hz), subjects could use one frequency band for horizontal and the other for vertical cursor control. So far this study remains the

only one to demonstrate two-dimensional cursor control by means of a non-invasive BCI. One reason for this is the intensive training of up to 170 hours required by subjects to learn modulating their EEG rhythms. This prolonged training might be due to the fact that Wolpaw et al. could not provide instructions for subjects how to control their EEG-rhythms. Instead, each subject had to explore different strategies to discover suitable ones. Not surprisingly, this led to different subjects utilizing different strategies.

Laboratory of Brain-Computer Interfaces, Technische Universität Graz

In 1997, G. Pfurtscheller et al. published a seminal study on non-invasive BCIs also utilizing volitional modulation of EEG rhythms [PNFP97]. While they did not obtain better results than Wolpaw et al. in 1991 [WMNF91], the significance of their work was that they provided specific instructions how to modulate the EEG mu-rhythm. They instructed subjects to perform haptic imagination of left and right hand movements, and showed that haptic motor imagery of one hand resulted in a decrease in variance in the EEG mu-rhythm measured over the contralateral motor cortex. While in the study of Wolpaw et al. extensive training was required for subjects to gain control over their EEG-rhythms, the use of haptic motor imagery almost eliminated the need for subject training. This was not the only important contribution of Pfurtscheller's group to non-invasive BCIs. Another seminal study introduced the concept of optimal spatial filtering to non-invasive BCIs [RMGP00]. They showed how to combine measurements of the electric field at different scalp locations to extract those components of the EEG suitable for inferring the subject's intention, thereby significantly improving classification accuracies.

1.3 Contributions and Outline of this Thesis

The work presented in this thesis only concerns non-invasive BCIs. In this context, the main obstacle to a significant increase in ITR is identified as the lack of sophisticated methods for feature extraction. Consequently, most of this thesis deals with the development of algorithms for feature extraction in the context of non-invasive BCIs.

Before these algorithms can be presented, it is necessary to establish a framework for the analysis and evaluation of BCIs. This is done in Chapter 2, in which it is argued that BCIs constitute communication channels that can be investigated with the powerful tools provided by information theory as initiated by C. Shannon in 1948 [Sha48]. After introducing the framework of BCIs as communication channels in Section 2.1, Sections 2.2 and 2.3 discuss how to measure the performance of BCIs and address a common misconception about the meaning of ITR in BCIs. This leads to a discussion why feature extraction is of central importance to increasing the ITR of BCIs in Section 2.4. The rest of Chapter 2 addresses the control of an unstable dynamic system solely by use of a BCI (Section 2.5).

Chapters 3 - 6 cover the main contributions of this thesis. Three new algorithms for feature extraction in non-invasive BCIs are presented and compared with each other as well as with existing approaches. The experimental evaluation of these algorithms is largely carried out using signals recorded by EEG. This is done for the simple reason that of all non-invasive recording modalities for brain signals EEG is the most readily available. In terms of a future widespread dissemination of BCIs, EEG is thus the current method of choice. It should be pointed out, however, that all algorithms presented here are not limited to EEG, and can be adapted to other modalities with relative ease.

In Chapter 3, the feasibility of source localization for feature extraction in non-invasive BCIs is investigated. It is shown that it is possible to infer whether a subject is performing imaginary tapping movements of the left or the right index finger from estimates of the current density in left and right motor cortex. Estimates of the current density are obtained by performing Independent Component Analysis (ICA) on the available data, and localizing the sources of the obtained independent components (ICs) by single current dipoles in a four-shell spherical head model. Since ICA can not separate multiple Gaussian sources, a new procedure is derived that identifies correctly reconstructed (non-Gaussian) sources, and incorrectly reconstructed (Gaussian) noise.

Chapter 4 develops a supervised method for feature extraction using concepts of information theory. A procedure for spatial filtering is designed that extracts those components of the recorded EEG data that provide a maximum of information on the BCI-user's intention. This is achieved by deriving an analytic approximation of mutual information of class labels, i.e., BCI-user's intention and extracted EEG components under assumptions valid in the context of non-invasive BCIs. Using this approximation, it is shown that Common Spatial Patterns (CSP), an algorithm frequently used for feature extraction in BCIs, is optimal in terms of maximizing (an approximation of) mutual information for two-class paradigms but not for multi-class paradigms. The approximation of mutual information is then used to derive a procedure for spatial filtering, termed multi-class Information Theoretic Feature Extraction (ITFE), that is optimal in terms of maximizing mutual information for multi-class paradigms. Multi-class ITFE is then applied to experimental data from a motor imagery paradigm, and is shown to perform superior to multi-class CSP.

In Chapter 5, ICA is investigated in more detail in the context of EEG/MEG analysis and non-invasive BCIs. It is argued that the mixing model usually assumed in complete ICA, i.e., assuming an equal number of sensors and sources, is unrealistic in the context of EEG/MEG analysis. This serves as the motivation for a theoretical investigation of the behavior of complete ICA for arbitrary mixture models, i.e., including overcomplete mixture models with more sources than sensors. Necessary and sufficient conditions for separability and identifiability of complete ICA for arbitrary mixture models are derived. These results serve to argue that in EEG/MEG analysis a mixture model with more sources than sensors but less non-Gaussian sources than sensors should be assumed. The implications of this mixture model

for EEG/MEG analysis by ICA are discussed, and testable predictions are formulated. A new approach for improving the SNR of ICA in EEG/MEG analysis based on linearly constrained minimum variance (LCMV) spatial filtering is presented, and used to validate the predictions resulting from the proposed mixture model. This new method is then applied to feature extraction in the context of BCIs based on motor imagery paradigms, and used to provide an explanation for the success of complete ICA in EEG/MEG analysis in spite of an overcomplete mixture model.
In Chapter 6, an unsupervised method for feature extraction is developed. Spatial filters are derived that optimally extract all EEG sources that originate in a chosen region of interest within the brain. By utilizing neuro-physiological a-priori knowledge these regions of interest can be chosen to correspond to those locations within the brain that provide most information on the BCI-user's intention for a given paradigm. This concept, similar in spirit to beamforming in array signal processing, leads to very robust feature extraction, since all artifacts that do not originate in the chosen regions of interest are optimally attenuated. The efficacy of the proposed method is demonstrated on experimental data from a two-class motor imagery paradigm. It is shown that it outperforms established algorithms for feature extraction, and that it reduces the amount of required training data. Furthermore, an online implementation of this algorithm is presented that allows real-time control of a cursor in one dimension.
In the final Chapter 7 the relevance of the contributions of this thesis are discussed, and directions for future research are delineated. Section 7.1 provides a critical evaluation of the capabilities and limitations of the algorithms presented in Chapters 3 - 6. Future research directions addressing these limitations are delineated in Section 7.2. In Section 7.3, a framework for discovering the effective connectivity structure within the brain, termed Network Information Transfer Analysis (NITA), is proposed, and implications of this framework for feature extraction in non-invasive BCIs are discussed. In the final Section 7.4 of this thesis, the question of causal relevance of the electric field of the brain, as measured by EEG, is discussed. The prevalent belief that the electric field of the brain is an epiphenomenon, i.e., does not play a role in information processing within the brain, is criticized. Finally, an approach is delineated to investigate the relevance of the electric field of the brain for information processing within the brain building upon the framework of NITA.

Chapter 2

Information Transfer in Brain-Computer Interfaces

In this chapter, the framework of information- and statistical learning theory for the analysis of BCIs is introduced. This serves several purposes. First, it establishes the background for theoretical work presented in later chapters. While most information theoretic concepts can be applied to BCIs in a straightforward manner, there are some assumptions inherent to classical information theory that are not applicable in this context. These differences have to be taken into account when applying information theoretic concepts to BCIs in order to avoid drawing incorrect conclusions. However, the primary purpose of this chapter is to present a conclusive argument why feature extraction constitutes the main bottleneck in the performance of BCIs. This argument is carried out in the framework of information- and statistical learning theory, and results in a mathematical definition of the main objective of this thesis (Definition 2.13).
Basic knowledge of the concepts of information theory, e.g., as presented in [CT06], is assumed. In Section 2.1, BCIs are modeled as memoryless communication channels. This serves as the basis for Sections 2.2 and 2.3, in which the problem of measuring the performance of BCIs and a common misconception about the meaning of the information transfer rate (ITR) in BCIs are addressed. In Section 2.4, it is argued that feature extraction constitutes the main challenge in developing high-performance BCIs. This section thereby serves as the theoretical motivation for Chapters 3 - 6. The chapter concludes with a discussion of the control of unstable dynamic systems solely by use of a BCI in Section 2.5.

2.1 The BCI Communication Channel

A communication channel is a description of a process that transmits information. A model of a communication channel usually consists of the three components shown in Fig. 2.1. The central element in each model of communication is the channel

Figure 2.1: A communication channel.

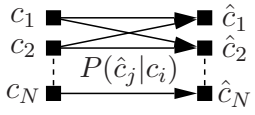

Figure 2.2: Graph representation of a discrete memoryless communication channel.

itself, i.e., the medium over which the information is to be transmitted. Here, only the discrete memoryless channel is considered. This implies that the information consists of symbols c that take on values in a finite set $\mathcal{C}$, and that the output of the channel $\hat{c} \in \mathcal{C}$ does not depend on past inputs to or outputs of the channel. The communication channel can then be described by the graph depicted in Fig. 2.2. Each arrow in Fig. 2.2 has an associated conditional probability $P(\hat{c} = c_j|c = c_i)$, that describes the probability of receiving symbol c_j given symbol c_i has been sent. The expression $P(\hat{c} = c_j|c = c_i)$ is subsequently abbreviated as $P(\hat{c}_j|c_i)$. The actual channel is complemented by an encoding and a decoding procedure that serve two purposes. The first purpose is to map the input symbols in the set $\mathcal{C}$ into a set $\mathcal{Y}$, which consists of symbols that can be send over the channel. The channel then answers to each transmitted symbol in $\mathcal{Y}$ with a received symbol in $\mathcal{X}$. In the decoding procedure, the received symbols in $\mathcal{X}$ are then mapped back to $\mathcal{C}$. Note that the set of received symbols $\mathcal{X}$ does not have to coincide with the set of transmitted symbols $\mathcal{Y}$, and that $\mathcal{X}$ and $\mathcal{Y}$ may or may not coincide with $\mathcal{C}$. The second purpose of the encoding/decoding procedure is to minimize the probability of receiving an incorrect symbol while maximizing the number of symbols sent over the channel. This problem is discussed in Section 2.3.

In the context of BCIs, the symbols $c \in \mathcal{C}$ transmitted over the channel are the BCI-user's intentions. The set $\mathcal{C}$ hence consists of the possible intentions the user can choose from. The actual channel of the BCI is the brain itself, i.e., the central nervous system (CNS). As a consequence, the encoding procedure in BCIs is represented by the experimental paradigm. The paradigm determines which thoughts should be carried out by the user to express a certain intention, thereby mapping the user's intention $c \in \mathcal{C}$ into a not further specified set $\mathcal{Y}$. The CNS then answers to each intention expressed through the experimental paradigm with a symbol $x \in \mathcal{X}$. The set $\mathcal{X}$ represents all possible signals that can be recorded from the CNS, e.g., the electric field of the brain as measured by EEG. In the decoding procedure, the received symbol is then used to reconstruct the user's intention. This model of a BCI as a communication channel is summarized in Fig. 2.3.

2.2 Measuring Performance of BCIs

There are two principal ways of measuring the performance of a communication channel. The first is the error probability of the channel, defined as follows:

Definition 2.1 (Error probability)**.** *The error probability of a communication channel with input $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ and output $\hat{c} \in \mathcal{C}$ is defined as*

$$P_e := \sum_{i=1}^{N} P(c_i)\left(1 - P(\hat{c}_i|c_i)\right), \tag{2.1}$$

with $P(c_i)$ the prior probability of symbol c_i and $P(\hat{c}_i|c_i)$ the probability of receiving symbol $\hat{c}_i$ if symbol c_i was transmitted.

The error probability hence equals the average probability of receiving an incorrect symbol. In the context of BCIs, it is desirable to minimize the error probability in order to minimize the instances in which the BCI does not react according to the user's intention.
The second performance measure is the mutual information.

Definition 2.2 (Mutual information)**.** *The mutual information of the input $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ and the output $\hat{c} \in \hat{\mathcal{C}} = \{\hat{c}_1, \ldots, \hat{c}_M\}$ of a communication channel is defined as*

$$I(c, \hat{c}) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(c_i, \hat{c}_j) \log \frac{P(c_i, \hat{c}_j)}{P(c_i)P(\hat{c}_j)}, \tag{2.2}$$

with $P(c_i, \hat{c}_j)$ the probability of jointly observing input/output symbols c_i and $\hat{c}_j$, and $P(c_i)$ and $P(\hat{c}_j)$ the marginal probabilities of symbols c_i and $\hat{c}_j$.

Mutual information can also be expressed in terms of the (class-conditional) Shannon entropy as $I(c, \hat{c}) = H(c) - H(c|\hat{c}) = H(\hat{c}) - H(\hat{c}|c)$ (cf. [CT06]). Note that while the definition of error probability requires the input and output of the channel to take values in the same set, mutual information can be computed for random variables that take values in arbitrary sets. In terms of generality, it is hence beneficial to also consider $\hat{c} \in \hat{\mathcal{C}} \neq \mathcal{C}$. The significance of mutual information as a performance measure for communication channels is due to the famous channel coding theorem of C. Shannon [Sha48], which states that the mutual information corresponds to the maximum number of bits that can be send on average over a channel with arbitrarily

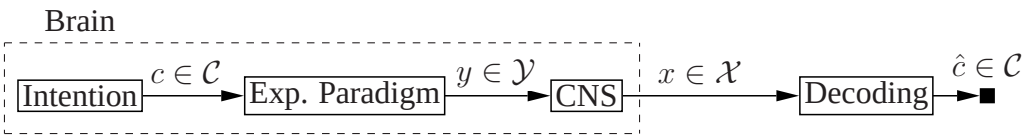

Figure 2.3: A BCI communication channel.

low error probability. This gives rise to the channel capacity of a communication channel and the concept of information transfer rate (ITR), which is discussed in the context of BCIs in Section 2.3. There are, however, more reasons to utilize mutual information as a performance measure for BCIs, or for communication channels in general, which are discussed now.

Mutual Information and the Minimum Bayes Error

First, mutual information provides upper and lower bounds on the minimum Bayes error.

Definition 2.3 (Minimum Bayes error)**.** *Let $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ and $\hat{c} \in \hat{\mathcal{C}} = \{\hat{c}_1, \ldots, \hat{c}_M\}$. The minimum Bayes error in estimating c from $\hat{c}$ is defined as*

$$P_{Bayes} := \sum_{j=1}^{M} P(\hat{c}_j) \left(1 - \max_{i \in \{1,\ldots,N\}} \{P(c_i|\hat{c}_j)\}\right). \tag{2.3}$$

The minimum Bayes error is the average probability of incorrectly inferring the transmitted symbol if always that symbol is selected that is the most probable one given the observed output of the channel. This constitutes an induction principle in machine learning. The minimal achievable average probability in inferring the value of one random variable from observation of another random variable is defined as the error that is obtained if the optimal Bayes classifier is employed.

Definition 2.4 (Optimal Bayes classifier)**.** *Let $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ and $\hat{c} \in \hat{\mathcal{C}} = \{\hat{c}_1, \ldots, \hat{c}_M\}$. The optimal Bayes classifier for inferring the value of c from observing $\hat{c}$ is defined as*

$$g_{Bayes}(\hat{c}) := \operatorname*{argmax}_{c \in \mathcal{C}} \{P(c|\hat{c})\}. \tag{2.4}$$

By construction, the optimal Bayes classifier achieves the minimum Bayes error.
As it is easy to see, the minimum Bayes error coincides with the error probability as defined in (2.1) if $\hat{\mathcal{C}} = \mathcal{C}$ and for each $c_i,\ i = 1, \ldots, N$ it holds that $P(\hat{c}_i|c_i) \geq P(\hat{c}_j|c_i)$ for all $j = 1, \ldots, N$ and $j \neq i$. If these conditions do not hold, the error probability may exceed the minimum Bayes error.
A lower bound on the minimum Bayes error in terms of mutual information was first given by R.M. Fano in his class notes on information theory in 1952.

Theorem 2.1 (Fano's inequality)**.** *Let $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ and $\hat{c} \in \hat{\mathcal{C}} = \{\hat{c}_1, \ldots, \hat{c}_M\}$. Then for the minimal Bayes error of estimating c from observation of $\hat{c}$ the following inequality holds:*

$$P_{Bayes} \geq \frac{H(c|\hat{c}) - H(P_{Bayes})}{\log |\mathcal{C}|} \geq \frac{H(c|\hat{c}) - 1}{\log |\mathcal{C}|} = \frac{H(c) - I(c, \hat{c}) - 1}{\log N}, \tag{2.5}$$

with $|\mathcal{C}|$ the number of elements in $\mathcal{C}$. If $\mathcal{C} = \hat{\mathcal{C}}$ the inequality can be further strengthened by replacing $\log N$ in the denominator by $\log(N-1)$.

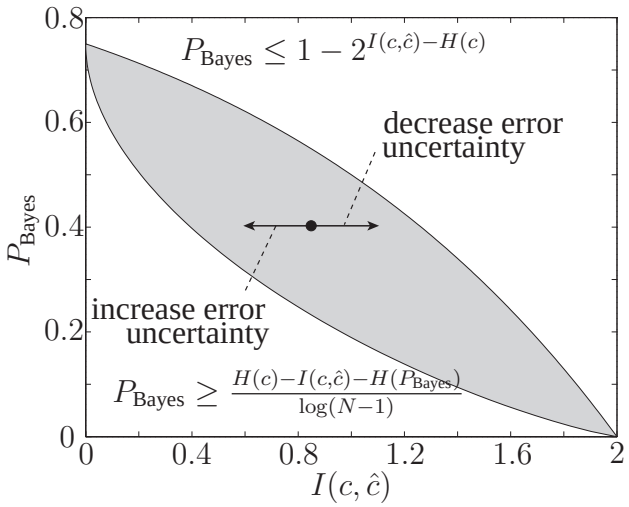

Figure 2.4: Relation of minimum Bayes error and mutual information.

A proof of Fano's inequality can be found in [CT06]. Fano's inequality is tight, i.e., there are probability distributions on c and $\hat{c}$ for which equality holds in (2.5). Note that tightness does not imply that equality in (2.5) holds for every distribution on c and $\hat{c}$.

An upper bound on the minimum Bayes error in terms of mutual information is given by Feder and Merhav in [FM94].

Theorem 2.2 (Feder & Merhav)**.** *Let $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ and $\hat{c} \in \hat{\mathcal{C}} = \{\hat{c}_1, \ldots, \hat{c}_M\}$. Then for the minimal Bayes error of estimating c from observation of $\hat{c}$ the following inequality holds:*

$$P_{Bayes} \leq 1 - 2^{I(c,\hat{c})-H(c)}. \tag{2.6}$$

Contrary to Fano's inequality, this bound is only tight at certain points.

Since $H(c)$ is constant, the two bounds (2.5) and (2.6) imply that maximizing mutual information of c and $\hat{c}$ minimizes the minimum Bayes error. Furthermore, $P_{\text{Bayes}} = 0$ if and only if $I(c, \hat{c}) = H(c)$, i.e., if the mutual information of c and $\hat{c}$ equals the entropy of c. The relationship of the minimum Bayes error and mutual information is illustrated in Fig. 2.4 for $\mathcal{C} = \hat{\mathcal{C}} = \{c_1, \ldots, c_4\}$ and $P(c) = 1/4$, with the area outside the shaded region corresponding to impossible combinations of minimum Bayes error and mutual information.

In summary, mutual information can be used, with some limitations, as a substitute for error probability. While this certainly is an interesting feature, it is unclear so far why mutual information should be used instead of or in addition to error probability. This is addressed next.

Mutual Information and Error Entropy

Since the mapping between mutual information and the minimum Bayes error is not one-to-one, it is instructive to investigate what gives rise to this ambiguity. Unfortunately, this is a difficult and so far poorly understood problem. Here, only the influence of the error uncertainty on the relation of minimum Bayes error and mutual information is discussed, and used to motivate the use of mutual information as a performance measure for BCIs.

Both, mutual information and minimum Bayes error, are fully determined by the probability distribution $P(c, \hat{c})$. Thus, any change in minimum Bayes error or mutual information has to be reflected in $P(c, \hat{c})$. As it is obvious from Fig. 2.4, $P(c, \hat{c})$ can be varied in order to alter mutual information while keeping the minimum Bayes error constant. The key to understanding why this is possible is the definition of the minimum Bayes error. Again, let $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ be the input and $\hat{c} \in \hat{\mathcal{C}} = \{\hat{c}_1, \ldots, \hat{c}_M\}$ the output of the communication channel, and $g_{\text{Bayes}}(\hat{c})$ the optimal Bayes classifier as defined in (2.4) for a given distribution $P(c, \hat{c})$. The minimum Bayes error can then be written as

$$\begin{aligned} P_{\text{Bayes}} &= \sum_{j=1}^{M} P(\hat{c}_j)(1 - \max_{i \in \{1,\ldots,N\}} \{P(c_i|\hat{c}_j)\}) \\ &= 1 - \sum_{j=1}^{M} P(\hat{c}_j) \max_{i \in \{1,\ldots,N\}} \{P(c_i|\hat{c}_j)\} \\ &= 1 - \sum_{j=1}^{M} P(\hat{c}_j) P(g_{\text{Bayes}}(\hat{c}_j)|\hat{c}_j) \\ &= 1 - \sum_{j=1}^{M} P(g_{\text{Bayes}}(\hat{c}_j), \hat{c}_j). \end{aligned} \tag{2.7}$$

As a consequence, those M elements of $P(c, \hat{c})$ that are indexed by $g_{\text{Bayes}}(\hat{c}_j)$ with $j = 1, \ldots, M$ fully determine the minimum Bayes error. Since $P(c, \hat{c})$ has a total of MN elements, $M(N-1)$ elements can be varied freely to alter the mutual information $I(c, \hat{c})$. Then note that mutual information can be written as [CT06]

$$I(c, \hat{c}) = H(c) - H(c|\hat{c}) = H(c) + H(\hat{c}) - H(c, \hat{c}). \tag{2.8}$$

Now define $\delta(\hat{c}) := \text{argmax}_{i \in \{1,\ldots,N\}} \{P(c_i, \hat{c})\}$, i.e., the index of the input symbol decoded by the minimum Bayes classifier for each output symbol. The joint entropy

of c and $\hat{c}$ can then be further decomposed into

$$\begin{aligned} H(c,\hat{c}) &= -\sum_{j=1}^{M}\sum_{i=1}^{N} P(c_i,\hat{c}_j)\log P(c_i,\hat{c}_j) \\ &= \underbrace{-\sum_{j=1}^{M}\sum_{i=1;i\neq\delta(\hat{c}_j)}^{N} P(c_i,\hat{c}_j)\log P(c_i,\hat{c}_j)}_{=:\tilde{H}_{\text{Error}}(c,\hat{c})} \\ &\quad \underbrace{-\sum_{j=1}^{M} P(g_{\text{Bayes}}(\hat{c}_j),\hat{c}_j)\log P(g_{\text{Bayes}}(\hat{c}_j),\hat{c}_j)}_{=:\tilde{H}_{\text{Bayes}}(c,\hat{c})}. \end{aligned} \tag{2.9}$$

Here, the term $\tilde{H}_{\text{Bayes}}(c,\hat{c})$ contains the elements of $P(c,\hat{c})$ that determine the minimum Bayes error and $\tilde{H}_{\text{Error}}(c,\hat{c})$ all other elements. Consequently, $\tilde{H}_{\text{Bayes}}(c,\hat{c})$ is a measure related to the entropy of the correctly classified symbols if the optimal Bayes classifier is used, and $\tilde{H}_{\text{Error}}(c,\hat{c})$ is a measure related to the error entropy, i.e., the uncertainty which type of error is being made. Note that both expressions are not real entropies since their probabilities do not add up to one. It is now assumed that the elements of $P(c,\hat{c})$ that determine $\tilde{H}_{\text{Bayes}}(c,\hat{c})$ are fixed, which implies that the minimum Bayes error is also held constant. If then the measure of error entropy $\tilde{H}_{\text{Error}}(c,\hat{c})$ is decreased while keeping $H(c)$ and $H(\hat{c})$ constant, this leads to an increase in mutual information $I(c,\hat{c})$ due to (2.8) and (2.9). The converse holds if $\tilde{H}_{\text{Error}}(c,\hat{c})$ is increased, i.e., this leads to a decrease in mutual information. This relation is indicated by the arrows in Fig. 2.4. The uncertainty which type of error is being made thus influences mutual information, with high mutual information correlating with low error uncertainty. This is further illustrated in the following example.

Example 2.1. *Consider two different BCIs a) and b) (Fig. 2.5) with input/output symbols* $c,\hat{c}\in\mathcal{C}=\{c_1,\ldots,c_4\}$*. For BCI a), let* $P(c_i,\hat{c}_i)=3/15$ *and* $P(c_i,\hat{c}_{j\neq i})=1/60$*, i.e., the probability of jointly observing the same input and output symbol equals* $3/15$ *for all symbols, and the joint probability of observing different input and output symbols equals* $1/60$ *for all combinations of symbols. This leads to an error probability of* $P_e=0.2$ *and a mutual information of* $I(c,\hat{c})=0.96$ *bits. Now consider BCI b). Here, the probability of jointly observing the same input and output symbol also equals* $3/15$*. As a result, the error probability of BCI b) is the same as that of BCI a):* $P_e=0.2$*. The joint probability of observing different input and output symbols however is not equal for all combinations of symbols. Instead, this probability is* $1/20$ *for combinations* $\{c_1,c_2\},\{c_2,c_1\},\{c_3,c_4\},\{c_4,c_3\}$*, and zero for all other symbol combinations (indicated by the missing arrows in Fig. 2.5). This constitutes a decrease in the error uncertainty, since each symbol can only be decoded incorrectly in one way. As a result, the mutual information of BCI b) equals*

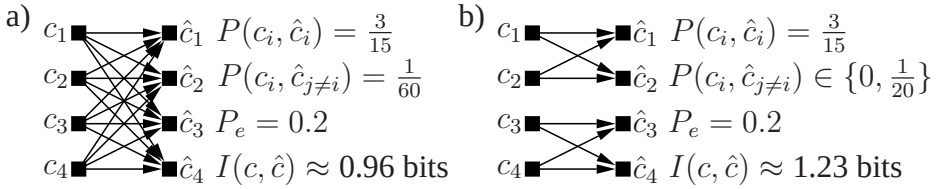

Figure 2.5: Two BCIs with equal error probability but a) lower and b) higher mutual information.

$I(c, \hat{c}) = 1.23$ *bits and hence exceeds the mutual information of BCI a) in spite of equal error probability.*

The relation of error entropy and mutual information is of high significance for BCIs. Consider again the two BCIs in Fig. 2.5. If these are used for control of a hand prosthesis, input symbols c_1 and c_2 could be used for moving the hand to the left or right, and input symbols c_3 and c_4 could be used for opening and closing the hand. If BCI a) is used for control of the hand, each type of error can occur. For example, instead of moving the hand to the left the user of the BCI might unintentionally open the hand and thus drop a previously picked up object. This type of error is not possible when using BCI b) for control of the neuro-prosthesis. In BCI b), the two sets of input symbols $\{c_1, c_2\}$ and $\{c_3, c_4\}$ are decoupled. As a consequence, errors can only occur within one set. Accidentally opening instead of moving the hand can not occur.
In summary, the exact relation of mutual information and minimum Bayes error is largely not yet understood. Nevertheless, a high mutual information of a BCI is desirable not only because of the relation to the minimum Bayes error, but also due to the relation to error entropy. Mutual information thus provides a measure for the performance of BCIs that should be used in addition to error probability.

Mutual Information of Random Variables from Arbitrary Sets

One further benefit of mutual information is that it can be computed for random variables from different sets. While at first glance this does not seem significant, it does provide an important advantage in comparison to the error probability defined in (2.1). As illustrated in Fig. 2.3, information transmission in BCIs is not confined to one set. Instead, at different stages of the information transmission process the BCI-user's intention is encoded in variables that take values in different sets. If the error probability is used to measure performance of a BCI only variables that take values in the same set can be evaluated. As a direct consequence, the BCI can be evaluated only as a whole. Mutual information, on the other hand, allows, at least in principle, to measure the performance of different components of a BCI by estimating the mutual information of the input to and output of a component. This enables the analysis and optimization of different components of a BCI independently of other components. While in principle this also holds true for the minimum Bayes

error, mutual information is in general easier to estimate. Intensive use of this property of mutual information is made in Chapter 4.

2.3 Channel Capacity and Information Transfer Rate

One performance measure frequently used in the BCI literature is the so called information transfer rate (ITR) briefly mentioned in Section 2.2. In this section, the ITR is discussed in more detail, and it is shown that in the context of BCIs it does not have the meaning usually attributed to it.
The ITR is defined as follows [WBH$^+$00].

Definition 2.5. *Let $c, \hat{c} \in \mathcal{C} = \{c_1, \ldots, c_N\}$ the input and output of a BCI communication channel. Furthermore, let P_e the error probability of the BCI as defined in (2.1). The information transfer rate is then defined as*

$$\textit{ITR}(c, \hat{c}) := \log N + P_e \log \frac{P_e}{N-1} + (1 - P_e)\log(1 - P_e). \tag{2.10}$$

It is easy to show that the ITR equals the mutual information $I(c, \hat{c})$ iff $P(c) = 1/N$, the error probability for each transmitted symbol is equal, and each possible error is equally likely. The relation of ITR and mutual information can be rendered more precise by the following theorem.

Theorem 2.3. *Let $c, \hat{c} \in \mathcal{C} = \{c_1, \ldots, c_N\}$ the input and output of a BCI communication channel. Furthermore, let P_e the error probability of the BCI as defined in (2.1), $P_{\textit{Bayes}}$ the minimum Bayes error as defined in (2.3), and let $P_e = P_{\textit{Bayes}}$. Then the ITR as defined in (2.10) constitutes a lower bound on the mutual information of the output and input of a communication channel, i.e.,*

$$I(c, \hat{c}) \geq \textit{ITR}(c, \hat{c}). \tag{2.11}$$

Proof. Recollect Fano's inequality in (2.5) for input and output of a channel taking values in the same set,

$$P_{\text{Bayes}} \geq \frac{H(c|\hat{c}) - H(P_{\text{Bayes}})}{\log(N-1)}. \tag{2.12}$$

Rearranging and using $P_e = P_{\text{Bayes}}$ results in

$$I(c, \hat{c}) \geq H(c) - H(P_e) - P_e \log(N-1). \tag{2.13}$$

Then note that for $P(c) = 1/N$ the entropy $H(c) = \log N$, and that $H(P_e) = -P_e \log P_e - (1 - P_e)\log(1 - P_e)$. Equation (2.13) then becomes

$$I(c, \hat{c}) \geq \log N + P_e \log \frac{P_e}{N-1} + (1 - P_e)\log(1 - P_e) = \text{ITR}(c, \hat{c}). \tag{2.14}$$

□

The use of ITR as a performance measure for BCIs hence derives from the fact that it provides a lower bound on the mutual information of the input and output of a BCI that, contrary to the actual mutual information, can be easily estimated. Since the mutual information equals the maximum number of bits that can be send on average over a communication channel with arbitrarily low error probability, the ITR is taken to provide a conservative measure of how much information can be send over a BCI. This is incorrect, as is shown now.
The basis of this argument is the famous channel coding theorem of C. Shannon [Sha48]. For a discussion of this theorem the following definitions (adapted from [CT06]) are required.

Definition 2.6 (Code)**.** *Consider a communication channel depicted in Fig. 2.1 with input c and output $\hat{c}$ with $c, \hat{c} \in \mathcal{C} = \{c_1, \ldots, c_N\}$ and a given probability mass function $P(c, \hat{c})$. A (N, n) code for this channel consists of*

1. *An encoding function $h_{enc}^{(n)} : \mathcal{C} \to \mathcal{Y}^{(n)}$ that maps each input symbol in $\mathcal{C}$ into a sequence of length n in $\mathcal{Y}$.*

2. *A decoding function $h_{dec}^{(n)} : \mathcal{X}^{(n)} \to \mathcal{C}$ that maps each sequence of n symbols in $\mathcal{X}$ into $\mathcal{C}$.*

In a communication channel with an encoding and decoding procedure the information is thus not directly transmitted over the channel. Instead, a sequence of n symbols in $\mathcal{Y}$ is sent over the channel for each input symbol in $\mathcal{C}$, and the corresponding sequence at the output of the channel in $\mathcal{X}$ is used to infer the original transmitted symbol in $\mathcal{C}$. Note that $\mathcal{C}$, $\mathcal{Y}$ and $\mathcal{X}$ may or may not coincide.

Definition 2.7 (Maximum error probability)**.** *The maximum error probability for a (N, n) code is defined as*

$$\lambda^{(n)} := \max_{i \in \{1,\ldots,N\}} \left\{ Pr\left(h_{dec}^{(n)}\left(x^{(n)}\right) \neq c_i | h_{enc}^{(n)}(c_i) \right) \right\}. \tag{2.15}$$

Definition 2.8 (Rate)**.** *The rate of a (N, n) code is defined as*

$$R := \frac{\log N}{n}. \tag{2.16}$$

The rate specifies the average number of bits per transmission that carry useful information, i.e., information that is to be transmitted over the channel.

Definition 2.9 (Achievable rates)**.** *A rate R is said to be achievable if there exists a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} \lambda^{(n)} = 0$.*

Definition 2.10 (Channel capacity)**.** *The channel capacity of a discrete memoryless channel is defined as*

$$C := \max_{P(y)} \{I(y, x)\}. \tag{2.17}$$

Note that the channel capacity refers to $I(y, x)$, while the ITR provides a lower bound on $I(c, \hat{c})$. This, however, does not affect the following argument. With these definitions the channel coding theorem can be stated [CT06].

Theorem 2.4 (Channel coding theorem)**.** *For a discrete memoryless channel, all rates R below capacity C are achievable. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} \lambda^{(n)} = 0$ must have $R \leq C$.*

An accessible proof of the theorem is provided in [CT06]. Here, only the first part of the theorem is of interest. It asserts that for every rate below capacity as defined in (2.17) there exists a coding scheme that achieves an arbitrarily low maximum error probability. The channel coding theorem thereby also specifies what precisely is meant by the term information transfer.

Definition 2.11 (Information transfer)**.** *Information transfer is understood as transmitting data over a channel with arbitrarily low maximum error probability.*

The crucial part in the statement of the channel coding theorem is that arbitrarily low maximum error probability requires arbitrarily long codes, i.e., that n may go to infinity in Definition 2.9. In ordinary communication channels this seldom poses problems, since here long codes, at least in principle, only imply a delay in the data transmission. In BCIs, however, this is different. Consider again the structure of a BCI communication channel in Fig. 2.3. Here, the encoding procedure is implemented by the experimental paradigm. It thus has to be carried out within the brain, i.e., by the user of the BCI. While this might still be feasible for short and simple codes, increasing the code length and/or code complexity will soon exhaust the intellectual capabilities of any BCI-user. The channel coding theorem, however, only applies if arbitrarily long codes are permitted. As a direct consequence, the channel coding theorem does not apply to BCIs. For this reason, the ITR can not be interpreted as the amount of information that can be transmitted over a BCI.

The results of this section can be summarized as follows. The ITR provides a lower bound on the mutual information of a BCI which is easy to compute. Since in ordinary communication channels mutual information equals the maximum number of bits that can be send on average over a channel with arbitrarily low maximum error probability, ITR is often used in the BCI literature in a way that implies that it specifies a lower bound on the information that can be transmitted over a BCI. This is incorrect, since the channel coding theorem does not apply to BCIs. Hence, the ITR does not have any theoretically justifiable meaning in the context of BCIs, and it does not provide any information on the performance of the BCI that is not already provided by the error probability in conjunction with the number of actions the user of the BCI can choose from. Its only use is the combination of these two properties of a BCI into a single expression.

2.4 The Significance of Feature Extraction

While so far only the problem of measuring performance was addressed, this section discusses the problems that arise when actually designing a BCI in order to optimize performance measures. In this context, it is argued that the problem of feature extraction constitutes the main challenge in designing high-performance BCIs. This section thereby provides the theoretical justification for the work presented in Chapters 3 - 6.

Recollecting the structure of a BCI communication channel in Fig. 2.3, there are two components of a BCI that can be engineered to optimize performance. These are the experimental paradigm and the decoding procedure. The experimental paradigm controls how much information on the user's intention is contained in the data recorded from the CNS. This amount of information can be expressed in terms of the mutual information $I(c, x)$ and determines, via (2.5) and (2.6), upper and lower bounds on the minimum Bayes error that can be achieved in estimating c from x. For this reason, one goal in designing experimental paradigms is to maximize mutual information of the BCI-user's intention and the recorded data. For now, it is assumed that the experimental paradigm and the recording procedure are fixed, and a signal x is recorded with a certain mutual information $I(c, x)$. This signal is then used in the decoding procedure to infer the BCI-user's intention. Here, the goal is to optimize the decoding procedure in terms of a certain performance measure, e.g., the error probability or the mutual information of original intention c and inferred intention $\hat{c}$.

Learning the Optimal Bayes Classifier

Drawing from the discussion of the previous section, the minimum error that can be achieved in estimating c from x is the minimum Bayes error. Hence, it seems sensible to employ the optimal Bayes classifier to infer c from x. For $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ and $x \in \mathcal{X} = \{x_1, \ldots, x_M\}$ the optimal Bayes classifier is given by

$$g_{\text{Bayes}}(x) := \underset{c \in \mathcal{C}}{\operatorname{argmax}}\{P(c|x)\} = \underset{c \in \mathcal{C}}{\operatorname{argmax}}\{P(c, x)\}. \tag{2.18}$$

Constructing the optimal Bayes classifier thus requires knowledge of the unknown distribution $P(c, x)$. This raises the question how the optimal Bayes classifier can be constructed. Assuming a set of training data $\mathcal{S} = \{(c_1, x_L), \ldots, (c_L, x_L)\}$ with L samples drawn i.i.d. from $P(c, x)$ is available, one way to obtain the optimal Bayes classifier is the following procedure. First, the distribution $P(c, x)$ is estimated from $\mathcal{S}$ as

$$\hat{P}(c_i, x_j) = \frac{\sharp\mathcal{S}\{c = c_i \wedge x = x_j\}}{L} \tag{2.19}$$

for $i = 1, \ldots, N$, $j = 1, \ldots, M$, and $\sharp\mathcal{S}\{.\}$ the number of occurrences of the expression in the bracket in the training set. Almost sure convergence of this estimate

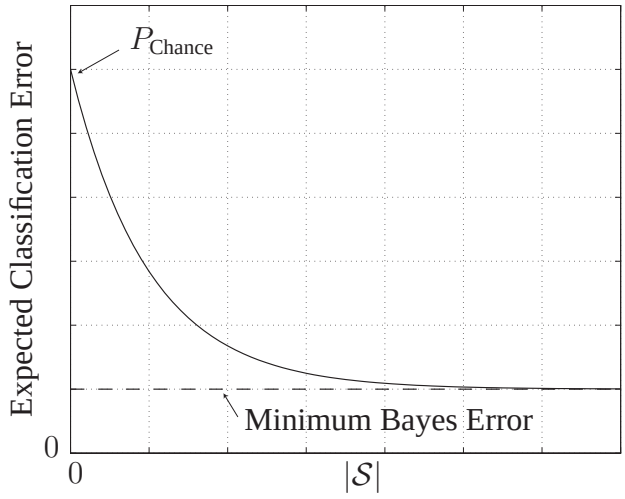

Figure 2.6: Illustration of the learning curve for the optimal Bayes classifier

to the real distribution $P(c, x)$ for $L \to \infty$ is guaranteed by the Bernoulli Theorem. An estimate of the optimal Bayes classifier is then constructed from $\hat{P}(c, x)$ as

$$\hat{g}_{\text{Bayes}}(x) := \underset{c \in \mathcal{C}}{\operatorname{argmax}}\{\hat{P}(c, x)\}. \tag{2.20}$$

The viability of this procedure depends on the amount of training data available. To see this, it is instructive to investigate the conditions under which the estimated optimal Bayes classifier and the true optimal Bayes classifier coincide, i.e., make the same decision for each $x \in \mathcal{X}$. Quite surprisingly, this does not require that $\hat{P}(c, x)$ is a good estimate of $P(c, x)$. Necessary and sufficient conditions for $g_{\text{Bayes}} \equiv \hat{g}_{\text{Bayes}}$ are that $\forall x \in \mathcal{X}$ it holds that

$$\underset{c \in \mathcal{C}}{\operatorname{argmax}}\{\hat{P}(c|x)\} = \underset{c \in \mathcal{C}}{\operatorname{argmax}}\{P(c|x)\}. \tag{2.21}$$

Upper and lower bounds on the probability that (2.21) holds for a certain $x \in \mathcal{X}$ can be calculated as a function of the amount of training data using Chernoff bounds. In general, the probability that (2.21) holds for a certain $x \in \mathcal{X}$ increases with the number of occurrences of x in $\mathcal{S}$. The elements in $\mathcal{X}$ for which (2.21) does not hold then determine by how much the error probability of the estimated Bayes classifier exceeds the minimum Bayes error. This gives rise to the learning curve, which illustrates the convergence of the expected classification error to the minimum Bayes error as a function of the size of the training set $\mathcal{S}$ (Fig. 2.6).
Now consider the set $\mathcal{X}$ which constitutes the feature space. As defined in Section 2.1, each element $x \in \mathcal{X}$ specifies one possible observation of recorded EEG data. Let T be the duration of the recorded data, f_s the sampling rate, d the quanitzation accuracy and N the number of electrodes. Then the number of elements in $\mathcal{X}$ equals $|\mathcal{X}| = d^{T \cdot f_s \cdot N}$. For example, assume that one second of EEG data is recorded from 128 channels at a sampling rate of 500 Hz and digitized with 16 bit. Then the

number of elements in $\mathcal{X}$ equals $|\mathcal{X}| = 16^{1\cdot500\cdot128}$. This is an incredibly large number. To make matters worse, constructing the optimal Bayes classifier in (2.20) from the training set $\mathcal{S}$ requires observing multiple instances of each element of $\mathcal{X}$ in $\mathcal{S}$ to obtain an estimate of (2.19) that fulfills the conditions in (2.21) with high probability. Accordingly, just recording the amount of training data necessary to obtain a sensible estimate of the optimal Bayes classifier in (2.18) is absolutely impossible. Conversely, for any practically feasible amount of training data the error probability of the estimated Bayes classifier will by far exceed the minimum Bayes error. It is thus apparent that the optimal Bayes classifier can not be directly applied to x to infer c.

In the above discussion only the optimal Bayes classifier for discrete feature spaces is considered. This restriction is made due to the fact that the optimal Bayes classifier is the theoretically optimal classifier. As such, it is especially well suited to illustrate important concepts. It can be argued that other classifiers, such as support vector machines or logistic regression, can be employed that display significantly higher rates of convergence than the optimal Bayes classifier. This is indeed correct, and such classifiers are extensively employed in later chapters. However, the above discussion is similar for other types of classification algorithms and continuous feature spaces. For example, if support vector machines are considered instead of the discrete optimal Bayes classifier, the above argument can be carried out by investigating the VC-dimension of the separating hyper-plane, and demonstrating the slow convergence of the empirical to the expected risk using distribution independent bounds [Vap98]. In summary, even the most advanced classification algorithms fail if they are applied to feature spaces as large as those discussed here.

Feature Extraction and the Rate of Convergence

The above discussion raises the question how the rate of convergence of the estimated Bayes classifier to the minimum Bayes error can be increased. In general, it is impossible to derive distribution independent bounds on the rate of convergence for the estimated Bayes classifier [DGL96]. This implies that, not surprisingly, the rate of convergence of the expected error probability to the minimum Bayes error depends on the properties of the distribution $P(c, x)$. Unfortunately, it is largely unknown exactly which properties of $P(c, x)$ influence the rate of convergence. The only obvious property that adversely affects the rate of convergence is $|\mathcal{X}|$, the size of the feature space. Given a fixed amount of training data, decreasing the size of the feature space leads to a better estimate of $\hat{P}(c, x)$, and thereby $\forall\epsilon > 0$ to a higher probability that the error probability of the estimated Bayes classifier does not exceed the minimum Bayes error by more than ϵ. It is thus desirable to find a transformation $T : \mathcal{X} \mapsto \hat{\mathcal{X}}$ with $|\hat{\mathcal{X}}| < |\mathcal{X}|$ and use $\hat{x} = T(x)$ instead of x to infer c. However, as the following theorem shows, not every T with $|\hat{\mathcal{X}}|$ fixed is equally suited.

Theorem 2.5 (Transformations of x can not decrease the minimum Bayes error)**.** *Let $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$, $x \in \mathcal{X} = \{x_1, \ldots, x_M\}$, and $P_{Bayes}^{(x \to c)}$ the minimum Bayes error in inferring c from x as defined in (2.18). Then for all transformations $T : \mathcal{X} \mapsto \hat{\mathcal{X}}$ it holds that $P_{Bayes}^{(T(x) \to c)} \geq P_{Bayes}^{(x \to c)}$.*

Proof. The proof of Theorem 2.5 is easiest to understand if $P(c, x)$ is seen as a matrix, with the rows corresponding to the N elements of $\mathcal{C}$ and the columns to the M elements of $\mathcal{X}$. The minimum Bayes error in estimating c from x is defined as

$$P_{\text{Bayes}}^{(x \to c)} := 1 - \sum_{j=1}^{M} \max_{i \in \{1,\ldots,N\}} \{P(c_i, x_j)\}, \tag{2.22}$$

i.e., as the error that is obtained if for each column of $P(c, x)$ the row with the maximum entry is chosen. The joint probability mass function of c and $\hat{x} = T(x)$ is denoted by $P_{T(x)}(c, \hat{x})$, and the corresponding minimum Bayes error is defined as

$$P_{\text{Bayes}}^{(\hat{x} \to c)} := 1 - \sum_{j=1}^{M} \max_{i \in \{1,\ldots,N\}} \left\{P_{T(x)}\left(c_i, \hat{x}_j\right)\right\}. \tag{2.23}$$

Now, any transformation $T : \mathcal{X} \mapsto \hat{\mathcal{X}}$ can either be one-to-one and onto, one-to-one but not onto, not one-to-one and onto, or not one-to-one and not onto.

1. *T is one-to-one and onto*
   In this case, T is an invertible transformation and $|\mathcal{X}| = |\hat{\mathcal{X}}|$. This implies that each column of $P_{T(x)}(c, \hat{x})$ corresponds to exactly one column of $P(c, x)$, i.e., the columns are permuted. This does not affect the minimum Bayes error, since in (2.22) and (2.23) the sum over all columns is taken.

2. *T is one-to-one but not onto*
   This implies that $|\mathcal{X}| < |\hat{\mathcal{X}}|$, since in addition to those M elements in $\hat{\mathcal{X}}$ that are hit by T exactly once there are elements in $\hat{\mathcal{X}}$ that are not hit by T. However, these elements do not enter into the minimum Bayes error since their probability is zero. Since T is one-to-one, all columns of $P_{T(x)}(c, \hat{x})$ with $P_{T(x)}(\hat{x}) > 0$ correspond to exactly one column of $P(c, x)$. This again does not alter the minimum Bayes error, since in (2.22) and (2.23) the sum over all columns is taken.

3. *T is not one-to-one but onto*
   This implies that $|\mathcal{X}| > |\hat{\mathcal{X}}|$, since every element in $\hat{\mathcal{X}}$ is hit (T is onto), and at least one element in $\hat{\mathcal{X}}$ is hit at least twice (T is not one-to-one). First consider all elements in $\hat{\mathcal{X}}$ that are hit exactly once. Each of the corresponding columns of $P_{T(x)}(c, \hat{x})$ corresponds to exactly one column of $P(c, x)$, which does not alter the contribution of these columns to the minimum Bayes error. Now consider all elements in $\hat{\mathcal{X}}$ that are hit at least twice. Denote this set by

$\hat{\mathcal{X}}^*$, and let $\mathcal{X}_j = \{x \in \mathcal{X} : \hat{x}_j = T(x)\}$, i.e., all elements of $\mathcal{X}$ that hit a certain element $\hat{x}_j \in \hat{\mathcal{X}}$. Then note that $\forall \hat{x}_j \in \hat{\mathcal{X}}^*$ it holds that

$$\begin{aligned} \max_{i\in\{1,\ldots,N\}} \{P_{T(x)}(c_i, \hat{x}_j)\} &= \max_{i\in\{1,\ldots,N\}} \left\{ \sum_{x\in\mathcal{X}_j} P(c_i, x) \right\} \\ &\leq \sum_{x\in\mathcal{X}_j} \max_{i\in\{1,\ldots,N\}} \{P(c_i, x)\} \end{aligned} \tag{2.24}$$

by the triangle inequality. Plugging (2.24) into (2.23) then leads to $P_{\text{Bayes}}^{(\hat{x}\to c)} \geq P_{\text{Bayes}}^{(x\to c)}$.

4. *T is not one-to-one and not onto*
   First note that all elements in $\hat{\mathcal{X}}$ that are not hit by T do not enter into the computation of the minimum Bayes error due to zero probability. Then apply the argument for T one-to-one and onto.

In summary, transformations that are one-to-one do not alter the minimum Bayes error, and transformations that are not one-to-one can only increase the minimum Bayes error. This completes the proof. □

The above theorem shows that any transformation of x that reduces the size of the feature space can at best not affect the minimum Bayes error, while in practice it very likely increases it. It is hence desirable to find a transformation of the observed data that reduces the dimension of the feature space in order to increase the rate of convergence while not affecting the minimum Bayes error. This is the goal of feature extraction.

Definition 2.12 (Feature Extraction)**.** *The goal of feature extraction is to find a transformation $T : \mathcal{X} \mapsto \hat{\mathcal{X}}$ with $|\hat{\mathcal{X}}| < |\mathcal{X}|$ and $P_{Bayes}^{(T(x)\to c)} = P_{Bayes}^{(x\to c)}$.*

This goal might be overly optimistic, since reducing the dimensionality of the feature space can be expected to almost always increase the minimum Bayes error. On the other hand, a small increase of the minimum Bayes error might be irrelevant as long as insufficient training data is available to actually get close to the minimum Bayes error. In practice, the goal of feature extraction is to find a transformation of the data that minimizes the *expected* error probability for a given set of training data. This done by trying to find a transformation that achieves an optimal trade-off between increasing the rate of convergence of the learning curve and not increasing the minimum Bayes error.

Implementing Feature Extraction in BCIs

After demonstrating the necessity of feature extraction for BCIs, it is now discussed how feature extraction can be approached. First, recall that any transformation T :

$\mathcal{X} \mapsto \hat{\mathcal{X}}$ with $|\hat{\mathcal{X}}| < |\mathcal{X}|$ is a possible feature extraction algorithm. Consequently, the set $\hat{\mathcal{X}}$ can be any subset of the power set of $\mathcal{X}$ with $|\hat{\mathcal{X}}| < |\mathcal{X}|$, i.e., $\hat{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$. For example, $\hat{\mathcal{X}}$ can be the set of possible variances at a certain electrode, the set of maximum amplitudes at a certain electrode, or even more abstract sets such as the set of all possible values of mutual information of the EEG signals at multiple electrodes. In fact, the notation used here is general enough for $\hat{\mathcal{X}}$ to represent any property of the observed data x. This raises the question which set $\hat{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$ with $|\hat{\mathcal{X}}| < |\mathcal{X}|$ should be chosen as the new feature space. One way to approach this is to fix the dimension of the feature space, e.g., let $|\hat{\mathcal{X}}| = d$, and then develop a sophisticated algorithm that determines the subset of $\mathcal{P}(\mathcal{X})$ with the lowest (estimated) minimum Bayes error under the constraint that $|\hat{\mathcal{X}}| = d$. Unfortunately, it can be proved that this requires an exhaustive search over all possible subsets of $\mathcal{P}(\mathcal{X})$ with $|\hat{\mathcal{X}}| = d$ [Cv78]. Considering the enormous size of $\mathcal{P}(\mathcal{X})$, i.e., all possible subsets of $\mathcal{X}$, this is impossible to realize.
This finally leads to what is regarded in this work as the main challenge in the design of high-performance BCIs. The original feature space of BCIs is by far too large to be used directly for training a classification algorithm. This requires a feature extraction algorithm that maps the original feature space into a lower dimensional feature space, on which it is feasible to train a classifier given a limited amount of training data. However, in the context of BCIs, the size of the class of possible feature spaces is enormous. Consequently, any algorithm that does not restrict the class of possible feature spaces is impossible to realize. How then can the class of possible feature spaces be restricted? Such a restriction has to be specific enough to decrease the number of allowed feature spaces to a computationally feasible point, while being general enough to ensure that feature spaces with a low minimum Bayes error are included. The only possible procedure to restrict the class of allowed feature spaces in a sensible way is to incorporate a-priori information. This a-priori information has to reflect our knowledge on how the brain processes information, and which properties of signals recorded from the CNS can provide information on the BCI-user's intention. Given such a restriction on the class of possible feature spaces, powerful algorithms have to be developed that determine the in some way optimal element of the class of admitted feature spaces. This can be summarized in a mathematical way as follows.

Definition 2.13 (Feature extraction in BCIs)**.** *Let $c \in \mathcal{C}$ the BCI-user's intention and $x \in \mathcal{X}$ the data recorded from the central nervous system. The goal of feature extraction in BCIs is to solve the optimization problem*

$$T^* = \underset{T:\mathcal{X}\mapsto\hat{\mathcal{X}}}{\operatorname{argmin}} \{f(c, T(x))\} \text{ s.t. } \hat{\mathcal{X}} \in \mathcal{P}^* \subset \mathcal{P}(\mathcal{X}), \tag{2.25}$$

with $f : \mathcal{C} \times \hat{\mathcal{X}} \mapsto \mathbb{R}$ some cost function related to the expected error probability of inferring c from $T(x)$, and $\mathcal{P}^*$ some subset of the power set $\mathcal{P}(\mathcal{X})$ that encodes a-priori knowledge on how the brain processes information, i.e., which properties of the data x provide information on c.

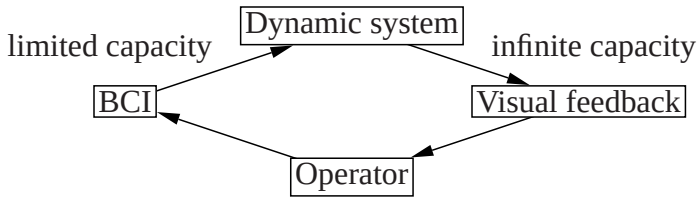

Figure 2.7: Control of a dynamic system by a BCI.

Note that in order to solve the optimization problem (2.25) a subset $\mathcal{P}^*$ and a cost function f have to be specified in advance. The process of developing a sophisticated method for feature extraction in non-invasive BCIs can thus be summarized in the following three tasks:

1. Determine $\mathcal{P}^*$ by specifying assumptions on how information on the user's intention is encoded in the recorded data.

2. Find a suitable cost function f that estimates the expected error probability.

3. Find a way to efficiently solve (2.25).

Finding suitable subsets $\mathcal{P}^*$ and good estimators of the expected error probability f constitute the main contributions of this thesis in Chapters 3 - 6. It should be emphasized again that finding a suitable subset $\mathcal{P}^*$, specifying f, and solving (2.25) for a certain choice of $\mathcal{P}^*$ and f are problems from different domains. The first problem of determining $\mathcal{P}^*$ pertains to how information is processed by our brain, and how this is reflected in data that can be recorded from the CNS. This is usually considered to be the domain of neuroscience. The problems of determining f and solving (2.25), on the other hand, lie within the domain of signal processing and machine learning. Designing high-performance feature extraction algorithms requires a good understanding of both domains, which stresses the importance of interdisciplinary research in the context of BCIs.

2.5 Control of Dynamic Systems by BCIs

Currently, BCIs are only used for controlling simple devices such as a cursor on a screen or a spelling device. The goal of research on BCIs, however, is controlling more complex systems such as robotic devices. These systems are often unstable, which leads to the following question: Can an unstable dynamic system be stabilized by control through a BCI (Fig. 2.7)? The limitation that is imposed here due to the presence of a BCI is the limited bandwidth in the feedback loop between the operator and the dynamic system. The problem of controlling a dynamic system through a BCI can thus be formulated as a control problem with bandwidth constraints.

Research on control with bandwidth constraints on the communication channel between plant and controller has been initiated about ten years ago. A recent overview of the state-of-the-art in this field is given in [NFZE07]. Most research on bandwidth limited control assumes that the bandwidth of the feedback from the sensors to the controller is limited, but data can be transmitted with zero error probability. While the obtained results differ depending on which notion of stability is adopted (e.g., whether asymptotic stability or only boundedness of the state is required), it has been shown that depending on the dynamic system there exists a lower bound on the rate of the communication channel that must be met to allow stabilizability. Hence, the amount of information that can be transfered by the communication channel determines the class of dynamic systems that can be stabilized.
The use of a BCI in place of the controller of a dynamic system differs from the problem usually considered in the control literature. Here, the bandwidth is limited only between the controller, i.e., the BCI, and the dynamic system. Feedback from the system to the BCI on the other hand is provided by visual feedback, and can thus be considered, at least in practice, as obtained with infinite capacity. This setting is considered in [MS07], in which it is proved that if a communication channel between controller and dynamic system has zero zero-error capacity no unstable system can be stabilized almost surely. The concept of zero-error capacity has been introduced by C. Shannon in [Sha56].

Definition 2.14 (Zero-error capacity)**.** *The zero-error capacity C_0 of a noisy channel is defined as the least upper bound of rates at which it is possible to transmit information with zero probability of error.*

In general, the problem of establishing the zero-error capacity of an arbitrary noisy channel remains unsolved [KO98]. If, however, feedback of the received symbols back to the sender is allowed, which is the case in the setting considered here, C. Shannon provided a sufficient condition for zero zero-error capacity using the concept of adjacency [Sha56].

Definition 2.15 (Adjacency)**.** *Let $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ the input and $\hat{c} \in \hat{\mathcal{C}} = \{\hat{c}_1, \ldots, \hat{c}_M\}$ the output of a discrete memoryless communication channel. Two symbols c_i and c_j with $i, j \in \{1, \ldots, N\}$ and $i \neq j$ are called adjacent if there is an output symbol $\hat{c}_k$, $k \in \{1, \ldots, M\}$ that can be caused by either of these two.*

Theorem 2.6 (Zero-error capacity of memoryless discrete channels with feedback)**.** *In a memoryless discrete channel with complete feedback of received symbols to the transmitting point, the zero-error capacity C_0 is zero if all pairs of input symbols are adjacent.*

For this reason, a BCI can only have a zero-error capacity greater than zero if there exist at least two intentions of the user that are never confused with each other by the BCI. At present, there is no BCI that meets this requirement, and it is unclear how such a BCI could be constructed. On the other hand, there appear to be no

reasons why this should be impossible, at least in principle. Nevertheless, due to [MS07] and Theorem 2.6 at present no BCI can be used to stabilize any unstable dynamic system. This conclusion can be illustrated by the following example.

Example 2.2. *Consider the discrete-time scalar dynamic system*

$$x[t+1] = ax[t] + bu[t], \tag{2.26}$$

with $x[t]$ the state of the system at time t, $u[t]$ the input to the system at time t, and $a, b \in \mathbb{R}$. It is assumed that $|a| > 1$, i.e., the system is unstable, and $b > 0$. If $u[t] \equiv 0$ and $x[t_0] \neq 0$, then $\lim_{t\to\infty} |x[t]| = \infty$, i.e., the state is unbounded. The state of the system can be bounded if a control law

$$u[t] = -\mathit{sign}\{x[t]\} \tag{2.27}$$

is chosen. Then

$$x[t+1] = ax[t] - b \cdot \mathit{sign}\{x[t]\} = \begin{cases} x[t+1] = ax[t] - b & ; \quad x[t] \geq 0 \\ x[t+1] = ax[t] + b & ; \quad x[t] < 0 \end{cases} . \tag{2.28}$$

For $x[t] \geq 0$, $x[t+1] < x[t] \Leftrightarrow x[t] < \frac{b}{a-1}$, and for $x[t] < 0$, $x[t+1] > x[t] \Leftrightarrow x[t] > -\frac{b}{a-1}$. Consequently, $\limsup_{t\to\infty} |x[t]| < \frac{b}{a-1}$ if $|x[t_0]| < \frac{b}{a-1}$, i.e., the state of the dynamic system is bounded. If however $|x[t]| > \frac{b}{a-1}$ for any t, the state of the system grows without bounds since the control input is not powerful enough to drive the state of the system back to its stable region $|x[t]| < \frac{b}{a-1}$ (see Fig. 2.8).
Now consider the control law (2.27) to be carried out by an operator using a binary BCI. The control law then becomes stochastic, since errors might be introduced by the BCI. Hence $P(u[t] = -\mathit{sign}\{x[t]\}) = 1 - P_e$, and $P(u[t] = \mathit{sign}\{x[t]\}) = P_e$, with $P_e > 0$ the error probability of the binary BCI. Independently of the desired output of the controller, the probability that the control sequence $u[t_i] = 1,\ i = 0, \ldots, T$, occurs is hence greater zero. Since the state of the system at time T is given by

$$x[T] = a^t x[t_0] + \sum_{i=1}^{T} a^{i-1} b u[t_i], \tag{2.29}$$

this control sequence leads to the state $x[T] = a^t x[t_0] + b\sum_{i=1}^{T} a^{i-1}$. This is an increasing function of T with $\lim_{T\to\infty} x[T] = \infty$. Consequently, there is some T such that $x[T] > \frac{b}{a-1}$, which leads to the state of the system becoming unbounded even if the correct control signals are transmitted by the BCI for $t > T$. This illustrates why no unstable dynamic system can be stabilized almost surely by a BCI with zero zero-error capacity.

While in practical situations a low error probability of the BCI might lead to a very low probability of the state of the system exceeding some bound in finite time, the

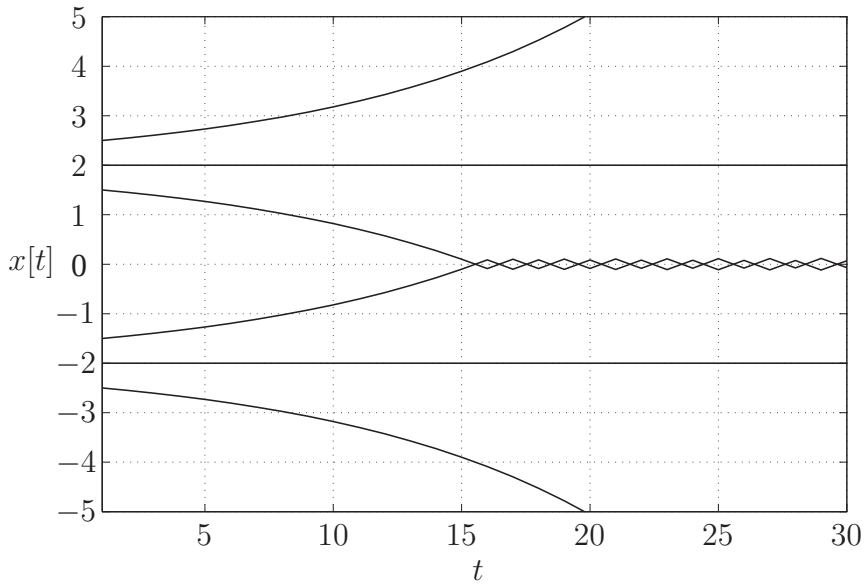

Figure 2.8: State evolution for the dynamic system (2.28) for $a = 1.1, b = 0.2$ and different initial conditions.

boundedness can not be guaranteed as long as BCIs have zero zero-error capacity. Consequently, if an unstable system is controlled by a BCI, measures have to be taken that ensure stability of the system independently of the control signals received from the BCI. This is a non-trivial control theoretic problem, which could be approached by methods such as invariance control that ensure that the state of the system never leaves an invariance region (cf. [WB05], [WB07]).

Chapter 3

Feature Extraction via Source Localization

3.1 Introduction

In this chapter, the feasibility of source localization as a method for feature extraction in non-invasive BCIs is investigated. This is motivated by the following considerations. As discussed in Section 2.4, it is necessary in BCIs to restrict the class of allowed feature spaces, denoted by $\mathcal{P}^*$, in order to construct viable feature extraction algorithms. The class $\mathcal{P}^*$ determines which properties of the data recorded from the CNS are allowed as possible features. It hence represents the a-priori knowledge that is available on how the BCI-user's intention is encoded in the recorded data. This is the problem of deciphering the neural code. Most research on neural coding deals with action potentials of single neurons or small networks of neurons (cf. [DA01] for an introduction to this topic). For the recording modalities employed in this thesis, i.e., EEG and MEG, the question of the neural code is a largely open problem (cf. [NS05]). In traditional neuropsychology the main tool for the analysis of EEG/MEG data is averaging event related potentials (ERPs), i.e., averaging responses of the electric or magnetic field of the brain to external stimuli over many trials. In recent years, measures of event related synchronization/desynchronization (ERS/ERD), i.e., changes in the power of the electric/magnetic field in specific frequency bands, have been increasingly used for investigating neural processes [PL99]. Considering the complexity of the human brain, these are relatively simple measures. In general, the problem of how information on cognitive processes is encoded within EEG/MEG data remains unsolved. For this reason, it is also unclear how the class of allowed feature spaces $\mathcal{P}^*$ could be restricted to properties of the recorded data that provide most information on the BCI-user's intention.

The idea behind using source localization for feature extraction in BCIs is not to decipher the neural code, but rather to largely circumvent this problem. The aim of source localization in EEG/MEG is to detect areas within the brain that are active

during a certain cognitive task. Here, active brain areas are defined as areas with a high current density, since the spatial distribution of current density within the brain gives rise to the electric/magnetic field that can be measured on the scalp [NS05]. It is well known that information processing within the brain, at least for low-level processes such as the first stages in processing of visual, auditory, or sensorimotor stimuli, is spatially localized, i.e., that certain brain areas are specialized for certain tasks. It thus seems sensible to infer the user's intention, or in general any cognitive task, from measures of activity of certain brain areas. This has got the advantage that the specific nature of how the information is temporally encoded within the electric/magnetic field becomes irrelevant.

Employing source localization methods for feature extraction in non-invasive BCIs has been first proposed almost simultaneously by this author in [GGWB05], by Qin et al. in [QDH04], and by Grave de Peralta Menendez et al. in [GGP+05]. More recent studies include [KLH05] and [LLA07]. A comparison of these studies with the work presented in this chapter is carried out in Section 3.4, along with a critical evaluation of the efficiency of source localization for feature extraction.

In this chapter, source localization is combined with Independent Component Analysis (ICA) to obtain estimates of the current density in specific brain regions. ICA decomposes the measured EEG data into statistically independent components (ICs). Using ICA as a preprocessing step before source localization has got the advantage that often simple source models for each IC can be used for which the EEG inverse problem, i.e., estimating the current density within the brain from measurements on the scalp, is well defined. A disadvantage of using ICA is that it is necessary to identify which ICs constitute meaningful components, and which ICs represent noise. This is solved here by proving that ICs representing noise are not invariant with respect to initial conditions of the ICA algorithm. Using ICA multiple times on the same data set with randomized initial conditions hence allows the identification and exclusion of ICs that represent noise. The origin of meaningful ICs is then localized by modeling each IC as a single current dipole within a four-shell spherical head model. This methodology is applied to EEG data obtained during real and imaginary tapping movements of the right and left index finger, and it is shown that it constitutes a viable option for feature extraction in non-invasive BCIs. The contribution of this chapter is therefore twofold. First, it establishes the viability of source localization for non-invasive BCIs, and second, it develops a methodology how in ICA components representing noise can be reliably identified and excluded from further analysis.

The structure of this chapter is as follows. In Section 3.2, the concept of ICA and the specific algorithm used in this work are introduced. It is then shown how the origin of each IC can be localized, and how an estimate of the spatial current density distribution within the brain can be obtained. The main contributions of this Section are a theorem on the behavior of ICA in presence of multiple Gaussian sources, and a methodology for identifying and excluding ICs representing noise. In Section 3.3, preliminary experimental results from one subject are presented. The

chapter concludes with a critical evaluation of the efficiency of source localization for feature extraction in BCIs. Parts of the work in this chapter have been presented in [GGWB05] and [GB06].

3.2 Methods

As before, let $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$ be the BCI-user's intention, and $x \in \mathcal{X}$ the recorded EEG/MEG data. Since the experimental evaluation in Section 3.3 is performed using EEG recordings, only this modality is considered here. In Chapter 2, $\mathcal{X}$ was defined as a discrete space in which each element uniquely determines the recorded EEG data. For the purpose of this chapter, it is beneficial to let $\mathcal{X} = \mathbb{R}^{M \times T}$ with M the number of recording electrodes and T the number of recorded samples. Distinguishing the variable T, describing the number of samples, and the mapping T, representing the feature extraction algorithm, should be clear from the context. Consequently, the matrix $X \in \mathbb{R}^{M \times T}$ refers to one block of recorded data, and the vector $\boldsymbol{x}(t) \in \mathbb{R}^M$ refers to the recorded data at all electrodes at one sample point. Note that while this is the convention usually employed in the analysis of EEG data, it is in fact an approximation if EEG data is digitally recorded.
Before going into details of the methodology, it is important to clearly state the main assumptions that are being made in this chapter on the class of allowed feature spaces $\mathcal{P}^*$ to solve the feature extraction problem (cf. Definition 2.13).

1. Only the activity of brain areas, defined as the spatial distribution of current density within the brain, provides information on the BCI-user's intention.

2. Distinct brain areas produce electric fields that that are statistically independent.

3. The activity of a distinct brain area can be modeled by a single current dipole.

These assumptions warrant some further explanations. The first assumption has already been motivated in Section 3.1. The second assumption is required in order to apply ICA to EEG data, and there is considerable experimental evidence that it is indeed justified (reviewed in [HKO01]). Assumption three is based on experimental evidence [VSJ$^+$00], and constitutes the main reason for using ICA as a preprocessing step in source localization as discussed in Section 3.2.2.
It is furthermore necessary to specify the exact form of the desired feature extraction algorithm T. Since the goal of this chapter is to infer the user's intention from activity of certain brain areas, $T(X)$ should return an estimate of the current density at certain locations within the brain. For simplicity, it is assumed that for each class $c \in \mathcal{C}$ the activity in one region of the brain is sufficient. The desired feature extraction algorithm is thus a mapping $T : \mathcal{X} \mapsto \mathbb{R}^N_+$, i.e., given some data the mapping T returns an estimate of the current density at N distinct locations within the brain. The optimal feature extraction algorithm is then found by solving (2.25)

with $f(c, T(x)) = \sum_{i=1}^{N} |T(X_{c_i})|$ with X_{c_i} denoting the recorded EEG data during condition c_i. The resulting mapping $T^*(X)$ thus returns estimates of the current density at those N locations in the brain with maximum activity for each condition. These areas are assumed to be optimal for inferring the BCI-user's intention. The implementation of this approach is now presented.

3.2.1 Independent Component Analysis

In ICA, a generative model of the data $X = [\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(T)] \in \mathbb{R}^{M \times T}$ measured at M electrodes is assumed,

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t). \tag{3.1}$$

Here, $\boldsymbol{s}(t) \in \mathbb{R}^M$ describes the electric field of the original sources within the brain, and each column of the full rank matrix $A \in \mathbb{R}^{M \times M}$ describes the projection strength of a source to each of the electrodes. Throughout this chapter, all variables and matrices are assumed to be real. For now, it is assumed that at most one source has got a Gaussian distribution. Furthermore, it is assumed that each of the M sources has got zero mean and unit variance. This is no loss of generality, since the mean can be subtracted and added again at any point due to the linearity of the model, and the variance of each source can be arbitrarily traded between the source and the corresponding column of the mixing matrix A. The crucial assumption in ICA is that for the probability density function of the source vector it holds that $p(\boldsymbol{s}) = \prod_{i=1}^{M} p(s_i)$, i.e., that the original sources are mutually statistically independent. This also defines what is meant by the term source in this context.

Definition 3.1 (ICA sources in EEG analysis)**.** *In ICA applied to EEG data, a source is defined as a spatial current density distribution within the brain with identical temporal dynamics that is statistically independent of all other sources.*

Consequently, a source does not have to be spatially confined to one region of the brain or even consist of one connected region. This is discussed in the context of source localization in Section 3.2.2. Besides the assumption of mutual statistical independence of the sources, which is supported by experimental evidence reviewed in [HKO01], four other assumptions on the EEG data X are made by the model in (3.1). These are a) linearity of the mixing process, b) instantaneous propagation of the sources to the sensors, c) at most one source has got a Gaussian distribution, and d) equal number of sensors and sources. The first two assumptions are justified in the context of EEG analysis as discussed extensively in [NS05]. The third assumption is discussed in Section 3.2.3. The fourth assumption is questionable. Typically, EEG is recorded with up to 128 electrodes. This is contrasted with an estimated number of several million cortical columns within the brain, which are believed to constitute the main current sources within the brain giving rise to the electric field on the scalp [NS05]. In spite of this apparent contradiction, ICA has been applied with great success to EEG data. This is addressed in detail in Chapter

5. For now, this assumption is adopted as a working assumption, keeping in mind that its validity is questionable.

The goal of ICA is to reconstruct the original sources $\boldsymbol{s}(t)$ and the mixing matrix A only from observations of $\boldsymbol{x}(t)$ and the assumption on the sources of mutual statistical independence. The general approach to this problem is to formulate an optimization problem

$$W^* = \underset{W \in \mathbb{R}^{M \times M}}{\operatorname{argmin}} \{F(WX)\} \tag{3.2}$$

such that $W^* = P\Lambda A^{-1}$ with $P \in \mathbb{R}^{M \times M}$ a permutation matrix and $\Lambda \in \mathbb{R}^{M \times M}$ a diagonal matrix. Consequently, it then holds that $\boldsymbol{y}(t) = W^*\boldsymbol{x}(t) = P\Lambda\boldsymbol{s}(t)$. The elements of $\boldsymbol{y}(t)$ are called the independent components (ICs). Different choices of the cost function F, sometimes also called a contrast function [Com94], lead to different algorithms. An excellent review of algorithms for ICA and the statistical principles underlying the construction of their contrast functions is [Car98].

Here, only the extended Infomax algorithm ([BS95],[LGS99]) is considered, which has been shown to perform well in the context of EEG analysis (reviewed in [JMB$^+$01]). Results obtained with other algorithms might slightly vary. The Infomax algorithm is based on minimizing mutual information of the reconstructed ICs, i.e.,

$$F(\boldsymbol{y}) = I(y_1, \ldots, y_M) = \sum_{i=1}^{M} H(y_i) - H(y_1, \ldots, y_M). \tag{3.3}$$

Contrary to Chapter 2, here $H(.)$ refers to the differential entropy (cf. [CT06]). Equation (3.3) is an adequate cost function for ICA due to the following theorem (cf. [Com94]).

Theorem 3.1. *Let $\boldsymbol{y} \in \mathbb{R}^M$ be a vector of random variables. Then it holds that $I(y_1, \ldots, y_M) \geq 0$ with equality if and only if the elements of $\boldsymbol{y}$ are mutually statistically independent.*

Proof. Mutual information can be expressed as

$$I(y_1, \ldots, y_M) = \int_{-\infty}^{\infty} p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{\prod_{i=1}^{M} p(y_i)} \mathrm{d}\boldsymbol{y} = D\left(p(\boldsymbol{y}) \| \prod_{i=1}^{M} p(y_i)\right), \tag{3.4}$$

with $D\left(p(\boldsymbol{y}) \| \prod_{i=1}^{M} p(y_i)\right)$ the Kullback-Leibler divergence of $p(\boldsymbol{y})$ and $\prod_{i=1}^{M} p(y_i)$. Due to Gibbs's inequality, it holds that $D\left(p(\boldsymbol{y}) \| \prod_{i=1}^{M} p(y_i)\right) \geq 0$ with equality if and only if $p(\boldsymbol{y}) = \prod_{i=1}^{M} p(y_i)$. □

Finding a transformation W^* with $F(W^*\boldsymbol{x}) = 0$ thus results in mutually statistically independent components. Under the assumptions on the source model (3.1) it then holds that $\boldsymbol{y}(t) = W^*\boldsymbol{x}(t) = P\Lambda\boldsymbol{s}(t)$ and $W^* = P\Lambda A^{-1}$ [Com94]. Details on how the optimization problem (3.2) can be solved for this contrast function are given in [LGS99].

3.2.2 Source Localization and ICA

Excellent reviews of different source localization methods for EEG data are given in [BML01] and [MML$^+$04]. For this reason, only those aspects of source localization important in this context are presented here.

There are two reasons for using ICA as a preprocessing step before source localization, which was first suggested in [ZWJ00]. The first reason is that the inverse of the obtained unmixing matrix constitutes an estimate of the original mixing matrix up to scaling and permutation. Scaling and permutations are neglected from here on, since these are irrelevant in the context of source localization. It is thus assumed that $(W^*)^{-1} = A$. Recall that each column $\boldsymbol{a}_i$ of A describes the projection strength of source s_i to each of the electrodes. Since the temporal evolution of the source s_i only constitutes a scaling of the topography $\boldsymbol{a}_i$, all information required to localize the specific source is already contained in $\boldsymbol{a}_i$. Source localization of ICs can thus be completely decoupled from their temporal evolution. The second reason for using ICA before source localization is empirical evidence that ICs can often be accurately modeled by a single current dipole [VSJ$^+$00]. Since a current dipole has only got six degrees of freedom (three for its position, two for its orientation, and one for its strength), the inverse problem of determining the parameters of a current dipole that best explain the topography of an IC is well-defined. This is in contrast to source localization of raw EEG data, which usually requires multiple current dipoles to explain the data. If more than $M/6$ current dipoles are employed, the inverse problem is ill-posed. Consequently, additional assumptions on the parameter space, such as minimum variance or sparsity constraints, have to be imposed to obtain an unique solution (cf. [BML01, MML$^+$04]). This is usually unnecessary when performing source localization of ICs.

It should be noted that increasing the complexity of a model, i.e., increasing the number of current dipoles, always leads to a more accurate representation of the observed data. Choosing a model that achieves an optimal trade-off between model accuracy and model complexity is an intricate problem that requires a prior on the class of allowed models (see [HB01] for a general introduction and [KBJP98] for a comparison of model selection techniques in EEG analysis). The assumption that one IC can be modeled by a single current dipole is thus not to be understood as meaning that the source of an IC does indeed correspond to a single current dipole, or that a single current dipole is the best model in terms of some trade-off between model accuracy and model complexity. Instead, it only asserts that a single current dipole, which constitutes the most simple model in EEG source localization, is

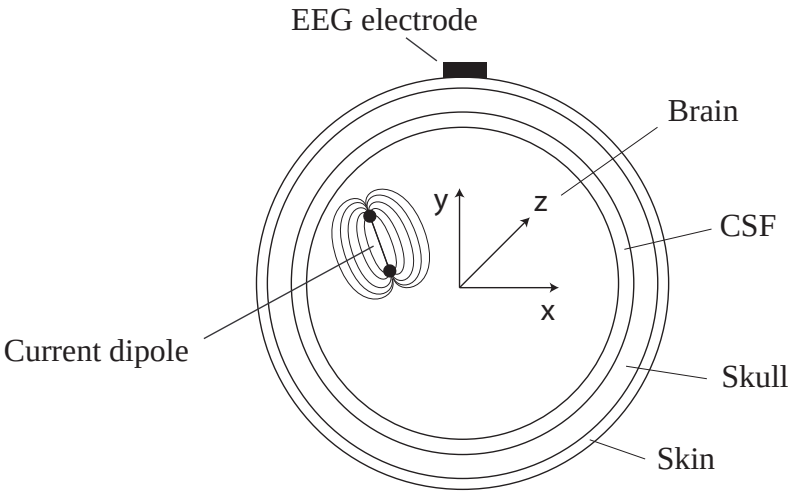

Figure 3.1: The four-shell spherical head model.

	Radius [cm]	Conductivity [S/m]
Brain	7.1	0.33
CSF	7.2	1
Skull	7.9	0.0042
Skin	8.5	0.33

Table 3.1: Radii and conductivities used in the four-shell spherical head model

sufficiently accurate to model the topography of one IC for the purpose of feature extraction.

In addition to a source model, source localization also requires specification of a suitable head model. Here, only a four-shell spherical head model is considered (see Fig. 3.1). This head model consists of four nested spheres, with the shells representing the brain, the cerebrospinal fluid (CSF), the skull, and the scalp. Each sphere is assumed to be isotropic with a certain conductivity value. The radius and conductivity value of each sphere shown in Tab. 3.1 is chosen as in the publicly available toolbox EEGLAB [DM04]. It should be noted that the conductivity values and radii of the spheres vary between subjects. However, an accurate choice of these parameters is not required for feature extraction, as is discussed at the end of this section.

Using Legendre-polynomials, an analytic solution for the electric potential at any position on the scalp due to a current dipole inside the innermost sphere, i.e., the brain, can be computed [RD69]. Let $\boldsymbol{r}_{\text{dip}} \in \mathbb{R}^3$ be the position of the dipole and $\boldsymbol{\theta} \in \mathbb{R}^3$ the orientation of the dipole. The electric potential at the i^{th} electrode with coordinates $\boldsymbol{r}_i$ can then be computed as

$$\hat{x}_i = \boldsymbol{l}(\boldsymbol{r}_{\text{dip}}, \boldsymbol{r}_i)^{\text{T}}\boldsymbol{\theta}, \tag{3.5}$$

with $\boldsymbol{l} \in \mathbb{R}^3$ the so called leadfield vector. The position of the dipole enters non-

linearly into the computation, while the influence of the orientation is linear. The electric potential at all electrodes can then be computed as

$$\hat{\boldsymbol{x}} = L(\boldsymbol{r}_{\text{dip}})\boldsymbol{\theta}, \tag{3.6}$$

with $L(\boldsymbol{r}_{\text{dip}}) \in \mathbb{R}^{M\times 3}$ the leadfield matrix. Note that for simplicity the dependence of the leadfield matrix on the positions of the EEG electrodes is dropped. It should be pointed out that it is incorrect to speak of an electric potential without specifying a reference. In EEG source localization, all electric potentials are computed with respect to a common average reference. For this reason, before performing source localization, the recorded EEG data has to be transformed such that the mean electric potential across all electrodes equals zero for every sample point.
The goal of source localization in this context is then to solve the optimization problem

$$[\boldsymbol{r}^*_{\text{dip}}, \boldsymbol{\theta}^*] = \underset{\boldsymbol{r}_{\text{dip}},\boldsymbol{\theta}}{\operatorname{argmin}} \{\|\boldsymbol{a}_i - L(\boldsymbol{r}_{\text{dip}})\boldsymbol{\theta}\|_2\} \tag{3.7}$$

for each IC, i.e., $i = 1, \ldots, M$. Depending on the topography of an IC, this can constitute a non-convex optimization problem. Consequently, (3.7) is solved in a two-step procedure for each IC. First, a dipole grid, covering the volume of the innermost sphere, is constructed, and the optimal orientation of each dipole on the grid is computed. The grid position that minimizes (3.7) is then chosen as the initial position for a standard numerical optimization procedure (see [NW06] for a good introduction to numerical optimization). The result of the optimization procedure is an optimal dipole position $\boldsymbol{r}^*_i$ and dipole orientation $\boldsymbol{\theta}^*_i$ for each of the M ICs. The locations of the M dipoles thereby identify the regions of the brain that can be considered active for the observed data.
Considering the complex geometry of the brain and the skull, the use of an isotropic spherical head model might appear questionable. Furthermore, the source localization accuracy might be impaired by imprecise conductivity values and sphere radii. Both concerns are indeed justified if the goal of source localization is to identify active brain areas with maximum accuracy. In this case, more complex head models such as boundary element models (BEM) or finite element models (FEM) should be employed (cf. [BML01]). For the purpose of feature extraction, however, this is irrelevant. The goal of feature extraction is not to localize active brain areas with maximum precision, but rather to map the observed data, i.e., the original feature space, into another feature space in which classification is simplified. All that is required of feature extraction by source localization is that the observed features in the new feature space are separable, i.e., that different intentions lead to distinct dipole locations. The physiological validity of the dipole locations is irrelevant.

3.2.3 Signal Subspace Identification by ICA

It is well known that if ICA is applied repeatedly to the same data set some ICs are stable while others vary [JMB+01]. Those ICs that vary are termed unstable, since

they depend on the initial conditions of the ICA algorithm. As such, these ICs do not solely depend on the observed data and should be excluded from further analysis. In this section, it is shown that in the framework considered here unstable components represent mixtures of Gaussian sources. This is in contrast to the assumption usually made in ICA in order to ensure separability and identifiability that at most one of the original sources may have a Gaussian distribution [Com94]. For real-world applications, this is an unrealistic assumption. However, dropping this assumption requires an analysis whether the mixing matrix and the original sources can still be reconstructed in the framework of ICA. It is shown here that this is indeed the case for those sources with a non-Gaussian distribution. Gaussian sources, on the other hand, can not be reconstructed. Furthermore, it is shown how the incorrectly reconstructed Gaussian sources can be excluded from further analysis without requiring the a-priori specification of any cut-off criterion. This can be seen as a method for subspace identification, with the signal subspace defined as the space spanned by sources with a non-Gaussian distribution.

ICA and Multiple Gaussian Sources

So far, it was assumed in the mixing model (3.1) that at most one source has a Gaussian distribution. This ensured separability and identifiability of the ICA model as discussed in [Com94]. This assumption is now dropped. More specifically, it is assumed that $L < M$ of the sources have a Gaussian distribution. Without loss of generality, it is assumed that these are the first L sources, i.e., $p(s_i) = \mathcal{N}(0,1),\ i = 1,\ldots,L$. The following argument requires a famous theorem derived independently by Darmois and Skitovic (cf. [Com94]).

Theorem 3.2 (Darmois-Skitovic)**.** *Define two random variables*

$$y_1 = \sum_{i=1}^{M} a_i s_i, \quad y_2 = \sum_{i=1}^{M} b_i s_i, \tag{3.8}$$

with s_i statistically independent random variables. If y_1 and y_2 are statistically independent, then all variables s_i for which $a_i b_i \neq 0$ are Gaussian.

Put differently, Theorem 3.2 states that two different sums of M statistically independent random variables can only be statistically independent if the M variables are Gaussian. Now consider the unmixing matrix W^*, obtained by running ICA on the data X, with $F(W^*\boldsymbol{x}) = 0$. Then the elements of $\boldsymbol{y} = W^*\boldsymbol{x}$ are mutually statistically independent due to Theorem 3.1. Now write

$$\boldsymbol{y} = W^*\boldsymbol{x} = W^*A\boldsymbol{s} =: C\boldsymbol{s}. \tag{3.9}$$

It is then instructive to consider the possible class of matrices $C = W^*A \in \mathbb{R}^{M\times M}$ that are in accord with the requirements of the elements of $\boldsymbol{s}$ as well as the elements of $\boldsymbol{y}$ being mutually statistically independent. Note that a different proof of this theorem is given in [CL96].

Theorem 3.3. *Let $\boldsymbol{s} \in \mathbb{R}^M$ be a random variable with mutually statistically independent elements, and let $\boldsymbol{y} = C\boldsymbol{s}$ with $C \in \mathbb{R}^{M\times M}$ full rank. Furthermore, let $p(s_i) = \mathcal{N}(0,1)$ for $i = 1,\ldots,L < M$, and s_i not Gaussian but also with zero mean and unit variance for $i = L+1,\ldots,M$. Then a necessary and sufficient condition for mutual statistical independence of the elements of $\boldsymbol{y} \in \mathbb{R}^M$ is that C is of the form*

$$C = \begin{bmatrix} Q & 0 \\ 0 & P \end{bmatrix}, \tag{3.10}$$

with $Q \in \mathbb{R}^{L\times L}$ an orthogonal matrix, and $P \in \mathbb{R}^{M-L\times M-L}$ a permutation matrix.

Proof. Sufficiency is proved first, i.e., it is shown that for C of the form in (3.10) the elements of $\boldsymbol{y}$ are mutually statistically independent. Consider the first L elements of $\boldsymbol{y}$, denoted by $\boldsymbol{y}^{(1)}$. These are a mixture of the L original Gaussian sources $s_i,\ i = 1,\ldots,L$. It then holds that

$$E\left\{\boldsymbol{y}^{(1)}\boldsymbol{y}^{(1)^{\mathrm{T}}}\right\} = E\left\{Q\boldsymbol{s}^{(1)}\boldsymbol{s}^{(1)^{\mathrm{T}}}Q^{\mathrm{T}}\right\} = CC^{\mathrm{T}} = I. \tag{3.11}$$

Here, the second equality is due to unit variance and statistical independence of the original sources, and the third equality due to Q orthogonal. The elements of $\boldsymbol{y}^{(1)}$ are hence uncorrelated. Since they are also jointly Gaussian, being a sum of Gaussian random variables, this implies mutual statistical independence. Next, consider the last $M - L$ elements of $\boldsymbol{y}$, denoted by $\boldsymbol{y}^{(2)}$. Each element of $\boldsymbol{y}^{(2)}$ corresponds to exactly one scaled non-Gaussian source variable. By assumption, the elements of $\boldsymbol{y}^{(2)}$ are hence mutually statistically independent. Now consider $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$. Their joint probability function can be written as

$$\begin{aligned} p\left(\boldsymbol{y}^{(1)},\boldsymbol{y}^{(2)}\right) &= p\left(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(2)}\right)p\left(\boldsymbol{y}^{(2)}\right) = \frac{1}{\det|Q|}p\left(\boldsymbol{s}^{(1)}|\boldsymbol{y}^{(2)}\right)p\left(\boldsymbol{y}^{(2)}\right) \\ &= p\left(\boldsymbol{s}^{(1)}|\boldsymbol{y}^{(2)}\right)p\left(\boldsymbol{y}^{(2)}\right) = p\left(\boldsymbol{s}^{(1)}\right)p\left(\boldsymbol{y}^{(2)}\right) \\ &= p\left(\boldsymbol{y}^{(1)}\right)p\left(\boldsymbol{y}^{(2)}\right) \end{aligned} \tag{3.12}$$

since $\det|Q| = 1$ due to orthogonality and $\boldsymbol{y}^{(2)}$ corresponds to the (scaled and permuted) statistically independent non-Gaussian source variables. This establishes mutual statistical independence of the elements of $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$, which completes the proof of sufficiency.

To prove necessity, it is shown that any deviation of C from the form in (3.10) leads to a contradiction. First, assume that the elements of $\boldsymbol{y}^{(1)}$ are mutually statistically independent and that Q is not orthogonal (the trivial case of Q being diagonal is neglected here). Then it holds that

$$E\left\{\boldsymbol{y}^{(1)}\boldsymbol{y}^{(1)^{\mathrm{T}}}\right\} = E\left\{Q\boldsymbol{s}^{(1)}\boldsymbol{s}^{(1)^{\mathrm{T}}}Q^{\mathrm{T}}\right\} = CC^{\mathrm{T}} \neq I. \tag{3.13}$$

The elements of $\boldsymbol{y}^{(1)}$ are hence correlated, which is a contradiction to the assumption of mutual statistical independence of the elements of $\boldsymbol{y}^{(1)}$. Next, assume that

the elements of $\boldsymbol{y}$ are mutually statistically independent, and consider the upper right block of zeros in C. If at least one of these elements is not equal to zero, there is at least one element in $\boldsymbol{y}$ that is a mixture of at least one Gaussian source and one non-Gaussian source (denoted by $s^{(*)}$). Since the elements of $\boldsymbol{y}$ are assumed mutually statistically independent, it can be concluded that $s^{(*)}$ is Gaussian by Theorem 3.2. This is a contradiction to the assumptions. A similar argument can be applied if P is not a permutation matrix. In this case, there is at least one element of $\boldsymbol{y}$ which is a mixture of at least two non-Gaussian sources. Due to the assumed mutual independence of $\boldsymbol{y}$ it can be concluded by Theorem 3.2 that the non-Gaussian sources are Gaussian, which is again a contradiction. Finally, consider the lower left block of zeros in C. Assume this block contains one row $\boldsymbol{b} \in \mathbb{R}^L$ with at least one non-zero element. Then the corresponding element of $\boldsymbol{y}^{(2)}$ can be written as $y^{(b)} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{s}^{(1)} + \lambda s^{(*)}$, with $\lambda \in \mathbb{R}$ and $s^{(*)}$ one of the non-Gaussian sources. Now consider the covariance of the elements of $\boldsymbol{y}^{(1)}$, denoted by $y_i^{(1)},\ i = 1, \ldots, L$, and $y^{(b)}$. It then holds that

$$E\left\{y_i^{(1)} y^{(b)}\right\} = E\left\{\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{s}^{(1)}\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{s}^{(1)} + \lambda s^{(*)}\right)\right\} = \boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{b}, \tag{3.14}$$

due to the assumptions of unit variance and statistical independence of the original sources. Now there exists at least one $i \in \{1, \ldots, L\}$ for which $\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{b} \neq 0$, since the rows $\boldsymbol{q}_i$ of Q form a complete orthogonal basis. For this row of Q it hence holds that $E\left\{y_i^{(1)} y^{(b)}\right\} \neq 0$. Consequently, $y^{(b)}$ is correlated with at least one element of $\boldsymbol{y}^{(1)}$, which is a contradiction to mutual statistical independence of $\boldsymbol{y}$. This concludes the proof. □

Due to (3.9) and theorem 3.3, possible solutions of an ICA algorithm in the presence of multiple Gaussian sources are given by

$$\boldsymbol{y}(t) = W^*\boldsymbol{x}(t) = \begin{bmatrix} Q & 0 \\ 0 & P \end{bmatrix} A^{-1}A\boldsymbol{s}(t) = \begin{bmatrix} Q & 0 \\ 0 & P \end{bmatrix} \boldsymbol{s}(t). \tag{3.15}$$

It can thus be concluded that in the presence of multiple Gaussian sources the non-Gaussian sources are still correctly reconstructed by ICA. The Gaussian sources, on the other hand, are arbitrarily mixed together. For the reconstructed mixing matrix, solutions are given by

$$\hat{A} = W^{*-1} = AC^{-1} = A\begin{bmatrix} Q^{\mathrm{T}} & 0 \\ 0 & P^{-1} \end{bmatrix}. \tag{3.16}$$

Since P^{-1} is also a permutation matrix, the columns of A corresponding to non-Gaussian sources are correctly reconstructed (up to the usual permutation and scaling). The columns associated with Gaussian sources, however, are arbitrarily mixed together. This result is illustrated in the following simple example.

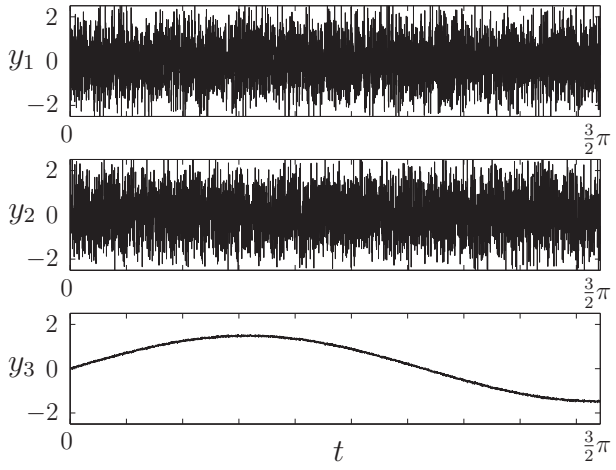

Figure 3.2: Reconstructed sources

Example 3.1. *Consider the case $M = 3$ with one non-Gaussian source with sub-Gaussian distribution*

$$s_1(t) = \sin(t) \quad , \quad t \in \left[0, \tfrac{3}{2}\pi\right] \tag{3.17}$$

and two Gaussian sources with zero mean and unit variance,

$$s_2, s_3 \sim N(0, 1), \tag{3.18}$$

each sampled with 5000 data points. The sources are mixed according to

$$\boldsymbol{x} = A\left[s_1, s_2, s_3\right]^T \tag{3.19}$$

with a randomly generated full rank non-orthogonal matrix

$$A = \begin{pmatrix} -0.1735 & 0.7240 & -0.1545 \\ 0.3621 & 0.4088 & 0.7137 \\ 0.9158 & -0.5556 & 0.6832 \end{pmatrix}. \tag{3.20}$$

The original sources are then reconstructed as

$$\boldsymbol{y}(t) = W^* \boldsymbol{x}(t), \tag{3.21}$$

with W^* obtained with the extended Infomax algorithm [LGS99]. The reconstructed signals are shown in Fig. 3.2 with normalized variance to remove scaling indeterminacies. As can be seen in the third panel, signal s_1 is reconstructed despite the presence of two sources with Gaussian distribution.
Then, fifty reconstructions of $\boldsymbol{y}$ using the extended Infomax algorithm with uniformly distributed initial conditions W_0 are carried out. Inverting the resulting unmixing matrices delivers the representations $(A_i, \boldsymbol{s}_i), i = 1, \ldots, 50$ of $\boldsymbol{x}$. Normalizing the columns of all matrices A_i to remove scaling indeterminacies and

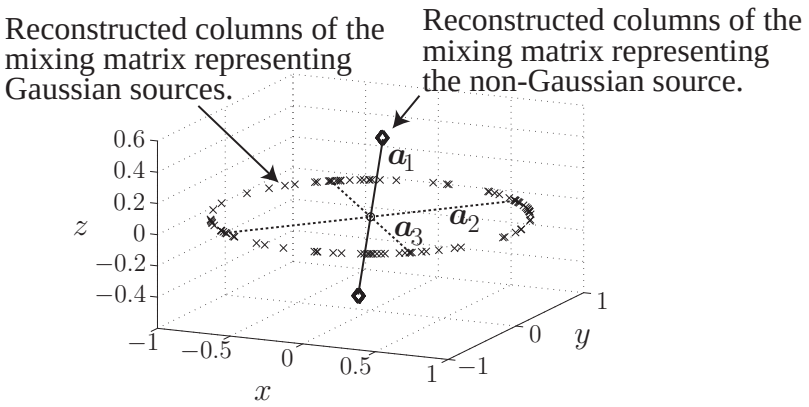

Figure 3.3: Original ($\boldsymbol{a}_i, i = 1 \ldots 3$) and reconstructed columns of the mixing matrix

plotting these together with the original columns of (3.20) results in Fig. 3.3. As expected from (3.16), the column $\boldsymbol{a}_1$ of the non-Gaussian source is consistently reconstructed, while the reconstructed columns associated with the Gaussian sources are randomly distributed in the subspace spanned by the two columns $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ of (3.20) associated with the Gaussian sources s_2 and s_3. Note that for better visualization the data has been rotated such that the subspace spanned by $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ coincides with the xy - plane.

Application to Source Localization

The above discussion shows that ICs with a Gaussian distribution constitute a mixture of originally Gaussian sources. If the associated topography of such an IC is used for source localization, as described in Section 3.2.2, the location of the obtained current dipole does not correspond to the location of a any true source within the brain due to (3.16). Consequently, Gaussian ICs should be excluded from the analysis. This requires an identification of those ICs that have a Gaussian distribution. The most straight-forward way to do this is to estimate the deviation of every IC from a Gaussian distribution with identical variance, and specify a lower bound on this deviation. If the deviation from Gaussianity of an IC falls below this bound, it is excluded from further analysis. This, however, requires the lower bound to be matched to the capability of an ICA algorithm to correctly reconstruct sources close to a Gaussian distribution. Moreover, for finite data no IC can have an exact Gaussian distribution. It would hence be desirable to have a methodology that excludes only those ICs that are too close to a Gaussian distribution in order to be consistently reconstructed by a certain algorithm. Such a methodology is now presented.
Consider K representations of the observed data $X = A_i S_i,\ i = 1, \ldots, K$ obtained by running ICA K times on the observed data X with randomized initial conditions. For each original source with a non-Gaussian distribution there are K linearly dependent columns in the set $\{A_1, \ldots, A_K\}$, all representing the topography of this

source. The topographies of the original sources that are too close to Gaussianity for consistent reconstruction, on the other hand, are not represented multiple times in the set $\{A_1, \ldots, A_K\}$. Instead, if there are L sources that can not be consistently reconstructed, KL columns of $\{A_1, \ldots, A_K\}$ are randomly distributed in a L-dimensional subspace of $\mathbb{R}^M$. If then every column of $\{A_1, \ldots, A_K\}$ is localized by a single current dipole, as described in Section 3.2.2, there is a spatial accumulation of dipoles at those positions within the brain that correspond to the origin of non-Gaussian ICs. Dipoles corresponding to Gaussian ICs are randomly distributed within the brain.
This density of current dipoles can be estimated by the Parzen window method. Let $\hat{\boldsymbol{r}}_i,\ i = 1, \ldots, KM$, be the locations of the KM current dipoles obtained by solving (3.7) for all columns $\boldsymbol{a}_i$ of the set $\{A_1, \ldots, A_K\}$. The density of the the current dipoles at a certain location $\boldsymbol{r}$, termed the Activation Density Function (ADF), is then estimated as

$$\text{ADF}(\boldsymbol{r}) = \frac{1}{KM} \sum_{i=1}^{KM} h(\boldsymbol{a}_i) g(r, \hat{r}_i), \tag{3.22}$$

with

$$g(r, \hat{r}_i) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{\|r - \hat{r}_i\|_2^2}{2\sigma_k^2} \right) \tag{3.23}$$

the Gaussian kernel with variance σ_k^2,

$$h(\boldsymbol{a}_i) = c_h(1 - \tanh(a_h \text{rv}(\boldsymbol{a}_i) - b_h)), \tag{3.24}$$

and $\text{rv}(\boldsymbol{a}_i)$ the normalized residual variance of approximating $\boldsymbol{a}_i$ by a single current dipole as obtained by solving (3.7). The function $h(\boldsymbol{a}_i)$ hence ensures that ICs with a high residual variance, i.e., ICs that can not be represented reasonably well by a single current dipole, do not contribute to the evaluation of the ADF-function.
The peaks of the ADF thereby identify the origins of ICs that can be modelled reasonably well by a single current dipole and can be consistently reconstructed, i.e., that represent non-Gaussian sources. The ADF thus constitutes an estimate of the spatial distribution of non-Gaussian sources within the brain.

3.3 Experimental Results

In this section, the methodology for feature extraction described in the previous section is used to classify imaginary tapping movements of the left and right index finger using EEG data.
EEG signals caused by real and imaginary movements of the left and right index finger were recorded from one subject (age 26, normal vision, no known neurological disorders and no prior experience with BCIs or imaginary movements). The subject sat in a shielded and dimly lit room in front of a computer screen, and was instructed to perform real and imaginary tapping movements with the left or right

M	K	σ_k^2	a_h	b_h	c_h
60	50	20	30	10	0.5

Table 3.2: Parameters used for the ADF.

index finger. These tapping movements were to be performed in synchrony with a centrally displayed grey box, flashing with a frequency of 1.33 Hz on a black background. A control condition was added in which the subject passively had to watch the flashing box. Each of the five blocks (real movement right (MR), real movement left (ML), imaginary movement right (IR), imaginary movement left (IL), no movement (NG)) consisted of 100 movements/flashes, and was repeated ten times in pseudo-randomized order. Each block was followed by a break of five seconds in which the instructions for the next block were displayed. EEG was recorded continuously with BrainAmp-Amplifiers (BrainProducts Inc.) with $M = 60$ channels according to the extended 10-20 system at 5 kHz sampling rate. Additionally, vertical and horizontal eye movements were monitored. The data was recorded with FPz as reference, and re-referenced offline to common average reference.

To ensure that no covert muscle activation took place during the imaginary conditions, EMG activity was recorded bipolarly using standard forearm flexor placement [Lip67]. EMG recordings were then band-pass filtered with 4 Hz and 100 Hz cut-off frequencies and half-rectified. Trials of imaginary movements were chosen to be rejected if the mean EMG activity during the trial exceeded 10% of the maximal EMG activity of the corresponding real movement [VMMW98]. No trials had to be rejected.

Ocular correction was performed [GCD83], and trials with onset of flashing boxes were averaged separately for each condition. For conditions MR and ML, the grand average of all 1000 trials for each condition was taken. For conditions IR and IL, the average was computed for each block of 100 trials separately. This resulted in one data set per condition MR and ML, and ten data sets per condition IR and IL.

The following steps were then applied to each of the data sets with the parameters shown in Tab. 3.2. First, the grand average of condition NG was subtracted from each data set to eliminate task irrelevant activity (e.g., visual evoked responses). Subsequently, ICA was applied K times to the data set by using the extended Infomax-algorithm as implemented in EEGLAB [DM04]. This resulted in KM ICs, each of which was then localized as described in Section 3.2.2. In a fourth step, the locations of all ICs were used to compute the ADF as given in (3.22). This resulted in one ADF for each of the conditions MR and ML, and ten ADFs for each of conditions IR and IL.

The actual classification was then performed in the following way. In a first step, the location of maximal activity for conditions MR and ML was determined as given

a) b)

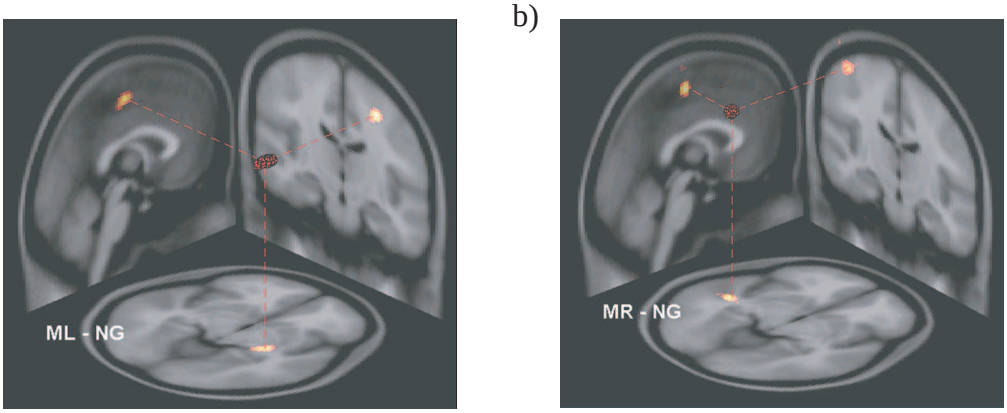

Figure 3.4: Maxima of the ADF for tapping movements of a) the left and b) the right index finger.

by the corresponding ADFs, i.e.,

$$\boldsymbol{r}_{\mathrm{MR}} = \underset{\boldsymbol{r}}{\operatorname{argmax}} \left\{ \mathrm{ADF}(\boldsymbol{r}|\mathrm{MR}) \right\}, \tag{3.25}$$

$$\boldsymbol{r}_{\mathrm{ML}} = \underset{\boldsymbol{r}}{\operatorname{argmax}} \left\{ \mathrm{ADF}(\boldsymbol{r}|\mathrm{ML}) \right\}. \tag{3.26}$$

The resulting maxima, shown in Fig. 3.4 superimposed on MRI images from the publicaly available MNI database, are located in neuro-physiologically plausible areas, i.e., in vicinity of the hand areas of the left and right motor cortex.
To determine the correct class of a data set caused by an imaginary movement, its respective ADF was evaluated at the positions of maximal activation for real movements $\boldsymbol{r}_{\mathrm{MR}}$ and $\boldsymbol{r}_{\mathrm{ML}}$. If $\mathrm{ADF}(\boldsymbol{r}_{\mathrm{MR}}) \geq \mathrm{ADF}(\boldsymbol{r}_{\mathrm{ML}})$, the data set was classified as being caused by an imaginary movement of the right index finger and vice versa. This very simple classification procedure is based on evidence that real and imaginary movements are indeed correlated with activity in overlapping brain regions [PL99].
This procedure was used to classify all 20 data sets of imaginary movements, and resulted in nine out of ten correct classifications for condition IR, and eight out of ten correct classifications for condition IL. Thus a total of 17 out of 20 data sets (85%) were correctly classified.

3.4 Discussion

In this chapter, the viability of source localization for feature extraction in non-invasive BCIs was investigated. By combining ICA with source localization, it was shown that without any knowledge on the temporal aspects of neural coding, i.e., how cognitive states are temporally encoded in the electric field of the brain, a low-dimensional feature space providing information on the BCI-user's intention

can be extracted. This feature space was shown to provide enough information to classify imaginary tapping movements of the left and right index finger with an accuracy of 85% in one untrained subject. The main theoretical contribution of this chapter is Theorem 3.3, characterizing the behavior of ICA in presence of multiple Gaussian sources. This result was further used to derive a methodology for identifying unstable ICs without explicitly estimating their Gaussianity.

While the work presented in this chapter establishes the viability of source localization for non-invasive BCIs, the results can only be considered preliminary due to several factors. First, the proposed methodology was only tested experimentally on one subject. Clearly, a sound evaluation and assessment of the capabilities of the proposed methodology require validation on experimental data from multiple subjects and substantially more test data. Second, the classification results were obtained using averaged data of multiple trials. Ideally, a BCI should be capable of processing single-trial data in order to realize online control of an effector such as a computer cursor. Finally, some aspects of the employed procedure are very simplistic and surely sub-optimal. The proposed methodology is based on the assumptions that a) the activity of only two areas within the brain provide all information on the BCI-user's intention, b) real and imaginary tapping movements lead to peak activations at the same spatial locations within the brain, and c) imaginary tapping movements of one finger are accompanied by larger contralateral than ipsilateral activity in areas of the motor cortex. In terms of the general procedure of feature extraction for non-invasive BCIs (Definition 2.13), assumption a) amounts to a restriction on the dimension of the new feature space $\hat{X}$, and assumption b) specifies the cost function f that forms an estimate of the expected Bayes error. Assumption three is not related to feature extraction, but specifies the actual classification procedure. All three assumptions are questionable, and could be easily eliminated by combining the proposed methodology for feature extraction with state-of-the-art classification algorithms.

Interestingly though, the classification accuracies reported in two other studies on non-invasive BCIs utilizing source localization for feature extraction are within the same range as the accuracy reported here [QDH04, KLH05], while in a third study classification accuracies above 95% were obtained [GGP$^+$05]. Comparing these studies with each other and the results presented in this chapter, one crucial difference can be found. While in the results reported here as well as in [QDH04] and [KLH05] only information on the spatial distribution of brain activity was employed, in [GGP$^+$05] spatial *and* temporal features were used for classification. While there is very little knowledge on how cognitive states are temporally encoded within the electric field of the brain, it is known that motor imagery is accompanied by frequency specific changes in variance of the electric field originating in motor areas [PL99]. Even though this is a very simple measure, the results reported in [GGP$^+$05] demonstrate that using this information results in a significant increase in classification accuracy.

In summary, the lesson learned from using source localization for feature extrac-

tion in non-invasive BCIs is the following. Source localization alone can provide information on the BCI-user's intention. However, if information on how cognitive states are encoded in temporal properties of the electric field is available, no matter how limited it may be, this information can and should be used.

Chapter 4

Information Theoretic Feature Extraction

4.1 Introduction

In the previous chapter, only the spatial distribution of current density within the brain was used for inferring the BCI-user's intention. While this was shown to be a viable option, it is demonstrated in [GGP+05] that additionally making use of information on how cognitive states are temporally encoded within the electric field results in a significant increase in classification accuracy, even if the available information on temporal coding is rather limited. In this chapter, the available information on temporal coding of EEG/MEG signals is used to design a feature extraction algorithm that (under some assumptions) is optimal in terms of maximizing an approximation of mutual information of class labels, i.e., the BCI-user's intention and extracted features.

As briefly discussed in Sections 1.2 and 3.4, it is known that some information on cognitive states is encoded in the power of specific frequency bands of the electric/magnetic field in specific brain regions (reviewed in [PL99]). An increase in the power of the EEG/MEG is usually termed event related synchronization (ERS), and a decrease is referred to as event related desynchronization (ERD). This is due to the fact that an increase in EEG/MEG power is caused by temporal synchronization of the electric field across cortical columns [NS05]. Quite contrary to intuition, ERD is usually associated with increased activity of a certain brain region, while ERS can often be observed during rest. The significance of this process for non-invasive BCIs was first realized by the group of J. Wolpaw [WMNF91]. The work presented in this chapter is primarily based on the seminal work in [PNFP97], in which it is demonstrated that imaginary movements of different limbs lead to strong ERD over the contralateral motor cortex. These changes can be used to infer the BCI-user's intention without or with only little subject training. However, in [PNFP97] average classification accuracies of only about 80% were obtained, which can be attributed to a lack of sophisticated feature extraction algorithms.

As demonstrated in [GGP+05], it is possible to combine feature extraction by source localization with ERD/ERS caused by motor imagery to construct non-invasive BCIs with a high classification accuracy. However, source localization is a computationally intensive procedure. If information is available on which temporal properties of the recorded EEG/MEG data provide information on the BCI-user's intention, it should be possible to derive less computationally intensive algorithms that selectively extract those components of the recorded data that are optimal for inferring the BCI-user's intention. This was already realized by Ramoser et al. in [RMGP00], in which for two-class paradigms spatial filters are devised that extract those components of the recorded data which variances maximally vary between conditions. Using this algorithm, termed Common Spatial Patterns (CSP), it was shown that in a two-class paradigm classification accuracies close to 100% could be achieved. CSP has become one of the most frequently used algorithms for feature extraction in BCIs, and was also used in the winning entry of the BCI competition 2003 [BB04]. Its improvement, especially its extension to the spectral domain, is an active field of research (cf. [LBCM05, DBCM04, TDN+06, FHLS06] and the references therein).

In this chapter, a conceptually different approach to spatial filtering is taken. Under the assumption that the user's intention is encoded in variance changes of components of the EEG/MEG data, spatial filters are derived that maximize (an approximation of) mutual information of the user's intention and extracted EEG/MEG components. This approach, termed Information Theoretic Feature Extraction (ITFE), has got the advantage that maximizing mutual information provides a direct link to minimizing the minimum Bayes error as discussed in Section 2.2. It is proved that for two-class paradigms the obtained spatial filters are identical to those obtained by CSP, thereby establishing the optimality of CSP in terms of maximizing an approximation of mutual information. An extension of the CSP algorithm for multi-class paradigms proposed in [DBCM04], on the other hand, is shown to be suboptimal. This deficiency of multi-class CSP is resolved by showing how this algorithm can be rendered optimal in the framework of ITFE. To support the theoretical results, multi-class CSP and multi-class ITFE are applied to experimental EEG data from a four-class motor imagery paradigm provided by the Laboratory of Brain-Computer Interfaces at the Technische Universität Graz for the third BCI competition, and it is shown that multi-class ITFE leads to an average increase in classification accuracy of 23.4% in comparison to multi-class CSP.

The structure of this chapter is as follows. In Section 4.2, the assumptions made on $\mathcal{P}^*$ and f for the CSP and ITFE algorithms are specified (cf. Section 2.4), and the CSP algorithm is presented for two-class and multi-class paradigms. Then, it is shown how ITFE can be realized in this context by deriving an approximation of mutual information of class labels and extracted EEG/MEG components. This approximation of mutual information is then used to prove the optimality of two-class CSP, and to show how multi-class CSP can be rendered optimal. After demonstrating some experimental results in Section 4.3, the chapter concludes with

a discussion of the limitations of CSP and ITFE in Section 4.4.

4.2 Methods

In this chapter, the same notation as in Chapter 3 is used. The BCI-user's intention is again denoted by $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$, and the recorded EEG/MEG data by $X \in \mathbb{R}^{M \times T}$ for a block of data and $\boldsymbol{x}(t) \in \mathbb{R}^M$ for a single sample point recorded at M electrodes. If the time index t is dropped, $\boldsymbol{x}$ is considered as a M-dimensional random variable with probability density function $p(\boldsymbol{x})$. All results equally apply to EEG and MEG signals. Since the experimental evaluation is carried out using EEG signals, only this modality is subsequently referred to. The precise assumptions made on the class of allowed features $\mathcal{P}^*$ (cf. Definition 2.13) are as follows:

1. The BCI-user's intention is encoded in variance changes of the recorded EEG data.

2. For each subject and paradigm, those components of the electric field of the brain that provide information on the subject's intention originate in spatially invariant brain regions.

It should be emphasized again that the first assumption is a working assumption. While there is certainly more to neural coding in the electric field of the brain than variance changes, these provide a basis for developing useful feature extraction algorithms. The second assumption, which was already employed in Chapter 3, expresses our knowledge on localized information processing in the brain, i.e., that (at least for low-level information processing) certain brain areas are specialized for certain tasks. Since propagation of the electric field of a certain brain region to the EEG electrodes is linear (cf. Section 3.2.1), and brain regions relevant for a certain task are assumed invariant, this limits the class of transformations that have to be considered to time-invariant linear spatial filters.
The desired feature extraction algorithm hence takes the form $T : \mathbb{R}^{M \times T} \mapsto \mathbb{R}_+^{KL}$, $T(X) = \text{Var}\left\{W^{\mathrm{T}} X\right\}$, with $W \in \mathbb{R}^{M \times L}$ the matrix of $L << M$ spatial filters and $K \in \mathbb{N}$ the number of analyzed frequency bands of each component. It then remains to specify f of Definition 2.13 in this context, i.e, the cost function used to estimate the expected error probability.

4.2.1 Two-class Common Spatial Patterns

In this section, a two-class paradigm is assumed, i.e., $\mathcal{C} = \{c_1, c_2\}$. The CSP algorithm then solves the optimization problem [PSGS05]

$$\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathbb{R}^M}{\operatorname{argmax}} \left\{ \frac{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w}} \right\}, \tag{4.1}$$

with $R_{\boldsymbol{x}|c_1}, R_{\boldsymbol{x}|c_2}$ the covariance matrices of $\boldsymbol{x}$ given c_1, c_2 respectively. Since (4.1) is in the form of the well-known Rayleigh quotient, solutions to (4.1) are given by eigenvectors of the generalized eigenvalue problem

$$R_{\boldsymbol{x}|c_1}\boldsymbol{w} = \lambda R_{\boldsymbol{x}|c_2}\boldsymbol{w}. \tag{4.2}$$

The eigenvectors of (4.2) thus correspond to the desired spatial filters. Furthermore, for a given eigenvector $\boldsymbol{w}^*$ the corresponding eigenvalue determines the value of the cost function:

$$\lambda^* = \frac{\boldsymbol{w}^{*\mathrm{T}} R_{\boldsymbol{x}|c_1}\boldsymbol{w}^*}{\boldsymbol{w}^{*\mathrm{T}} R_{\boldsymbol{x}|c_2}\boldsymbol{w}^*}. \tag{4.3}$$

The eigenvalues thus are a measure for the quality of the obtained spatial filters, i.e., the eigenvalue associated with a spatial filter expresses the ratio of the variance between conditions of the component of the EEG data extracted by the spatial filter. Feature extraction is then usually done by combining the L eigenvectors of (4.2) with the smallest/largest eigenvalues to form $W \in \mathbb{R}^{M\times L}$ and computing $T(X) = \mathrm{Var}\left\{W^{\mathrm{T}}X\right\}$. Note that L has to be specified in advance and determines the dimension of the new feature space.
The cost function maximized in the CSP algorithm in order to minimize the expected error probability is thus given by

$$f_{\mathrm{CSP}}(c, T(X)) = \sum_{i=1}^{L} \max\left\{\frac{\boldsymbol{w}_i^{\mathrm{T}} R_{\boldsymbol{x}|c_1}\boldsymbol{w}_i}{\boldsymbol{w}_i^{\mathrm{T}} R_{\boldsymbol{x}|c_2}\boldsymbol{w}_i}, \frac{\boldsymbol{w}_i^{\mathrm{T}} R_{\boldsymbol{x}|c_2}\boldsymbol{w}_i}{\boldsymbol{w}_i^{\mathrm{T}} R_{\boldsymbol{x}|c_1}\boldsymbol{w}_i}\right\}, \tag{4.4}$$

with $\boldsymbol{w}_i$ the i^{th} column of W. While choosing spatial filters that extract those components of the EEG with maximum ratio of variance between conditions seems sensible, it is in fact an open question whether this is optimal in terms of minimizing the optimal Bayes error or the expected error probability.

4.2.2 Multi-class Common Spatial Patterns

Extending CSP to multi-class paradigms, i.e., again letting $\mathcal{C} = \{c_1, \ldots, c_N\}$, is either done by performing two-class CSP on different combinations of classes (e.g., by computing CSPs for all combinations of classes or by computing CSP for one class versus all other classes), or by joint approximate diagonalization (JAD) (cf. [DBCM04] and the references therein). Since the first approach is conceptually identical to CSP for two-class paradigms, the focus here is on CSP by JAD.
Given EEG data of N different classes, the goal of CSP by JAD is to find a transformation $W \in \mathbb{R}^{M\times M}$ that diagonalizes the covariance matrices $R_{\boldsymbol{x}|c_i}$, i.e.,

$$W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W = D_{c_i},\ i = 1, \ldots, N, \tag{4.5}$$

with $D_{c_i} \in \mathbb{R}^{M\times M}$ diagonal matrices. There are several approaches to this problem (discussed in [ZLNM04]), the details of which are not of interest here. The

idea behind using JAD for multi-class CSP lies in the fact that CSP for two classes can be understood as diagonalizing two covariance matrices. More precisely, if the eigenvectors of the generalized eigenvalue problem (4.2) are combined in a matrix W, then $W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W = D_{c_i},\ i = 1, \ldots, 2$. It then seems plausible to extend CSP to multi-class paradigms by finding a transformation W that approximately diagonalizes multiple covariance matrices. A total of L columns of the obtained matrix W are then taken as the desired spatial filters.
There are, however, two caveats. First, this approach is motivated heuristically and lacks a firm theoretical foundation. Second, it remains unclear which columns of W provide the optimal spatial filters. Or, as it is put in [DBCM04], *as opposed to the two-class problem, there is no canonical way to choose the relevant CSP patterns for multi-class CSP*. In [DBCM04], the following heuristic is proposed to choose the L optimal spatial filters: Given a matrix W obtained by JAD, compute the eigenvalues of all covariance matrices, i.e., compute $\boldsymbol{\lambda}_i = \mathrm{diag}\{W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W\},\ i = 1, \ldots, N$. Then map all $j = 1, \ldots, M$ eigenvalues of each class $i = 1, \ldots, N$ to $\lambda_{i,j} = \max\{\lambda_{i,j}, 1/(1 + (N-1)^2 \lambda_{i,j}/(1 - \lambda_{i,j}))\}$, and select the L/N eigenvectors with the largest transformed eigenvalues of each class as spatial filters. If one eigenvector is selected more than once, replace it by the eigenvector with the next highest transformed eigenvalue.
One disadvantage of this extension of CSP to multiple classes is that the cost function f that is optimized is not specified. It is hence unclear whether multi-class CSP is optimal in any sense. In this chapter, it is shown that multi-class CSP by JAD is equivalent to ICA. This allows treating multi-class CSP in the framework of ITFE, which can be used to derive a methodology to select those spatial filters of multi-class CSP that are optimal in terms of maximzing (an approximation of) mutual information of class labels and extracted EEG components. A theoretical foundation for multi-class CSP by JAD is thereby provided, and the need for heuristics in choosing spatial filters is eliminated.

4.2.3 Information Theoretic Feature Extraction

In this section, the framework of Information Theoretic Feature Extraction (ITFE) for feature extraction is introduced. ITFE has recently received considerable attention in the machine learning community, mostly in a non-parametric setting (cf. [PXZF00, Tor03]). The general idea of ITFE is to find a transformation that directly maximizes mutual information of class labels and extracted features. This is desirable, since maximizing mutual information minimizes bounds on the optimal Bayes error as discussed in Section 2.2. The cost function that is maximized in this context is hence given by

$$f_{\mathrm{ITFE}}\left(c, W^{\mathrm{T}} X\right) = I\left(c, W^{\mathrm{T}} X\right) \tag{4.6}$$

with $W \in \mathbb{R}^{M \times L}$.

To find the spatial filters that maximize (4.6) the optimization problem

$$\boldsymbol{w}_i^* = \underset{w\in\mathbb{R}^M}{\operatorname{argmax}} \left\{ I\left(c, \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}\right)\right\}, \tag{4.7}$$

with $\boldsymbol{w}_i$ the i^{th} column of W and $i = 1, \ldots, L$ is considered. Note that (4.7) implies that the desired spatial filters are derived sequentially. While it is also possible to extract several components simultaneously, this is equivalent to extracting components sequentially in the setting considered here as is shown later. Furthermore, note that (4.7) requires computing mutual information of a discrete and a continuous variable. To make this expression well defined, it is necessary to assume a quantization that discretizes the continuous variable $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}$. This quantization scheme, however, has negligible influence on the mutual information, since the entropy of a n-bit quantization of a continuous random variable is approximately the entropy of the continuous variable plus n [CT06]. Since the entropy enters twice with different sign into the computation of mutual information, the terms due to the quanitzation cancel out. The quantization scheme is thus disregarded in the sequel and only the differential entropy is employed.
To the best of the author's knowledge no analytic expression of $I(c, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$ for the assumptions made in this chapter exists. Hence, an analytic approximation of $I(c, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$ is first derived. Then, the solution of (4.7) based on this approximation is discussed for two-class paradigms. Finally, the extension to multi-class paradigms is presented.

Approximation of Mutual Information

First note that the mutual information of c and $\hat{x} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ can be written as

$$\begin{aligned} I(c, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) &= H(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) - H(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}|c) = H(\hat{x}) - H(\hat{x}|c) \\ &= H(\hat{x}) - \sum_{i=1}^{N} P(c_i) H(\hat{x}|c_i). \end{aligned} \tag{4.8}$$

Since differential entropy is not scale invariant, it is assumed that $\sigma_{\hat{x}}^2 = 1$. This is no loss of generality, since $\boldsymbol{w}$ can always be scaled to meet this assumption. Now recollect that it is assumed that all information on the BCI-user's intention is contained in variance changes of the EEG. Hence, no information is lost if it is assumed that $p(\boldsymbol{x}|c) = \mathcal{N}(\boldsymbol{0}, R_{\boldsymbol{x}|c})$. Since $\hat{x}$ is a linear combination of the elements of $\boldsymbol{x}$ it also follows a (now one-dimensional) Gaussian distribution with zero mean, i.e., $p(\hat{x}|c) = \mathcal{N}(0, \sigma_{\hat{x}|c}^2)$. It is then possible to express the entropy of $\hat{x}$ given class c_i as

$$H(\hat{x}|c_i) = \log\sqrt{2\pi e \sigma_{\hat{x}|c_i}^2} = \log\sqrt{2\pi e \boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i}\boldsymbol{w}}. \tag{4.9}$$

The marginal distribution $p(\hat{x})$, however, does not follow a Gaussian distribution since

$$p(\hat{x}) = \sum_{i=1}^{N} P(c_i) p(\hat{x}|c_i) = \sum_{i=1}^{N} P(c_i)\mathcal{N}(0, \sigma_{\hat{x}|c_i}), \tag{4.10}$$

which is a sum of N Gaussian distributions and thus not itself Gaussian. To the best of the author's knowledge there is no analytical solution to the entropy of a sum of Gaussian distributions, and thus no closed form solution of $H(\hat{x})$. It is possible, however, to approximate $H(\hat{x})$ in the following manner. First, note that the entropy of $\hat{x}$ can be expressed as

$$H(\hat{x}) = H_g(\hat{x}) - J(\hat{x}), \tag{4.11}$$

with $H_g(\hat{x})$ the entropy of a Gaussian random variable with the same variance as $\hat{x}$ and $J(\hat{x})$ the (always positive) negentropy of $\hat{x}$. The negentropy of $\hat{x}$ can then be approximated as

$$J(\hat{x}) \approx \frac{1}{12}\kappa_3(\hat{x})^2 + \frac{1}{48}\kappa_4(\hat{x})^2, \tag{4.12}$$

with the third- and fourth-order cumulants $\kappa_3(\hat{x}) = E\{\hat{x}^3\}$ and $\kappa_4(\hat{x}) = E\{\hat{x}^4\} - 3$ [Com94]. Since $p(\hat{x})$ is a sum of Gaussian distributions with zero mean it is symmetric, and hence $\kappa_3(\hat{x}) = 0$. Furthermore, $\kappa_4(\hat{x}) = 3\sum_{i=1}^{N} P(c_i)\left(\sigma^4_{\hat{x}|c_i} - 1\right)$ since the fourth moment of a Gaussian distribution with zero mean and unit variance equals three and $\kappa_4(\alpha x) = \alpha^4\kappa_4(x)$ (see any textbook on advanced statistics). Combining (4.11) and (4.12) yields

$$H(\hat{x}) \approx \log\sqrt{2\pi e} - \frac{3}{16}\left(\sum_{i=1}^{N} P(c_i)\left(\sigma^4_{\hat{x}|c_i} - 1\right)\right)^2. \tag{4.13}$$

Combining (4.8), (4.9) and (4.13) an estimate of the mutual information of c and $\hat{x}$ is obtained as

$$I(c,\hat{x}) \approx -\sum_{i=1}^{N} P(c_i)\log\sqrt{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i}\boldsymbol{w}} - \frac{3}{16}\left(\sum_{i=1}^{N} P(c_i)\left((\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i}\boldsymbol{w})^2 - 1\right)\right)^2. \tag{4.14}$$

Note that in terms of the observed data this expression only depends on the variance conditioned on class labels, as required by the assumptions on the class of allowed features employed in this chapter.

Validating the Approximation of Mutual Information

It then remains to investigate the accuracy of this approximation of mutual information. The only approximation used in deriving (4.14) is the approximation of negentropy in (4.12). This approximation is based on an Edgeworth expansion up to order four of the true probability density function (4.10) about its best Gaussian approximation. As such, (4.14) is exact if $p(\hat{x})$ is Gaussian distributed, and the quality of the approximation deteriorates with deviation of $p(\hat{x})$ from Gaussianity. To quantitatively evaluate the accuracy of the approximation of mutual information, the true mutual information in (4.8) was computed by numerical integration (using recursive adaptive Lobatto quadrature as implemented in Matlab) for $\mathcal{C} = \{c_1, c_2\}$

and $\sigma_{\hat{x}|c_1} \in]0, 1]$. Note that this covers the whole range of $\sigma_{\hat{x}|c_i},\ i \in \{1, 2\}$ due to symmetry of (4.8) with respect to $\sigma_{\hat{x}|c_i}$ and the assumption of unit variance of $\hat{x}$. The error of the approximation of mutual information in (4.14) was then evaluated for different prior class probabilities by subtracting the numerically computed true mutual information from the approximation of mutual information. The resulting error (in per cent of the true mutual information) is shown in Fig. 4.1. Note that $\sigma_{\hat{x}|c_1} = 1$ implies $\sigma_{\hat{x}|c_2} = 1$ and hence $p(\hat{x}) = \mathcal{N}(0, 1)$. As expected, the error between the approximated and true mutual information is zero for $\sigma_{\hat{x}|c_1} = 1$ and small for $\sigma_{\hat{x}|c_1}$ close to one. In fact, the error of the approximation is below one per cent for $\sigma_{\hat{x}|c_1} \in [0.84, 1]$. As long as $\sigma_{\hat{x}|c_1} > 0.36$ the error stays below ten per cent. However, for even smaller values of $\sigma_{\hat{x}|c_1}$ the error grows large, limiting the usefulness of the approximation. Qualitatively, this behavior of the approximation is independent of the number of classes, i.e., if $p(\hat{x})$ is close to Gaussianity a small error can be expected also for $M > 2$. Quantitatively, the goodness of the approximation varies as a function of the number of classes. The validity of the approximation in (4.14) for multiple classes is experimentally validated in Section 4.3.

The applicability of the approximation of mutual information in the context of non-invasive BCIs thus depends on by how much EEG/MEG sources that provide information on the user's intention deviate from Gaussianity, i.e., how much their variances vary across conditions. In general, such sources can be expected to be rather close to Gaussianity, and thus the approximation to be accurate, for the simple reason that inferring a BCI-user's intention is a hard task. If variances of EEG/MEG sources providing information on the user's intention would vary significantly across conditions, inferring the user's intention could be expected to be substantially easier than it is the case. This claim is experimentally validated in Section 4.3.

Two-class ITFE and Optimality of Two-class CSP

Solutions to (4.7) based on the above approximation of mutual information are now discussed for two-class paradigms, i.e., it is again assumed that $\mathcal{C} = \{c_1, c_2\}$. Equation (4.14) then reduces to

$$\begin{aligned} I(c, \hat{x}) \approx & -P(c_1) \log \sqrt{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w}} - P(c_2) \log \sqrt{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w}} \\ & - \frac{3}{16} \left(P(c_1) \left(\sigma^4_{\hat{x}|c_1} - 1 \right) + P(c_2) \left(\sigma^4_{\hat{x}|c_2} - 1 \right) \right)^2 . \end{aligned} \tag{4.15}$$

From here on this expression is referred to as mutual information, keeping in mind that it is in fact an approximation thereof. Taking the derivative of (4.15) with

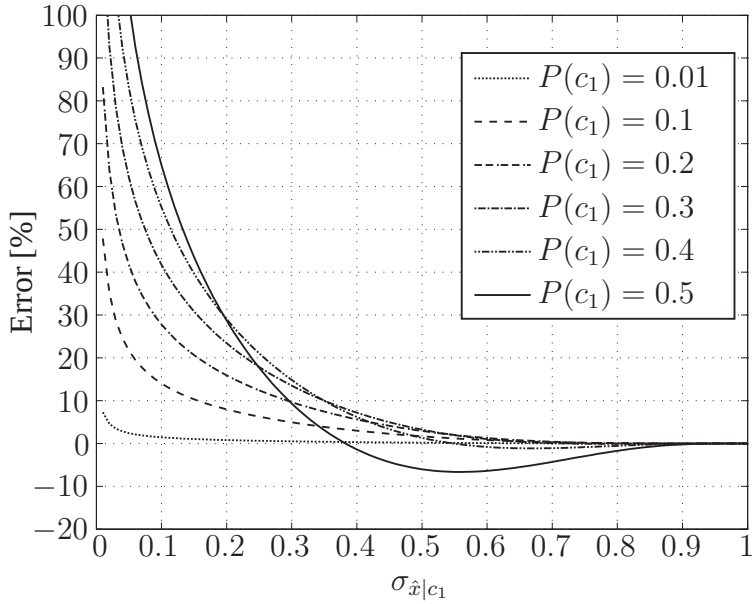

Figure 4.1: Error of the approximation of mutual information (4.14) in per cent for $\mathcal{C} = \{c_1, c_2\}$ as a function of $\sigma_{\hat{x}|c_1}$ for different prior class probabilities.

respect to $\boldsymbol{w}$ then yields

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{w}} I(c, \hat{x}) \;=\; & -\frac{P(c_1)}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w}} R_{\boldsymbol{x}|c_1} \boldsymbol{w} - \frac{P(c_2)}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w}} R_{\boldsymbol{x}|c_2} \boldsymbol{w} \\ & -\frac{3}{2} \left(P(c_1)(\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w})^2 + P(c_2)(\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w})^2 - 1 \right) \\ & \left(P(c_1) \boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w} R_{\boldsymbol{x}|c_1} \boldsymbol{w} + P(c_2) \boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w} R_{\boldsymbol{x}|c_2} \boldsymbol{w} \right) \end{aligned} \tag{4.16}$$

Letting $\alpha_i := -\frac{P(c_i)}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}}$, $\beta_i := -\frac{3}{2} \left(\sum_{j=1}^{2} P(c_j)(\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_j} \boldsymbol{w})^2 - 1 \right) \boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}$, and setting (4.16) to zero results in

$$(\alpha_1 + \beta_1) R_{\boldsymbol{x}|c_1} \boldsymbol{w} + (\alpha_2 + \beta_2) R_{\boldsymbol{x}|c_2} \boldsymbol{w} = 0. \tag{4.17}$$

Rearranging and letting $\lambda := -\frac{\alpha_2 + \beta_2}{\alpha_1 + \beta_1}$ then finally yields

$$R_{\boldsymbol{x}|c_1} \boldsymbol{w} = \lambda R_{\boldsymbol{x}|c_2} \boldsymbol{w}. \tag{4.18}$$

In the case of two-class paradigms and the stated assumptions, solutions to (4.7) are thus given by the eigenvectors of the generalized eigenvalue problem (4.18). Comparing the solutions obtained by ITFE (4.18) and CSP (4.2) shows that for two-class paradigms both methods yield identical spatial filters. Furthermore, if equal class probabilities are assumed and the obtained spatial filters are ranked in terms of the ratio of the variance between conditions (CSP) and in terms of mutual information (ITFE) the ordering is the same. This can be seen by the following argument. For spatial filters obtained by CSP, the maximum ratio of the

variance between conditions of every spatial filter $\boldsymbol{w}^*$ is given by $\max\{\lambda^*, 1/\lambda^*\} = \max\{\sigma^2_{\hat{x}|c_1}/\sigma^2_{\hat{x}|c_2}, \sigma^2_{\hat{x}|c_2}/\sigma^2_{\hat{x}|c_1}\}$. For $\sigma^2_{\hat{x}} = 1$ this is a symmetric and convex function. This also holds true for (4.15), as is easy to check. Hence, if for two spatial filters $\boldsymbol{w}^*_{1/2}$ it holds that $\max\{\lambda^*_1, 1/\lambda^*_1\} > \max\{\lambda^*_2, 1/\lambda^*_2\}$, then also $I(c, \boldsymbol{w}^{*\mathrm{T}}_1 \boldsymbol{x}) > I(c, \boldsymbol{w}^{*\mathrm{T}}_2 \boldsymbol{x})$ and vice versa. As a result, choosing L eigenvectors of (4.18) or (4.2) with maximum ratio of variance between conditions is identical to choosing L eigenvectors with maximum mutual information. Note, however, that this does not hold anymore for unequal class probabilities. In this case $I(c, \hat{x})$ becomes asymmetric, and choosing L spatial filters with maximum ratio of variance between conditions is not identical to choosing L spatial filters with maximum mutual information.

Summarizing the results of this section, it was shown that for equal class probabilities, conditionally Gaussian distributed EEG data, and linear transformations feature extraction by CSP and ITFE lead to the same spatial filters. Under the given assumptions, two-class CSP thus maximizes an approximation of mutual information of extracted EEG components and class labels.

Multi-class Information Theoretic Feature Extraction

Possible solutions of (4.7) for multi-class paradigms, i.e., for $\mathcal{C} = \{c_1, \ldots, c_N\}$, are now discussed. In principle, taking the derivative of (4.14) with respect to $\boldsymbol{w}$ and setting it to zero gives an implicit solution for the spatial filters that correspond to local extrema of (4.14). However, due to the presence of multiple covariance matrices $\partial I(c, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})/\partial \boldsymbol{w} = 0$ can not be formulated as a generalized eigenvalue problem anymore. Furthermore, to the best of the author's knowledge, no analytic solution to this expression exists. This leaves the possibility of deriving a gradient descent rule for finding a solution to (4.7). While this is a straightforward procedure, (4.7) does not constitute a convex optimization problem. Consequently, gradient descent is not an efficient approach for finding all local extrema of (4.14).

Due to these difficulties a different approach is considered. It is assumed that the observed EEG data follows the standard mixing model of Independent Component Analysis (ICA) as discussed in Section 3.2.1, i.e.,

$$\boldsymbol{x} = A\boldsymbol{s}, \tag{4.19}$$

with $\boldsymbol{s} \in \mathbb{R}^M$ a random vector with zero mean representing the original EEG current sources inside the cortex, and $A \in \mathbb{R}^{M \times M}$ a full-rank mixing matrix with each column $\boldsymbol{a}_j,\ j = 1, \ldots, M$ describing the projection strength of source s_j to each of the M EEG electrodes. It is furthermore assumed that $p(\boldsymbol{s}) = \prod_{j=1}^{M} p(s_j)$, i.e., it is assumed that the elements of $\boldsymbol{s}$ are mutually statistically independent. For a discussion of the validity of this model in the context of EEG analysis cf. Section 3.2.1. In addition, it is also assumed that there are only L sources that provide information on the BCI-user's intention. Without loss of generality, these are assumed to be the first L sources, i.e., $I(c, s_i) = 0,\ i = L+1, \ldots, M$.

It is now shown how for this model multi-class ITFE can be realized by JAD, thereby also establishing a connection to multi-class CSP as discussed in Section 4.2.2. First note that the covariance matrix of $\boldsymbol{x}$ given condition c_i is now given by

$$R_{\boldsymbol{x}|c_i} = AR_{\boldsymbol{s}|c_i}A^{\mathrm{T}}, \tag{4.20}$$

with $R_{\boldsymbol{s}|c_i}$ the covariance matrix of $\boldsymbol{s}$ given condition c_i. If JAD is performed, it is obvious that $W^{\mathrm{T}} = A^{-1}$ is a solution of the JAD procedure that diagonalizes all covariance matrices:

$$W^{\mathrm{T}}R_{\boldsymbol{x}|c_i}W = R_{\boldsymbol{s}|c_i} = D_{c_i} \tag{4.21}$$

for $i = 1, \ldots, N$. Note that $R_{\boldsymbol{s}|c_i} = D_{c_i}$ are diagonal matrices because of the mutual independence of the elements of $\boldsymbol{s}$. In this case it holds that

$$\hat{\boldsymbol{x}} = W^{\mathrm{T}}\boldsymbol{x} = W^{\mathrm{T}}A\boldsymbol{s} = \boldsymbol{s}, \tag{4.22}$$

and the obtained spatial filtering matrix W applied to the EEG data results in estimates of the underlying independent components (ICs) of the observed data. It remains to be established if, or under which conditions, $W^{\mathrm{T}} = A^{-1}$ is the only matrix that jointly diagonalizes all covariance matrices. This question of uniqueness has been addressed for orthogonal mixing matrices A (or for sphered data) in [BAMCM97] and for arbitrary mixing matrices in [ten02]. It turns out that in the context considered here a necessary and sufficient condition for $W^{\mathrm{T}} = A^{-1}$ to be the unique joint diagonalizer (up to scaling and permutations) of $R_{\boldsymbol{x}|c_i},\ i = 1, \ldots, N$, is that the matrix

$$S := \begin{bmatrix} \sigma^2_{s_1|c_1} & \cdots & \sigma^2_{s_M|c_1} \\ \vdots & \ddots & \vdots \\ \sigma^2_{s_1|c_N} & \cdots & \sigma^2_{s_M|c_N} \end{bmatrix} \tag{4.23}$$

has no pair of proportional columns, i.e, that for no pair of ICs the variances covary across conditions. Under these conditions any JAD procedure that converges, i.e., that jointly diagonalizes all covariance matrices, returns a matrix W that, if applied to the observed EEG data, returns (scaled and permuted) estimates of the underlying ICs according to (4.22). While it is not possible to ensure a-priori that the variances of no pair of ICs covary across conditions, this can be considered highly unlikely. Consequently, JAD of the EEG covariance matrices conditioned on the class labels, and thus multi-class CSP as discussed in Section 4.2.2, can be considered an implementation of ICA.

It then remains to be established which columns of the unmixing matrix W should be chosen as spatial filters. While so far this choice was based on heuristics such as the one presented in Section 4.2.2, the framework of ITFE suggests to choose those spatial filters that maximize mutual information of extracted EEG components and class labels. Now note that if the ICA model (4.19) and the uniqueness condition hold, a matrix W obtained by JAD that diagonalizes all EEG covariance matrices

conditioned on class labels implies that

$$I(c, \boldsymbol{x}) = I\left(c, W^{\mathrm{T}}\boldsymbol{x}\right) = I(c, \boldsymbol{s}) = \sum_{i=1}^{M} I(c, s_i) = \sum_{i=1}^{L} I(c, s_i) = \sum_{i=1}^{L} I\left(c, \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}\right). \tag{4.24}$$

with $\boldsymbol{w}_i$ some column of W. Here, the first equality follows from the fact that mutual information is invariant under invertible transformations [CT06], the second equality follows from (4.22), the third equality follows from the mutual independence of the elements of $\boldsymbol{s}$, the fourth equality from the assumption that only the first L sources provide information on the BCI-user's intention, and the fifth equality again from (4.22). Hence, all information in $\boldsymbol{x}$ on c is contained in the first L ICs, and consequently the L spatial filters that maximize mutual information are simply those L columns of W with the highest mutual information $I(c, \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x})$. This term can be easily evaluated, and thus the optimal spatial filters identified, according to the approximation of mutual information (4.14) derived in Section 4.2.3.

To summarize the results of this section, it was shown that JAD of the EEG covariance matrices conditioned on class labels can be considered equivalent to ICA. By deriving an analytic approximation of the mutual information of class labels and EEG components a procedure was provided to choose the optimal spatial filters in terms of maximization of (an approximation of) mutual information. The need for heuristics in choosing optimal spatial filters obtained by multi-class CSP is thereby eliminated, and a sound theoretical foundation for spatial filtering in the context of BCIs using multi-class paradigms is provided. Finally, multi-class ITFE, as derived here, allows incorporating unequal class probabilities by choosing those spatial filters that maximize mutual information in (4.14). For convenience, the complete procedure of multi-class ITFE is summarized in Fig. 4.2.

4.3 Experimental Results

Experimental results from a four-class motor imagery paradigm supporting the results of the previous section are now presented. The purpose of this section is to compare pre-processing by multi-class ITFE with multi-class CSP, i.e., comparing the effect of choosing spatial filters that maximize mutual information versus choosing spatial filters according to the heuristic presented in Section 4.2.2.

The data used was recorded in the Laboratory of Brain-Computer Interfaces at the Technische Universität Graz for the third BCI Competition (data set IIIa), and is available at *http://ida.first.fraunhofer.de/ projects/bci/competition_iii/*. A detailed description of the recording procedure can be found in [BMK$^+$06]. Three subjects (k3b, k6b, and l1b) were asked to perform motor imagery of the left/right hand, one foot, or tongue. Each trial lasted for seven seconds, with the motor imagery performed during the last four seconds of each trial. During the experiment EEG was recorded at 60 channels, using the left mastoid as reference and the right mastoid as

Input: Covariance matrices $R_{\boldsymbol{x}|c_i},\ i = 1, \ldots, N$

1. Perform joint approximate diagonalization s.t. $W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W = D_{c_i},\ i = 1, \ldots, N$ (e.g., with the FFDiag-algorithm [ZLNM04]).

2. For each column $\boldsymbol{w}_j,\ j = 1, \ldots, M$, of W scale $\boldsymbol{w}_j$ s.t. $\boldsymbol{w}_j^{\mathrm{T}} R_{\boldsymbol{x}} \boldsymbol{w}_j = 1$ and estimate mutual information according to

$$\begin{aligned} I(c, \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}) \;\approx\; & -\sum_{i=1}^{N} P(c_i) \log \sqrt{\boldsymbol{w}_j^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}_j} \\ & -\frac{3}{16} \left(\sum_{i=1}^{N} P(c_i) \left((\boldsymbol{w}_j^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}_j)^2 - 1 \right) \right)^2 . \end{aligned}$$

3. Choose the L columns of W with highest mutual information.

Output: Spatial filtering matrix $W \in \mathbb{R}^{M \times L}$

Figure 4.2: Multi-class Information Theoretic Feature Extraction

ground. The sampling rate was 250 Hz, and the data was filtered between 1 and 50 Hz with a notchfilter on. For subjects k6b and l1b a total of 60 trials per condition were recorded, and for subject k3b 90 trials per condition were recorded. Four trials of subject k6b had to be discarded due to missing data. Otherwise no trials were rejected and no artifact correction was performed.

For each subject, the following evaluation procedure was performed. First, all data was filtered with a fifth-order butterworth filter with cut-off frequencies 5 and 35 Hz. Then, the four seconds of each trial in which motor imagery was performed were extracted. Afterwards, the data was randomly partitioned into a training and a test set. The size of the training set was varied between 10 and 50 trials in steps of ten trials for subjects k6b and l1b, and between 10 and 80 trials for subject k3b. The covariance matrices of all four conditions were computed using only data of the training set. JAD was performed on the obtained covariance matrices using the algorithm presented in [ZLNM04], and the L optimal spatial filters were chosen according to a) the heuristic presented in Section 4.2.2 (multi-class CSP), b) the procedure described in Fig. 4.2 (multi-class ITFE), and c) multi-class ITFE with evaluation of the mutual information of class labels and extracted EEG components by numerical integration as described in Section 4.2.3. Note that while procedure c) is feasible due to the knowledge of $p(\hat{x})$ in (4.10), it is undesirable from a practical point of view due to increased computational complexity. For multi-class ITFE, equal class probabilities were assumed. Note that the choice of L is a problem of model identification that is beyond the scope of this work. Here, $L = 8$ was arbi-

trarily chosen. The spatial filters obtained by procedures a) - c) were then applied to the training- and test data sets. This resulted in eight-dimensional signals for each trial of the test and training data set. Features were then computed by extracting 15 frequency bands of 2 Hz width ranging from 5 to 35 Hz using a fifth-order butterworth filter, and computing the sample variance in each frequency band for each of the extracted EEG/MEG components. This resulted in a 120-dimensional feature vector for each trial. The feature vectors of the training set were then used to train four logistic regression classifiers with L1-regularization, since this classifier is known to perform well in the presence of many irrelevant features [Ng04]. Each classifier was trained on one versus all other conditions, with a regularization parameter chosen manually as 0.1. To infer the class label of trials in the test data set the continuous output of each classifier was computed for all trials. The output of each logistic regression classifier ranges from zero to one, representing the probability of a certain class. Then, the class label attached to each trial was chosen as the index of the classifier with maximum output for that trial. For each partitioning of the data in a test- and training set this procedure was repeated 20 times.

The resulting classification accuracies for all subjects and evaluation procedures a) and b) are shown in Tab. 4.1 and Fig. 4.3, with the thin horizontal line indicating chance level. Results of evaluation procedure c) are not shown, since on average these differed from procedure b) by only 0.4%. This experimentally validates the accuracy of the derived approximation of mutual information (4.14) in the context of non-invasive BCIs. While the classification accuracies vary significantly across subjects, it is evident that multi-class ITFE outperforms multi-class CSP by far, with a mean increase in classification accuracy of 23.4%. This increase is especially significant for subject l1b, for which multi-class CSP performs only slightly above chance. With spatial filters chosen according to multi-class ITFE, subject k3b even achieves classification accuracies of about 95%.

It should be pointed out that the classification accuracies achieved here do not, with the exception of subject k3b, compare favorably with the best entries to the BCI competition III for the same data set [Sch05]. This is attributed to the fact that while the algorithms submitted to the third BCI competition were extensively tuned, there are several parameters in the procedure presented here that were determined arbitrarily. For example, it is well known that computing spatial filters in narrow frequency bands, tuned according to the most reactive frequency bands for each subject, significantly improves classification accuracy as opposed to selecting a rather broad frequency band as done here. Furthermore, the number of spatial filters retained was chosen arbitrarily as eight for all subjects and training sets, and the regularization parameter of the classification procedure was also determined manually and constant for all subjects. All of these parameters could be tuned using methods such as cross-validation on the training set to achieve higher classification accuracies. This, however, is not the point of this section. A rather simple classification procedure was chosen to emphasize the importance of choosing the optimal spatial filters: while the total set of spatial filters is identical for

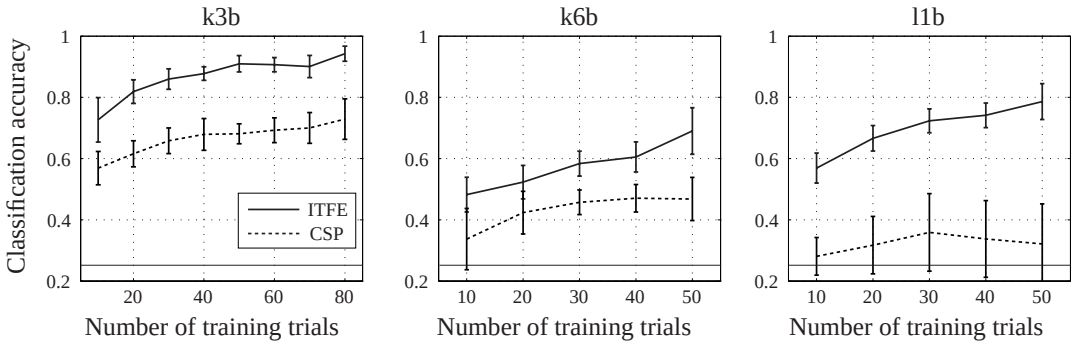

Figure 4.3: Classification accuracies of subjects k3b, k6b, and l1b as a function of the number of training trials for multi-class ITFE and multi-class CSP. The thin horizontal line indicates chance level.

Training trials per cond.	10	20	30	40	50	60	70	80
Subject k3b (ITFE)	72.7	81.9	86.0	87.7	91.0	90.7	90.1	94.2
Subject k3b (CSP)	56.9	61.6	65.8	67.9	68.1	69.3	70.0	72.9
Subject k6b (ITFE)	48.2	52.3	58.4	60.5	69.0	-	-	-
Subject k6b (CSP)	33.7	42.3	45.7	47.0	46.8	-	-	-
Subject l1b (ITFE)	56.9	66.6	72.3	74.1	78.6	-	-	-
Subject l1b (CSP)	28.1	31.7	35.9	33.8	32.1	-	-	-

Table 4.1: Mean classification results in percent for multi-class ITFE and multi-class CSP.

multi-class CSP and multi-class ITFE, choosing a subset of filters that maximize mutual information, according to the procedure of multi-class ITFE summarized in Fig. 4.2, as opposed to the procedure proposed in [DBCM04], leads to a significant increase in classification accuracy.

4.4 Discussion

In this chapter, the knowledge that variance changes in the electric field of the brain provide information on the BCI-user's intention was used to design linear spatial filters that extract those components of the EEG that maximize (an approximation) of mutual information of class labels and extracted features. Using mutual information to design spatial filters was shown to be beneficial for two reaons. First, mutual information provides a direct link to the optimal Bayes error as discussed in Section 2.2. Second, an analytic approximation of mutual information could be derived that a) allowed investigating optimality of the popular CSP algorithm, and b) is easy to evaluate from a computational point of view. It could be shown that the popular two-class CSP algorithm [RMGP00] is optimal in terms of maximiz-

ing (an approximation) of mutual information of class labels and extracted features. An extension of CSP to multiple classes as proposed in [DBCM04], on the other hand, was shown not to be optimal. This deficiency of multi-class CSP could be resolved by showing how optimal spatial filters can be selected in the framework of multi-class ITFE. The theoretical results were supported by experimental evidence from a four-class paradigm, demonstrating a significant increase in classification accuracy for multi-class ITFE in comparison to multi-class CSP. In summary, the results presented in this chapter underline the importance of utilizing any available information on how cognitive states are encoded in the electric field of the brain.

While the results in this chapter demonstrate that learning spatial filters from the available data enables high classification accuracies in multi-class paradigms, the CSP and ITFE algorithms both suffer from a tendency of overfitting. Inspecting the topographies of spatial filters learned by CSP/ITFE reveals that some of the extracted filters focus on artifactual components of the data such as eye blinks or muscle activity. This can be attributed to the fact that muscle artifacts cause electric fields which variances usually exceed the variance of electric fields generated by the CNS. Since both CSP and ITFE are based on variance measures only, they are especially prone to focussing on artifactual components only present in the EEG data of one condition. The effects of overfitting can be alleviated by increasing L, i.e., by selecting more spatial filters for feature extraction. However, increasing L decreases the rate of convergence of the subsequent classifier to its minimum classification error (cf. Section 2.4). It would thus be desirable to develop an algorithm that extracts those components of the EEG that provide most information on the user's intention in an unsupervised manner, i.e., without using labeled training data, since such an algorithm would be robust against artifactual components in the recorded data. This is the topic of Chapter 6.

Chapter 5

Complete Independent Component Analysis in EEG/MEG Analysis

5.1 Introduction

In Chapters 3 and 4, it has been shown that Independent Component Analysis (ICA) (introduced in Section 3.2.1) can be used to construct powerful algorithms for feature extraction in non-invasive BCIs. While in Chapter 3 ICA was utilized as a preprocessing step in order to simplify a subsequent source localization procedure and exclude irrelevant noise sources, it was shown in Chapter 4 that the multi-class Common Spatial Patterns (CSP) algorithm proposed in [DBCM04] is also based on an implementation of ICA. By deriving an approximation of mutual information applicable in the context of non-invasive BCIs, it was further shown how ICA can be used to compute spatial filters that are, under some assumptions, optimal in terms of maximizing an approximation of mutual information of class labels and extracted features. This algorithm, termed multi-class Information Theoretic Feature Extraction (ITFE), was shown to enable classification accuracies above 90% in a four-class motor imagery paradigm.

ICA has been considered for feature extraction in other studies on non-invasive BCIs as well. In [BNL+07], different ICA algorithms are compared with each other and with the CSP algorithm in terms of the quality of the obtained spatial filters in a four-class motor imagery paradigm based on EEG. In [HLS+06], ICA is compared with CSP in a two-class motor imagery paradigm using EEG, MEG and ECoG recordings. While these two studies differ in how the independent components (ICs) that provide most information on the user's intention are identified, both conclude that ICA outperforms CSP.

The success of ICA as a tool for feature extraction in non-invasive BCIs seems surprising in light of the restrictive assumptions incorporated in this framework. While different algorithms require divers assumptions, in this chapter only the assumptions required by the extended Infomax algorithm is considered [LGS99]. This is done for three reasons. First, the extended Infomax algorithm is based on mutual informa-

tion and is thus closely related to a wider class of ICA algorithms [HKO01, Car97]. Second, extended Infomax is shown to outperform other ICA algorithms in the context of computing spatial filters for non-invasive BCIs in [BNL+07], and third, it is one of the most frequently employed ICA algorithms in the analysis of EEG/MEG data due to its implementation in the open-source toolbox EEGLab [DM04]. The assumptions incorporated in this context are

1. Linearity of the mixture model.

2. Mutual statistical independence of the original sources.

3. At most one Gaussian distributed source.

4. At least as many sensors as sources.

The first assumption is justified in the context of non-invasive BCIs based on EEG or MEG (cf. Section 3.2.1 and [NS05]). The second assumption might appear questionable. However, this hinges on what is considered to constitute a source in the context of ICA applied to EEG/MEG data. If an EEG/MEG source is not identified with a certain region within the brain but rather seen as a spatial current distribution with identical dynamics (as defined in Section 3.2.1), then the second assumption does not necessarily constitute a restriction on the applicability of ICA. Instead, it should be understood as a restriction on the interpretability of an IC. While the goal of ICA is to reconstruct statistically independent sources, the physiological relevance of an IC can not be determined a-priori but has to be inferred from the topography and dynamics of each reconstructed source. Notwithstanding this argument, EEG/MEG sources reconstructed by ICA can indeed often be identified with a single brain region (cf. Section 3.2.1). This, however, should not be seen as an empirical justification for identifying EEG sources reconstructed by ICA with certain brain regions, but rather as a special (although frequent) case. Furthermore, note that in the above discussion it is assumed that ICA is capable of reconstructing a set of independent sources for a given data set. In practice, this should be checked by running ICA repeatedly on the same data set with randomized initial conditions. Only if ICA is capable of reconstructing a set of identical ICs independently of the initial conditions of the algorithm the reconstructed ICs should be considered meaningful [MZKM02]. The third assumption has already been discussed in Section 3.2.3. In brief, this assumption is only necessary if all original sources are to be reconstructed. If only non-Gaussian sources are of interest assumption three can be neglected. The fourth assumption, however, is highly questionable. In EEG/MEG, the continuous current distribution in the brain that gives rise to the electric/magnetic field on the scalp is mapped by a linear transformation onto a finite number of EEG/MEG electrodes. As such, the measurements of the electric/magnetic potential on the scalp constitute an underdetermined representation of the spatial current distribution within the brain. Accordingly, the fourth assumption

is justified only if the continuous current distribution within the brain can be partitioned into M distinct sets with identical dynamics, where M refers to the number of EEG/MEG electrodes. This can be considered as highly unlikely, and hence the fourth assumptions has to be rejected in the context of EEG/MEG analysis. Instead, it is maintained here that this assumption has been adopted as a working assumption for a lack of better alternatives.

Since the introduction of ICA with complete bases in [Com94], several algorithms have been developed that address the problem of ICA with overcomplete bases, i.e., with more sources than sensors. In [LLGS99] and [ZP01], sparsity constraints are imposed to obtain a unique reconstruction of sources, while in [TLP04] a geometric approach to overcomplete ICA is proposed. Interestingly, algorithms for overcomplete ICA have found virtually no application in the EEG/MEG community so far. This can be attributed to the success of applying complete ICA to EEG/MEG data. Since the obtained results are in accord with physiological expectations, it is generally assumed that complete ICA is sufficient for the analysis of EEG/MEG signals [OWTM06].

This view is challenged in this chapter. Instead of accepting the adequacy of complete ICA in the context of EEG/MEG analysis, the behavior of ICA designed for complete bases is investigated if the assumption of completeness is violated, i.e., if more sources than sensors are assumed. This serves several purposes. The first is to establish whether, or under which conditions, ICA designed for complete bases can be applied to underdetermined problems, and to investigate how this affects the reconstruction of the original sources. This is a prerequisite for the second purpose, which is to address the question which type of mixture model can realistically be assumed in EEG/MEG analysis. Finally, it is investigated how adverse effects on reconstructed ICs resulting from underdetermined problems can be alleviated without resorting to additional constraints on the reconstructed sources as in [LLGS99] and [ZP01].

To address these issues, a linear mixture model with an arbitrary number of non-Gaussian and Gaussian sources is assumed. Then, necessary and sufficient conditions for solutions of complete ICA for this mixture model are derived. While identifiability and separability of ICA have already been considered in [Com94, EK03], and [CL96], to the best of the author's knowledge the theorem presented here is the first one providing necessary and sufficient conditions for solutions of complete ICA for a mixture model with an arbitrary number of non-Gaussian and Gaussian sources. This theorem is then used to investigate source separability and model identifiability for different relations of sensors, non-Gaussian-, and Gaussian sources. The conclusions of this investigation are used to argue that in EEG/MEG analysis it is valid to assume that less non-Gaussian sources than sensors but more Gaussian sources than sensors are present. This reconciles the apparent contradiction of the underdetermined nature of EEG/MEG recordings and the physiological plausibility of results obtained by complete ICA. It is shown that this mixture model further implies that while the topographies of non-Gaussian sources are recon-

structed correctly by complete ICA in spite of an overcomplete mixture model, the reconstructed dynamics of non-Gaussian sources are arbitrarily mixed with Gaussian sources. The consequences of this result for the analysis of EEG/MEG data are discussed, and testable predictions for further validation of the assumed mixture model in the context of EEG/MEG analysis are formulated.

It is then shown how the adverse effect of applying complete ICA to overcomplete mixture models can be alleviated by linearly constrained minimum variance (LCMV) spatial filtering [VvYS97]. Since the topographies of non-Gaussian sources are correctly reconstructed by complete ICA in spite of the presence of an arbitrary number of Gaussian sources, this knowledge can be used to construct spatial filters that minimize the interference of Gaussian sources in the reconstruction of the dynamics of non-Gaussian sources.

The combined procedure of complete ICA and LCMV spatial filtering is then applied to experimental MEG and EEG data. First, it is shown that in the reconstruction of auditory event related fields (ERFs) recorded by MEG combining ICA with LCMV spatial filtering significantly outperforms ordinary ICA. Then, combined ICA and LCMV spatial filtering is employed to construct spatial filters for a four-class motor imagery BCI based on EEG. Interestingly, it is shown that ICA and LCMV spatial filtering does not outperform ordinary ICA. The reasons for this observation are discussed in light of the theoretical results of this chapter, thereby further validating the proposed overcomplete mixture model in the context of EEG/MEG analysis. Also, an explanation is provided for the success of complete ICA in constructing feature extraction algorithms for non-invasive BCIs.

The remainder of this chapter is structured as follows. As in previous chapters, Section 5.2 begins with introducing the notation and stating the assumptions made in this chapter on the class of allowed feature spaces $\mathcal{P}^*$. Then, the ICA mixture model used throughout this chapter is introduced. In Section 5.2.2, necessary and sufficient conditions for solutions of ICA applied to overcomplete mixture models are derived. This constitutes the primary theoretical contribution of this chapter. It also serves as the basis for Section 5.2.3, in which the validity of different source models for EEG/MEG are discussed. Section 5.2 concludes with a discussion on how the performance of complete ICA can be improved by LCMV spatial filtering in Section 5.2.4. In Section 5.3, experimental results from the reconstruction of auditory ERFs recorded by MEG as well as from a four-class motor imagery BCI based on EEG are presented. The chapter concludes with a discussion of the theoretical and experimental results and their relevance for EEG/MEG analysis by complete ICA in Section 5.4. Parts of the work in this chapter already have been presented in [GWB07].

5.2 Methods

The same notation as in previous chapters is also employed in this chapter. The BCI-user's intention is again denoted as $c \in \mathcal{C} = \{c_1, \ldots, c_N\}$, and the EEG/MEG data recorded at M electrodes as $\boldsymbol{x}(t) \in \mathbb{R}^M$. If the time index is dropped $\boldsymbol{x}$ refers to a M-dimensional random variable with probability density function $p(\boldsymbol{x})$. Again, $X \in \mathbb{R}^{M \times T}$ refers to a block of data with T samples. The assumptions on the class of allowed feature spaces $\mathcal{P}^*$ employed in this chapter are:

1. The BCI-user's intention is encoded in variance changes of the electric/magnetic field of the brain originating in spatially invariant brain regions.

2. The electric/magnetic field of the brain can be decomposed into statistically independent components.

Note that the second assumption does not make any statement on the number of statistically independent sources within the brain. It just asserts that the spatial current distribution within the brain can be decomposed into statistically independent sources (cf. Definition 3.1 for what is considered to constitute a source in this context). The first assumption again expresses our knowledge on how cognitive states are temporally encoded in the electric/magnetic field of the brain (cf. Chapter 4). It also provides a justification for using linear time-invariant spatial filters for extracting those components of the electric/magnetic field of the brain that provide most information on the user's intention.

5.2.1 The ICA Model

The source model assumed in this chapter is similar to the one in Section 3.2.1. The EEG/MEG data is assumed to obey the generative model

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

with $\boldsymbol{s}(t) \in \mathbb{R}^K$ the original EEG/MEG sources with probability density function $p(\boldsymbol{s})$ and the matrix $A \in \mathbb{R}^{M \times K}$ the full row-rank mixing matrix, describing the projection strength of each source to each of the M electrodes. Without loss of generality, it is assumed that each source has got zero mean and unit variance. Furthermore, it is assumed that the first L sources are non-Gaussian distributed, while the last $K - L$ sources follow a Gaussian distribution, i.e., that $p(s_i) = \mathcal{N}(0, 1)$ for $i = L+1, \ldots, K$. The special case of only one Gaussian source, i.e., $L+1 = K$, is disregarded. Finally, it is assumed that $p(\boldsymbol{s}) = \prod_{i=1}^{K} p(s_i)$, i.e., that the sources are mutually statistically independent. Note that at this point no assumption is made on the relation of M, L, and K, i.e., complete as well as overcomplete models are considered.

5.2.2 Identifiability and Separability of Complete ICA for Arbitrary Mixture Models

The goal in ICA is to reconstruct the original sources $\boldsymbol{s}$ and the mixing matrix A only from observations of $\boldsymbol{x}$ and using the assumption of mutual statistical independence of the original sources. One way to approach this problem in complete ICA is to construct a full-rank unmixing matrix $W \in \mathbb{R}^{M\times M}$ that solves the optimization problem (cf. Section 3.2.1)

$$W = \underset{W\in\mathbb{R}^{M\times M}}{\operatorname{argmin}} \{I(y_1, \ldots, y_M)\} \tag{5.2}$$

with $\boldsymbol{y} = W\boldsymbol{x}$ and $I(y_1, \ldots, y_M)$ the mutual information of the elements of $\boldsymbol{y}$ (cf. [CT06]). The elements of $\boldsymbol{y}$ are called the independent components (ICs). This approach is due to Theorem 3.1, establishing that the mutual information of the elements of $\boldsymbol{y}$ is zero if and only if they are mutually statistically independent. Other (largely equivalent) approaches to ICA are discussed in [HKO01]. While it seems sensible to assume that reconstructing mutually statistically independent sources results in the original independent sources, it remains to be investigated which form W may take such that the elements of $\boldsymbol{y}$ are mutually statistically independent, and whether this does indeed result in the elements of $\boldsymbol{y}$ corresponding to the original sources in $\boldsymbol{s}$. This is referred to as the problem of source separability. Furthermore, it has to be investigated whether taking the inverse of the unmixing matrix W reconstructs the mixing matrix A, i.e., if $\hat{A} = W^{-1} = A$. This is referred to as the problem of model identifiability.
Toward these goals, first note that $\boldsymbol{x}$ can always be sphered, i.e., subjected to a transformation $P = R_{\boldsymbol{x}}^{-\frac{1}{2}}$ such that for the covariance matrix $R_{\hat{\boldsymbol{x}}}$ of $\hat{x} = P\boldsymbol{x}$ it holds that

$$R_{\hat{\boldsymbol{x}}} = \langle P\boldsymbol{x}, P\boldsymbol{x}\rangle = PR_{\boldsymbol{x}}P^{\mathrm{T}} = I_{M\times M}. \tag{5.3}$$

The sphering transformation is subsequently neglected, assuming that $\boldsymbol{x}$ has already been sphered. Note that this implies that

$$R_{\boldsymbol{x}} = \langle \boldsymbol{x}, \boldsymbol{x}\rangle = AR_{\boldsymbol{s}}A^{\mathrm{T}} = AA^{\mathrm{T}} = I_{M\times M}. \tag{5.4}$$

Without loss of generality, the rows of A can hence be considered mutually orthogonal. Then define $C := WA$, such that

$$\boldsymbol{y} = W\boldsymbol{x} = WA\boldsymbol{s} = C\boldsymbol{s}. \tag{5.5}$$

Now note that any solution of (5.2) requires the elements of $\boldsymbol{y}$ to be mutually statistically independent. This implies uncorrelatedness, and hence

$$R_{\boldsymbol{y}} = \langle \boldsymbol{y}, \boldsymbol{y}\rangle = CR_{\boldsymbol{s}}C^{\mathrm{T}} = I_{M\times M}. \tag{5.6}$$

Hence the rows of C have to be mutually orthogonal. Since this also holds for A and furthermore $WA = C$, the class of matrices that have to be considered for solutions of (5.2) can be constrained to orthogonal matrices.

The following theorem extends the results in [Com94] and [EK03] to mixture models with an arbitrary number of Gaussian and non-Gaussian sources. Note that a similar theorem is given in [CL96], but without proving sufficiency and only for pairwise independence of the elements of $\boldsymbol{y}$.

Theorem 5.1 (Separability of complete ICA for arbitrary mixture models)**.** *Let $\boldsymbol{s} \in \mathbb{R}^K$ and $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{y} = WA\boldsymbol{s} = C\boldsymbol{s}$ with full rank $W \in \mathbb{R}^{M\times M}$ and full rank $A \in \mathbb{R}^{M\times K}$, and the elements of $\boldsymbol{s}$ as well as the elements of $\boldsymbol{y}$ mutually statistically independent. Furthermore, let $p(s_i) = \mathcal{N}(0,1)$ only for $i = L+1,\ldots,K$. Then it holds that*

$$C = [B \mid Q], \tag{5.7}$$

with $B \in \mathbb{R}^{M\times L}$ a matrix with only one non-zero entry in each column, at most $M-L$ zero rows, and at most $M-1$ zero entries in each row. Furthermore, $Q \in \mathbb{R}^{M\times(K-L)}$ is a matrix with orthogonal rows.

Proof. Proving Theorem 5.1 amounts to showing that for C of the form in (5.7) the elements of $\boldsymbol{y}$ are mutually statistically independent (sufficiency), and that any deviation of C from this form leads to a contradiction to the elements of $\boldsymbol{y}$ being mutually statistically independent (necessity). Sufficiency is proved first.
To prove sufficiency, it is necessary and sufficient to show that $I(y_1,\ldots,y_M) = 0$ for C of the form in (5.7), since this is a necessary and sufficient condition for mutual statistical independence of the elements of $\boldsymbol{y}$ due to Theorem 3.1. Then note that $I(y_1,\ldots,y_M) = D(p(\boldsymbol{y})||\prod_{i=1}^M p(y_i))$ with $D(.||.)$ the Kullback-Leibler divergence. Since $I(y_1,\ldots,y_M) \geq 0$, with equality if and only if the elements of $\boldsymbol{y}$ are mutually statistically independent, and the Kullback-Leibler divergence is convex as well as continuously differentiable (cf. [CT06]), $I(y_1,\ldots,y_M)$ has a unique global minimum at zero, and consequently $I(y_1,\ldots,y_M) = 0 \Leftrightarrow \frac{\partial}{\partial W} I(y_1,\ldots,y_M) = 0$. Then note that

$$\begin{aligned}
\frac{\partial}{\partial W} I(y_1,\ldots,y_M) &= \frac{\partial}{\partial W}\left\{\sum_{i=1}^M H(y_i) - H(\boldsymbol{y})\right\} \\
&= \frac{\partial}{\partial W}\left\{\sum_{i=1}^M H(y_i) - \log(|W|) - H(\boldsymbol{x})\right\} \\
&= \frac{\partial}{\partial W}\sum_{i=1}^M H(y_i), \qquad (5.8)
\end{aligned}$$

since W orthogonal and $H(\boldsymbol{x})$ independent of W. The following derivations extend the results of [BPR02] to the overcomplete case. Define $F(W) := \sum_{i=1}^M H(y_i)$. The gradient of $F(W)$ under the orthogonality constraint on W is given by [EAS98]

$$\nabla_{\text{ortho}} F(W) = \nabla F(W) - W\nabla F(W)^T W. \tag{5.9}$$

Since $WW^T = I_{M\times M}$, solutions to $\frac{\partial}{\partial W} I(y_1, \ldots, y_M) = \nabla_{\text{ortho}} F(W) = 0$ are given by

$$\nabla F(W) W^T = W \nabla F(W)^T. \tag{5.10}$$

Denoting $h(\boldsymbol{w}_i) := H(y_i)$ with $\boldsymbol{w}_i$ the ith row of W, (5.10) becomes

$$\nabla h(\boldsymbol{w}_k) \cdot \boldsymbol{w}_l^T = \nabla h(\boldsymbol{w}_l) \cdot \boldsymbol{w}_k^T \tag{5.11}$$

for $k, l = 1, \ldots M;\ k \neq l$. With

$$\frac{\partial h(\boldsymbol{w}_i)}{\partial w_{i,j}} = -\int_{-\infty}^{\infty} (\log p_{y_i}(u) + 1) \frac{\partial p_{y_i}(u)}{\partial w_{i,j}} \mathrm{d}u, \tag{5.12}$$

(5.11) results in

$$\begin{aligned} &\int_{-\infty}^{\infty} (\log p_{y_k}(u) + 1) \left[\tfrac{\partial p_{y_k}(u)}{\partial w_{k,1}} w_{l,1} + \ldots + \tfrac{\partial p_{y_k}(u)}{\partial w_{k,M}} w_{l,M} \right] \mathrm{d}u \\ &\quad = \int_{-\infty}^{\infty} (\log p_{y_l}(u) + 1) \left[\tfrac{\partial p_{y_l}(u)}{\partial w_{l,1}} w_{k,1} + \ldots + \tfrac{\partial p_{y_l}(u)}{\partial w_{l,M}} w_{k,M} \right] \mathrm{d}u \end{aligned} \tag{5.13}$$

for $k, l = 1, \ldots, M;\ k \neq l$. A sufficient condition for (5.13) to hold is

$$\frac{\partial p_{y_k}(u)}{\partial w_{k,1}} w_{l,1} + \ldots + \frac{\partial p_{y_k}(u)}{\partial w_{k,M}} w_{l,M} = 0 \tag{5.14}$$

for $k, l = 1, \ldots, M;\ k \neq l$. Recalling (5.5), the elements of $\boldsymbol{y}$ can be written as

$$y_i = c_{i,1} s_1 + \ldots + c_{i,K} s_K, \tag{5.15}$$

with $c_{i,j}$ denoting the element of C in the i^{th} row and the j^{th} column. The probability density function of y_i is then given by

$$p_{y_i}(u) = \frac{1}{c_{i,1}} p_{s_1}\left(\frac{u}{c_{i,1}}\right) * \ldots * \frac{1}{c_{i,K}} p_{s_K}\left(\frac{u}{c_{i,K}}\right). \tag{5.16}$$

The analysis of (5.14) is simplified in the frequency domain. With $\varphi_{y_i}(\omega)$ the characteristic function of p_{y_i}, (5.16) becomes

$$\varphi_{y_i}(\omega) = \prod_{j=1}^{K} \varphi_{s_j}(c_{i,j}\omega). \tag{5.17}$$

Transforming (5.14) into the frequency domain as well, substituting (5.17), and dividing by $\prod_{j=1}^{K} \varphi_{s_j}(c_{i,j}\omega)$ results (after some tedious algebraic manipulations) in

$$\frac{\omega \varphi'_{s_1}(c_{k,1}\omega) c_{l,1}}{\varphi_{s_1}(c_{k,1}\omega)} + \ldots + \frac{\omega \varphi'_{s_K}(c_{k,K}\omega) c_{l,K}}{\varphi_{s_K}(c_{k,K}\omega)} = 0 \tag{5.18}$$

for $k, l = 1, \ldots, M;\ k \neq l$. Now only if s_i has a Gaussian distribution it holds that

$$\varphi'_{s_i}(\alpha\omega) = -\alpha\omega\varphi_{s_i}(\alpha\omega). \tag{5.19}$$

Since the sources $s_i,\ i = L+1, \ldots, K$ are assumed to be Gaussian, (5.18) simplifies to

$$\begin{aligned} &\tfrac{\omega\varphi'_{s_1}(c_{k,1}\omega)c_{l,1}}{\varphi_{s_1}(c_{k,1}\omega)} + \ldots + \tfrac{\omega\varphi'_{s_L}(c_{k,L}\omega)c_{l,L}}{\varphi_{s_M}(c_{k,L}\omega)} \\ &\quad -\omega^2(c_{k,L+1}c_{l,L+1} + \ldots + c_{k,K}c_{l,K}) = 0 \end{aligned} \tag{5.20}$$

for $k, l = 1, \ldots, M;\ k \neq l$. Note that $\varphi'_{s_i}(c_{i,j}\omega)|_{c_{i,j}=0} = 0$ because all sources have zero mean. Hence, the first term in (5.20) is zero if in the first column of C for every pair of elements only one of them is non-zero. This in turn implies that only one element of the first column of C may be non-zero. The same holds for every up to and including the L^{th} column. Considering the last term of (5.20), this term is zero if the rows of C, starting with the $(L+1)^{\text{th}}$ element, are mutually orthogonal. It hence follows that C of the form in (5.7) implies that (5.20) holds and consequently $I(y_1, \ldots, y_M) = 0$. Since $I(y_1, \ldots, y_M) = 0$ implies mutual statistical independence of the elements of $\boldsymbol{y}$ this completes the proof of sufficiency. The proof of necessity consists of three steps. The first step follows directly from Theorem 3.2 (Darmois-Skitovic). Assume that column $\boldsymbol{b}_k$, $k \in \{1, \ldots, L\}$, of B in (5.7) has got more than one non-zero entry, and further assume the elements of $\boldsymbol{y}$ to be mutually statistically independent. Then the original source s_k is Gaussian distributed by Theorem 3.2. This is a contradiction to the assumptions, which proves that each column of B may have at most one non-zero entry. Second, note that

$$\boldsymbol{w}_i^{\mathrm{T}}\,[\boldsymbol{a}_1,\ \ldots, \boldsymbol{a}_L] = \boldsymbol{b}_i^{\mathrm{T}} = \boldsymbol{0}^{\mathrm{T}}, \tag{5.21}$$

with $\boldsymbol{w}_i^{\mathrm{T}}$ the i^{th} row of W, $\boldsymbol{a}_j$ the j^{th} column of A, and $\boldsymbol{b}_i^{\mathrm{T}}$ the i^{th} row of B, implies that $\boldsymbol{w}_i \in \mathbb{R}^M$ lies in a $(M-L)$-dimensional subspace of $\mathbb{R}^M$. Since W is assumed to have full rank, (5.21) can hold for at most $M - L$ rows of W. This is turn proves that B can have at most $M - L$ zero rows. Furthermore, note that (5.21) can only hold if $M > L$. Conversely, this shows that each row of B may have at most $M - 1$ zero elements. Finally, the requirement of orthogonality of Q follows from orthogonality of C. This completes the proof of necessity. □

It should be noted that in Theorem 5.1 no assumptions are made on the relation of M, K, and L. The theorem thus applies to mixture models with an arbitrary number of non-Gaussian and Gaussian sources. The implications of Theorem 5.1 for different relations of M, K, and L are now discussed. The case of less sources than sensors is neglected, since this case can be reduced to the problem of an equal number of sources and sensors by disregarding some sensors.

Equal Number of Sources and Sensors ($M = K > L$)

If $M = K$, it follows from Theorem 5.1 that

$$\boldsymbol{y} = C\boldsymbol{s} = P_{M\times M}\begin{bmatrix} \Lambda_{L\times L} & 0_{M-L\times L} \\ 0_{M-L\times L} & Q_{M-L\times M-L} \end{bmatrix}\boldsymbol{s}, \tag{5.22}$$

with $P \in \mathbb{R}^{M\times M}$ a permutation matrix, $\Lambda \in \mathbb{R}^{L\times L}$ a diagonal matrix, and $Q \in \mathbb{R}^{M-L\times M-L}$ an orthogonal matrix. The permutation matrix P is subsequently disregarded, since it does not have any qualitative influence on the source reconstruction. The lower left block of zeros follows from the fact that B in (5.7) may have only one non-zero entry in each column. Furthermore, since Q forms a $(M-L)$-dimensional complete orthogonal basis and C is orthogonal, the upper right hand block of C has to consist of zeros. Finally, rows with only zero elements (that could result in non-Gaussian sources mixed with each other or with Gaussian sources) are not allowed due to full rank of A and W. Consequently, if $K = M$, all sources with non-Gaussian sources are correctly reconstructed by complete ICA, while all Gaussian sources are arbitrarily mixed together. The non-Gaussian sources are hence separable, while the Gaussian sources are non-separable.
In terms of identifiability of the mixing matrix A, note that

$$\hat{A} = W^{-1} = AC^{-1} = A \begin{bmatrix} \Lambda_{L\times L}^{-1} & 0_{M-L\times L} \\ 0_{M-L\times L} & Q^{\mathrm{T}}_{M-L\times M-L} \end{bmatrix}. \tag{5.23}$$

The inverse of the unmixing matrix W hence correctly reconstructs the columns of A corresponding to topographies of non-Gaussian sources (up to scaling and permutations), while the columns of A corresponding to Gaussian sources are arbitrarily mixed together. Hence, the topographies of non-Gaussian sources are identifiable by complete ICA, while the topographies of Gaussian sources are non-identifiable. Note that these results are in agreement with the results on complete ICA in Section 3.2.3.

More non-Gaussian Sources than Sensors ($K > L > M$)

If more non-Gaussian sources than sensors are present in the data set, then $B \in \mathbb{R}^{M\times L}$ has got more columns than rows. Then note that Theorem 5.1 states that the matrix B may have at most $M-1$ zero entries in each row. Since there are more columns than rows in B, this is a contradiction to the requirement of each column of B having at most one non-zero entry. Consequently, for $K > L > M$ it is impossible to construct a matrix C that is in agreement with Theorem 5.1. Thus, separation of the original sources by complete ICA is not possible.
Regarding the identifiability of the mixing model for $K > L > M$, it should be noted that in general an overcomplete mixing model is identifiable [CL96]. Consider the following example.

Example 5.1 (Block-independent reconstructions)**.** *If $M = 3, L = 6$ and $K > L$, then one possible source reconstruction with mutually independent elements is given by*

$$\boldsymbol{y} = \left[\begin{array}{cccccc|c} 1 & 1 & 0 & 0 & 0 & 0 & \boldsymbol{q}_1^T \\ 0 & 0 & 1 & 1 & 1 & 0 & \boldsymbol{q}_2^T \\ 0 & 0 & 0 & 0 & 0 & 1 & \boldsymbol{q}_3^T \end{array}\right] \boldsymbol{s}, \tag{5.24}$$

with $\boldsymbol{q}_i \in \mathbb{R}^{K-L}$ mutually orthogonal vectors.

Note that indeed any partitioning of the original sources into distinct sets results in mutually statistically independent reconstructions. However, in Example 5.1 the requirement of at most $M-1$ zero elements in each row of B is violated. Example 5.1 hence does not provide an admissible solution for *complete* ICA applied to overcomplete mixture models. In general, the form of C obtained by applying complete ICA to an overcomplete mixture model with more non-Gaussian sources than sensors depends on the specific algorithm. It is easy to show that since C can not achieve source separation, the columns of the reconstructed mixing matrix $\hat{A} = W^{-1}$ do not correspond to the columns of the original mixing matrix A (using the same argument that is used below for the case of $K > M > L$). For $K > L > M$ the mixing model is thus not identifiable. Note that this is in agreement with previous studies on complete ICA [Com94, EK03].

More Sources than Sensors, but less non-Gaussian Sources than Sensors ($K > M > L$)

If more sources than sensors are present in the data set, but the number of non-Gaussian sources is smaller than the number of sensors, possible source reconstructions in agreement with Theorem 5.1 are given by

$$\boldsymbol{y} = C\boldsymbol{s} = \left[\begin{array}{c|c} \Lambda_{L\times L} P_{L\times L} & Q_{M\times K-L} \\ 0 & \end{array} \right] \boldsymbol{s}, \tag{5.25}$$

with $\Lambda \in \mathbb{R}^{L\times L}$ a diagonal matrix, $P \in \mathbb{R}^{L\times L}$ a permutation matrix, and $Q \in \mathbb{R}^{M\times K-L}$ a matrix with mutually orthogonal rows. This implies that while the Gaussian sources are non-separable, the set of non-Gaussian sources is separable. However, Gaussian sources are arbitrarily mixed into the non-Gaussian sources.
In terms of the identifiability of the mixing matrix A by complete ICA for $K > M > L$, it is shown now that the columns of A corresponding to the non-Gaussian sources can indeed be reconstructed, i.e., that the non-Gaussian part of the mixing model is identifiable by complete ICA. Without loss of generality, it is assumed that the reconstructed sources are given by

$$\boldsymbol{y} = C\boldsymbol{s} = \left[\begin{array}{c|c} I_{L\times L} & Q_{M\times K-L} \\ 0 & \end{array} \right] \boldsymbol{s}. \tag{5.26}$$

This implies that

$$\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{a}_j = \left\{ \begin{array}{ccc} 1 & ; & i = j \ \wedge \ i,j \in \{1,\ldots,L\} \\ 0 & ; & i \neq j \ \wedge \ i \in \{1,\ldots,M\}, j \in \{1,\ldots,L\} \\ \neq 0 & ; & i \in \{1,\ldots,M\}, j \in \{L+1,\ldots,K\} \end{array} \right. , \tag{5.27}$$

with $\boldsymbol{w}_i^{\mathrm{T}}$ the i^{th} row of W and $\boldsymbol{a}_i$ the i^{th} column of A. For the reconstructed columns

$\hat{\boldsymbol{a}}_i$ of $\hat{A}$ it holds that

$$\begin{array}{cccc} \boldsymbol{w}_1^{\mathrm{T}}\hat{\boldsymbol{a}}_1 = 1 & \boldsymbol{w}_1^{\mathrm{T}}\hat{\boldsymbol{a}}_2 = 0 & \cdots & \boldsymbol{w}_1^{\mathrm{T}}\hat{\boldsymbol{a}}_M = 0 \\ \boldsymbol{w}_2^{\mathrm{T}}\hat{\boldsymbol{a}}_1 = 0 & \boldsymbol{w}_2^{\mathrm{T}}\hat{\boldsymbol{a}}_2 = 1 & \cdots & \boldsymbol{w}_2^{\mathrm{T}}\hat{\boldsymbol{a}}_M = 0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_M^{\mathrm{T}}\hat{\boldsymbol{a}}_1 = 0 & \boldsymbol{w}_M^{\mathrm{T}}\hat{\boldsymbol{a}}_2 = 0 & \cdots & \boldsymbol{w}_M^{\mathrm{T}}\hat{\boldsymbol{a}}_M = 1, \end{array} \tag{5.28}$$

since $\hat{A} = W^{-1}$ and hence $W\hat{A} = I_{M\times M}$ by construction. Now consider $\hat{\boldsymbol{a}}_1$, which has to be jointly orthogonal to $\boldsymbol{w}_i^{\mathrm{T}}$ with $i = \{2, \ldots, M\}$. Equation (5.27) implies that $\boldsymbol{w}_i^{\mathrm{T}}\hat{\boldsymbol{a}}_1 = 0$ if and only if

$$\hat{\boldsymbol{a}}_1 = \sum_{j=1, j\neq i}^{L} \alpha_j \boldsymbol{a}_j, \tag{5.29}$$

with $\alpha_i \in \mathbb{R}$. Since it is required in (5.28) that $\boldsymbol{w}_i^{\mathrm{T}}\hat{\boldsymbol{a}}_1 = 0$ for all $i \in \{2, \ldots, L\}$, it follows that $\hat{\boldsymbol{a}}_1 = \alpha_1 \boldsymbol{a}$, and the first column of $\hat{A}$ is a scaled version of the first column of A. The same argument applies for the remaining first L columns of $\hat{A}$. Hence, the columns of A corresponding to the non-Gaussian sources are correctly reconstructed up to scaling and possible permutations.
With regard to the last $K - L$ columns of $\hat{A}$, consider the following example.

Example 5.2 (Non-identifiability of Gaussian mixture models)**.** *Let $M = 4, L = 2$ and $K = 8$. Then one possible source reconstruction in agreement with Theorem 5.1 is given by*

$$\boldsymbol{y} = WA\boldsymbol{s} = C\boldsymbol{s} = \left[\begin{array}{cc|cccccc} 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \end{array}\right] \boldsymbol{s}. \tag{5.30}$$

Now consider the third column $\hat{\boldsymbol{a}}_3$ of the reconstructed mixing matrix $\hat{A} = W^{-1}$, and let $\hat{\boldsymbol{a}}_3 = \alpha\boldsymbol{a}_3 + \beta\boldsymbol{a}_4$ with $\alpha, \beta \in \mathbb{R}$. Then due to the third and fourth column of C in (5.30) it holds that

$$\begin{array}{l} \boldsymbol{w}_1^T\hat{\boldsymbol{a}}_3 = 0 \\ \boldsymbol{w}_2^T\hat{\boldsymbol{a}}_3 = 0 \\ \boldsymbol{w}_3^T\hat{\boldsymbol{a}}_3 = 1 \\ \boldsymbol{w}_4^T\hat{\boldsymbol{a}}_3 = 0 \end{array}, \tag{5.31}$$

with α, β chosen such that $\boldsymbol{w}_3^T(\alpha\boldsymbol{a}_3 + \beta\boldsymbol{a}_4) = 1$. This is in agreement with $WA = I_{M\times M}$, and hence $\hat{\boldsymbol{a}}_3 = \alpha\boldsymbol{a}_3 + \beta\boldsymbol{a}_4$ constitutes an admissible reconstruction of a column of the mixing matrix A.

This example establishes that in general the columns of A corresponding to Gaussian sources are not correctly reconstructed by complete ICA.

Separability of sources	$M = K > L$	$K > L > M$	$K > M > L$
Non-Gaussian	✓	-	✓
Gaussian	-	-	-
Gaussian from non-Gaussian	✓	-	-
Model identifiability	$M = K > L$	$K > L > M$	$K > M > L$
Non-Gaussian	✓	-	✓
Gaussian	-	-	-

Table 5.1: Identifiability and separability of mixture models with K sources, L non-Gaussian sources, and M sensors by complete ICA.

In summary, complete ICA correctly reconstructs the columns of the mixing matrix A corresponding to non-Gaussian sources if $K > M > L$. The columns of A corresponding to the Gaussian sources, however, are not correctly reconstructed. Hence, only the non-Gaussian part of the mixing model is identifiable by complete ICA for the case of $K > M > L$.
For convenience, all results on identifiability and separability of arbitrary mixture models by complete ICA obtained in this Section are summarized in Tab. 5.1.

5.2.3 Validity of Mixture Models in EEG/MEG Analysis

In this section, the plausibility of different source models is discussed in the context of EEG/MEG analysis. As pointed out in the introduction of this chapter, the continuous spatial current distribution within the brain gives rise to the electric potential / magnetic field measurable on the scalp [NS05]. If this potential / field is sampled at M electrodes, this constitutes a mapping from an infinite to finite dimensional space. As such, this mapping can only fully describe the continuous current distribution within the brain if the current distribution can be partitioned into at most M distinct sets with identical dynamics. In general, this can be considered highly unlikely, and hence the assumption of less sources than sensors has to be rejected in EEG/MEG analysis.
This conclusion is usually not accepted by the EEG/MEG community, since it appears in contradiction to the apparent success of complete ICA in the analysis of EEG/MEG recordings. Instead, it is argued that only a few EEG/MEG sources are strong enough to be picked up by ICA, and that hence, at least from a practical point of view, less sources than sensors can be assumed (cf. [OWTM06]). This argument is supported by empirical evidence that source dynamics and topographies constructed by complete ICA are physiologically plausible [JMB+01, MWJ+02, MDOD04].
In contrast, it is maintained here that the available empirical and theoretical evidence suggests that the reason for the success of complete ICA in EEG/MEG analysis is not that only a few sources are strong enough to be picked up by ICA, but rather that only a few sources are *non-Gaussian* enough to be picked up by ICA. This claim

is based on the following argument. First, assuming a mixture model with only a few non-Gaussian sources is in agreement with the physiological plausibility of results obtained by complete ICA, since in this case the non-Gaussian sources are separable and the non-Gaussian part of the mixture model is identifiable (cf. Section 5.2.2). Second, this model is not in contradiction to the underdetermined nature of EEG/MEG recordings, since an arbitrary number of additional Gaussian sources can be assumed. Thirdly, consider the following definition.

Definition 5.1 (Stable Independent Component)**.** *An independent component is called stable if it is contained (up to possible scaling) in all source reconstructions obtained by complete ICA, i.e., if it is independent of the type of algorithm used for complete ICA and the initial conditions of the algorithm.*

If only a few non-Gaussian and multiple Gaussian sources are present in a data set, Theorem 5.1 asserts that that there exists a subset of unstable ICs that correspond to mixtures of Gaussian sources. This is indeed supported by empirical evidence, with unstable ICs usually observed in EEG/MEG analysis [JMB$^+$01].
It is hence maintained that a mixture model with less non-Gaussian sources than sensors but more Gaussian sources than sensors constitutes a realistic assumption in EEG/MEG analysis. Note that due to Theorem 5.1 this mixture model implies that reconstructed dynamics of non-Gaussian sources are arbitrarily mixed with Gaussian sources. This prediction can be used to further validate the proposed mixture model. In terms of the analysis of event related potentials/fields (ERPs/ERFs) by ICA, the inclusion of Gaussian sources results in a lower signal-to-noise ratio (SNR) of the reconstructed ICs. Note, however, that Gaussian sources can be temporally white as well as correlated, i.e., in general no statement can be made on whether the inclusion of Gaussian sources distorts the temporal structure of reconstructed ICs. Regarding the analysis of event related synchronization/desynchronization (ERD/ERS) (cf. [PL99] and [NS05]), note that ERS/ERD measures dynamic changes in variance of reconstructed sources. As such, only non-stationary sources can contribute to the temporal structure of ERS/ERD. Since Gaussianity of a source over the whole temporal range of the recorded EEG/MEG implies stationarity, the inclusion of Gaussian sources in reconstructed ICs amounts to only raising the baseline of ERS/ERD measurements. This has got no adverse effects on the temporal structure or the significance level of ERS/ERD measurements. Consequently, the proposed mixture model predicts adverse effects on reconstructed ERPs/ERFs, and no adverse effects on the analysis of ERS/ERD. In the next section a methodology is presented that allows testing these predictions.

5.2.4 Overcomplete ICA via LCMV Spatial Filtering

If a mixture model with a small number of non-Gaussian and a very large amount of Gaussian sources applies, the temporal reconstruction of non-Gaussian sources is arbitrarily mixed with Gaussian sources. It is then desirable to derive a methodology

with which this adverse effect can be minimized without having to resort to overcomplete ICA, which raises the complexity of source reconstruction and requires additional constraints on the reconstructed sources (cf. [LLGS99, ZP01]).
Towards this goal, note that the topographies of the non-Gaussian sources are correctly reconstructed by complete ICA in spite of the presence of an arbitrary number of Gaussian sources. This can be used to improve the SNR of the reconstructed non-Gaussian sources in the following way. Assume that a reconstructed source topography $\hat{\boldsymbol{a}}_i$ correctly represents the topography of a non-Gaussian source $\boldsymbol{s}_i$, i.e., that $\hat{\boldsymbol{a}}_i = \boldsymbol{a}_i$. Without loss of generality possible scaling is disregarded here. To estimate the temporal evolution of $\boldsymbol{s}_i$, it is desirable to design a spatial filter $\boldsymbol{v}_i$ that extracts the source $\boldsymbol{s}_i$ from the available measurements $\boldsymbol{x}$ while optimally attenuating all other sources. If optimal attenuation is defined in terms of the variance of interfering sources, this can be formulated mathematically as

$$\boldsymbol{v}_i = \underset{\boldsymbol{v}\in\mathbb{R}^M}{\operatorname{argmin}} \left\{ \boldsymbol{v}^{\mathrm{T}} R_{\boldsymbol{x}} \boldsymbol{v} \right\} \quad \text{s.t.} \quad \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{a}_i = 1. \tag{5.32}$$

This is the problem of linearly constrained minimum variance (LCMV) spatial filtering, which has been originally proposed and solved in [VvYS97]. The solution to (5.32) is given by

$$\boldsymbol{v}_i = \left(\boldsymbol{a}_i^{\mathrm{T}} R_{\boldsymbol{x}}^{-1} \boldsymbol{a}_i \right)^{-1} \boldsymbol{a}_i^{\mathrm{T}} R_{\boldsymbol{x}}^{-1}. \tag{5.33}$$

Accordingly, $\boldsymbol{y}_i = \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{x}$ is an optimal estimate of $\boldsymbol{s}_i$ in so far that the variance of the interference of all other sources is minimized. Note that this minimization of interference includes other Gaussian as well as non-Gaussian sources. Consequently, statistical independence of the non-Gaussian sources is traded here against minimization of the variance of the interference of all sources.
In terms of the predictions formulated in Section 5.2.3, in order to validate the proposed mixture model, note that a reduction of the SNR of the reconstructed ICs by LCMV spatial filtering can only be expected for an overcomplete mixture model. If $K = M$, and hence $A \in \mathbb{R}^{M\times M}$, it is easy to show that indeed $\boldsymbol{y}_i = \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{x} = \boldsymbol{s}_i$ by plugging in (5.33). The proposed mixture model thus predicts that complete ICA in conjunction with LCMV spatial filtering outperforms complete ICA in the reconstruction of ERPs/ERFs, while a complete mixture model leads to no improvements. This is tested in the next section. In studies using measures of ERS/ERD, as in feature extraction for non-invasive BCIs based on motor imagery, LCMV spatial filtering can not be expected to affect the results since only the baseline is altered. Consequently, the proposed mixing model predicts that using complete ICA in conjunction with LCMV spatial filtering in feature extraction for non-invasive BCIs does not alter results in comparison to using complete ICA alone.

5.3 Experimental Results

In this section, complete ICA in conjunction with LCMV spatial filtering is applied to auditory evoked ERFs and to EEG data from a four-class motor imagery

paradigm. The primary purpose of this section is to test the predictions made in Section 5.2.3, i.e., to validate the assumption of a mixture model with more sources than sensors but less non-Gaussian sources than sensors in the context of EEG/MEG analysis. This is achieved by showing that complete ICA in conjunction with LCMV spatial filtering achieves results superior to complete ICA alone in the reconstruction of auditory ERFs in Section 5.3.1, and by showing that in the context of non-invasive BCIs based on motor imagery constructing spatial filters by complete ICA and LCMV spatial filtering does not perform better than complete ICA alone (Section 5.3.2).
Besides the validation of the proposed mixture model, this also serves to illustrate the efficacy of complete ICA in conjunction with LCMV spatial filtering in the reconstruction of the dynamics of non-Gaussian sources, and to establish why ICA constitutes a powerful tool for feature extraction in the context of non-invasive BCIs in spite of the overcomplete mixture model.

5.3.1 Denoising of Event Related Fields

In this section, complete ICA combined with LCMV spatial filtering is employed for denoising of ERFs recorded by MEG. In general, data denoising by ICA is based on the assumption that only a small number of ICs reconstructed from a given data set are relevant for the considered experimental setup, i.e., belong to the signal subspace, while all other ICs constitute noise. Only the ICs belonging to the signal subspace are then reprojected onto the observation space, resulting in a rank-reduced signal with improved signal-to-noise ratio (SNR). It should be noted that the identification of ICs relevant for a given experimental setup is not trivial, and hence mostly done manually. In the context of the source model considered here, it is assumed that only the L non-Gaussian sources belong to the signal subspace. The deviation from Gaussianity of the reconstructed sources is hence considered as a criterion for the identification of relevant ICs (cf. [BGUB06]).
As MEG data Event Related Fields (ERFs) are chosen. ERFs typically have a very low SNR, and are difficult to detect in single trial data. For this reason numerous trials are recorded, and the ERF is estimated by taking the ensemble average of all trials. Based on the assumption that only the ERF component of the MEG is invariant in every trial, this results in an unbiased estimator of the ERF (termed the *grand average* ERF). In complex experimental setups, or if subjects with a short attention span such as small children are under investigation, the recording of numerous trials is not feasible. The goal of ERF denoising by ICA is then to reconstruct the grand average ERF from only a small number of trials. This application is well suited for evaluating the approach presented in the Section 5.2.4, because a data set can be used for which the grand average ERF actually is available. This allows an objective evaluation of the obtained denoising results, and thus a validation of the predictions formulated in Section 5.2.4.
The test data set consists of Auditory Evoked Fields (AEFs), recorded during an

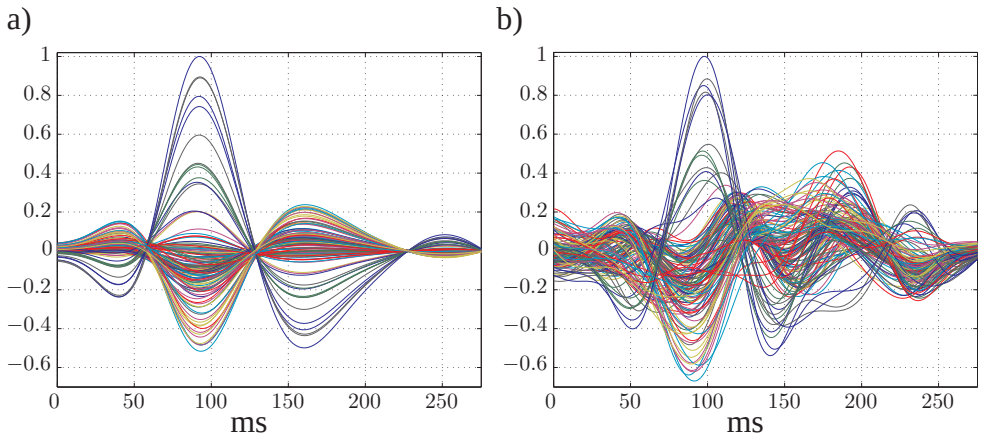

Figure 5.1: Grand average ERF $\boldsymbol{y}^*$ (a) and ERF average of ten randomly chosen trials $\boldsymbol{y}^{\text{raw}}$ (b).

auditory oddball task at the Biomagnetic Imaging Laboratory of the University of California, San Francisco. Auditory stimuli were applied to the left ear, while MEG was recorded at a sampling rate of 4 kHz with $M = 132$ sensors covering the right hemisphere. A total of 250 trials were recorded, with each trial lasting from -275 to 275 ms and the stimulus being applied at 0 ms (see [NAHS06] for a detailed description of the recording procedure). Out of the total number of 250 trials ten trials were chosen randomly for estimation of the raw average ERF. The grand average $\boldsymbol{y}^*$ was computed by taking the average time course of all 250 trials, and filtering the resulting average sequentially with a low- and high-pass filter with cut-off frequencies 2 Hz and 16 Hz respectively (for all temporal filtering procedures in this section a third order Butterworth filter was used). The resulting temporal activity at all channels is shown in Figure 5.1 (a). The same temporal filtering procedure was applied to the average of the randomly chosen ten trials, resulting in the temporal activity $\boldsymbol{y}^{\text{raw}}$ shown in Figure 5.1. Note that only the post-stimulus period is shown in both figures. For a quantitative comparison of the data sets, the SNR is defined as

$$\text{SNR}(\hat{\boldsymbol{y}}) := 10 \log_{10} \left(\frac{1}{M} \sum_{i=1}^{M} \frac{\sum_{t=1}^{T} y_i^* [t]^2}{\sum_{t=1}^{T} \left(y_i^* [t] - \hat{y}_i [t] \right)^2} \right) \text{(dB)}, \tag{5.34}$$

with samples $t = 1 \ldots T$ corresponding to the post-stimulus period of the data. Each data sets was first normalized to the maximum value of all channels before computing the SNR. This resulted in a SNR of -0.09 dB for the data set $\boldsymbol{y}^{\text{raw}}$.
To evaluate the denoising capabilities of ICA, the extended Infomax-algorithm as implemented in EEGLab [DM04] was applied to the concatenated ten trials that were randomly chosen as test data (from here on referred to as the data vector $\boldsymbol{x}$), resulting in estimated source topographies $\hat{\boldsymbol{a}}_i$ and temporal source estimates y_i with

$i = 1, \ldots, M$. Note that for simplicity the time index is dropped. Four different evaluation schemes were then investigated:

1. *Ordinary ICA.* The reconstructed sources $\boldsymbol{y}_i$ are sorted in descending order according to the variance of the original data explained by each source. Only the first L sources with the highest explained variance are reprojected onto the observation space,

$$\hat{\boldsymbol{x}}^{(1)} = \sum_{i=1}^{L} \hat{\boldsymbol{a}}_i y_i. \tag{5.35}$$

2. *ICA with LCMV spatial filtering.* The temporal source activity of each source is estimated using the LCMV spatial filtering approach, and the resulting source estimates are again sorted in descending order according to the amount of variance of the original data explained by each source. The first L sources explaining the highest amount of variance are reprojected onto the observation space, resulting in

$$\hat{\boldsymbol{x}}^{(2)} = \sum_{i=1}^{L} \hat{\boldsymbol{a}}_i (\hat{\boldsymbol{a}}_i^{\mathrm{T}} R_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{a}}_i)^{-1} \hat{\boldsymbol{a}}_i^{\mathrm{T}} R_{\boldsymbol{x}}^{-1} \boldsymbol{x}. \tag{5.36}$$

   Note that in this and the fourth evaluation scheme diagonal loading is used to obtain numerically stable estimates of the inverse of the covariance matrix $R_{\boldsymbol{x}}$.

3. *Ordinary ICA with identification of relevant non-Gaussian sources.* The sources are reconstructed with complete ICA, but not sorted in descending order according to the amount of variance explained by each IC. Instead, the deviation from Gaussianity of each source y_i is estimated in multiple stages. First, the average temporal activity of each source across the ten trials is computed. Then, the probability density function (pdf) of each averaged source is estimated for the post-stimulus period using a non-parametric kernel approach (cf. [BA97]). A Gaussian kernel is used, which is optimal for Gaussian distributions. Then, the Kullback-Leibler distance (cf. [CT06]) of the estimated pdf to a Gaussian distribution with equal variance is calculated by numerical integration. Finally, the sources are sorted from highest to lowest Kullback-Leibler distance, i.e., from least to most Gaussian. The data set $\hat{\boldsymbol{x}}^{(3)}$ is then calculated in the same way as in (5.35), but by reprojecting the L most non-Gaussian sources.

4. *ICA with LCMV spatial filtering and identification of relevant non-Gaussian sources.* The temporal source activity of each source is again estimated using complete ICA in conjunction with LCMV spatial filtering. The estimated sources are sorted in descending order according to their deviation from Gaussianity as for evaluation scheme three. The data set $\hat{\boldsymbol{x}}^{(4)}$ is then calculated

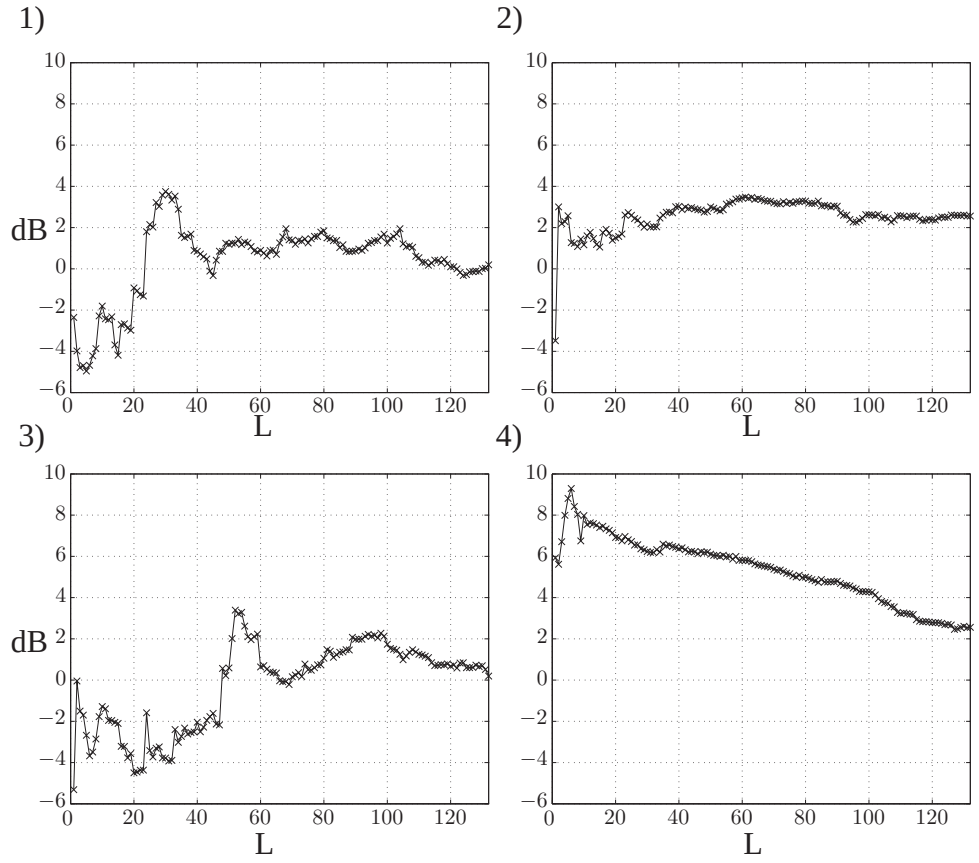

Figure 5.2: SNR of the evaluation schemes 1-4.

in the same way as in (5.36), but by reprojecting the L most non-Gaussian sources.

The denoised data sets $\boldsymbol{y}^{(j)},\ j = 1, \ldots, 4$, are calculated from the data sets $\hat{\boldsymbol{x}}^{(j)}$ by taking the average across the ten trials of $\boldsymbol{x}^{(j)}$, applying the same temporal filtering procedure as for the grand average data set, and normalizing to the maximum value across all channels of each $\boldsymbol{y}^{(j)}$. Note that determining the parameter L, corresponding to the dimension of the signal subspace, is a non-trivial issue related to model identification. This is beyond the scope of this work. The resulting SNRs for all four schemes applied to the ten randomly chosen trials are shown in Figure 5.2 as a function of $L \in \mathbb{N}$. The maximum SNR achieved for each evaluation scheme is summarized in Table 5.2 with Figure 5.3 showing the corresponding time series. As can be seen from Table 5.2, the best SNR of 9.29 dB is achieved for ICA with LCMV spatial filtering and sorting of the estimated sources by their deviation from Gaussianity. The SNRs for the other three evaluation schemes are roughly equal at about 3.5 dB. Note that the best SNR for evaluation scheme four is obtained for $L = 6$, while the optimal SNRs for the other evaluation schemes are obtained for much higher dimensions of the signal subspace (cf. Figure 5.2). As it can be

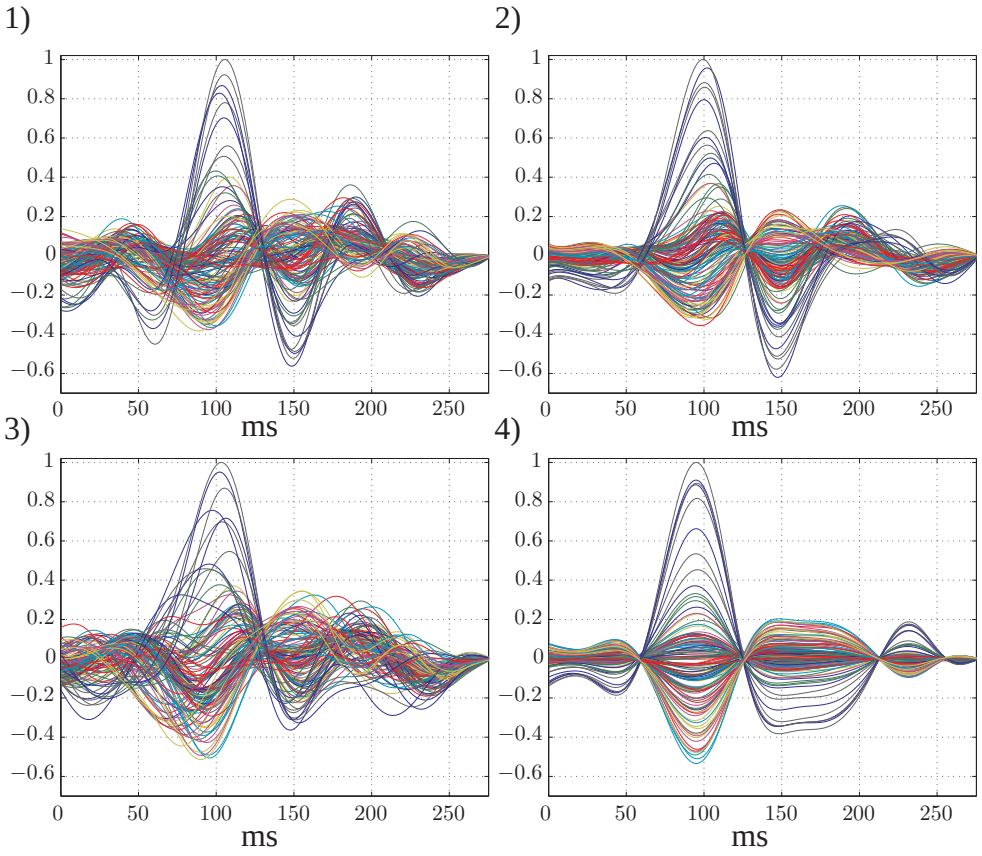

Figure 5.3: Denoised ERFs with optimal L for evaluation schemes 1-4.

Evaluation Scheme	1	2	3	4
Maximum SNR	3.75 dB	3.48 dB	3.32 dB	9.29 dB
$L_{\max}$	30	61	51	6

Table 5.2: Maximum SNR for each of the four denoising schemes.

Training trials per cond.	10	20	30	40	50	60	70	80
Subject k3b (ICA)	73.7	82.0	86.4	89.1	88.8	91.2	91.2	93.4
Subject k3b (LCMV)	73.6	82.0	86.5	89.1	89.3	91.4	91.8	93.4
Subject k6b (ICA)	45.8	52.0	56.7	59.8	62.9	-	-	-
Subject k6b (LCMV)	45.9	51.8	56.5	59.4	62.9	-	-	-
Subject l1b (ICA)	59.8	67.3	71.2	74.0	78.9	-	-	-
Subject l1b (LCMV)	59.6	67.6	71.1	74.2	78.4	-	-	-

Table 5.3: Mean classification results in percent for multi-class ITFE with complete ICA and complete ICA in conjunction with LCMV spatial filtering.

expected from the SNRs, the temporal activities at the recording channels for the optimum SNR of each evaluation scheme differ significantly (cf. Figure 5.3). While the fourth evaluation scheme correctly reconstructs all major peaks of the grand average ERF (compare Figure 5.1), for the other three evaluation schemes only the major peak around 100 ms is clearly discernible.
In summary, the experimental results presented in this section establish that combining ICA with LCMV spatial filtering significantly improves the performance of ICA in the reconstruction of ERFs. Note that this is in agreement with the expected results for an overcomplete mixture model with more sources than sensors but less non-Gaussian sources than sensors as formulated in Section 5.2.3. Furthermore, note that the results are in contradiction to the assumption of a complete mixture model, for which LCMV spatial filtering would not result in improved performance (cf. Section 5.2.4).

5.3.2 Feature Extraction in BCIs

To investigate the efficacy of complete ICA in conjunction with LCMV spatial filtering for feature extraction in non-invasive BCIs, the same experimental data and evaluation procedure as in Section 4.3 are employed (multi-class ITFE). However, instead of performing ICA by joint approximate diagonalization (JAD) [ZLNM04], spatial filters are computed using a) the extended Infomax algorithm as implemented in EEGLab [DM04], and b) using the extended Infomax algorithm in conjunction with LCMV spatial filtering as described in Section 5.2.4. The obtained classification results for subjects k3b, k6b, and l1b are shown in Table 5.3. Comparing these results with those obtained in Section 4.3 reveals only minor differences

in mean classification accuracy between using JAD and extended Infomax for ICA. More importantly, however, the mean classification accuracies differ on average by less than 0.2% between using the extended Infomax algorithm alone and combining it with LCMV spatial filtering.
Note that this is in agreement with expectations due to the fact that interference from Gaussian sources, which can be alleviated by LCMV, only corresponds to a shift of the baseline of ERS/ERD measurements as discussed in Section 5.2.3. If ERS/ERD measurements, i.e., variance changes, are used as features for inferring the BCI-user's intention, a baseline shift corresponds to a translation of the feature space. A translation has got no qualitative influence on the training process of a linear classifier, and thus does not affect classification accuracies.

5.4 Discussion

Motivated by the apparent contradiction between the success of complete ICA and the implausibility of a complete mixture model in EEG/MEG analysis, in this chapter the performance of complete ICA was theoretically investigated for arbitrary mixture models. Necessary and sufficient conditions for solutions of complete ICA for arbitrary mixture models could be provided (Theorem 5.1), resulting in the characterization of separability of sources and model identifiability of complete ICA for arbitrary mixture models (summarized in Table 5.1).
These results were then used to argue that empirical evidence on complete ICA in the analysis of EEG/MEG data is in agreement with an overcomplete mixture model with less non-Gaussian sources than sensors but an arbitrary number of Gaussian sources. Under the assumption of this mixture model, predictions were formulated on the behavior of complete ICA in the reconstruction of ERPs/ERFs and in the analysis of ERS/ERD. By combining complete ICA with LCMV spatial filtering, a methodology was presented that enables testing these predictions. It was then shown that the performance of complete ICA with and without LCMV spatial filtering in the reconstruction of ERFs does indeed conform to the predictions derived from the proposed mixture model, and that this empirical evidence is in contradiction to a complete mixture model. Furthermore, it was shown that experimental results on feature extraction for non-invasive BCIs by complete ICA and LCMV spatial filtering also agree with an overcomplete mixture model. Note, however, that the results on feature extraction are not in contradiction to a complete mixture model and thus, in comparison to the results on ERFs, provide no compelling evidence for an overcomplete mixture model. In summary, it is concluded that the theoretical and empirical evidence is in favor of the proposed mixture model. This argument thereby provides an explanation for the success of complete ICA in the analysis of EEG/MEG recordings without resorting to a physiologically implausible complete mixture model.
It should be noted that the validity of an overcomplete mixture model with few

non-Gaussian and arbitrary many Gaussian sources has got several implications for the analysis of EEG/MEG recordings by complete ICA. First, it ensures that the topographies of non-Gaussian sources are correctly reconstructed in spite of an overcomplete mixture model. Second, Gaussian sources are arbitrarily mixed into reconstructed Gaussian sources. While this has got no qualitative effect on the analysis of ERS/ERD measures, it does degrade the SNR of the dynamics of reconstructed ICs. This adverse effect can be alleviated by combining complete ICA with LCMV spatial filtering. However, this trades statistical independence of reconstructed sources against minimization of the variance of interference from unwanted sources. This has to be taken into account when deriving physiological conclusions from reconstructed ICs.

Finally, the presented theoretical and experimental results provide an explanation for the success of complete ICA in the design of feature extraction algorithms for non-invasive BCIs. As long as only variance changes are used for inferring the user's intention, as it is usually done in BCIs based on motor imagery paradigms, the inclusion of Gaussian sources in reconstructed ICs amounts to a translation of the feature space. Since this does not affect the performance of linear classifiers, this adverse effect of complete ICA in EEG/MEG analysis can at present be disregarded in the context of non-invasive BCIs based on motor imagery.

Chapter 6

Feature Extraction via Beamforming

6.1 Introduction

In Chapter 3, only the spatial distribution of current density within the brain was used for inferring the BCI-user's intention, while in Chapters 4 and 5 algorithms for feature extraction were developed that were primarily based on a-priori information on temporal coding of cognitive states in the electric field of the brain. In this chapter, a-priori information on spatial as well as temporal coding of cognitive states is used to design a robust, computationally simple, and effective feature extraction algorithm. This algorithm can be interpreted as a special type of beamformer, a spatial filter usually associated with applications in radar technology or classical communication theory (cf. [VB88] for a review).

In EEG analysis, a beamformer is a linear spatial filter that optimally attenuates all EEG sources not originating from a specific location or region within the brain (cf. [GI99] for a review). Beamformers are applicable in the context of non-invasive BCIs, since some knowledge on which regions of the brain, termed regions of interest (ROIs), provide information on the user's intention is usually available. For example, for non-invasive BCIs based on motor imagery paradigms it is well known that haptic motor imagery of a limb leads to a decrease in power of the electric field of the brain originating in that part of the motor cortex representing the specific limb (cf. [PL99] and the experimental results presented in Chapter 3). Furthermore, the location of a certain region of the human brain within the skull, e.g., the motor cortex, does not vary significantly across subjects, and is thus approximately known. In conjunction, this information can be used to design spatial filters that selectively extract those components of the EEG that originate in the regions of the brain considered most relevant for inferring the user's intention.

In comparison to approaches discussed in the previous chapters this has got several advantages. First, beamforming is a very robust form of feature extraction, since any noise that does not originate within the ROI, e.g., that is caused by muscular or ocular artifacts, is optimally attenuated. Second, is demonstrated in Section 6.3 and discussed in Section 6.4, beamforming is computationally less demanding than

source localization, and thus applicable in online BCIs with real-time feedback. Finally, beamforming is completely unsupervised, i.e., it does not require any labeled training data. As such, it does not suffer from overfitting phenomena as the algorithms presented in Chapters 3 and 4, and enables a high rate of convergence of a subsequent classifier to its minimum expected error probability.
The structure of this chapter is as follows. In Section 6.2, the assumptions made in this chapter on the class of allowed feature spaces is specified. Then, a beamforming approach, similar to the concept of maximum SNR beamforming [MM80], is derived that selectively extracts EEG components from specific brain regions. In Section 6.3, experimental results based on EEG recordings from a two-class motor imagery paradigm are presented. First, offline classification results of ten healthy subjects are presented, and classification accuracies are compared with those obtained by using the CSP algorithm for feature extraction (cf. Section 4.2.1). Then, the feasibility of using the beamforming approach for realizing real-time control of a cursor in one dimension is demonstrated. The chapter concludes in Section 6.4 with a general discussion of the usability of beamforming for feature extraction in non-invasive BCIs. Some of the work in this chapter has already been presented in [GWGB07].

6.2 Methods

In this chapter, only two-class paradigms are considered. Hence, the BCI-user's intention is denoted by $c \in \mathcal{C} = \{c_1, c_2\}$. The recorded EEG data is again referred to as $X \in \mathbb{R}^{M \times T}$ for a block of T samples, and $\boldsymbol{x}(t) \in \mathbb{R}^M$ for a single sample point. If the time index is dropped, $\boldsymbol{x}$ is treated as a M-dimensional random variable. The assumptions made in this chapter to limit the class of allowed feature spaces $\mathcal{P}^*$ (cf. Definition 2.13) are:

1. The user's intention is encoded in variance changes of the recorded EEG data.

2. For motor imagery paradigms, only the EEG components originating in those parts of the motor cortex representing the involved limbs provide information on the BCI-user's intention. These regions are termed regions of interest (ROIs).

The first assumption is identical to Chapters 4 and 5, and expresses our knowledge on temporal coding of cognitive states in the electric field of the brain. The second assumption, on the other hand, incorporates more detailed spatial constraints on the class of allowed features than in previous chapters. Contrary to previously employed spatial constraints, those imposed here depend on the specific experimental paradigm being used. In general, different paradigms require different ROIs.
The desired feature extraction algorithm then is of the form $T : \mathbb{R}^{M \times T} \mapsto \mathbb{R}_+^{KN}$, $T(X) = \text{Var}\left\{W^{\text{T}} X\right\}$, with the columns of the matrix $W \in \mathbb{R}^{M \times N}$ containing the

N beamformers. Note that, as in previous chapters, the operator Var$\{.\}$ refers to the variances of the components in K specific frequency bands, and $N = 2$ due to the restriction to two-class paradigms.
In the context of EEG analysis, beamforming is frequently used for the purpose of source localization ([VvYS97],[GI99]). This is realized by specifying a three-dimensional grid within the brain, and designing a beamformer for every single grid point. The power of the obtained estimate of the electric field at a grid point is then taken as an estimate of the current density at this location within the brain. In this way, the whole brain can be scanned, resulting in a three-dimensional map of the estimated current density distribution. For this purpose, it is desirable to maximize the spatial resolution of the employed beamformer in order to ensure minimum interference from adjacent grid points in estimating the current density at a certain location within the brain. This is in contrast to the requirements of beamforming in the context of non-invasive BCIs. Here, the ROI can be considered as an extended region rather than a single point within the brain. Furthermore, the ROI is only approximately known. It is hence desirable to derive a beamformer that can be pointed at a whole region within the brain, and that optimally attenuates all sources not originating within this ROI. In the next section, the derivation of such a beamformer is presented.

6.2.1 Maximum SNR Beamforming in EEG

In general, it is desirable to derive a spatial filter that eliminates all electric activity that does not originate in a chosen ROI. This, however, is not possible due to the ill-posed nature of the inverse problem of EEG. In EEG recordings, electric activity originating from an infinite dimensional space (the continuous current distribution within the brain) is mapped onto a finite number of measurement electrodes. For this reason, estimating the electric field at a certain position inside the brain constitutes an underdetermined problem. The best one can do is to find a spatial filter that in some sense optimally attenuates all activity not originating in the chosen ROI. Motivated by the assumption that only variance changes provide information on the BCI-user's intention, optimal attenuation is defined here as maximizing the ratio of the variance of the electric field originating in a certain ROI and the total variance of the electric field. Such a linear spatial filter is now derived, and its properties are discussed.

Derivation of the Maximum SNR Beamformer

The electric potential generated by the brain and measured at a position $\boldsymbol{r}$ on the scalp is given by (cf. [NS05])

$$\Phi(\boldsymbol{r}, t) = \int_V L(\boldsymbol{r}, \boldsymbol{r}')^{\mathrm{T}} P(\boldsymbol{r}', t) \mathrm{d}V(\boldsymbol{r}'), \tag{6.1}$$

with V the volume of the brain, $P : \mathbb{R}^3 \times \mathbb{R} \mapsto \mathbb{R}^3$ the tissue dipole moment (source strength) at position $\boldsymbol{r}'$ and time t in x, y, and z - direction, and $L : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}^3$ the so called leadfield equation, describing the projection strength of a source with dipole moment in x, y, and z - direction at position $\boldsymbol{r}'$ to a measured electric potential at position $\boldsymbol{r}$. Note that the leadfield equation incorporates all geometric and conductive properties of the brain. Here, the electric field of the brain is spatially sampled at $i = 1, \ldots, M$ electrodes on the scalp with position $\boldsymbol{r}_i$, resulting in a measurement vector $\boldsymbol{x}(t)$ with the elements

$$x_i(t) = \int_V L(\boldsymbol{r}_i, \boldsymbol{r}')^{\mathrm{T}} P(\boldsymbol{r}', t) \mathrm{d}V(\boldsymbol{r}'),\ i = 1, \ldots, M. \tag{6.2}$$

The goal of maximum SNR beamforming is to find a linear transformation of the measured EEG

$$y(t) = \boldsymbol{w}^{*\mathrm{T}} \boldsymbol{x}(t) \tag{6.3}$$

that maximizes the ratio of the variance of the electric field originating in a certain region of the brain and the overall variance. For this, the component of the EEG originating in a certain ROI is defined as $\tilde{\boldsymbol{x}}(t)$, with the elements

$$\tilde{x}_i(t) = \int_{\mathrm{ROI}} L(\boldsymbol{r}_i, \boldsymbol{r}')^{\mathrm{T}} P(\boldsymbol{r}', t) \mathrm{d}V(\boldsymbol{r}'),\ i = 1 \ldots M. \tag{6.4}$$

Computing the spatial filter that maximizes the ratio of the variance of $\tilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ requires their respective covariance matrices. For $\boldsymbol{x}$, the electric field due to sources within the whole brain, the covariance matrix $R_{\boldsymbol{x}}(t)$ can be computed using the recorded EEG data. The covariance matrix of $\tilde{\boldsymbol{x}}$, however, has to be estimated in a different way. First note that the integral in (6.4) can be approximated in a very simplistic manner as

$$\tilde{x}_i(t) = \alpha \sum_{j=1}^{J} L(\boldsymbol{r}'_j, \boldsymbol{r}_i)^{\mathrm{T}} P(\boldsymbol{r}'_j, t), \tag{6.5}$$

with $\boldsymbol{r}'_j, j = 1, \ldots, J$ the locations of an equally spaced grid with J points within the ROI and α some numerical constant. The electric field at the M electrodes on the scalp can thus be approximated as

$$\tilde{\boldsymbol{x}}(t) = \alpha L \boldsymbol{p}(t), \tag{6.6}$$

with the leadfield matrix $L \in \mathbb{R}^{M \times 3J}$ describing the projection strength in x,y, and z-direction of the sources at the J grid points to the M electrodes, and $\boldsymbol{p}(t) \in \mathbb{R}^{3J}$ representing the dipole moments of the J sources. Without loss of generality, it is assumed that each element of $\boldsymbol{p}(t)$ has zero mean. The covariance matrix of $\tilde{\boldsymbol{x}}(t)$ can then be written as

$$R_{\tilde{\boldsymbol{x}}}(t) = \alpha^2 L R_{\boldsymbol{p}}(t) L^{\mathrm{T}}, \tag{6.7}$$

with $R_{\boldsymbol{p}}(t)$ the source covariance matrix. The desired spatial filter is then found by solving the optimization problem

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathbb{R}^M}{\operatorname{argmax}} \left\{ \frac{\boldsymbol{w}^{\mathrm{T}} R_{\tilde{\boldsymbol{x}}}(t) \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}}(t) \boldsymbol{w}} \right\}. \tag{6.8}$$

Since (6.8) is in the form of the well-known Rayleigh quotient, solutions to (6.8) are given by the eigenvectors of the generalized eigenvalue problem

$$R_{\tilde{\boldsymbol{x}}}(t)\boldsymbol{w} = \lambda R_{\boldsymbol{x}}(t)\boldsymbol{w}. \tag{6.9}$$

Since for an eigenvalue λ^* with associated eigenvector $\boldsymbol{w}^*$ it holds that

$$\lambda^* = \frac{\boldsymbol{w}^{*\mathrm{T}} R_{\tilde{\boldsymbol{x}}}(t) \boldsymbol{w}^*}{\boldsymbol{w}^{*\mathrm{T}} R_{\boldsymbol{x}}(t) \boldsymbol{w}^*}, \tag{6.10}$$

the eigenvector of (6.9) with the largest eigenvalue constitutes the desired beamformer. Then note that letting $\tilde{\lambda} := \lambda/\alpha^2$ and inserting (6.7) into (6.9) yields the generalized eigenvalue problem

$$L R_{\boldsymbol{p}}(t) L^{\mathrm{T}} \boldsymbol{w} = \tilde{\lambda} R_{\boldsymbol{x}}(t)\boldsymbol{w}. \tag{6.11}$$

Solving (6.11) requires knowledge of the leadfield matrix L and the covariance matrix $R_{\boldsymbol{p}}(t)$ for the sources within the ROI. The leadfield matrix for a given ROI can be estimated using models of EEG volume conduction discussed in Section 3.2.2 and [BML01]. In this chapter, as in Chapter 3, only the four-shell spherical head model is considered [RD69]. Each column of L describes the projection strength of a current dipole at a certain grid point within the ROI to all M electrodes due to its dipole moment in x, y, or z-direction. The columns of L thereby implicitly define the ROI and the orientation of sources within the ROI. The source covariance matrix $R_{\boldsymbol{p}}(t)$, on the other hand, has to be specified using a-priori knowledge. In absence of any a-priori knowledge, it is assumed that $R_{\boldsymbol{p}}(t) = I$, i.e., that all sources have equal variance and are mutually uncorrelated. However, more realistic assumptions, such as an exponential decrease of correlation of sources with geometric distance, could easily be implemented. Finally, it should be noted that any constant scaling of the source covariance matrix or the leadfield matrix has no effect on the eigenvectors of (6.11), and thus also no effect on the optimal spatial filter. The largest eigenvector of (6.11), and thus the optimal beamformer, can then be computed by standard numerical tools for generalized eigenvalue problems.

Properties of the Maximum SNR Beamformer

Several issues in the derivation of the beamformer warrant a further discussion. First, it is assumed that the covariance matrix of the EEG data can and should be estimated from available data. This is not imperative. Instead, the same model-based

approach used for estimating the covariance matrix $R_{\tilde{\boldsymbol{x}}}(t)$ could be employed to estimate $R_{\boldsymbol{x}}(t)$. This would result in a data-independent beamformer, i.e., a beamformer that does not depend on the observed EEG. There are two primary reasons, however, to prefer estimating $R_{\boldsymbol{x}}(t)$ from available data. First, a model-based approach for estimating $R_{\boldsymbol{x}}(t)$ introduces unnecessary uncertainties, i.e., the source and head model, into the evaluation process. This should be avoided if possible. Second, for real EEG data some regions of the brain can be expected to be more active than others, resulting in a non-uniform current distribution. If $R_{\boldsymbol{x}}(t)$ is estimated from real data, the beamformer can adapt to this non-uniform current distribution. The result is a spatial filter that focuses on attenuating those sources within the brain that interfere most with the electric field originating within the ROI. This is in contrast to a beamformer using a model-based approach for estimating $R_{\boldsymbol{x}}(t)$. Here, all sources within the brain are attenuated regardless of their actual contribution to the deterioration of the SNR. Hence, a higher SNR of the beamformer can be expected if $R_{\boldsymbol{x}}(t)$ is estimated from real data.

Up to this point, it has been assumed that $R_{\boldsymbol{x}}(t)$ can be easily estimated from available data. This is in indeed correct if $\boldsymbol{x}$ is a stationary (or quasi-stationary) random variable with independently distributed samples. In this case the standard unbiased estimator of a covariance matrix can be employed, i.e.,

$$R_{\boldsymbol{x}} = \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{x}(t)-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}(t)-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathrm{T}} \tag{6.12}$$

with $\boldsymbol{\mu}_{\boldsymbol{x}}$ the (sample) mean of $\boldsymbol{x}$. However, if $\boldsymbol{x}$ is non-stationary, which for EEG data unfortunately is indicated by empirical evidence [Pal96], estimation of $R_{\boldsymbol{x}}(t)$ becomes non-trivial. More specifically, estimation of a non-stationary covariance matrix requires, explicitly or implicitly, the definition of a time window in which the random variable is considered stationary (parametric methods for estimating $R_{\boldsymbol{x}}(t)$ in which the non-stationarity is explicitly modelled are disregarded here). The optimal length of this window, in terms of minimizing some error between estimated and real covariance matrix, is influenced by several factors. These include a) the extent of the non-stationarity, i.e., the speed with which the covariance matrix changes, b) the deviation of $\boldsymbol{x}$ from the assumption of independently distributed samples (temporally correlated samples of $\boldsymbol{x}$ provide less information on $R_{\boldsymbol{x}}(t)$ than uncorrelated samples), and c) the actual probability density function of $\boldsymbol{x}$ (note that the standard unbiased estimator for a covariance matrix is only optimal in terms of the Cramer-Rao lower bound if $\boldsymbol{x}$ is Gaussian distributed). The actual effect of varying the number of samples for computing (6.12) on the classification accuracy is demonstrated in Section 6.3.1.

Furthermore, the beamformer derived here differs from the maximum SNR beamformer usually considered in the literature (cf. [MM80]) in the choice of the denominator in (6.8). In the standard maximum SNR beamformer, $R_{\tilde{\boldsymbol{x}}}(t)$ refers to the covariance matrix of the signal subspace, while $R_{\boldsymbol{x}}(t)$ refers to the covariance

matrix of the noise subspace. Here, however, $R_{\boldsymbol{x}}(t)$ represents the covariance matrix of the recorded EEG data, and thereby includes the noise as well as the signal subspace. As such, it is not immediately evident that (6.8) indeed results in a desirable spatial filter, and it could be argued that instead of solving (6.8) it would be desirable to solve the optimization problem

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathbb{R}^M}{\operatorname{argmax}} \left\{ \frac{\boldsymbol{w}^{\mathrm{T}} R_{\tilde{\boldsymbol{x}}}(t) \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} R_{\text{Noise}}(t) \boldsymbol{w}} \right\} \tag{6.13}$$

with $R_{\text{Noise}}(t)$ describing the covariance matrix of all sources outside the ROI. Having to solve (6.13) would be disadvantageous, since it is in practice impossible to estimate $R_{\text{Noise}}(t)$ from recorded data. Using a model-based approach to estimate $R_{\text{Noise}}(t)$, on the other hand, is undesirable due to the same reasons as discussed above for estimating $R_{\boldsymbol{x}}(t)$. It is shown now that under some mild assumption solving (6.8) and (6.13) yield the same generalized eigenvalue problem, thereby establishing that solving (6.8) does indeed result in the desired spatial filter. First, consider again the linear EEG model

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) = \begin{bmatrix} A_{\boldsymbol{s}} & A_{\boldsymbol{n}} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}(t) \\ \boldsymbol{n}(t) \end{bmatrix}, \tag{6.14}$$

with $\boldsymbol{s}(t)$ denoting the sources within the ROI (the signal subspace) and $\boldsymbol{n}(t)$ denoting the signals outside the ROI (the noise subspace). Assuming $\boldsymbol{s}(t)$ and $\boldsymbol{n}(t)$ to be uncorrelated (and, without loss of generality, to have zero mean), the covariance matrix of $\boldsymbol{x}$ is given by

$$R_{\boldsymbol{x}}(t) = A_{\boldsymbol{s}} R_{\boldsymbol{s}}(t) A_{\boldsymbol{s}}^{\mathrm{T}} + A_{\boldsymbol{n}} R_{\boldsymbol{n}}(t) A_{\boldsymbol{n}}^{\mathrm{T}} = R_{\tilde{\boldsymbol{x}}}(t) + R_{\text{Noise}}(t). \tag{6.15}$$

Now, solutions to (6.13) are given by eigenvectors of the generalized eigenvalue problem

$$R_{\tilde{\boldsymbol{x}}}(t)\boldsymbol{w} = \lambda R_{\text{Noise}}(t)\boldsymbol{w}, \tag{6.16}$$

which can be rewritten using (6.15) as

$$R_{\tilde{\boldsymbol{x}}}(t)\boldsymbol{w} = \tilde{\lambda} R_{\boldsymbol{x}}(t)\boldsymbol{w} \tag{6.17}$$

with $\tilde{\lambda} = \lambda/(1+\lambda)$. Comparing (6.17) and (6.9) then shows that under the assumption of the EEG sources within and outside the ROI being uncorrelated the optimization problems (6.8) and (6.13) yield identical spatial filters.

6.3 Experimental Results

In this section, the beamformer derived in the previous section is applied to experimental EEG data from a two-class motor imagery paradigm. First, offline results of ten healthy subjects are presented, and the performance of the beamforming approach, using two beamformers with their ROIs centered within the left and

right motor cortex, are compared with that of the CSP algorithm (cf. Section 4.2.1). The CSP algorithm is chosen for comparison due to its optimality for two-class paradigms in terms of maximizing (an approximation of) mutual information of class labels and extracted features as proved in Section 4.2.3.
Two different procedures for computing the beamformers are evaluated. In the first procedure, termed static beamforming, the beamformers are computed using previously recorded EEG data and then kept invariant for the rest of the experiment. In static beamforming, the beamformers are hence not dynamically adapted to the observed EEG, but are optimized based on some imprint of brain activity as measured by the EEG covariance matrix of previously recorded data. In the second procedure, termed block-adaptive beamforming, the optimal beamformers are re-computed for each trial of EEG data, resulting in a time-discrete adaptation of the beamformers to the observed EEG data. This section concludes with the presentation of experimental results from a BCI with real-time feedback based on the static beamforming approach.

6.3.1 Offline Results

Experimental Setup

Ten healthy subjects (S1-S10) participated in the experimental evaluation. Of these two were female, eight were right handed, and their average age was 25.6 years with a standard deviation of 2.5 years. Subject S3 had already participated twice in a BCI experiment, while all other subjects were naive to BCIs.
Each subject was seated in a dimly lit and shielded room, approximately two meters in front of a silver screen. Each trial started with the central display of a white fixation cross. After three seconds, a white arrow was superimposed on the fixation cross, either pointing to the left or the right. Subjects were instructed to perform haptic motor imagery of the left or the right hand, as indicated by the direction of the arrow, while seeing the arrow. The conditions motor imagery of the left and right hand are subsequently referred to as conditions c_1 and c_2. After another seven seconds the arrow was again removed, indicating the end of the trial. While subjects were explicitly instructed to perform haptic motor imagery with the specified hand, the exact choice of which type of imaginary movement, i.e., moving their fingers up and down, gripping an object, etc., was left unspecified. A total of 150 trials per condition were carried out by each subject, with the trials presented in pseudo-randomized order.
During the experiment, EEG was recorded at $M = 128$ electrodes placed according to the extended 10-20 system. Data was recorded at 500 Hz with electrode Cz as reference. Four BrainAmp amplifier were used for this purpose, using a temporal analog high-pass filter with a time constant of 10 s. The data was re-referenced to common average reference offline. Electrode impedances were below 10 kHz for all electrodes and subjects. No trials were rejected and no artifact correction was

performed. For each subject, the locations of the 128 electrodes were measured in three dimensions using a Zebris ultrasound tracking system and stored for further offline analysis.

Common Spatial Patterns

To evaluate the classification accuracy using CSP for feature extraction the following procedure is adopted. First, the EEG data is filtered with a sixth-order butterworth filter with cut-off frequencies 7 and 30 Hz, since this is known to improve the quality of the obtained CSP filters [BDK$^+$07]. Then, the data is randomly partitioned into a training and test data set. While the number of trials included in the training set is systematically varied, always the same number of trials per condition are selected. The EEG covariance matrices of conditions c_1 and c_2 are then estimated according to (6.12), only using data from the training set and the last 6.5 s of each trial. This is done to ensure that visual evoked responses due to presentation of the arrow at 3 s have already decayed. CSPs are then computed as described in Section 4.2.1. The $L = 5$ spatial filters with maximum and minimum eigenvalues are used to form the spatial filtering matrix $W \in \mathbb{R}^{M \times 2L}$.
This matrix is then applied to each trial in the training and the test data set, resulting in a reduced data space of $2L$ EEG components for each recorded trial. For each trial and extracted EEG component, $K = 20$ frequency bands of 2 Hz width, ranging from 2 - 40 Hz, are extracted using a sixth-order butterworth filter. For each trial, the sample variance in the time window ranging from 3.5 - 10 s in each frequency band for all components then forms the 200-dimensional feature vector. The feature vectors of the trials included in the training set are then used to train a logistic regression classifier with L_1-regularization, with the regularization parameter tuned heuristically to 0.1. This linear classifier is chosen for two reasons. First, it is well known that considering non-linear classifiers does not significantly improve classification accuracy in non-invasive BCIs while needlessly increasing complexity [GPAT03, MAB03]. Second, it is also known that only some frequency bands provide information on the user's intention in motor imagery paradigms, and that these frequency bands vary across subjects [PL99]. It can thus be expected that most features of the 200-dimensional feature vector are irrelevant, but it is unknown which ones are relevant for a certain subject. For this class of classification problems, i.e., a high-dimensional feature space with many irrelevant features, it is proved in [Ng04] that logistic regression with L_1-regularization possesses a sample complexity that only grows logarithmically in the number of irrelevant features, while rotationally invariant classifiers, such as support vector machines, have a worst case sample complexity that grows linearly in the number of irrelevant features. Hence, for this class of problems logistic regression with L_1-regularization can be expected to display a faster convergence (in terms of the required amount of training data) to its minimum error than other state-of-the-art classification algorithms. The classifier trained on the training set is then used to infer the BCI-user's intention for the

trials in the test set. The number of training trials per condition are systematically varied between 10 and 100 trials, and the above evaluation procedure is carried out 10 times for each subject and amount of training data to obtain sensible estimates of the mean and standard deviation of the classification accuracy for each subject as a function of the number of training trials.

Static Beamforming

To evaluate feature extraction by static beamforming, the recorded EEG data is again first randomly partitioned into a training and a test data set. The EEG data in the time window ranging from 3.5 - 10 s of all trials in the training data set is then used to estimate the EEG covariance matrix according to (6.12). Note that the EEG covariance matrix is not estimated separately for each condition. Instead, trials from both conditions are combined to obtain an imprint of subject specific EEG patterns as manifested in the EEG covariance matrix. Then, two beamformers are computed, with their respective ROIs chosen as spheres of 1 cm radius centered 1.9 cm radially below electrodes C3 and C4. Electrodes C3 and C4 are chosen due to their location over the left and right motor cortex according to the 10-20 system for electrode placement. The leadfield matrices L for each ROI, required in (6.11), is computed by placing a radially oriented current dipole at every position of an equally spaced grid with 2 mm grid point distance within each ROI, and computing the contribution of each current dipole to the electric potential at the M electrodes on the scalp according to the four-shell spherical head model (cf. Section 3.2.2 and [BML01]). For each subject, the employed electrode positions are obtained by radially projecting the measured electrode positions onto the outermost sphere of the four-shell spherical head model. The beamformers are then finally obtained by computing the eigenvector with the largest eigenvalue of (6.11) for each of the two leadfield matrices, assuming a unit source covariance matrix $R_{\boldsymbol{p}}(t)$. This results in a spatial filtering matrix $W \in \mathbb{R}^{M\times 2}$. The actual computation of the feature vectors and evaluation of the classification accuracy is then carried out as for feature extraction by CSP. Note, however, that since only two spatial filters are employed the feature vector for static beamforming is of dimension 40.

Block-adaptive Beamforming

For evaluation of block-adaptive beamforming, the same procedure as for static beamforming is adopted. However, instead of computing the EEG covariance matrix from the training data, an EEG covariance matrix is computed for every single trial. These covariance matrices are then used to compute two beamformers for every trial according to (6.11), resulting in a time-discrete adaptation of the beamformers to actually observed data. The computation of the feature vectors and estimation of the classification accuracies is then again performed as for feature extraction by CSP and static beamforming.

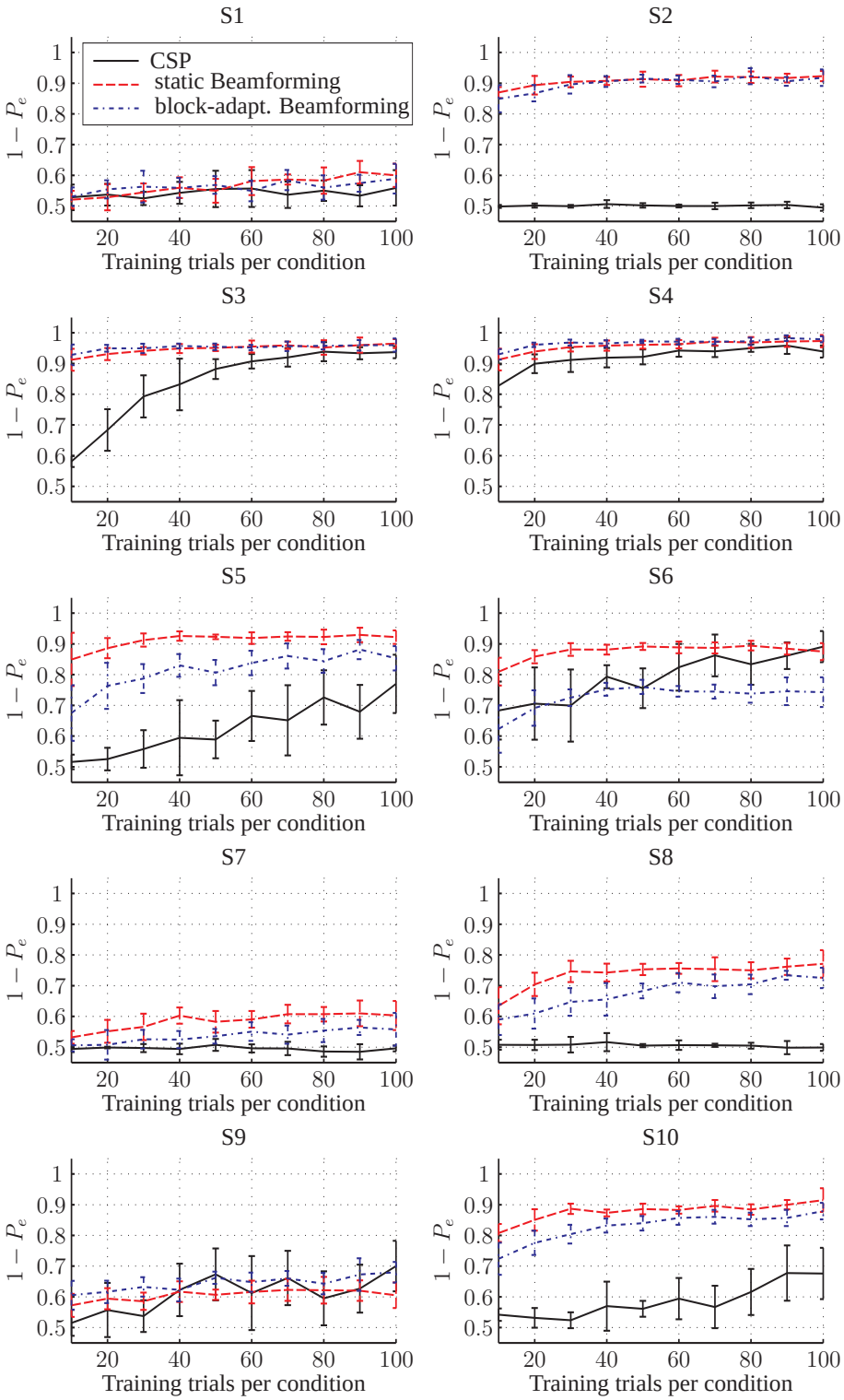

Figure 6.1: Estimated mean and standard deviation of classification accuracies for subjects S1-S10 as a function of the number of training trials.

Trial #	10	20	30	40	50	60	70	80	90	100
S1 (CSP)	52.8	53.7	52.5	54.3	55.6	55.7	53.7	55.0	53.3	55.9
S1 (SB)	51.1	52.9	54.5	56.0	55.0	58.1	58.6	58.3	61.0	60.0
S1 (BB)	53.0	55.4	56.3	55.9	56.9	54.9	58.4	56.0	57.4	58.9
S2 (CSP)	49.9	50.2	49.9	50.6	50.2	50.0	50.0	50.2	50.3	49.5
S2 (SB)	87.0	89.4	90.4	90.7	91.3	90.8	92.1	91.9	91.7	92.3
S2 (BB)	84.9	86.7	89.6	90.5	91.5	91.1	90.6	92.3	90.6	91.8
S3 (CSP)	58.1	68.3	79.3	83.2	88.2	90.7	92.0	93.9	93.3	93.7
S3 (SB)	91.2	93.1	94.1	94.9	95.1	95.5	95.9	95.2	96.0	96.5
S3 (BB)	92.8	95.0	94.9	95.7	95.5	95.2	95.6	95.8	95.9	96.0
S4 (CSP)	82.8	89.9	91.2	91.9	92.1	94.2	93.9	95.0	95.8	93.9
S4 (SB)	91.3	93.9	95.3	95.8	96.1	96.3	97.1	96.9	97.2	97.3
S4 (BB)	93.0	96.1	96.8	96.5	97.2	97.1	97.1	97.2	98.3	97.9
S5 (CSP)	51.6	52.5	55.8	59.5	58.9	66.6	65.1	72.6	67.9	77.0
S5 (SB)	84.9	88.6	91.3	92.6	92.3	91.9	92.4	92.3	92.9	92.2
S5 (BB)	67.5	76.3	78.7	83.0	80.7	83.8	86.1	84.4	88.2	85.3
S6 (CSP)	68.3	70.6	69.9	79.3	75.6	82.4	86.3	83.4	86.2	89.1
S6 (SB)	81.0	85.9	88.2	88.1	89.2	88.8	88.7	89.4	88.4	87.5
S6 (BB)	62.3	69.2	72.5	75.2	76.0	74.6	74.5	73.8	74.6	74.3
S7 (CSP)	49.4	49.9	49.7	49.4	50.8	49.6	49.6	48.6	48.5	49.6
S7 (SB)	53.2	55.2	56.6	60.3	58.3	59.1	60.8	60.7	61.0	60.3
S7 (BB)	50.6	50.8	52.6	52.5	53.6	55.1	54.1	55.4	56.4	55.8
S8 (CSP)	50.8	50.7	50.8	51.6	50.5	50.7	50.6	50.5	49.8	49.9
S8 (SB)	63.5	70.4	74.6	74.3	75.3	75.6	75.4	75.0	76.2	77.1
S8 (BB)	59.0	60.9	64.7	65.5	68.3	71.0	69.8	70.4	73.4	72.5
S9 (CSP)	51.5	55.7	53.7	62.2	67.3	61.2	66.1	59.5	62.7	70.0
S9 (SB)	57.2	59.4	58.5	61.7	60.7	61.6	62.3	62.1	62.0	60.5
S9 (BB)	60.5	61.6	63.3	62.2	66.2	64.8	66.0	64.3	67.3	68.0
S10 (CSP)	54.2	53.2	52.4	57.0	56.1	59.4	56.7	61.6	67.8	67.6
S10 (SB)	80.8	85.0	88.7	87.3	88.6	88.3	89.6	88.4	90.0	91.5
S10 (BB)	72.4	77.6	80.4	83.2	83.9	85.7	86.1	85.2	85.7	87.9

Table 6.1: Mean classification results in percent as a function of the number of training trials per condition for feature extraction by CSP, static beamforming (SB), and block-adaptive beamforming (BB).

Results

The resulting classification accuracies of all subjects and evaluation methods are shown in Fig. 6.1 and Tab. 6.1. As can be seen from the figure, the mean and the standard deviation of the classification accuracies vary significantly across subjects, number of training trials, and algorithm used for feature extraction. It should be pointed out again that the evaluation procedures differ only in the choice of the algorithm used for spatial filtering, and not in the classification procedure itself. Any differences in the classification accuracy for a subject can thus solely be attributed to the different algorithms used for extracting relevant EEG components.

Classification results obtained with the CSP algorithm vary significantly across subjects. While three subjects (S3, S4, and S6) achieve an accuracy close to or even above 90%, classification accuracy is not (or only hardly) above chance for four other subjects (S1, S2, S7, and S8). As a result, the mean classification accuracy obtained with the CSP algorithm if averaged across all subjects and number of training trials equals only 64.2%. In comparison to CSP, the beamforming approaches display a considerable higher mean classification accuracy if averaged across all subjects and number of training trials of 79.2% for static- and 76.1% for block-adaptive beamforming. In fact, static beamforming achieves classification accuracies above 90% for five out of ten subjects, with only two subjects displaying accuracies not significantly above 60%. Notably, there are two subjects (S2 and S8) for which CSP does not perform above chance, while both beamforming approaches display classification accuracies of above 90% and above 70%, respectively. In summary, the best mean classification results are observed for static beamforming, outperforming block-adaptive beamforming by 3.6% and CSP by 15.5%.

It should be pointed out again that these rather low mean classification results are due to computing classification accuracies across different amounts of training data, with few training trials naturally resulting in low classification accuracies. The maximum classification results, usually obtained for the largest amount of training data, are significantly higher (cf. Tab. 6.1), with subjects S3 and S4 even achieving classification accuracies close to 100%. However, the quality of a feature extraction algorithm is determined not only by the maximum classification accuracy that is achieved, but also by the amount of training data required to achieve a desired classification accuracy. For this reason, mean classification results taking into account different amounts of training data are considered more meaningful.

The above remark naturally leads to the question of the rate of convergence of the classification results to the maximum classification accuracy for a given feature extraction algorithm. Here, the CSP algorithm displays a rather low rate of convergence. Even though excellent maximum classification results are obtained using CSP for subjects S3 and S4, about 80 training trials are required until the classification accuracy approximately converges. This observation is even more pronounced for subjects S5, S6, S9, and S10, for which even 100 trials do not suffice for convergence. Considering that 100 trials per condition correspond to a training time

of over 30 minutes, this is a rather significant limitation of the CSP algorithm. In contrast, the beamforming approaches display a much higher rate of convergence. Using the static beamformer, a mean classification accuracy above 90% is obtained for subjects S3 and S4 using only ten training trials per condition, corresponding to a training time of less than three and a half minutes. While not all subjects display such a fast rate of convergence, it is nevertheless evident in Fig. 6.1 that the beamforming approaches require much less trials to converge to their maximum classification accuracy than the CSP algorithm.

Another important issue in the evaluation of feature extraction methods is the standard deviation of the obtained classification results for a given amount of training data, i.e., how much the classification accuracy varies for different sets of training data of equal size. In general, it is desirable to have a low standard deviation to increase the probability that for a given amount of training data the resulting mean classification accuracy is close to the expected one. As can be seen in Fig. 6.1, the standard deviation is rather large for the CSP algorithm, with a mean standard deviation across all subjects and amounts of training data of 6.5%. The beamforming approaches, on the other hand, result in a standard deviation of only 3.5% (block-adaptive beamforming) and 3.0% (static beamforming), i.e., roughly half the standard deviation of the CSP algorithm.

In summary, the proposed beamforming approaches outperform the CSP algorithm considerably in terms of mean classification accuracy, rate of convergence, and standard deviation of classification accuracy for a given amount of training data.

6.3.2 Online Results

To establish the feasibility of beamforming for BCIs with real-time feedback, the experimental setup of Section 6.3.1 is adapted in the following way. First, a certain number of training trials are recorded with an equal number of trials per condition presented in pseudo-randomized order. This training data set is then used to compute two static beamformers and train a logistic regression classifier with L_1-regularization. Up to this point, the experimental setup, the computation of the static beamformers, and the training of the classifier are (including all parameters) identical to the procedures in Section 6.3.1. After training, however, real-time feedback is provided to the BCI-user. To achieve this, the following procedure is implemented in Matlab/Simulink. First, the recorded EEG data is sent via TCP/IP to Matlab/Simulink running at 500 Hz. The two static beamformers are then applied to every new data sample, and the resulting two extracted EEG components are band-pass filtered with a sixth-order butterworth filter in 20 frequency bands of 2 Hz width ranging from 2 to 40 Hz. The variances of the temporally and spatially filtered time series are then calculated recursively for every sample step according to

$$\mathrm{Var}(y_i)[t+1] = \frac{t-1}{t}\mathrm{Var}(y_i)[t] + \frac{1}{t-1}y_i[t+1]^2, \tag{6.18}$$

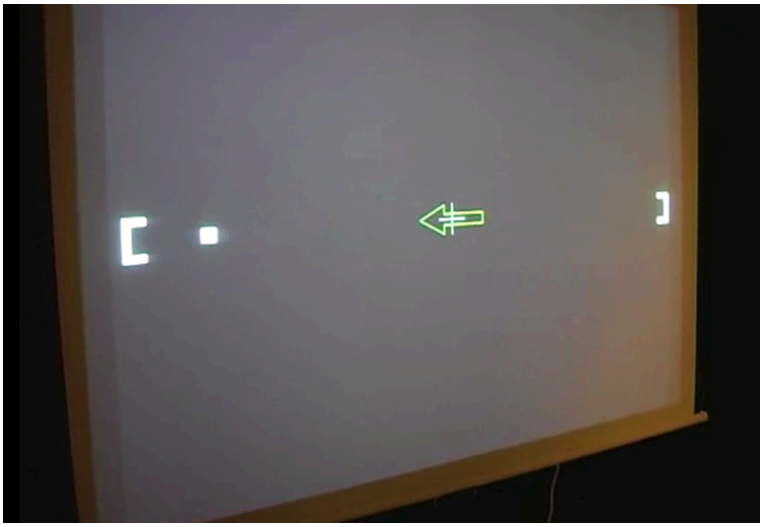

Figure 6.2: Setup of the feedback experiment.

with $i = 1, \ldots, 40$, $t = t_0, \ldots, T$ (with t_0 designating the time of display of the arrow indicating the class of the trial and T the length of each trial), and $\mathrm{Var}(y_i)[t_0] = 0$. Note that due to the band-pass filtering the elements of $\boldsymbol{y}$ have approximately zero mean, which is hence neglected in (6.18). The estimates of the variances of the extracted EEG components in the different frequency bands are then fed into the previously trained logistic regression classifier. The output of the classifier at each sample point, ranging from zero to one, is then fed back to the subject by drawing a white filled square on the screen. The output of the classifier is linearly mapped to the horizontal position of the square, with an output of zero mapped to the left border and an output of one mapped to the right border of the screen. The horizontal position of the square thus informs the BCI-user of the certainty of the classifier about his intention (with the left border of the screen indicating 100% certainty of an imaginary movement of the left hand and the right border of the screen indicating 100% certainty of an imaginary movement of the right hand). To further motivate the subject, two white boxes are drawn at the left and right borders of the screen into which the subject has to move the white square. Also, the color of the centrally displayed arrow is set to green or red, depending on whether the output of the classifier leads to a correct decision or an error. The complete setup is shown in Fig. 6.2. Each trial ends after a preset time, or if a certain threshold of certainty of the classifier is achieved. Note that the threshold criterion is only checked after a certain minimum time into each trial to ensure sensible estimates of the variances of the EEG components, and that each trial begins with a pause of 3 s.

Due to the excellent performance in the offline experiment, subject S4 was asked to perform again in the online experiment. Twenty-five trials per condition were recorded as training data, corresponding to a training time of eight minutes and

Block #	Min trial length	Max trial length	Thresholds	$(1 - P_e)$
1	9.99 s	10 s	[0.1 0.9]	92.5%
2	9.99 s	10 s	[0.1 0.9]	87.5%
3	6 s	10 s	[0.1 0.9]	87.5%
4	6 s	30 s	[0.1 0.9]	92.5%
5	6 s	30 s	[0.1 0.9]	90.0%

Table 6.2: Results of the online experiment for subject S4.

twenty seconds. Then five blocks of twenty trials per condition were carried out with feedback provided, with a break of approximately two minutes between each block. The obtained classification results are shown in Tab. 6.2, along with the minimum and maximum trial lengths and the thresholds for termination of a trial. The mean classification accuracy across all blocks was 90.0%, which is in accord with the classification accuracy obtained by subject S4 in the offline experiment using the static beamforming approach (cf. Tab. 6.1). A video recording of this experiment can be downloaded at *http://www.lsr.ei.tum.de/fileadmin/multimedia/videos/TUM_BCI.avi*.

6.4 Discussion

6.4.1 Comparison of CSP and Beamforming

In Section 6.3.1, it has been shown that beamforming enables a higher mean classification accuracy, a higher rate of convergence, and a lower standard deviation of the classification accuracy than the CSP approach. This raises the question why the CSP algorithm performs so poorly in this study in spite of its popularity within the BCI community. The mediocre performance of the CSP algorithm can be attributed primarily to the choice of the eigenvectors of (4.2) used as spatial filters. According to (4.3), the eigenvectors with the smallest and largest eigenvalue of (4.2) correspond to the spatial filters that maximize the ratio of the class-conditional variances, and are thus optimal in terms of maximizing an approximation of mutual information of class labels and extracted EEG components (cf. Section 4.2.3). However, the variance of artifactual components frequently present in EEG data usually exceeds the variance of endogenous EEG components. If a certain artifact, e.g., an eye blink, is only present in the training data of one class, then the CSP algorithm focuses on optimally extracting the artifactual EEG component. Since this component is unrelated to the actual motor imagery, this results in a poor classification accuracy. This overfitting phenomenon is illustrated in Fig. 6.3, showing ten typical spatial filters with maximum/minimum eigenvalues as obtained by CSP for subject S2 using 20 trials of each condition for training. Subject S2 is chosen for this purpose since the recorded EEG data is very noisy, but the subject is capable of operating the BCI

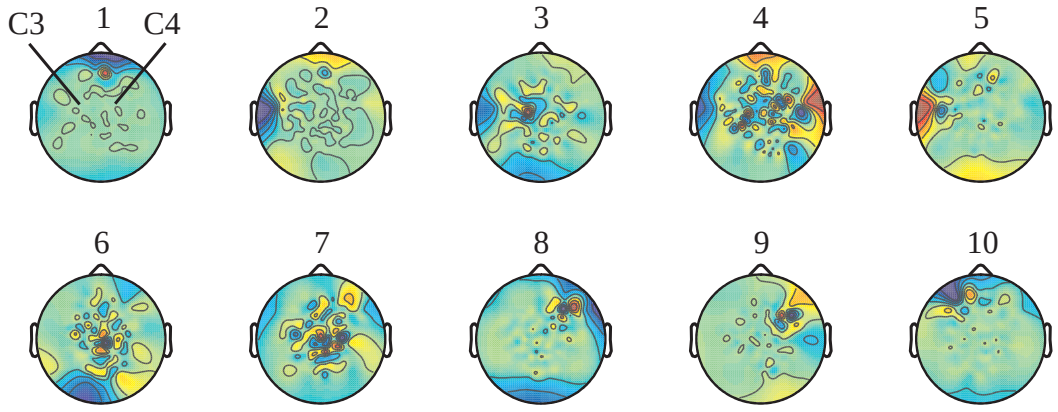

Figure 6.3: Typical spatial filters obtained by CSP for subject S2.

as indicated by mean classification accuracies of above 90% for feature extraction by block-adaptive beamforming. As can be seen in Fig. 6.3, of all spatial filters obtained by CSP only the third one focuses on the vicinity of the left motor cortex (electrode C3), albeit some activity in the left temporal lobe is also picked up. All other filters focus on artifactual components not related to EEG signals originating in the motor cortex. Consequently, the spatial filters do not pick up those components that provide information on the user's intention, resulting in a classification accuracy not above chance. This is in contrast to subject S4, for which ten typical spatial filters, obtained by CSP using 20 trials of each condition for training, are shown in Fig. 6.4. Here, all spatial filters except the second, fourth and fifth one focus on areas over the left and right motor cortex (electrodes C3 and C4). As a result, the spatial filters extract EEG components that are related to motor imagery, and provide sufficient information on the user's intention to achieve a mean classification accuracy of about 90%. For comparison, typical spatial filters obtained by block-adaptive beamforming for subjects S2 and S4 are shown in Fig. 6.5. Here, it is evident that, as expected, the beamformers focus on areas over the left and right motor cortex. Furthermore, for subject S4 the spatial filters obtained by beamforming resemble those obtained by CSP, indicating that both approaches extract similar EEG components if applied to data sets with few artifactual components. The results of subject S2 demonstrate that CSP breaks down for noisy data with many artifactual components, while the beamforming approach still extracts meaningful components.

In principle, there are three ways to alleviate overfitting phenomena observed when using CSP for feature extraction. The first is to increase the amount of training data. Since the probability that one special type of artifact is present in the training data of only one condition decreases with the amount of training data, overfitting phenomena are attenuated by increasing the number of trials in the training set. However, it is in general desirable to minimize the amount of training data to minimize training

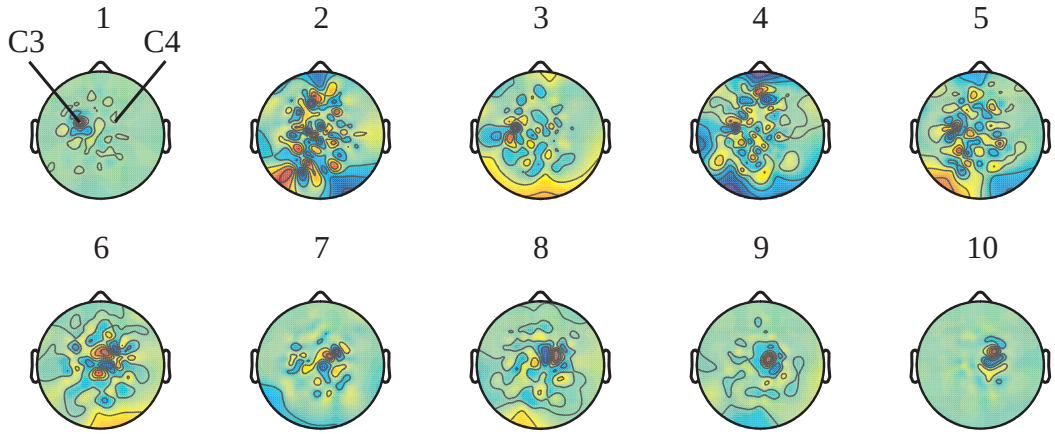

Figure 6.4: Typical spatial filters obtained by CSP for subject S4.

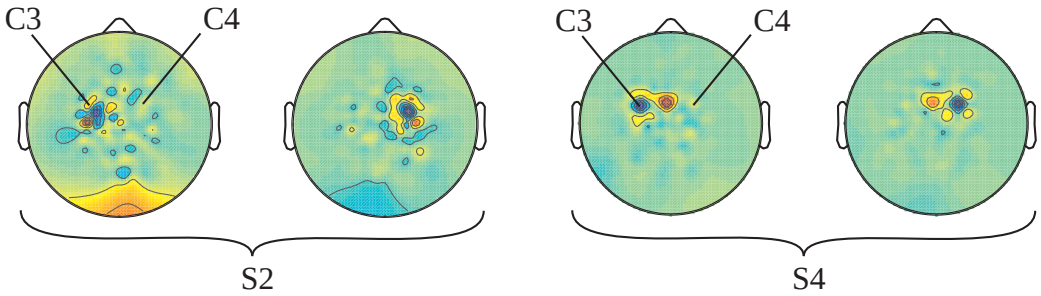

Figure 6.5: Typical spatial filters obtained by block-adaptive beamforming for subjects S2 and S4.

time for each subject. Another approach to alleviate overfitting phenomena is to increase the number of eigenvectors of (4.2) used as spatial filters. This increases the probability of including spatial filters that focus on motor areas and thus provide information on the user's intention. There is, however, an inherent trade-off between increasing the number of spatial filters and the rate of convergence of the subsequent classification algorithm. This is illustrated in Fig. 6.6, showing the mean and standard deviation of the classification accuracy of subject S3 for different numbers of spatial filters per condition as obtained by CSP. Using only one spatial filter per condition (Fig. 6.6.a) excellent classification results are obtained for 80 or more training trials per condition. However, if the number of trials used for training is decreased, the standard deviation of the classification accuracy increases. More specifically, if 50 or less trials per condition are used for training, which still corresponds to a training time of over 15 minutes, the standard deviation becomes so large that classification accuracies not above chance as well as close to 100% become rather likely. This undesired large dependence of the classification accuracy on the specific training set can be significantly reduced by increasing the number of spatial filters. Unfortunately, this also results in a slower rate of convergence of the mean classification accuracy, as can be seen in Fig. 6.6.e. Furthermore, this dependence of the classification results on the number of spatial filters varies across subjects. In this study, five spatial filters per condition have been chosen for each subject to achieve an acceptable trade-off between a fast rate of convergence and small overfitting effects. Due to the difficulties of choosing the correct spatial filters in order to alleviate overfitting phenomena, CSPs are frequently manually selected by an experienced researcher. This is the third approach to reducing overfitting phenomena. By only selecting spatial filters that focus on motor areas excellent classification results can be obtained, and the effects of overfitting can be significantly reduced. However, manual selection of spatial filters introduces subjectivity into the analysis and thus prevents an objective evaluation of the power of different feature extraction algorithms. Furthermore, having to select spatial filters manually is clearly undesirable if BCIs are to be employed by subjects without expert supervision.

In summary, CSP is a feature extraction algorithm that enables excellent results if a large amount of training data is available and the recorded EEG does not contain many artifacts, or if it is feasible to have an experienced user manually selecting the spatial filters that provide most information on the BCI-user's intention by visual inspection. However, if expert supervision is undesirable, long training periods are unfeasible, or the recorded data is very noisy classification results obtained by using CSP for feature extraction are unsatisfactory. The beamforming approach, on the other hand, enables a high mean classification accuracy with low standard deviation and a high rate of convergence. Importantly, the problem of selecting a subset of optimal spatial filters, as it is necessary when using the CSP algorithm, is absent in the beamforming approach. This is due to the fact that the signal subspace of the data, i.e., the subspace of the recorded data providing information on the user's

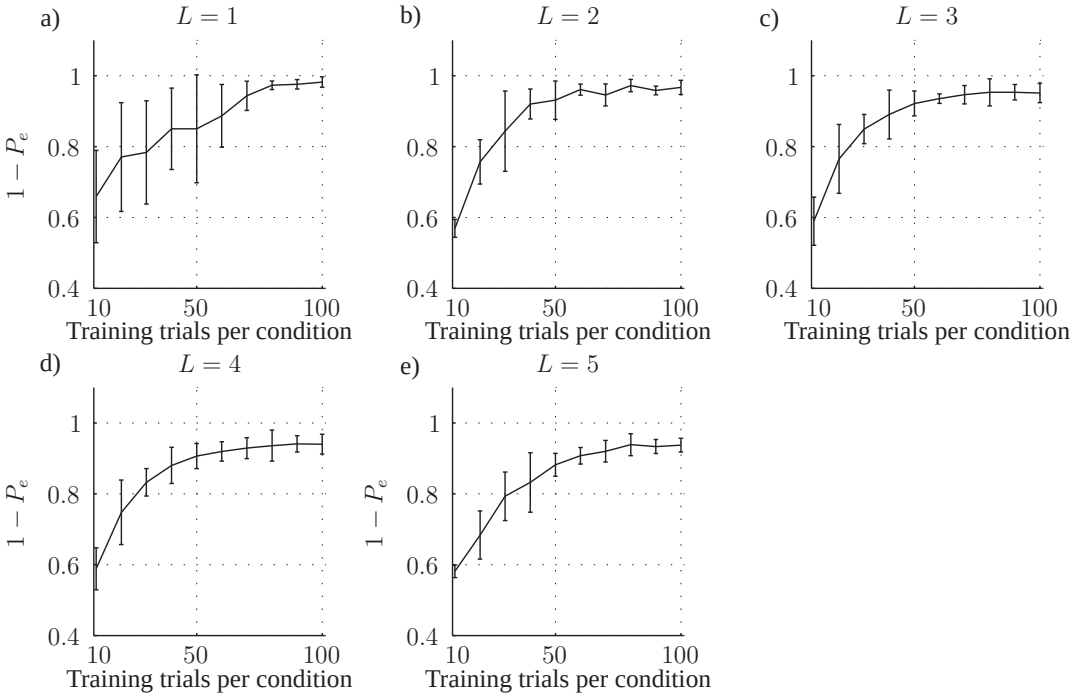

Figure 6.6: Classification accuracies (mean and standard deviation) of subject S3 for different numbers of CSPs.

intention, is already determined by the a-priori knowledge used in specifying the ROIs for a certain paradigm. Once the ROIs have been selected, the unsupervised adaptation of the beamformers only concerns the noise subspace, i.e., the space spanned by sources which are not to be included in the extracted EEG components (the denominator in (6.8)). As such, the beamforming approach does not suffer from overfitting phenomena. On the contrary, any artifacts present in the EEG data and not originating within the ROI can be optimally attenuated. The unsupervised nature of the beamforming approach also provides an explanation for the high rate of convergence and the small standard deviation of the classification results. The beamformers are essentially independent of the specific data in the training set, since no class-related information is utilized. Consequently, the rate of convergence and standard deviation of the classification results can be primarily attributed to the logistic regression classifier.

6.4.2 Beamformer Optimization

Evidently, the performance of the beamforming approach depends on the accuracy of the incorporated a-priori knowledge. This includes the choice of the ROIs, the model used for volume conduction, the orientation of current sources within a ROI, and the source covariance matrix of sources within a ROI. Considering this mul-

titude of parameters, the obtained classification results might seem surprising, especially since parameters have not been optimized for each subject but have been specified a-priori.

In terms of the chosen ROIs, it is indeed unlikely that spheres located radially below electrodes C3 and C4 are optimal in terms of the expected classification error. A marginal misspecification of the ROIs, on the other hand, is unlikely to result in a large decrease of the expected classification error. This is again due to the underdetermined nature of the inverse problem of EEG (cf. Section 6.2.1). Due to the impossibility of extracting only sources within the ROI, the spatial edges of the obtained beamformers are not sharp, i.e., attenuation of sources is low directly outside the ROI and increases with distance to the ROI. In practice, the beamformers can thus be expected to extract sources from a rather large region within the brain, alleviating adverse effects of misspecifying the ROIs. On the other hand, this effect naturally also leads to a lower SNR if the ROIs are correctly centered within the hand areas of the left and right motor cortex. In [LGWGB07], the optimal centers of the ROIs for the beamforming approach presented in this chapter are determined for four subjects using a source localization approach, and the resulting classification performances are compared with those obtained if the ROIs are chosen as in this study. For two subjects, optimizing the centers of the ROIs resulted in a mean increase in classification accuracy of 4.4%, with the optimal centers of the ROIs located on average 1.9 cm away from the positions chosen in this study. This indicates that a rather large misspecification of the location of the ROIs of almost 2 cm (taking into account that the human head has got a radius of only approximately 8.5 cm) results in only a moderate decrease in mean classification accuracy. For the other two subjects evaluated in [LGWGB07], classification accuracies obtained with optimized ROIs decreased. This illustrates another important issue. One of the primary advantages of the beamforming approach is its unsupervised nature, rendering it robust to artifactual components in the EEG data. Any optimization procedure carried out on the data dilapidates this advantage. It should be noted that the classification procedure in [LGWGB07] differs from the one employed here. Absolute classification accuracies can thus not be compared.

A further issue in the derivation of the beamforming approach is the head model used for computing the leadfield matrices. In this study, one of the most simplistic head models available in the literature has been employed. The beamforming approach can be easily combined with more complex head models (cf. [BML01] and Chapter 3) by altering the methodology for computing the leadfield matrices of the ROIs used in (6.11). This can be done without significantly increasing the computational complexity of the beamforming approach, since the leadfield matrices only have to be computed once for each subject and electrode configuration. An assessment of the effects of more realistic head models on the classification accuracy, however, is beyond the scope of this work.

Regarding the orientation of sources within a ROI, it might be expected that improved classification results can be obtained by specifying three-dimensional dipole

moments at every grid point within the ROI, since this allows extracting EEG sources from the ROI with arbitrary dipole orientation. This does not appear to be the case. Increasing the dimensionality of the dipole moments increases the rank of $R_{\tilde{\boldsymbol{x}}}$, the covariance matrix of sources within the ROI, and thus the rank of the signal subspace. As a direct result, less dimensions of the spatial filters are available for attenuating sources outside the ROI, leading to decreased classification accuracies. Best results were obtained in this study by only using radially oriented dipoles, which is attributed to the physiological accuracy of this assumption. Finally, the choice of the source covariance matrix $R_{\boldsymbol{p}}(t)$ is a parameter whose effect on the classification accuracy has to be investigated in future work.

6.4.3 Static- vs. Block-adaptive Beamforming

So far the discussion of the beamforming approach has neglected the differences between static- and block-adaptive beamforming. Interestingly, static beamforming outperforms block-adaptive beamforming in all but one subject (S9) in terms of mean classification accuracy. This is rather surprising, since it could have been expected that a trial-wise adaptation of beamformers to recorded data results in a higher SNR. The converse observation suggests, that in a trial wise adaptation of the beamformers the available data is not sufficient to obtain good estimates of the EEG covariance matrix. On the other hand, the classification accuracies obtained with static beamforming suggest that non-stationarities in EEG data do not prohibit excellent classification results. This is fortunate from a practical point of view, since static beamforming is computationally less intensive than block-adaptive beamforming, and can be applied directly in BCIs with real-time feedback as demonstrated in Section 6.3.2.

6.4.4 Source Localization and Beamforming

It could be argued that source localization methods, as discussed in Chapter 3, should enable identical classification results as the beamforming approaches presented in this chapter. If in source localization identical ROIs are chosen as in beamforming, the estimated EEG components should, at least in principle, provide identical information on the BCI-user's intention. This is indeed correct. However, the computational complexity of beamforming methods is significantly lower than that of most methods for source localization. Once the leadfield matrix for sources in the ROI has been computed, which only has to be done once for each subject and electrode configuration, the actual beamformer can be computed by solving a single generalized eigenvalue problem. In static beamforming, this eigenvalue problem has to be solved only once for every ROI, while in block-adaptive beamforming it has to be solved once for every ROI and block of EEG data. For every sample point, the desired EEG components can then be estimated by a simple linear transformation. Source localization methods, on the other hand, usually possess a

larger computational complexity [BML01]. This renders source localization methods unfeasible for BCIs with real-time feedback.

6.5 Summary and Outlook

In this chapter, it has been shown that beamforming provides a viable alternative to supervised spatial filtering in non-invasive BCIs. This holds true especially if the goal is to design a BCI that can operate without expert supervision and with little training data. For these specifications, the beamforming approach was shown to outperform the CSP algorithm in terms of mean classification accuracy, standard deviation of the classification accuracy, and rate of convergence of the classifier. Also, beamforming was shown to be feasible in BCIs with online feedback.
While feature extraction via beamforming is completely unsupervised, the classification procedure employed in this chapter still requires labeled training data. It should be pointed out that this is not necessary, and a completely unsupervised BCI can, at least in theory, be devised by resorting to clustering approaches in feature space. First results using block-adaptive beamforming with non-supervised classification procedures are reported in [EGWB07]. While in this paper the feasibility of a completely unsupervised BCI based on beamforming is established, the obtained classification results are not yet satisfactory and require further work.
In this work, only motor-imagery paradigms have been considered. It should be pointed out, however, that beamforming approaches can be applied to BCIs based on other experimental paradigms as well. This requires knowledge on the brain regions involved in a certain experimental paradigm. As discussed in Section 6.1, this is the case for motor imagery paradigms. Other paradigms might require source localization studies to identify relevant ROIs prior to utilizing beamforming approaches.
Finally, only two-class paradigms have been considered here. However, beamforming approaches can be extended in a straight-forward manner to multi-class paradigms. If motor imagery of further limbs, e.g., a foot and the tongue, are considered, new ROIs have to be specified for those parts of the motor cortex representing the specific limbs. It remains to be experimentally established if beamforming approaches also display the advantageous properties demonstrated in this chapter if they are applied to multi-class paradigms, with ROIs possibly buried deeper within the cortex.

Chapter 7

Conclusions and Open Problems

In this chapter, the most important contributions of this thesis are summarized and critically evaluated (Section 7.1). It is discussed why, in spite of the progress already made, in non-invasive BCIs inferring the user's intention still is a hard task. Limitations of current methods for feature extraction in non-invasive BCIs are discussed, and possible future research directions are delineated. More specifically, in Section 7.2 it is argued that it is necessary to acknowledge the fact that the brain forms a complex network with time-varying functional connectivity patterns in order to significantly enhance the capabilities of future non-invasive BCIs. In Section 7.3, a possible approach to this problem is outlined. Finally, this thesis concludes in Section 7.4 with a few comments on the possibly underestimated significance of the electric/magnetic field for information processing within the human brain.

7.1 Summary

The motivation for the work presented in this thesis is the conviction that the lack of sophisticated feature extraction methods constitutes the main performance bottleneck of non-invasive BCIs. This was explicated in Section 2.4, in which it was argued that the high dimensionality of the feature space in non-invasive BCIs prohibits training any type of classifier directly on the original features, i.e., without a prior dimensionality reduction. Further, it was shown that the class of possible feature spaces in non-invasive BCIs is so large that the application of any automated algorithm for dimensionality reduction is not feasible. Taken together, it was argued that this implies that a-priori knowledge on how cognitive states are encoded in signals recorded from the CNS has to be incorporated into the process of feature extraction in order to restrict the class of allowed feature spaces in a sensible way. This resulted in Definition 2.13, summarizing the main topic of this thesis.
As further discussed in Section 2.4, for recording modalities considered in this thesis the high dimensionality of the original feature space is determined by two factors: the large number of EEG/MEG electrodes, used to sample the electric/magnetic

field on the scalp, and the duration and sampling rate of the recordings. Accordingly, dimensionality reduction can be achieved by incorporating a-priori knowledge on coding of cognitive states that a) acts on the spatial domain by fusing the data recorded at multiple channels, b) acts on the temporal domain by fusing multiple observations recorded at the same electrode, or c) acts simultaneously on both domains.

In Chapter 3, it was argued that the activity of brain regions during a certain cognitive task provides information on the BCI-user's intention and can thus be used as a feature space in BCIs. It was further argued that the primary advantage of this source localization approach to feature extraction is that no information on how cognitive states are temporally encoded in the EEG/MEG recordings is required, thereby bypassing the largely unsolved problem of temporal coding of cognitive states in the electric/magnetic field of the brain. This approach, which only incorporates a-priori knowledge acting on the spatial domain, was realized by combining ICA with source localization in a four-shell spherical head model, and developing a procedure to identify and exclude EEG/MEG sources representing (Gaussian) noise. While the viability of this procedure for feature extraction in non-invasive BCIs could be established in a preliminary study based on a two-class motor imagery paradigm, the reported classification result did not compare favorably with a recent study combining source localization with temporal a-priori knowledge on coding of cognitive states for feature extraction in BCIs [GGP+05]. This led to the conclusion that while the incorporation of spatial a-priori knowledge only does indeed constitute a viable option, considering all spatial and temporal a-priori knowledge available on coding of cognitive states allows constructing superior feature extraction algorithms for BCIs.

This conclusion was further pursued in Chapter 4. In this chapter, the CSP algorithm (initially proposed in [RMGP00]) for feature extraction in BCIs was investigated theoretically for two-class as well as multi-class paradigms. By making use of a-priori knowledge available on temporal coding of cognitive states in the electric/magnetic field of the brain, the CSP algorithm computes spatial filters that aim to optimally extract those components of the EEG/MEG providing most information on the BCI-user's intention. The CSP algorithm thus utilizes temporal a-priori knowledge to achieve a dimensionality reduction acting on the spatial domain. However, while excellent classification results have been reported using the CSP algorithm for feature extraction, its optimality in terms of the minimum Bayes error (as discussed in Section 2.2) remained unsolved. Here, it could be shown in the framework of information theoretic feature extraction that the two-class CSP algorithm is optimal in terms of maximizing (an approximation of) mutual information of class labels and extracted EEG/MEG components. This provided a previously unknown link between the CSP algorithm and the minimum Bayes error. Note that while optimality in terms of maximizing mutual information is highly desirable (cf. the discussion in Section 2.2), it rules out optimality in terms of the minimum Bayes error. The extension of CSP to multi-class paradigms proposed in

[DBCM04], on the other hand, was shown to be suboptimal in terms of maximizing mutual information of extracted EEG/MEG components and class labels. This deficiency could be resolved by proving that computation of potential spatial filters by multi-class CSP is equivalent to ICA, and using the framework of information theoretic feature extraction for identifying those ICs providing most information on the user's intention. This algorithm, termed multi-class information theoretic feature extraction, was then shown to outperform multi-class CSP in a four-class motor imagery paradigm by on average 23.4%.

Motivated by the success of ICA in Chapters 3 and 4, Chapter 5 was devoted to an investigation of complete ICA in the context of EEG/MEG analysis. The goal of this analysis was to provide a theoretically and experimentally founded explanation for the apparent success of complete ICA in EEG/MEG analysis in spite of the physiologically unrealistic assumption of at most as many sources as sensors (as required by complete ICA). This was approached by theoretically investigating the behavior of complete ICA, i.e., ICA designed for an equal number of sensors and sources, in the context of overcomplete mixture models, i.e., for models with more sources than sensors. A general theorem (Theorem 5.1) could be proved, establishing necessary and sufficient conditions for solutions of complete ICA for arbitrary mixture models. This theorem was then used to argue that complete ICA performs well in EEG/MEG analysis not due to the fact that only a few EEG/MEG sources are strong enough (cf. [OWTM06]), but rather because only a few sources are non-Gaussian enough to be picked up by ICA. Testable predictions were formulated for this hypothesis and experimentally validated. In summary, an explanation for the success of complete ICA in EEG/MEG analysis (including feature extraction for BCIs) could be provided that dissolves the apparent contradiction between the requirement of at most as many sources as sensors and the physiological doubtfulness of this assumption.

In Chapters 4 and 5, only supervised feature extraction algorithms were considered, i.e., algorithms that require labeled training data. While algorithms that are theoretically optimal in terms of maximizing (an approximation of) mutual information of extracted features and class labels could be provided for two-class as well as multi-class paradigms in Chapter 4, these algorithms often perform poorly if only noisy training data is available. In Chapter 6, it was argued that this practical limitation of supervised feature extraction algorithms is caused by overfitting phenomena. To obtain a more robust feature extraction algorithm, that can also be applied to noisy EEG/MEG recordings, a spatial filtering approach incorporating a-priori information on the spatial position of relevant brain regions was designed. This algorithm, closely related to traditional beamforming methods, allows extracting EEG/MEG sources from pre-defined regions within the brain while optimally (in terms of the SNR) attenuating all sources outside these regions. In spite of the manifold and possibly inaccurate a-priori information incorporated in this feature extraction method, it could be shown that in a two-class motor imagery paradigm the proposed beamforming approach outperforms the CSP algorithm in terms of

classification accuracy and rate of convergence of the subsequent classifier. Indeed, classification accuracies above 96% could be obtained with a training time of less than seven minutes. This success was primarily attributed to the unsupervised nature of the beamforming approach, rendering it robust towards artifacts commonly encountered in EEG/MEG data. Finally, a BCI, based on motor imagery and the proposed beamforming approach, was realized, enabling online control of a cursor in one dimension.

In summary, in this thesis three new algorithms for feature extraction in non-invasive BCIs could be presented and experimentally validated. It could be shown that the proposed beamforming approach outperforms the CSP algorithm, which is one of the most powerful feature extraction algorithms for two-class paradigms. In the context of multi-class paradigms, the proposed algorithm, termed multi-class Information Theoretic Feature Extraction, was shown to outperform multi-class CSP, thereby contributing to the development multi-class BCIs with high classification accuracies. Furthermore, a framework for investigating the optimality of two-class CSP was presented, and an explanation for the success of complete ICA in EEG/MEG analysis could be provided.

7.2 Open Problems

In spite of this progress, inferring a BCI-user's intention still is a hard task. While in two-class paradigms classification accuracies close to 100% can be achieved, so far accurate classification has not been demonstrated for more than four classes. Carrying out more complex tasks by non-invasive BCIs, such as online control of an endeffector in multiple dimensions, hence still represents a long term rather than a short term goal. This raises the question of the causes of this limitation of current non-invasive BCIs. In general, it can not be ruled out that the electric/magnetic field of the brain does not provide full information on the user's intention, i.e., that (at least for paradigms with multiple classes) $I(\boldsymbol{x}, c) < H(c)$. However, it is the conviction of this author that the significance of the electric/magnetic field of the brain is generally underestimated, and that what is required in order to realize powerful non-invasive BCIs is a better understanding of how cognitive states are encoded in the electric/magnetic field of the brain.

As pointed out at the beginning of this chapter, feature extraction algorithms can act on the spatial as well as on the temporal domain of EEG/MEG recordings. Analyzing the feature extraction algorithms covered in this thesis, it is noteworthy that they all focus on the spatial domain. More specifically, all feature extraction algorithms designed in this thesis aim to extract EEG/MEG components from those regions of the brain most relevant for inferring the BCI-user's intention. While it has been shown that this is a viable approach, it is important to realize the inherent restrictions.

First, the only a-priori information on temporal coding of cognitive states utilized

in this thesis is that variance changes in specific frequency bands provide useful information. This restriction reflects the limited understanding of and the established assumptions on coding of cognitive states in EEG/MEG recordings. In general, it is an entirely open question whether measures other than variance changes, such as higher-order moments, do provide more information on cognitive states. Quite surprisingly, this question is hardly addressed in neuro-psychological research. This can be attributed to the inherent theoretical difficulty of developing signal processing methods departing from the assumption of Gaussianity, which hence have found very little dissemination in the neuro-scientific community. This point again emphasizes the importance of interdisciplinary research in this field.

Second, all approaches presented in this thesis can be classified as *localized*, i.e., inferences are made from intentionally induced pattern changes of signals originating in individual brain regions. However, neurons within the brain form complex networks with time-varying functional connectivity patterns [LS03, von99]. Consequently, the assumption of localized information processing, implicit to all feature extraction algorithms investigated in this thesis, might be too constrictive. Delocalized approaches would take this into account, making inferences from class-conditional functional connectivity patterns between brain regions. However, as pointed out in [DCF04], uncovering functional connectivity patterns from experimental data is a challenging problem in itself. Algorithms developed for this purpose are either based on linear models (reviewed in [ACM+07]), non-linear measures such as mutual information (reviewed in [DCF04]), or rather simple measures such as phase synchronization [RPK96, RP01] and amplitude coupling.

Until now, only phase synchronization and amplitude coupling have been employed as feature spaces in non-invasive BCIs [GC04, WWGG07]. This restriction can be primarily attributed to the computational complexity of other approaches and the rather large amount of training data required by these algorithms. Both studies report comparable classification rates for using phase synchronization and variance based measures as features, thereby establishing the viability of measures of functional connectivity for feature extraction in BCIs. Interestingly, both studies also report enhanced classification accuracies for combining connectivity- and variance based measures, indicating a complementarity of both domains.

While the increase in classification accuracy reported in [GC04] and [WWGG07] is rather small, it is nevertheless very promising considering the dimensionality of the employed feature space. In both studies, functional connectivity measures are computed for recordings obtained from different electrodes. Now note that even for a modest number of electrodes, say $M = 64$, the number of possible connectivity measures (even when neglecting directionality) already sums to $\sum_{i=1}^{M-1} i = 2016$. Since training a classifier on a feature space of this dimension requires a substantial amount of training data, in [GC04] and [WWGG07] only a small subset of electrodes is considered. In both studies, the selection of this subset is based on rather limited prior knowledge on the involvement of miscellaneous brain regions in the respective experimental paradigms. Furthermore, note that these results are obtained

without spatial filtering. As has been demonstrated in this thesis, spatial filtering can improve classification accuracy in two-class motor imagery paradigms based on variance measures from about 70% to almost 100%. In contrast, it has been shown in [WWGG07] that even without spatial filtering functional connectivity measures enable classification accuracies comparable to those obtained with variance based measures in combination with spatial filtering. The two most promising approaches for enhancing feature extraction by means of functional connectivity measures thus appear to be improvement of the prior knowledge on functional connectivity within the brain and inclusion of spatial filtering.

7.3 Network Information Transfer Analysis

In principle, functional connectivity measures can easily be combined with beamforming [GHT+01] or source localization approaches [ACM+07], which could be used for feature extraction in BCIs. Note, however, that this requires (usually unavailable) a-priori knowledge on which brain regions display class-conditional functional connectivity changes. While a complete evaluation of connectivity patterns for a set of recordings from different electrodes already constitutes a formidable task, a complete evaluation of the interactions between all possible regions of the brain clearly is impractical. For this reason, all research on functional connectivity in EEG/MEG analysis is currently exploratory: a hypothesis is formulated, expressing expected functional connectivity patterns between certain brain regions (termed regions of interest - ROIs), source localization or beamforming is performed to extract EEG/MEG signals originating in the ROIs, and functional connectivity measures are computed to validate or falsify the proposed hypothesis.

It would clearly be desirable to develop a data driven approach for the analysis of functional connectivity within the human brain. Given a multi-variate time series, e.g., EEG/MEG recordings, and a (possibly linear) mixture model, the goal of such a procedure would be to estimate those EEG/MEG sources that display maximum functional connectivity (or maximum functional connectivity changes) during a certain cognitive task. Such an algorithm, termed Network Information Transfer Analysis (NITA), would be similar in spirit to ICA. However, instead of estimating statistically independent sources, the goal would be to uncover the dynamic network structure of information transfer within the brain.

While such an algorithm could be expected to enable significant progress in understanding how cognitive states are encoded in the electric/magnetic field of the brain, it is far from trivial to realize. One promising approach to this problem might be the concept of transfer entropy, initially proposed in [Sch00]. Here, information transfer between two random processes is defined as the reduction in entropy of one process due to knowledge of the other. In contrast to model-based connectivity measures (cf. [ACM+07]), this concept can be used to define a metric for information transfer within the human brain and is amenable to a data driven optimization pro-

cedure in the context of a generative mixing model. However, considerable research is required to establish the viability of this approach for the proposed NITA.

7.4 Causality of the EM Field of the Brain

In this thesis, the electric/magnetic (EM) field of the brain was successfully used to infer the BCI-user's intention. It is hence trivial to point out that the EM field of the brain does provide information on cognitive states. Due to the pervasiveness of EEG/MEG recordings in neuro-scientific research, this rather surprising fact is rarely scrutinized. Why does the brain create an EM field that does provide information on cognitive states? As pointed out in [WL96], the EM field of the brain is traditionally seen as an epiphenomenon, a byproduct of neural processes within the brain. This argument is challenged by several authors [WL96, NS05, Fre01], arguing for a causal role of the EM field for information processing within the brain. Indeed, in [NS05, Fre01] it is argued that the EM field of the brain is essential for consciousness, while in [McF02] it is even proposed that it is the physical substrate of conscious awareness.

While a detailed presentation and discussion of the arguments for and against a causal role of the EM field for information processing within the brain is beyond the scope of this work, it is important to point out that indeed there is some empirical evidence in favor of a causal role. As reviewed in [Jef95], it is known that externally applied electric fields with a smaller field strength than endogenous electric fields alter cortical activity. Furthermore, it is shown in [MHMB06] that applying weak external electric fields to the skull of human subjects during sleep can have significant positive effects on declarative memory. Leaving aside the philosophical issues regarding consciousness and the EM field of the brain, the available empirical evidence suggests that the relevance of the EM field of the brain is probably underestimated in current research.

In conclusion, future research on EEG/MEG should consider the possibility of a causal role of the EM field of the brain. However, it is questionable whether established methodologies for the analysis of EEG/MEG recordings are powerful enough to reveal a causal role of the fields generated by the brain. One possible strategy to prove a causal role of the EM field might be to combine measures of functional connectivity, as discussed in the previous section, with single-cell recordings of neuronal activity. If it can be shown that there exists an information flow from one neuron to another neuron via the EM field of the brain, this would provide strong empirical evidence for a direct causal role of the EM field for information processing within the brain.

List of Figures

Bibliography

[ACM+07] L. Astolfi, F. Cincotti, D. Mattia, M.G. Marciani, L.A. Baccala, F. de Vico Fallani, S. Salinari, M. Ursino, M. Zavaglia, L. Ding, J.C. Edgar, G.A. Miller, B. He, and F. Babiloni. Comparison of different cortical connectivity estimators for high-resolution EEG recordings. *Human Brain Mapping*, 28:143–157, 2007.

[BA97] A.W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, 1997.

[BAMCM97] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[BB04] G. Blanchard and B. Blankertz. BCI competition 2003 - data set IIa: Spatial patterns of self-controlled brain rythm modulations. *IEEE Transactions on Biomedical Engineering*, 51(6):1062–1066, 2004.

[BDK+07] B. Blankertz, G. Dornhege, M. Krauledat, K.R. Mueller, and G. Curio. The non-invasive berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 27(2):539–550, 2007.

[BGH+99] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kuebler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398:297–298, 1999.

[BGUB06] P. Breun, M. Grosse-Wentrup, W. Utschick, and M. Buss. Robust MEG source localization of event related potentials: Identifying relevant sources by non-gaussianity. In *Lecture Notes in Computer Science*, pages 394–403. Springer, Berlin/Heidelberg, 2006.

[BMK+06] B. Blankertz, K.R. Mueller, D. Krusienski, G. Schalk, J.R. Wolpaw, A. Schloegl, G. Pfurtscheller, J.R. Millan, M. Schroeder, and N. Birbaumer. The BCI competition III: Validating alternative approaches to actual BCI problems. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):153–159, 2006.

[BML01] S. Baillet, J.C. Mosher, and R.M. Leahy. Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, 18(6):14–30, 2001.

[BNL+07] C. Brunner, M. Naeem, R. Leeb, B. Graimann, and G. Pfurtscheller. Spatial filtering and selection of optimized components in four class motor imagery EEG data using independent components analysis. *Pattern Recognition Letters*, 28:957–964, 2007.

[BPR02] R. Boscolo, H. Pan, and V.P. Roychowdhury. Beyond Comon's identifiability theorem for independent component analysis. *Artificial Neural Networks - ICANN 2002 Lecture Notes in Computer Science*, 2415:1119–1124, 2002.

[BS95] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[Car97] J.F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114, 1997.

[Car98] J.F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 9(10):2009–2025, 1998.

[CL96] X.-R. Cao and R.-W. Liu. General approach to blind source separation. *IEEE Transactions on Signal Processing*, 44(3):562–571, 1996.

[CLC+03] J.M. Carmena, M.A. Lebedev, R.E. Crist, J.E. O'Doherty, D.M. Santuccil, D.F. Dimitrov, P.G. Patil, C.S. Henriquez, and M.A.L. Nicolelis. Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biology*, 1(2):193–208, 2003.

[Com94] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[CT06] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley & Sons, 2006.

[Cv78] T.M. Cover and J.M. van Campenhout. On the possible orderings in the measurement selection problem. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(9):657–661, 1978.

[DA01] P. Dayan and L.F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2001.

[DBCM04] G. Dornhege, B. Blankertz, G. Curio, and K.R. Mueller. Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Transactions on Biomedical Engineering*, 51(6):993–1002, 2004.

[DCF04] O. David, D. Cosmelli, and K.J. Friston. Evaluation of different measures of functional connectivity using a neural mass model. *NeuroImage*, 21:659–673, 2004.

[DGL96] L. Devroye, L. Gyoerfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[DM04] A. Delorme and S. Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[EAS98] A. Edelman, T.A. Arias, and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998.

[EGWB07] S. Eren, M. Grosse-Wentrup, and M. Buss. Unsupervised classification for non-invasive brain-computer interfaces. *Tagungsband der Automed 2007, VDI-Fortschrittsberichte*, 17(267):65–66, 2007.

[EK03] J. Eriksson and V. Koivunen. Identifiability and separability of linear ICA models revisited. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 23–27, Nara, Japan, April 2003.

[FD88] L.A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70:510–523, 1988.

[FHLS06] J. Farquhar, N.J. Hill, T.N. Lal, and B. Schoelkopf. Regularised CSP for sensor selection in BCI. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course*, pages 14–15. Verlag der Technischen Universitaet Graz, Graz, 2006.

[FM94] M. Feder and N. Merhav. Relations between entropy and error-probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

[Fre01] W.J. Freeman. *How Brains Make Up Their Minds*. Columbia University Press, 2001.

[GB06] M. Grosse-Wentrup and M. Buss. Subspace identification through blind source separation. *IEEE Signal Processing Letters*, 13(2):100–103, 2006.

[GC04] E. Gysels and P. Celka. Phase synchronization for the recognition of mental tasks in a brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 12(4):406–415, 2004.

[GCD83] G. Gratton, M.G.H. Coles, and E. Donchin. A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, 55(4):468–484, April 1983.

[GGP+05] R. Grave de Peralta Menendez, S. Gonzalez Andino, L. Perez, P.W. Ferrez, and J. del R. Millan. Non-invasive estimation of local field potentials for neuroprosthesis control. *Cognitive Processing*, 6:59–64, 2005.

[GGWB05] M. Grosse-Wentrup, K. Gramann, E. Wascher, and M. Buss. EEG source localization for brain-computer interfaces. In *Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering*, pages 128–131, Arlington, Virginia, March 2005.

[GHT+01] J. Gross, M. Hamalainen, L. Timmermann, A. Schnitzler, and R. Salmelin. Dynamic imaging of coherent sources: Studying neural interactions in the human brain. *Proceedings of the National Academy of Sciences*, 98(2):694–699, 2001.

[GI99] J. Gross and A.A. Ioannides. Linear transformations of data space in MEG. *Physics in Medicine and Biology*, 44(8):2081–2097, 1999.

[GPAT03] D. Garrett, D.A. Peterson, C.W. Anderson, and M.H. Thaut. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):141–144, 2003.

[GWB07] M. Grosse-Wentrup and M. Buss. Overcomplete independent component analysis via linearly constrained minimum variance spatial filtering. *Journal of VLSI Signal Processing*, 48(1-2):161–171, 2007.

[GWGB07] Moritz Grosse-Wentrup, Klaus Gramann, and Martin Buss. Adaptive spatial filters with predefined region of interest for EEG based brain-computer-interfaces. In B. Schoelkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 537–544. MIT Press, Cambridge, MA, 2007.

[HB01] M.H. Hansen and Y. Bin. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[HKO01] A. Hyvaerinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley & Sons, 2001.

[HLS+06] N.J. Hill, T.N. Lal, M. Schroeder, T. Hinterberger, G. Widman, C.E. Elger, B. Schoelkopf, and N. Birbaumer. Classifying event-related desynchronization in EEG, ECoG and MEG signals. In *Lecture Notes in Computer Science*, pages 404–413. Springer, Berlin/Heidelberg, 2006.

[HSF+06] L.R. Hochberg, M.D. Serruya, G.M. Friehs, J.A. Mukand, M. Saleh, A.H. Caplan, A. Branner, D. Chen, R.D. Penn, and J.P. Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442:164–171, 2006.

[HWCM06] D.A. Heldman, W. Wang, S.S. Chan, and D.W. Moran. Local field potential spectral tuning in motor cortex during reaching. *IEEE Transactions on Neural Systems And Rehabilitation Engineering*, 14(2):180–183, 2006.

[Jef95] J.G.R. Jefferys. Nonsynaptic modulation of neuronal-activity in the brain - electric currents and extracellular ions. *Physiological reviews*, 75(4):689–723, 1995.

[JMB+01] T. Jung, S. Makeig, M.J. McKeown A.J. Bell, T. Lee, and T.J. Sejnowski. Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7):1107–1122, 2001.

[KBJP98] T.R. Knosche, E.M. Berends, H.R.A. Jagers, and M.J. Peters. Determining the number of independent sources of the EEG: A simulation study on information criteria. *Brain Topography*, 11(2):111–124, 1998.

[KLH05] B. Kamousi, Z. Liu, and B. He. Classification of motor imagery tasks for brain-computer interface applications by means of two equivalent dipoles analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(2):166–171, 2005.

[KO98] J. Koerner and A. Orlitsky. Zero-error information theory. *IEEE Transactions on Information Theory*, 44(6):2207–2229, 1998.

[LBCM05] S. Lemm, B. Blankertz, G. Curio, and K.R. Mueller. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Transactions on Biomedical Engineering*, 52(9):1541–1548, 2005.

[LGS99] T. Lee, M. Girolami, and T.J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11:417–441, 1999.

[LGWGB07] C. Liefhold, M. Grosse-Wentrup, K. Gramann, and M. Buss. Comparison of adaptive spatial filters with heuristic and optimized region of interest for EEG-based brain-computer interfaces. In *Lecture Notes in Computer Science*, pages 274–283. Springer, Berlin/Heidelberg, 2007.

[Lip67] O.C.J. Lippold. Electromyography. In P. H. Venables and I. Martin, editors, *Manual of Psychophysiological Methods*. Wiley & Sons, 1967.

[LLA07] F. Lotte, A. Lecuyer, and B. Arnaldi. FuRIA: A novel feature extraction algorithm for brain-computer interfaces using inverse models and fuzzy regions of interest. In *Proceedings of the 3rd International IEEE EMBS Conference on Neural Engineering*, pages 175–178, Kohala Coast, HI, USA, May 2007.

[LLGS99] T.W. Lee, M.S. Lewicki, M. Girolami, and T.J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 6(4):87–90, 1999.

[LNF+06] E. Lew, M. Nuttin, P.W. Ferrez, A. Degeest, A. Buttfield, G. Vanacker, and J.R. Millï¿½n. Non-invasive brain computer interface for mental control of a simulated wheelchair. In G.R. Mueller-Putz, C. Brunner, R. Leeb, R. Scherer, A. Schloegl, S. Wriessnegger, and G. Pfurtscheller, editors, *Proceedings of the 3rd International Brain-Computer Interface Workshop & Training Course*. TU-Gratz, 2006.

[LS03] S.B. Laughlin and T.J. Sejnowski. Communication in neuronal networks. *Science*, 301:1870–1874, 2003.

[MAB03] K.R. Mueller, C.W. Anderson, and G.E. Birch. Linear and nonlinear methods for brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):165–169, 2003.

[McF02] J. McFadden. Synchronous firing and its influence on the brain's electromagnetic field. *Journal of Consciousness Studies*, 9(4):23–50, 2002.

[MDOD04] S. Makeig, S. Debener, J. Onton, and A. Delorme. Mining event-related brain dynamics. *Trends in Cognitive Sciences*, 8(5):204–210, 2004.

[MHMB06] L. Marshall, H. Helgadottir, M. Molle, and J. Born. Boosting slow oscillations during sleep potentiates memory. *Nature*, 444(7119):610–613, 2006.

[MM80] R. Monzingo and T. Miller. *Introduction to Adaptive Arrays*. Wiley & Sons, 1980.

[MML+04] C.M. Michel, M.M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. Grave de Peralta. EEG source imaging. *Clinical Neurphysiology*, 115:2195–2222, 2004.

[MRV+03] C. Mehring, J. Rickert, E. Vaadia, S. Cardoso de Oliveira, A. Aertsen, and S. Rotter. Inference of hand movements from local field potentials in monkey motor cortex. *Nature Neuroscience*, 6(12):1253–1254, 2003.

[MS07] A.S. Matveev and A.V. Savkin. Shannon zero error capacity in the problems of state estimation and stabilization via noisy communication channels. *International Journal of Control*, 80(2):241–255, 2007.

[MWJ+02] S. Makeig, M. Westerfield, T.P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T.J. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295(5555):690–694, 2002.

[MZKM02] F. Meinecke, A. Ziehe, M. Kawanabe, and K.R. Mueller. A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. *IEEE Transactions on Biomedical Engineering*, 49(12):1514–1525, 2002.

[NAHS06] S.S. Nagarajan, H.T. Attias, K.E. Hild II, and K. Sekihara. A graphical model for estimating stimulus-evoked brain responses from magnetoencephalography data with large background brain activity. *Neuroimage*, 30:400–416, 2006.

[NFZE07] G.N. Nair, F. Fagnani, S. Zampieri, and R. J. Evans. Feedback control under data rate constraints: An overview. *Proceedings of the IEEE*, 95(1):108–137, 2007.

[Ng04] A.Y. Ng. Feature selection, L_1 vs.L_2 regularization, and rotational invariance. In Carla E. Brodley, editor, *Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004), Banff, Alberta, Canada, July 4-8*. ACM, 2004.

[NS05] P.L. Nunez and R. Srinivasan. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, 2005.

[NW06] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer, 2006.

[OWTM06] J. Onton, M. Westfield, J. Townsend, and S. Makeig. Imaging human EEG dynamics using independent component analysis. *Neuroscience and Biobehavioral Reviews*, 30(6):808–822, 2006.

[Pal96] M. Palus. Nonlinearity in normal human EEG: cycles, temporal asymmetry, nonstationarity and randomness, not chaos. *Biological Cybernetics*, 75(5):389–396, 1996.

[PL99] G. Pfurtscheller and F.H. Lopes da Silva. Even-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110:1842–1857, 1999.

[PNFP97] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer. EEG-based discrimination between imagination of right and left hand movement. *Electroencephalography and clinical Neurophysiology*, 103:642–651, 1997.

[PSGS05] L.C. Parra, C.D. Spence, A.D. Gerson, and P. Sajda. Recipes for linear analysis of EEG. *Neuroimage*, 28:326–341, 2005.

[PXZF00] J.C. Principe, D. Xu, Q. Zhao, and J.W. Fisher III. Learning from examples with information theoretic criteria. *Journal of VLSI SIgnal Processing*, 26(1-2):61–77, 2000.

[QDH04] L. Qin, L. Ding, and B. He. Motor imagery classification by means of source analysis for brain-computer interface applications. *Journal of Neural Engineering*, 1:135–141, 2004.

[RD69] S. Rush and D.A. Driscoll. EEG electrode sensitivity - an application of reciprocity. *IEEE Transactions on Biomedical Engineering*, 16:289–296, 1969.

[RGMA05] A. Rakotomamonjy, V. Guigue, G. Mallet, and V. Alvarado. Ensemble of SVMs for improving brain computer interface P300 speller performances. In *Artificial Neural Networks: Biological Inspirations - ICANN 2005*, volume 3696 of *Lecture Notes in Artificial Intelligence*, pages 45–50. Springer, 2005.

[RMGP00] H. Ramoser, J. Mueller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000.

[RP01] M.G. Rosenblum and A.S. Pikovsky. Detecting direction of coupling in interacting oscillators. *Physical Review E*, 64:0452021 – 0452024, 2001.

[RPK96] M.G. Rosenblum, A.S. Pikovsky, and J. Kurths. Phase synchronization of chaotic oscillators. *Physical Review Letters*, 76(11):1804–1807, 1996.

[Sch00] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

[Sch05] A. Schloegl. Results of the BCI competition 2005 for data set IIIa and IIIb. Technical report, Institute for Human-Computer Interfaces - BCI Lab, University of Technology Graz, Austria, 2005. available at `http://www.dpmi.tu-graz.ac.at/~schloegl/publications/TR_BCI2005_III.pdf`.

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 & 623–656, 1948.

[Sha56] C. Shannon. The zero error capacity of a noisy channel. *IEEE Transactions on Information Theory*, 2(3):8–19, 1956.

[SKM+07] G. Schalk, J. Kubanek, K.J. Miller, N.R. Anderson, E.C. Leuthardt, J.G. Ojemann, D. Limbrick, D. Moran, L.A. Gerhardt, and J.R. Wolpaw. Decoding two-dimensional movement trajectories using electrocorticographic signals in humans. *Journal of Neural Engineering*, 4:264–275, 2007.

[SRY+06] G. Santhanam, S.I. Ryu, B.M. Yu, A. Afshar, and K.V. Shenoy. A high-performance brain-computer interface. *Nature*, 442:195–198, 2006.

[SYTI05] H. Serby, E. Yom-Tov, and G.F. Inbar. An improved P300-based brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(1):89–98, 2005.

[TDN+06] R. Tomioka, G. Dornhege, G. Nolte, K. Aihara1, and K.R. Mueller. Optimizing spectral filters for single trial EEG classification. In *Lecture Notes in Computer Science*, pages 414–423. Springer, Berlin/Heidelberg, 2006.

[ten02] J.M.F. ten Berge. On uniqueness in CANDECOMP/PARAFAC. *Psychometrika*, 67(3):399–409, 2002.

[THS02] D.M. Taylor, S.I. Helms Tillery, and A.B. Schwartz. Direct cortical control of 3D neuroprosthetic devices. *Science*, 296:1829–1832, 2002.

[TLP04] F.J. Theis, E.W. Lang, and C.G. Puntonet. A geometric algorithm for overcomplete linear ICA. *Neurocomputing*, 56:381–398, 2004.

[Tor03] K. Torkkola. Feature extraction by non parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[Vap98] V.N. Vapnik. *Statistical Learning Theory*. Wiley & Sons, 1998.

[VB88] B.D. Van Veen and K.M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSAP Magazine*, 5(2):4–24, 1988.

[VMMW98] T.M. Vaughan, L.A. Miner, D.J. McFarland, and J.R. Wolpaw. EEG-based communication: analysis of concurrent EMG activity. *Electroencephalography and Clinical Neurophysiology*, 107(6):428–433, December 1998.

[von99] C. von der Malsburg. The what and why of binding: The modeler's perspective. *Neuron*, 24:95–104, 1999.

[VSJ+00] R. Vigario, J. Sarela, V. Jousmaki, M. Hamalainen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, 47(5):589–593, 2000.

[VvYS97] B.D. Van Veen, W. van Drongelen, M. Yuchtman, and A. Suzuki. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, 44(9):867–880, 1997.

[WB05] J. Wolff and M. Buss. Invariance control design for constrained nonlinear systems. In *Proceedings of the 16th IFAC World Congress*, Prague, Czech Republic, July 2005. Paper No. 04467.

[WB07] J. Wolff and M. Buss. On stability of invariance controlled linear systems. In *Proceedings of the European Control Conference*, pages 3281–3288, Kos, Greece, July 2007.

[WBH+00] J.R. Wolpaw, N. Birbaumer, W.J. Heetderks, D.J. McFarland, P.H. Peckham, G. Schalk, E. Donchin, L.A. Quatrano, C.J. Robinson, and T.M. Vaughan. Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*, 8(2):164–173, 2000.

[WBM+02] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

[WL96] J.J. Wright and D.T.J. Liley. Dynamics of the brain at global and microscopic scales: Neural networks and the EEG. *Behavioral and Brain Sciences*, 19(2), 1996.

[WM04] J.R. Wolpaw and D.J. McFarland. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the National Academy of Sciences*, 101:17849–17854, 2004.

[WMNF91] J.R. Wolpaw, D.J. McFarland, G.W. Neat, and C.A. Forneris. An EEG-based brain-computer interface for cursor control. *Electroencephalography and clinical Neurophysiology*, 78:252–259, 1991.

[WWGG07] Q. Wei, Y. Wang, X. Gao, and S. Gao. Amplitude and phase coupling measures for feature extraction in an EEG-based brain-computer interface. *Journal of Neural Engineering*, 4:120–129, 2007.

[ZLNM04] A. Ziehe, P. Laskov, G. Nolte, and K.R. Mueller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research*, 5:777–800, 2004.

[ZP01] M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, 2001.

[ZWJ00] L. Zhukov, D. Weinstein, and C. Johnson. Independent component analysis for EEG source localization. *IEEE Engineering in Medicine and Biology Magazine*, 19(3):87–96, 2000.