TECHNISCHE UNIVERSITÄT BERLIN

Advances in Neurotechnology for Brain Computer Interfaces

von
Siamac Fazli

Von der Fakultät IV,
Elektrotechnik und Informatik,
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

doctor rerum naturalium
- Dr. rer. nat. -

genehmigte Dissertation

Tag der wissenschaftlichen Aussprache: 28.November 2011

Berlin 2011

D 83

Promotionsausschuss:
Vorsitzender: Prof. Dr. Klaus Obermayer
Berichter: Prof. Dr. Klaus-Robert Müller
Berichter: Prof. Dr. Lucas C. Parra
Berichter: Prof. Dr. Gabriel Curio

To all the ones, who deserve it.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

ACKNOWLEDGMENTS

Firstly, I would like to thank my professor Klaus-Robert Müller, who taught me a great deal about machine learning, writing scientific papers, who motivated me and more importantly gave me the freedom and trust to pursue my scientific ideas. A special thanks goes to Benjamin Blankertz, who was always very patient with me and who always managed to generate a very open and warm atmosphere within the lab.

Furthermore, I would like to thank the members of the Brain2Robot team, namely Yakob Badower, Márton Danóczy and Cristian Grozea. They were not only valuable team members, with whom I enjoyed working everyday, but they also became real friends whom I could trust and build on. I would not want to miss them anymore. To this end I would also like to thank Florin Popescu for choosing such a great team.

I would like to thank Prof. Dr. Gabriel Curio and Prof. Dr. Lucas C. Parra for agreeing to act as referees and for their time and patience to the read and evaluate this thesis. Also, I would like to thank Dr. Cristian Grozea, Dr. Andreas Ziehe, Stefan Haufe and Sven Dähne for reading the manuscript and their valuable advice. Their constructive critisism helped to increase the quality of this manuscript

Additionally, I would like to thank all past and present lab members of IDA for generating an atmosphere, in which it is fun to be at. In particular I would like to mention Guido Dornhege, Matthias Krauledat, Stefan Haufe, Andreas Ziehe, Guido Nolte, Arne Ewald, Matthias Treder, Basti Venthur, Carmen Vidaurre, Steven Lemm, Dominik Kühne, Paul von Bünau, Felix Bießmann, Katja Hansen, Matthias Jugel, Marius Kloft, Frank Meinecke, Martijn Schreuder, Claudia Sanelli, Ryota Tomioka, Masashi Sugiyama, Imke Weitkamp, Andrea Gerdes, Sophie Schneiderbauer, Maria Kramarek, Rithwik Mutyala and all the others that I may have forgotten.

Finally, I would like to thank my family, who never pressured me into anything in particular, but rather always hoped I would someday end up doing something reasonable afterall. I am very greatful indeed for their everlasting support and for making me the person I have become. Last but not least I would like to thank Isabella, who I can always rely on and who reminds me of the bright side of life, whenever I seem to forget about it.

VITA

1979	Born, Torquay, Devon, UK
1999–2002	B.Sc. (Physics), University of Exeter, UK
2002–2005	M.Sc. (Medical Neurosciences), Charité, Humboldt Universität zu Berlin
2002–2004	Teaching Assistant, ITB, Humboldt Universität zu Berlin
2005–2008	PhD student, Fraunhofer First, Berlin
2009–today	PhD student, Technical University, Berlin

JOURNAL PUBLICATIONS / CONFERENCE PROCEEDINGS / ABSTRACTS

S. Fazli, J. Mehnert, G. Curio, A. Villringer, K.-R. Müller, J. Steinbrink, B. Blankertz (2012). Enhanced performance by a Hybrid NIRS-EEG Brain Computer Interface. NeuroImage, volume 59, issue 1, pages 519–529.

S. Fazli, M. Danoczy, J. Schelldorfer, K.-R. Müller (2011). ℓ_1-penalized Linear Mixed-Effects Models for high dimensional data with application to BCI. NeuroImage, volume 56, number 4, pages 2100–2108.

S. Fazli, M. Danóczy, J. Schelldorfer, K.-R. Müller (2011). ℓ_1-penalized Linear Mixed-Effects Models for BCI. ICANN 2011, Part I, LNCS 6791, pages 26–35. Springer, Heidelberg.

C. Grozea, C.D. Voinescu, and S. Fazli (2011). Bristle-sensors - Low-cost Flexible Passive Dry EEG Electrodes for Neurofeedback and BCI Applications. J. Neural Eng., volume 8, pages 025008.

G. Onose, C. Grozea, A. Anghelescu, C. Daia-Chendreanu, C. J. Sinescu, A. V. Ciurea, A. Mirea, I. Andone, A. Spînu, A.-S. Mihaescu, S. Fazli, M. Danoczy, F. Popescu (2010). EEG based brain-computer interface in chronic quadriplegics using robotic arm device as functional assistive technology. Spinal Cord. submitted

S. Fazli, J. Mehnert, J. Steinbrink, G. Curio, B. Blankertz (2010). NIRS signals predict SMR-based performance in EEG. 4th International BCI meeting.

J. Schelldorfer, S.Fazli, P. Bühlmann, K.-R. Müller (2010). Using Linear Mixed-Effects Models for subject-independent SMR-based BCI classification. 4th International BCI meeting.

B. Blankertz, M. Tangermann, C. Vidaurre, S. Fazli, C. Sannelli, S. Haufe, C. Maeder, L. Ramsey, I. Sturm, G. Curio, K.-R. Müller (2010). The Berlin Brain-Computer Interface: Non-Medical Uses of BCI Technology, Front Neuroscience, 4:198.

B. Blankertz, M. Tangermann, C. Vidaurre, T. Dickhaus, C. Sannelli, F. Popescu, S. Fazli, M. Danoczy, G. Curio, K.-R. Müller (2010), Detecting Mental States by Machine Learning Techniques: The Berlin Brain-Computer Interface. Brain-Computer Interfaces (Revolutionizing Human-Computer Interaction), Springer.

S. Fazli, C. Grozea and M. Danóczy and B. Blankertz and F. Popescu and K.-R. Müller (2009). Subject independent EEG-based BCI decoding. Advances in Neural Information Processing Systems 22 (NIPS), pages 513-521.

S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, C. Grozea (2009). Subject independent mental state classification in single trials. Neural Networks, volume 22(9), pages 1305-1315.

S. Fazli, M. Danóczy, F. Popescu, B. Blankertz, K.-R. Müller: Using Rest Class and Control Paradigms for Brain Computer Interfacing. IWANN (1) 2009: pages 651-665.

B. Blankertz, M. Tangermann, F. Popescu, M. Krauledat, S. Fazli, M. Danóczy, G. Curio, and K.-R. Müller (2008). The Berlin Brain-Computer Interface. In Jacek M. Zurada, Gary G. Yen, and Jun Wang, editors, WCCI 2008 Plenary/Invited Lectures, volume 5050 of LNCS, pages 79-101. Springer, Berlin Heidelberg.

S. Fazli, M. Danóczy, M. Kawanabe, and F. Popescu (2008). Asynchronous, adaptive BCI using movement imagination training and rest-state inference. In IASTED's Proceedings on Artificial Intelligence and Applications, pages 85-90.

S. Fazli, C. Grozea, M. Danóczy, B. Blankertz, K.-R. Müller, and F. Popescu (2008). Ensembles of temporal filters enhance classification performance for ERD-based BCI systems. In Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course 2008. Verlag der Technischen Universität Graz.

C. Grozea, S. Fazli, G. Nolte, M. Danóczy and F. Popescu (2008). A method for predicting the success of a BCI training session based on the classification of the CSP fil-

ters itself. In Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course 2008. Verlag der Technischen Universität Graz.

M. Danóczy, S. Fazli, C. Grozea, K.-R. Müller, and F. Popescu (2008). Brain2Robot: a grasping robot arm controlled by gaze and asynchronous EEG BCI. In Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course 2008. Verlag der Technischen Universität Graz.

F. Popescu, S. Fazli, Y. Badower, B. Blankertz, and K.-R. Müller (2007). Single trial classification of motor imagination using 6 dry EEG electrodes. PLoS ONE, volume 2, number 7, pages e637.

F. Popescu, Y. Badower, S. Fazli, G. Dornhege, and K.-R. Müller (2006). EEG-based control of reaching to visual targets. Dynamical Principles for neuroscience and intelligent biomimetic devices - Abstracts of the EPFL-LATSIS Symposium 2006, pages 123-124, Lausanne.

PATENTS

F. Popescu, S. Fazli, Y. Badower, K.-R. Müller (2008). Dry electrode cap for electroencephalography (2008). (WO/2008/067839), PCT/EP2006/011843.

DEMONSTRATIONS

Medica, Düsseldorf, Germany. $14^{th}-17^{th}$ November 2007. Brain2Robot: a grasping robot arm controlled by gaze and asynchronous EEG BCI.

NIPS, Vancouver, Canada. 9^{th} December 2008. Play Brain-Pong in 10 Minutes: Demonstration of a 6 electrode dry electrode cap for BCI.

ABSTRACT OF THE DISSERTATION

Advances in Neurotechnology for Brain Computer Interfaces

by

Siamac Fazli
Doctor of Philosophy in Computer Science
Technische Universität Berlin, 2011
Prof. Dr. Klaus Obermayer, Vorsitzender

Brain Computer Interfacing has witnessed a tremendous growth of scientific interest during the last 10 years. However, some downfalls have prevented this exciting technology to produce mainstream applications for the general public. Among those are long setup time, illiteracy of some subjects as well as non-stationarities within recording sessions.

This thesis introduces a number of hardware as well as software related neurotechnological developments, which address and alleviate these issues, thus making BCI a more compact, robust and ready-to-use technology. A patented dry electrode EEG cap with 6 channels is introduced and its capabilities demonstrated within a BCI environment. While this development certainly enhances BCI usability, also future EEG research will benefit from dry electrode technology. To further reduce setup time to essentially zero, an ensemble framework, consisting of a large number of BCI datasets, was developed and gated by a number of machine learning methods, to enable instantaneous feedback for users. In addition, a multimodal neuroimaging study was conducted and shown to reduce illiteracy among subjects as well as enabling basic neuroscientific insight.

ZUSAMMENFASSSUNG DER DISSERTATION

Advances in Neurotechnology for Brain Computer Interfaces

von

Siamac Fazli
Doktor der Naturwissenschaften
Technische Universität Berlin, 2011
Prof. Dr. Klaus Obermayer, Chair

Gehirn Computer Schnittstellen haben in den letzten 10 Jahren ein enormes wissenschaftliches Interesse hervorgerufen. Allerdings offenbart diese spannende Technology bei näherer Betrachtung noch einige Hürden, welche bisher die Entwicklung von massentauglichen Anwendungen verhindert haben. Unter Anderem eine lange Vorbereitungszeit eines BCI Systems, die fehlende Steuermöglichkeiten für manche Benutzer, sowie die nicht Stationaritäten innerhalb einer Aufnahme.

Diese Dissertation führt eine Reihe von neurotechnologischen Entwicklungen ein, welche diese Probleme addressieren. Dadurch wird BCI zu einer kompakteren, robusteren und praktikableren Technologie. Eine patentierte EEG Kappe mit sechs trockenen Elektroden wird vorgestellt und ihre Funktion innerhalb der BCI Umgebung demonstriert. Während diese Entwicklung für BCI von Nutzen ist, wird auch zukünftige EEG Forschung von dieser Technologie profitieren. Zur weiteren Reduzierungs der Vorbereitungszeit, wurde ein Ensemble Framework entwickelt, welches aus einer grossen Menge von BCI Daten besteht. Mit Hilfe der Methoden des maschinellen Lernens erlaubt dieses Framework damit ein instantanes Feedback. Weiterhin wurde eine multi-modale Studie durchgeführt, welche die Inoperabilität des Systems für einige Benutzer reduzieren konnte, und desweiteren zu neurophysiologischen Erkenntnissen geführt hat.

CHAPTER 1

Introduction

Oscillations abound all domains of nature and deciphering the characteristics of those oscillations is at the heart of most scientific fields of the past and today. The present work will deal with the oscillatory activity of the human brain and how voluntary modulation of those brain oscillations can be exploited to form meaningful communication channels.

A Brain Computer Interface (BCI) is a device that enables a subject to use her brain to communicate with an external device. In general one distinguishes two types of BCI: *invasive* and *non-invasive*. While invasive BCIs in humans require a craniotomy, the surgical removal of a section of the skull, in order to access the brain underneath, non-invasive BCIs measure brain activity without invading the integrity of the body. In electrocorticography (ECoG) electrodes are placed beneath the skull, directly onto the cerebral cortex. This is common practice in medical diagnosis for identifying epileptogenic zones in the cortex. Invasive BCIs, where electrodes are implanted into the grey matter, can measure single neurons or local field potentials (LFP) and thus yield the 'cleanest' signals. Animal and human experiments have shown that very accurate control of a cursor is possible in up to three dimensions [93, 76]. However, these implants also pose high (infectious) risks for the user. Another yet unsolved problem is the scaring of the brain tissue as a response to the implant, which in turn leads to progressively lower signal quality as time goes by. The following work will solely deal with non-invasive BCIs and uses the term 'BCI' interchangeably with 'non-invasive BCI'.

While recently there has been a surge of interest in non-invasive BCI, with many groups starting research in this area, impressive pioneering work had already begun in the early 70's [127, 128], which relied on visually evoked response potentials (VEPs). The early approaches to BCI primarily relied on electroencephalography (EEG) as a neuroimaging method [138, 35]. However, since then BCI technology has developed many variants and employed a large number of other neuroimaging methods, such as Magnetoencephalography (MEG) [133], Electrocorticography (ECoG) [110, 82, 47, 105, 90, 22], functional magnetic resonance imaging (fMRI) [136, 141, 116, 77] and near-infrared spectroscopy (NIRS) [126, 1, 44] among others.

This work will primarily focus on sensory motor rhythm (SMR)-based Brain Computer Interfaces, which exploit the suppression of motor related idle rhythms during

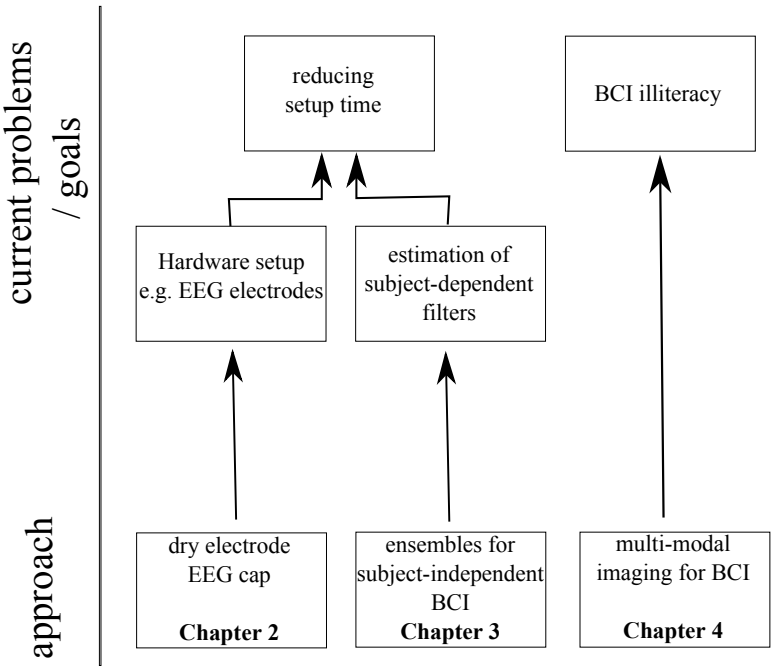

Figure 1.1: An illustration of current problems in BCI and where these are addressed within this thesis.

motor execution and *motor imagery*. Other physiological measures for successful communication are visual, auditory or sensory-motor evoked potentials as well as slow cortical potentials. For a detailed review of the different types and modes of current BCI research we would like to refer the reader to [35, 138].

While a large number of achievements have been made in SMR-based BCIs, there are still a number of problems, which hinder the introduction of this field of research to a wider community. The work presented here will show how various neurotechnological developments and their application help in making state-of-the-art Brain Computer Interfaces (BCI) more versatile, easier to use and more compact. A number of these problems are addressed and some of them alleviated within this thesis. The flowchart of Figure 1.1 gives an overview of the current problems and goals in present day BCI research and the approach that was developed to alleviate them.

One of the primary goals of BCI today is to reduce the setup time of a given BCI system. Setup time can be setup of hardware, such as applying conducting gel to EEG electrodes, as well as the estimation of subject-dependent filters. To reduce the hardware-related setup time a novel EEG cap, which is based on 6 dry electrodes is introduced. Furthermore, its successful operation within a BCI paradigm is demon-

strated.

Early BCI prototypes relied on operant conditioning of the subject [8]. With such a system it can take months for a subject to adapt his brain, such that she would be able to interact with the system. Through the statistical dissemination of motor-related EEG data it became apparent that it is possible to extract features from training data of a given subject and use the estimated subject-dependent model to form a stable high-speed BCI. It is now common practice to record such training sets with or without initial feedback [11, 99, 27, 24, 12, 35]. We show that being able to select from a very large database of experiments allows to construct a subject-independent classifier, which is comparable in quality to subject-dependent ones. The approach is thus able to decrease the calibration time of SMR-based BCI's to practically zero.

EEG is the neuroimaging method with the highest temporal resolution and could therefore potentially provide the highest information transfer rates, which so far is indeed the case. A long-standing problem of BCI designs which detect EEG patterns related to some voluntarily produced brain state is that such paradigms work with varying success among subjects/patients. We distinguish mental task based BCI such as SMR-based BCI from paradigms based on involuntary stimulus related potentials such as P300, which are limited to very specific applications such as typing for locked-in patients and require constant focus on stimuli extraneous to the task at hand. The peak performance to be achieved even after multiple sessions, varies greatly among subjects. Using a recent study [17] and other unreported data by many research groups, we estimate that about 20% of subjects do not show strong enough motor related mu-rhythm variations for effective asynchronous motor imagery BCI, that for another 30% performance is slow (<20 bits/min) and for up to 50% it is moderate to high ($20-35$ bits/min). It is still a matter of debate as to why BCI systems exhibit *illiteracy* in a significant minority of subjects and what can be done about it in terms of signal processing and machine learning algorithms. Furthermore long-term usage of a BCI can lead to non-stationarities in the data. While both these issues have been addressed within the EEG itself [70, 15], also combinational approaches for EEG features from multiple domains [34] as well as combinations of EEG and peripheral parameters like electromyography [78] have been shown to robustify the classification. In this context we propose a multi-modal approach, consisting of EEG and NIRS and show that NIRS can not only help to elevate classification accuracies for most subjects, but also enables successful BCI operation for some subjects, who were previously not able to do so.

1.1 Outline

The following parts of the introduction deal with the basic ingredients, necessary for the successful operation of a Brain Computer Interface. A brief history and the physiological basis of some non-invasive neuroimaging methodologies is given in Section 1.2. Section 1.3 reviews the machine learning and statistical tools, that are used within this work and Section 1.4 gives a general introduction to Brain Computer Interfacing. In Chapter 2 a novel dry electrode cap is introduced, utilized in a BCI study. The following Chapter explores how ensemble methods can be beneficial for reducing calibration times in BCI. Two approaches are proposed and validated. The first deals with the problem of findings subject specific temporal filters from training data. It shows that commonly used heuristics for temporal filter estimation can be unstable for low numbers of training trials or for subjects, where the discriminability is low in general. In these cases it shows that multiple classifier systems, trained with an ensemble of temporal filters enhances the decoding performance for BCI. The second approach shows how ensemble methods may help in reducing the calibration time to zero. Ensembles of subject-dependent classifiers are generated, using a very large set of previous experiments of many individual subjects and a convex optimization problem is formulated to obtain a weight vector, that enables new subjects to start high-speed feedback session without the need of recording a calibration session. Chapter 4 examines how multi method imaging may help in robustifying EEG-based BCI. Chapter 5 concludes this work and gives an outlook of the possible future directions of BCI.

1.2 Non-invasive Neuroimaging for the Brain

1.2.1 Electroencephalogramm (EEG)

The discovery that the brain exhibits electrical activity was first discovered in the 19th century [25]. In the late 1920s Hans Berger measured electrical potentials on the surface of the skull [5] for the first time and thus created the basis for a new field of study: the EEG.

The EEG records oscillations of electrical potentials, measured by electrodes that are placed on the human scalp. While the EEG has the highest temporal resolution of all non-invasive neuroimaging methods, its spatial resolution is limited for reasons which are explained below. To understand the neurophysiological and physical basis of the EEG one needs to consider the electrical properties of individual neurons and their anatomical organization within the cortex. While the central nervous system consists of neurons and glia cells, most of the effects measured by EEG reflect the summated activity of postsynaptic potentials of large populations of

neurons. Glia cells have been found to contribute only very modestly to the surface EEG.

A neuron shares many characteristics of other cells in the body, but in addition a neuron can communicate with other neurons by means of its axonal processes. Neurons keep a high intracellular concentration of potassium (K^+) and chloride (Cl^-), while maintaining a low intracellular sodium (Na^+) and calcium (Ca^{2+}) concentration. These concentration differences result in a negative cellular potential of approx. -70 millivolts, relative to the extracellular space. Within the cell membranes there are a large number of ion channels, which serve to maintain the negative cellular potential in its resting state. However these ion channels are also responsible for the generation and initiation of action potentials. An influx of Na^+ (and in some cases Ca^{2+}) causes the cell to depolarize. The outflow of K^+ repolarizes the membrane by restoring the initial charge distribution.

An action potential traveling along the axon generates a very brief local current in the axon and thus a small potential field. At the nerve terminal various neurotransmitters are released. These neurotransmitters produce changes in membrane conductance and transmembrane potentials at the post-synaptic membrane. The neurotransmitter can have an excitatory or an inhibitory effect on the postsynaptic neuron. If the effect is of excitatory nature, it leads to a temporary depolarization of the postsynaptic membrane potential, caused by the inflow of negatively charges ions into the cell. This effect is called *excitatory postsynaptic potential (EPSP)*. An EPSP makes it easier for the cell to fire an action potential. While a single EPSP is generally not sufficient for the generation of an action potential, EPSPs are additive. The higher the number of EPSPs that arrive at a given cell, the higher the probability that the membrane potential will reach the threshold for firing an action potential. If the neurotransmitter has an inhibitory effect on the postsynaptic neuron, it leads to a hyperpolarization of the postsynaptic cell (a so-called *inhibitory postsynaptic potential (IPSP))* and thus reducing the probability of that cell firing an action potential.

The pyramidal neurons are the major projection neurons of the neocortex. Their dendrites receive a variety of synaptic inputs and are oriented perpendicular to the cell surface. The ion fluxes of these pyramidal neurons, associated with their respective EPSPs and IPSPs, generate extracellular field potentials (or local field potentials). A local field potential, measured within the brain will always represent the linear sum of a large number of overlapping fields generated by currents from the intracellular space to the extracellular space (so-called *current sources*) as well as currents from extracellular space to intracellular space (so-called *current sinks*). The effects of postsynaptic potentials propagate much further in the extracellular space, as compared to action potentials. While EPSPs and IPSPs are far smaller in amplitude as compared to action potentials, they last up to $100ms$ and therefore

the probability that they occur in a temporally overlapping manner is far higher than the very brief action potentials, which typically last only a few milliseconds. Also EPSPs and IPSPs have a higher contribution to the local field potential, since only a small minority of neurons will spike at a given time, but EPSPs and IPSPs contribute to changes of the local field potential, since these effects are displayed by many more neurons.

The electrical activity on the scalp is thus the result of extracellular current flows from summated activity of a large number of relatively synchronously activated neurons. The primary source of EEG activity is the synaptic activity of the mentioned pyramidal neurons of the neocortex. The surface EEG mainly consists of a spatially smoothed version of the local field potentials under a scalp surface on the order of 10cm^2, but has little discernible relationship with the specific patterns of activity of the neurons that generate it [95, 96]. Thus the electrical potentials measured on the scalp mostly represent the superficial layers of the cortex, while deep structures of the brain, such as hippocampus, thalamus or brain stem are very implicit in the surface EEG.

1.2.2 Near Infrared Spectroscopy (NIRS)

NIRS is a relatively recent noninvasive neuroimaging technique. It enables continuous monitoring of changes in blood oxygenation and blood volume with respect to human brain function. First evidence that activity of nerve cells cause changes of the optical properties of brain tissue was discovered in 1949 [60]. While researchers in the 70's began to record oxygenation parameters from the intact, human brain [63], only in the 90's local functional brain mapping with optical signals became possible [87, 26].

The high transparency of brain tissue to waves in the near infrared spectrum (approximately from 0.7 micrometers to 300 micrometers) allows the transmission of photons through the intact brain. A photon, which enters the tissue, undergoes two types of interaction. It is either *absorbed* or *scattered*. Absorption leads to radiationless loss of energy to the medium. Infrared (IR) light can be absorbed and emitted by molecules, which in course undergo molecular electronic transitions. These take place, if electrons in a molecule are excited from one energy level to another. Scattering can occur at unchanged frequency in stationary tissue or be accompanied by a Doppler shift due to scattering by moving particles in tissue [130].

To understand the physical properties of NIRS one needs to consider the Beer-Lambert law. The Beer-Lambert law describes the absorption of light as a function of the properties of the material the light is traveling through (see also Figure 1.2):

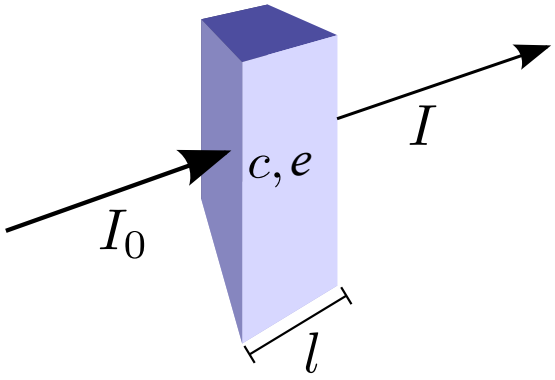

Figure 1.2: Illustration of the Beer-Lambert law.

$$A = \log_{10} \frac{I}{I_0} = e \cdot c \cdot l \tag{1.1}$$

where A is the Absorbance (or light extinction), I_0 the original light intensity, I the transmitted light intensity, e the molar absorbivity $[Lmol^{-1}cm^{-1}]$, l the path length of the light and c the concentration of the compound $[molL^{-}1]$.

The Beer-Lambert law holds as long as photons are either absorbed or transmitted in a straight line directly to the detector (see Figure 1.3: photons 2 and 3, respectively). Higher substance concentrations may lead to significant light scattering (photon 1) and Equation 1.1 needs to be modified, such that it takes into account the longer pathways of light (photon 1) and the loss of light due to scattering (photon 4). The modified Beer-Lambert law, as is given in Equation 1.2, therefore accounts for the increased pathlength with the term B, called the differential path length factor (DPF) and a term G, which represents the signal loss due to light scattering [130]:

$$A = e \cdot c \cdot l \cdot B + G \tag{1.2}$$

For some situations it may be sufficient to calculate the change of the concentration of the absorber (Δc). Assuming a constant light scattering loss Equation 1.2 reduces to: $\Delta A = e \cdot \Delta c \cdot l \cdot B$. Estimating the pathlength of light $l \cdot B$ enables to calculate absolute changes in concentration.

The fact that infrared light between 650 and 950 nm is only weakly absorbed by biological tissue and that the absorption spectra of oxyhemoglobin (HbO_2) and deoxyhemoglobin (HbR) differ substantially in this range [139], enable to measure (changes in) concentrations of HbO_2 and HbR *in vivo* [118].

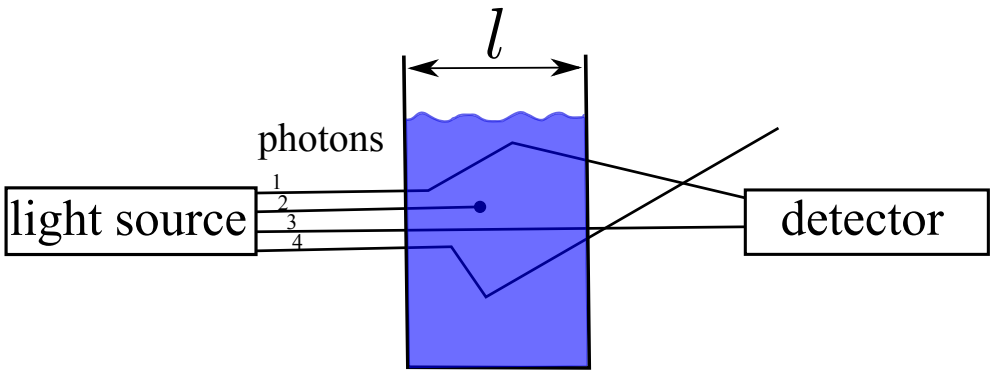

Figure 1.3: Illustration of the modified Beer-Lambert law (This figure is adopted from [130]).

For further literature we would like to refer the interested reader to the following classic NIRS papers [98, 130, 131].

1.3 Machine Learning, Signal Processing and Statistical Tools

The following sections will briefly introduce the most important analytical tools, which were used within this dissertation. This introduction focuses mostly on the intuitive understanding, rather than mathematical rigorosity or even completeness. More detailed descriptions and explanations of the only briefly mentioned tools would go beyond the scope of this document. However there is a variety of excellently written literature available, which we would like to refer the reader to [80, 57, 10, 37].

1.3.1 Statistical Tools

Given two random variables X and Y, with respective means

$$E[X] = \mu_X \qquad \text{and} \qquad E[Y] = \mu_Y \quad , \tag{1.3}$$

their standard deviations σ_X and σ_Y are defined as

$$\sigma_X = \sqrt{E\left[(X-\mu_X)^2\right]} \qquad \sigma_Y = \sqrt{E\left[(Y-\mu_Y)^2\right]} \quad . \tag{1.4}$$

E denotes the *expected value* of the random variable.

1.3.1.1 Independent two-sample t-test

The t-statistic was introduced by William Gosset in 1908 [119] and is one of the most popular statistical tests of today. Two-sample t-tests of independent samples are used if two separate sets of independent and identically distributed samples are available and one would like to test the null hypothesis that the means of two normally distributed populations are equal. It is defined by

$$t = \frac{\mu_X - \mu_Y}{\sqrt{\frac{1}{2}\left(\sigma_X^2 + \sigma_Y^2\right)}} \quad . \tag{1.5}$$

1.3.1.2 Covariance

The covariance between the two random variables is then defined by

$$\begin{aligned} \operatorname{cov}(X,Y) &= E\left[(X - E[X])(Y - E[Y])\right] \\ &= E\left[(X - \mu_X)(Y - \mu_Y)\right] \quad . \end{aligned} \tag{1.6}$$

1.3.1.3 Pearson's product-moment Correlation Coefficient

Pearson's correlation coefficient is defined as the covariance of X and Y divided by the product of their standard deviations

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad . \tag{1.7}$$

1.3.1.4 Point-biserial Correlation Coefficient (r-value)

The point-biserial correlation coefficient is a special case of the Pearson product-moment correlation coefficient and measures the association of a binary random variable and a continuous random variable. It is defined as

$$r_{pb} = \frac{(\mu_1 - \mu_2)}{\sigma} \sqrt{\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 + 1)}} \quad , \tag{1.8}$$

where $n_{1/2}$ are the number of examples in groups 1 and 2.

Using Fishers transformation these correlations can be transformed into unit variance z-scores for each subject j [61] and grand average z-scores can be obtained

by a weighted sum of individual z-scores over a number of subjects:

$$z_j = \frac{\tanh^{-1}(r_j)}{\sqrt{m_j - 3}} \qquad \text{and} \qquad \bar{z} = \frac{\sum_{j=1}^{N} z_j}{\sqrt{N}} \quad , \tag{1.9}$$

where m_j is the sample size of subject j and N the total number of subjects. p-values for the hypothesis of zero correlation in the grand average can now be computed by means of a two-sided z-test.

1.3.2 Classification and Regression

In neuroscience, brain imaging and in particular in BCI scientists are interested to find significant differences between two or more brain states within the recorded data due to some carefully chosen paradigm. These brain states need to be found within the spatial and temporal domains of the data. In recent years machine learning techniques greatly aided this search by estimating and identifying meaningful models, which lead to significant advances in the comprehension and detection of human brain function.

In *classification* the task is to find a rule, which assigns an N-dimensional data vector $\mathbf{x}$ to one of several classes. Given that only two classes exist, a classifier can be formalized as a decision function $f : \mathbb{R}^N \to \{-1, +1\}$. The decision function may be linear or non-linear. For the linear case f is a separating hyperplane. A separating hyperplane is parametrized by its vector $\mathbf{w}$ and a bias term b. The label y is thus predicted by: $y = f(\mathbf{x}; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T\mathbf{x} + b)$.

While in *classification* the label y to be predicted takes only discrete values, in *regression* the label y is continuously valued, such that $f : \mathbb{R}^N \to \{-\infty, +\infty\}$. As for the case of classification there also exist linear as well as non-linear regression functions. Within the following subsections we will briefly cover the most popular forms of regression, namely the classic least squares regression (LSR), ℓ_1-regularized least squares regression (ℓ_1-LSR) as well as logistic regression.

1.3.2.1 Linear Discriminant Analysis (LDA)

LDA assumes the classes to be normally distributed with different means μ_1 and μ_2 but identical covariance matrix Σ with full rank. Assuming these quantities to be known, the hyperplane, given by the normal vector $\mathbf{w}$, can be calculated by:

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2) \quad . \tag{1.10}$$

Given that these assumptions hold, the separating hyperplane is *Bayes optimal*. However in practice the true means and covariance matrices are not known and have to be approximated.

1.3.2.2 Linear Programming Machine

The linear programming machine (LPM) can be defined as follows:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}||w||_1 + \frac{C}{n}||\xi||_1 \\ \text{s.t.} \quad & y_i \cdot ((\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, ..., n \\ & \xi_i \geq 0 \end{aligned} \tag{1.11}$$

Due to the 1-norm optimization of $\mathbf{w}$ and ξ, the solution of the LPM is sparse. This sparsity can be very useful, since by identifying only most important features it leads to a compact model, which can in many cases lead to superior neurophysiological interpretations.

1.3.2.3 k-nearest neighbor

The decision k-nearest neighbor (knn) algorithm [30] is based on the distance of the closest training points within the feature space. The positive integer k defines how many of the nearest neighbors are considered for the decision. While the knn algorithm is one of the simplest machine learning algorithms, it represents an important baseline method for the valuation of more complex algorithms. Figure 1.4 gives an example of how the knn classification is made.

1.3.2.4 Support Vector Machines (SVM)

Support Vector Machines, originally invented by Vapnik in 1995 [29], are so-called *Large Margin Classifiers*. Based on the training data, the support vector machine constructs a hyperplane, such that the distance of the hyperplane to the nearest training data points is maximal. The training points closest to the hyperplane are the so-called *support vectors* and define the parameters of the model.

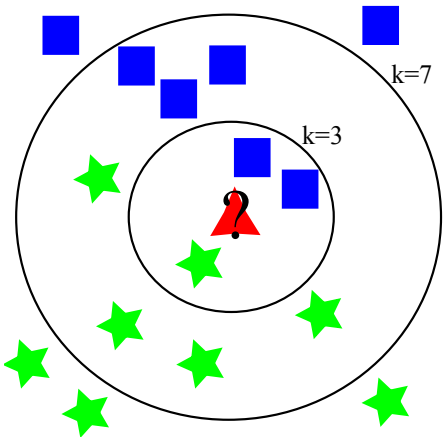

Figure 1.4: Illustration of the k-nearest neighbor algorithm: To classify the red triangle the number of nearest neighbors is to chosen to be $k = 3$. The red triangle would then be classified as blue (first circle). Note however, if $k = 7$, the red sample would be classifed as green (2^{nd} circle).

The optimization problem of the SVM is given as:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i \cdot ((\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1,...,n \\ & \xi_i \geq 0 \end{aligned} \tag{1.12}$$

While in their standard form SVMs are non-probabilistic binary linear classifiers, they can be adapted such that they also suit regression problems [36]. Figure 1.5 shows a sketch of the SVM.

1.3.2.5 Least Squares Regression

Let the input space be $X \in \mathbb{R}^{n\times p}$, where n denotes the number of trials/examples and p the number of parameters. Given that X^TX is nonsingular, the unique solution is

$$w = (X^TX)^{-1}X^Ty \quad . \tag{1.13}$$

Least squares regression (LSR) or ordinary least squares (OLS) models are estimated by minimizing the residual square error. The Gauss-Markov theorem states that the least-squares solution is the best linear unbiased estimate. OLS generally results in models known to have low bias, but a large variance. If one or more predictor vari-

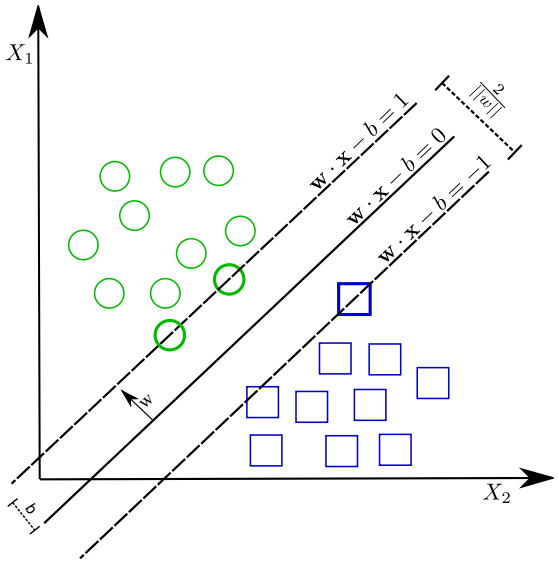

Figure 1.5: Sketch of an SVM. The *support vectors* are shown as bold.

ables are correlated (known as *near collinearity*), the determinant of $X^T X$ becomes almost singular, making w sensitive to random variations in y. The problem is then known to be 'ill-conditioned'.

1.3.2.6 ℓ_2-regularized Least Squares Regression

One possible method to reduce this sensitivity is called ℓ_2-regularized Least Squares Regression (ℓ_2-LSR) or *Ridge Regression*. The problem can be defined as a minimization problem:

$$\min_w \quad \frac{\lambda}{2}||w||_2^2 + \frac{1}{2}||Xw - y||_2^2 \tag{1.14}$$

In ridge regression a number δ is added to the diagonal elements, i.e. $X^T X + \delta I$. By doing so one sacrifices a little bias to reduce the variance of the predicted values and may therefore improve prediction accuracy.

1.3.2.7 ℓ_1-regularized Least Squares Regression

In 1996 Tibshirani proposed ℓ_1-regularized Least Squares Regression or *least absolute shrinkage and selection operator (lasso)* [125]. It is defined as:

$$\min_w \quad \frac{\lambda}{2}||w||_1 + \frac{1}{2}||Xw - y||_2^2 \tag{1.15}$$

By shrinking or setting some coefficients to 0 their prediction accuracy can sometimes be improved. Due to the sparse nature of the resulting models, the interpretability of results can be greatly improved, since sometimes it can be more desirable to identify as few as possible active coefficients having the strongest effects.

1.3.2.8 Logistic Regression

While the term logistic regression suggests otherwise, logistic regression is actually a classification technique. If formulated as an optimization problem it is given as:

$$\min_{w} \quad \frac{\lambda}{2}||w||_2^2 + l(Xw|y) \quad , \tag{1.16}$$

where $l(Xw|y)$ and $\sigma(w^T x_i)$ are defined as:

$$l(Xw|y) = \sum_{i=1}^{n} \{y_i \cdot \ln\left(\sigma(w^T x_i)\right) + (1-y_i) \cdot \ln\left(1-\sigma(w^T x_i)\right)\}$$
$$\sigma(w^T x_i) = \frac{1}{1+e^{-w^T x_i}}$$

1.3.3 Model Selection

1.3.3.1 Cross-validation

Cross-validation (CV) [67] is a method for evaluating how well an estimated predictive model generalizes to an independent dataset. Lets consider the first step of a cross-validation procedure: A given dataset is partitioned into two parts, a training and a test set. The training set is used to estimate the parameters of a predictive model, such that it fits the training data as well as possible. This model is then applied to the test set and its performance measured by an adequate *loss function*, such as mean squared error, or otherwise. To minimize the variance of this test error, this procedure is repeated a number of times. The mean loss and its standard deviation are usually reported as the result and are called *generalization error* or *expected risk*. Figure 1.6 shows a sketch of a cross-validation scheme with four splits.

While there are multiple ways of partitioning a given data set the most common method is *k-fold CV*: the data is split into k disjoint subsets of equal size. The model is trained then on all subsets except of one, on which it is tested. As before the procedure is then repeated k times, each time leaving out a different subset, such that each subset becomes the test set once. Another popular CV scheme is so-called *leave-one-out CV (LOO-CV)*. Here all data, except of one example (or trial) is used as the training set and the left-out trial is used as a test set. Also here the procedure

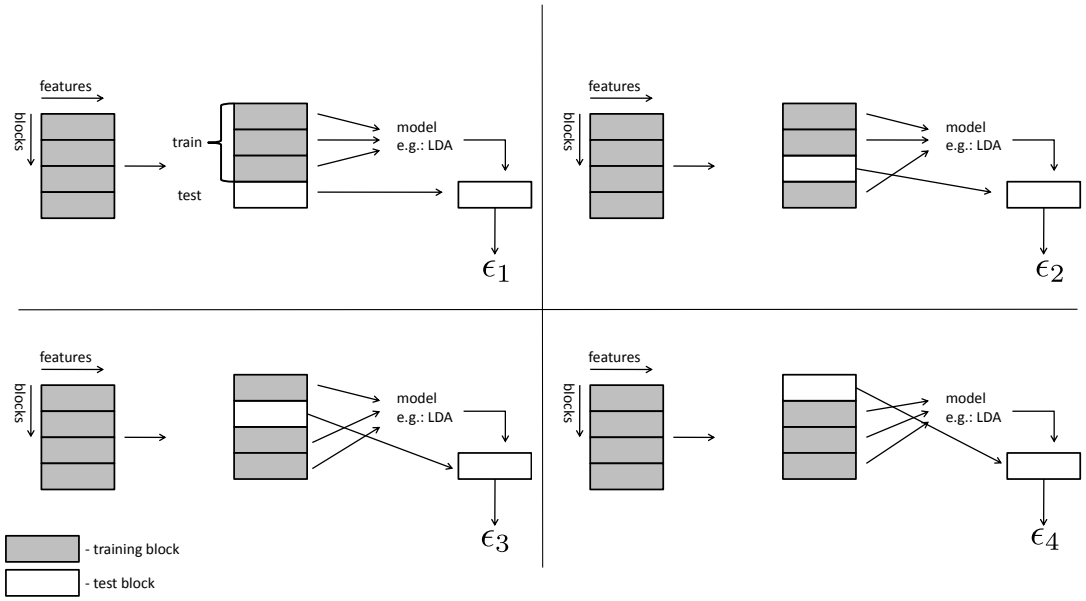

Figure 1.6: Chronological cross-validation with four blocks.

is repeated until each trial was used as a test set once. However, for large datasets LOO-CV becomes computationally expensive.

When choosing the appropriate cross-validation technique, one needs to take into account the exact form of the data, since there are some pitfall that need to be avoided. For example unbalanced class sizes need to be taken account for by choosing an appropriate loss function. Also non-stationarities in the data, which may stem from block-design - a common practice in neuroscientific experiments, needs to be addressed, since this non-stationarity can perturb the assumption of standard CV techniques that the data is independent and identically distributed (i.i.d.). For the case of possible non-stationarities a comparison of standard CV with *chronological CV* is proposed by [80]. For a recent review of common pitfall of applying machine learning techniques and their validation, we would like to refer the reader to [80].

1.4 The Berlin Brain Computer Interface (BBCI)

In the early 70's pioneers such as Vidal [127, 128] started the field of EEG-based BCI. Up to the mid 90's only very few groups were actively working in BCI. Today BCI now constitutes a diverse field with a large number of groups participating in this field

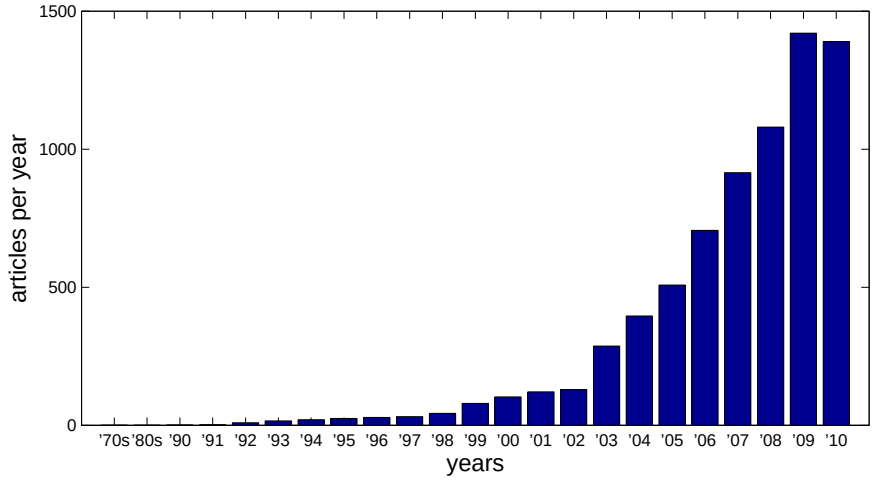

Figure 1.7: Number of articles containing the term Brain Computer Interface in the years from 1970 to today, according to Google scholar

of research (see Figure 1.7 for the growth of publications in the BCI field).

The BBCI group was formed in 1999 and has since then continuously grown. One of its main principles as compared to other groups at the time was to *let the machines learn* instead of the user. Due to this approach it was possible to reduce calibration times for individual users from many weeks to a only half an hour, before a BCI feedback session could be initiated. This approach has now been adopted by most other BCI groups and become a standard procedure in the field. The individual steps of the SMR-based BBCI setup are sketched below.

1.4.1 Calibration sessions

Before meaningful measurements of brain related potentials can be taken by means of EEG, the impedances between the electrodes of the EEG cap and the scalp must be reduced by means of a salt-containing gel. Depending on the number of channels this setup can take between 20 minutes and 1 hour. After the setup of the EEG cap, a training session is initiated, where typically three motor imagination tasks are cued, either by letters or arrows appearing on the screen. The three classes of motor imaginations were left hand (L), right hand (R) and right foot (F).

1.4.2 Outlier Removal

Excessive blinking, swallowing, clenching teeth, or severe tiredness may all be undesireable sources of noise that can interfere with the acquisition of 'clean' EEG data on a trial level during the calibration session and may therefore prevent successful estimation of covariance matrices and thus potentially harm the training of a BCI classifier. Furthermore an electric defect, or drying up of an electrode may render recordings of individual channels useless. It has therefore been an ongoing effort, within the BBCI to reduce these sources of noise in order to obtain a homogeneous set of training data. The methods considered include the Mahalanobis Distance of the variance of each trial and channel as measurement of the outlierness of the trials among others [56, 71, 70].

1.4.3 Temporal and Spatial filtering

For running any high-speed BCI system, it is of vital importance to identify features, which predict the intention of the user in a reliable and robust manner. In the context of BCI a high number of features is available and the choice of a small, but stable set is of paramount importance. Temporal and spatial filtering can help in reducing numbers of features significantly. We therefore briefly review a number of methods, many of which are used in the BBCI.

1.4.3.1 Finite Impulse Response Filter

The finite impulse response (FIR) filter is a digital filter and defined by the following difference equation:

$$y(t) = b_0 x(t) + b_1 x(t-1) + ... + b_{N_b} x(t-N_b) \tag{1.17}$$

where $x(t)$ is the input signal at time t, b the filter coefficients and N the order of the filter. An FIR filter is inherently stable, since all poles are located at the origin and thus within the unit circle.

1.4.3.2 Infinite Impulse Response Filter

While the FIR filter depends only on past value of the input signal, Infinite Impulse Response (IIR) Filters also depend on past output values. In its general form it can

be written as:

$$y(t) = \frac{1}{a_0}\Big(b_0 x(t) + b_1 x(t-1) + ... + b_{N_b} x(t-N_b) \tag{1.18}$$

$$- a_1 y(t-1) - a_2 y(t-2) + ... + a_{N_a} y(t-N_a)\Big) \tag{1.19}$$

or in a compacter form as:

$$y(t) = \frac{1}{a_0}\Big(\sum_{i=0}^{N_b} b_i x(t-i) - \sum_{j=1}^{N_a} a_j y(t-j)\Big) \tag{1.20}$$

One example of a IIR filter, among others is the butterworth filter [23], which we will use for temporal filtering in many of the following Chapters. The butterworth filter is designed, such that it does not have any ripples in the passband and thus particularly useful for the purpose of BCI.

1.4.3.3 Common Average Reference

The Common Average Reference [52] is a very simple method to get rid of the influence of having one particular reference. Subtracting from the potential V_i of each electrode i the mean potentials of all electrodes, results in the so-called *C*ommon Average Reference:

$$V_i^{\text{com}} = V_i - \frac{1}{N}\sum_{i=1}^{N} V_i \quad \forall i = 1...N \tag{1.21}$$

1.4.3.4 Weighted Local Average Reference

As proposed in [94] and [81] the potentials, of a given electrode are subtracted by a weighted sum of neighboring electrodes either *locally* or by all available electrodes *(weighted)*. The weights depend on the inverse linear distance to the electrode in question. For obvious reasons this method cannot be used for boundary electrodes, however it can be shown to yield beneficial results, when compared to simpler referencing methods, such as *C*ommon Average Reference or *L*aplace Filtering.

$$V_i^{\text{lap}} = V_i - \sum_{j \in S_i} g_{ij} V_j \quad \forall i = 1:N \quad \text{with} \quad g_{ij} = \frac{1/d_{ij}}{\sum_{j \in S_i} 1/d_{ij}} \tag{1.22}$$

and

$$V_i^{\text{lap}} = V_i - \sum_{j \in S_i} g_{ij} V_j \quad \forall i = 1:N \quad \text{with} \quad g_{ij} = \frac{1/d_{ij}}{\sum_{j \in S_i} 1/d_{ij}} \tag{1.23}$$

for the inclusion of all channels. A very early BCI study on the classification of movement onset in EEG showed superior results for local average referencing [48].

1.4.3.5 Common Spatial Patterns

The CSP algorithm (see e.g. [18, 79, 68]) searches for a matrix W and a vector of n values $0 \leq d_i \leq 1$ which achieves:

$$W\Sigma_1 W^\top = D \text{ and } W\Sigma_2 W^\top = I - D, \tag{1.24}$$

where n is the number of channels and D is a diagonal matrix with entries d_i. Using z-transform notation for digital signals, for any trial, the spatio-temporally demixed data is:

$$\boldsymbol{f}(z) = W\,H(z)\,\boldsymbol{s}(z) \tag{1.25}$$

Where $\boldsymbol{x}$ is the raw EEG signal and $H(z)$ is a diagonal matrix of identical bandpass filter transforms. The columns of the source to signal transform W^{-1} are called the Common Spatial Patterns (CSPs). The CSP decomposition can be thought of as a coupled decomposition of 2 matrices (for 2 classes) similar to a principal components analysis yielding eigenvectors and eigenvalues. As the eigenvalues d_i are equal to the power ratio of signals of class 1 by class 2 in the corresponding CSP filter (eigenvector in i-th column of matrix W), best discrimination is provided by filters with very high (i.e. near 1) or very low (i.e. near 0) eigenvalues. Accordingly CSP projections with the highest 2 and lowest 2 eigenvalues are generally chosen as features ($n = 4$).

CHAPTER 2

A novel dry electrode EEG cap

Electro-encephalography (EEG) is the oldest brain imaging technology, and among non-invasive methods it still offers the highest temporal resolution. Far from being a mere research aid, it promises an inexpensive, risk-free means of communication and neuroprosthetic control for the severely disabled [8, 138]. Recent advances in Brain Computer Interface (BCI) research have dramatically increased the amount of information we can extract from EEG over classical averaging and neurofeedback techniques [35]. Although EEG can monitor brain events very responsively in time, it suffers from high inter-trial variability and spatial mixing: numerous electrical sources active at any given time in the brain are superimposed onto the scalp across distances of over 5 cm [33]. These limitations have led to the assumption that many electrodes are necessary, and that one needs to average signal features across time or repeated trials to accurately discriminate mental states. However, as we will see, these assumptions do not necessarily hold for some paradigms we consider in the following.

Apart from intrinsic challenges of EEG signal analysis, one of the main obstacles precluding EEG-BCI from being used in patients' daily lives is setup encumbrance. Modern EEG practice, as part of the electrode application procedure known to specialists as montage, requires tedious application of conductive gel between electrodes and scalp (see left part of Figure 2.1). While recordings in certain clinical applications may last up to 72 hours, they progressively degrade as the gel dries leading to a failure of about a quarter of the electrodes within 24 hours and thus requires daily maintenance [40].

In this Chapter we introduce a new EEG cap design with a low number of electrodes and show that the much sought-after *dry electrode* technology can be surprisingly frugal and accurate enough for single trial discrimination. Dry electrodes have already been proposed since the early 90's [51, 121] and early pioneering work of capacitive electrodes had already begun in the early 70's [88]. Here we show the results of the first EEG-based BCI online study with dry electrodes.

Dry electrodes bypass gel application, thereby reducing set-up time. Fewer electrodes mean less time spent checking individual signal quality and adjusting the cap. Our new design can be seen in Figure 2.2c. It consists of only 6 dry unipolar electrodes and one dry reference electrode. The cap applies a moderate amount of

Figure 2.1: on the left: preparation of a gel cap, on the right: after the experiment

pressure upon the scalp via an array of gold-plated contacts which do not cause discomfort to the users as reported by our experimental subjects. The sparse electrode arrangement and slightly reduced 'dry' signal quality places the onus on robust signal processing for effective BCI.

The advent of machine learning in the field of BCI has led to significant advances in real-time EEG analysis. While early EEG-BCI efforts required neurofeedback training on the part of the user that lasted on the order of days [9] in current practice it suffices to collect data in which the patient is cued to perform one of a small set of mental tasks called classes. After setup and less than 30 minutes [12] of training data collection, a classification algorithm analyzes brief recordings and learns to discriminate mental tasks in less than 5 minutes of computation time, thereby relocating adaptation from the user to the computer. Robustness of BCI decoding algorithms, re-use of classifiers [72] and artifact removal have benefited from significant research effort [35].

As already discussed in Section 1.4, successful EEG analysis requires both temporal (filtering) and spatial (source-localizing) decomposition. The current Berlin Brain Computer Interface consists of a heuristic search of EEG frequency bands and time intervals which maximize class discrimination, as a temporal decomposition step. It is followed by an automatic, signal driven source localization algorithm termed Common Spatial Pattern (CSP) [69, 35] which correlates spatial activity within a class while concurrently discriminating this correlation pattern from that of another class. The final step is an algorithm which performs automatic discrimination (i.e. classification) based on features generated by the spatio-temporal decomposition. As has been shown [12], the frequency bands chosen, the time intervals and the spatial patterns are consistent with known neurophysiology of movement imagination, provide excellent discrimination and, as shown in this study, work well despite noise in the signal and sparse recording sites. Furthermore, the

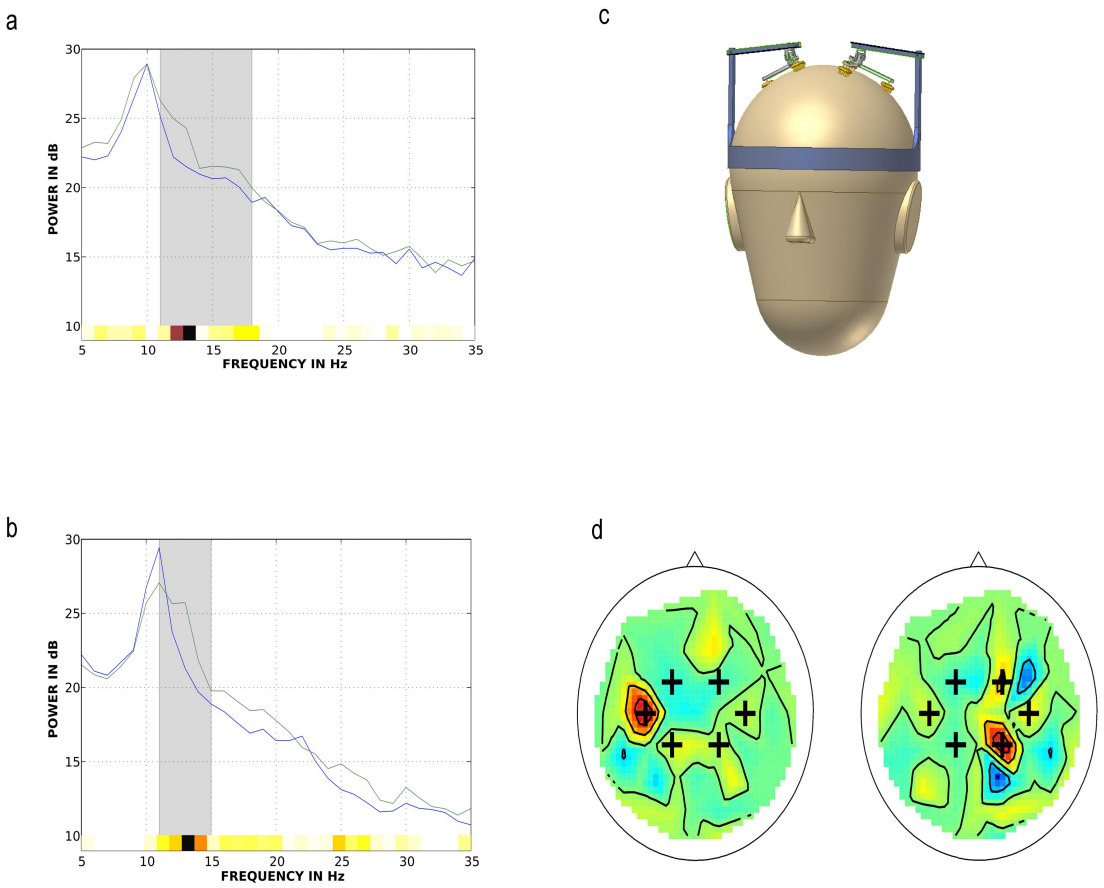

Figure 2.2: Signal spectra and electrode placement: a) Typical signal spectrum from proposed dry electrode (each trace corresponds to averaged spectra for each class). b) Comparable signal from conventional electrode with electrolyte gel (same subject, same conditions). c) Illustration of dry cap. d) Contralateral CSPs of left/right classes from full cap and location of 6 dry cap electrodes.

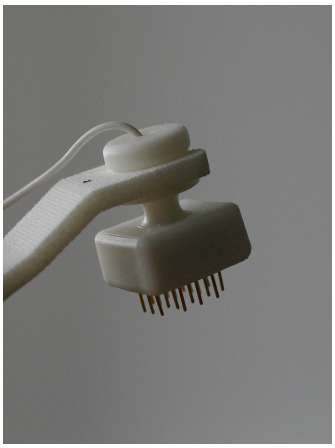

Figure 2.3: dry electrode prototype

analysis method required in order to maximize information gain from EEG, as evidenced by our investigative study, can be both straightforward and effective.

2.1 Development of dry electrode EEG cap prototypes

The impedances of dry electrodes are significantly higher than those of wet ones. Ensuring functionality of dry electrodes depends critically on the contact they make with the scalp surface. If impedances of individual electrodes are similar in magnitude, the external noise sources, that rise linearly with the impedances on individual electrode levels, can be minimized significantly by referencing electrodes with a common reference. This effect is due to the characteristics of the external measurement noise. It is instantaneous and global. It therefore cancels out completely by referencing, given that impedances are the same. While this assumption does not hold true completely, it is of vital importance that all electrodes maintain constant contact with similar pressures, while at the same time not hurt the wearer of the cap. Since the metal electrodes have sharp edges as can be seen in Figure 2.3 a complex mounting was necessary to be designed. A triangular arrangement of electrodes was found and combined a number of joints at various positions.

The first prototype fulfilling these requirements can be seen in Figure 2.4. A second, more advanced version was designed in collaboration with *fast part GmbH, Berlin, Germany*, as can be seen in Figure 2.5. The design and dry electrode technology resulted in an international patent [108], two journal publications [107, 55] and was demonstrated at the conference of *Neural Information Processing Conference* in 2009.

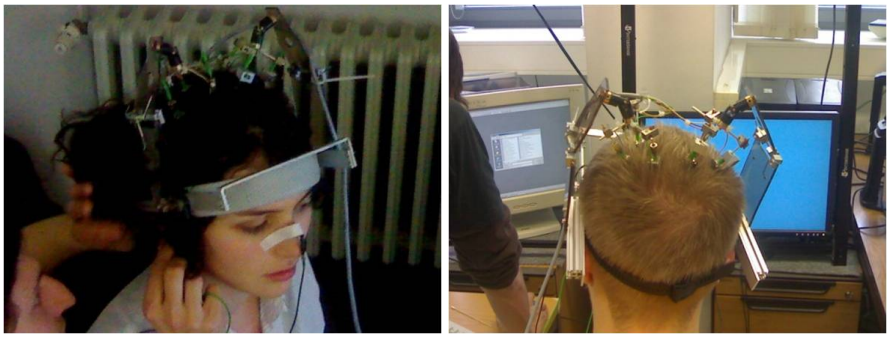

Figure 2.4: First prototype of the dry electrode cap

Figure 2.5: Second prototype of the dry electrode cap

2.2 High Speed BCI with dry electrodes

The results of our 1D cursor control paradigm [13], previously run with a full (64 gel electrode) cap [12], was repeated in this study such that dry cap performance could be compared for the same subjects. 5 healthy subjects (4 male, 1 female) participated. Two subjects were initially tested, however due to particularly thick and full hairstyle no continuously stable signal could be extracted, and thus they were excluded from the study. For 3 of the 5 selected subjects the previously collected data was used, while for the other 2 the paradigm was reproduced. All subjects were volunteers drawn from the members of the laboratory, and all had prior experience with the paradigm. As it was judged that through the use of dry electrodes there was minimal increase in physical, psychological and social risk to the subjects no further ethics board approval was needed than that already in use for gel electrodes (Charité - Universitätsmedizin Berlin Ethics Commission). As per our standard EEG procedure, which may involve skin preparation, in the unlikely case of a minor scratch, disinfectant and a first-aid kit were on hand. Subjects were instructed to end the session if they felt any discomfort. No injury of any kind occurred and no serious discomfort was reported. The subjects gave verbal consent to the eventual dissemination of results and are identified by randomized initials herein.

While EEG cap setup normally requires an attendant and about 30 minutes of preparation, the dry cap can be simply placed on the head and manually adjusted even by the subject herself in less than 2 minutes. For the 'dry cap' experiments a 14-channel DC amplifier set-up (BrainAmp128DC, Munich, Germany) was used (6 EEG channels and 4 bipolar artifact measure channels). In the first part of the experiment ('calibration session'), a sequence of 80 left/right cues was presented visually by means of a letter which appears in the middle of the computer screen. The subjects were asked to imagine the cued class without moving either limbs or the eye. All subjects used left/right hand movement imagination except one subject who used left hand/ right foot imagination since the earlier study [12] predicted this combination to be optimal for that subject. The cues were presented for 3.75 seconds with an inter-cue relax interval of 1.75 ± 0.5 seconds. Electro-oculo-grams (EOG) were measured using 2 standard (gel) electrodes per eye (one lateral to each eye, one above the left eye, one below the left eye) the difference between each pair being amplified as to obtain vertical and horizontal components, while surface bipolar electromyogram (EMG) electrodes where placed on the Flexor Carpi Radialis. As one of the subjects used right foot imagination for one class EMG was measured from the Gastrocnemius. Apart from off-line checks, electromyograms are monitored online and the maximal co-contraction EMG level recorded: no trials were excluded. The average dry electrode impedance measured was 78.6 ± 30.0 $K\Omega$.

The dry cap BCI system was thus ready for use after roughly 15 minutes: 2 for electrode preparation, 8 for calibration data collection and 5 minutes for the classifier algorithm to learn from the calibration data. For habitual use, calibration could be eliminated and classifiers reused [72]. In a second part of the experiment (*feedback session*) subjects were asked to move a dot displayed on the screen to a target represented by a bar on either the right or left side of the screen by imagining the corresponding class. The dot movement provided continuous performance feedback to the subjects. Each subject performed 400 trials divided into 4 sets allowing him/her a brief pause for mental relaxation.

A semi-automatic search was performed for the estimation of the event-related desynchronization (ERD) time interval and for the frequency band whose power discriminates most between classes. For each subject the heuristic generally selects the so-called mu- and beta- rhythms ($8-25$ Hz, Figure 2.7 a,b) in the motor cortex [12, 102]. The discriminating frequency band search determined a band-pass filter, which attenuated signal amplitude outside these bands and thereby accomplishing a temporal *demixing*.

The resulting filtered multivariate signals, segmented in the ERDs time interval, are used to compute two covariance matrices Σ_1 and Σ_2 from the calibration data. These are then fed to the CSP algorithm (see Section 1.4.3.5).

The decomposed time-varying multivariate signal y(t) can be easily transformed into an n-vector of log-variances, by estimating $\tilde{y} = \ln\left(\mathrm{var}\left(y(t)\right)\right)$ over a desired time window. The elements of this vector are the *features* that the classifier learns to associate with a given class. The classifier used was Linear Discriminant Analysis (LDA), which assigns linear weights to features as to provide a separating hyper-plane between classes in feature space. In the *feedback sessions* the time window length used was adjusted to subject preference for cursor responsiveness and ranged from 600 to 1000 msec. The speed of the cursor is proportional to the continuous linear weighted sum of features as computed by the LDA output.

In order to rule out that the reported ITRs are due to muscle artifact, we analyze whether a classifier based on EOG or EMG alone could achieve a significant ITR. For this, unfiltered EOG and EMG signals were segmented into 5 windows, each 500 msec long, starting after cue presentation for feedback data. The log variance of these segments provided features (i.e. 5 segments of 2 EOG resp. EMG channels = 10 features) that were classified by LDA in a leave-one-out fashion, i.e. each segmented feedback trial is labeled by a classifier trained on all other trials.

2.3 Online BCI feedback results with dry electrodes

The main object of the study was to compare the Information Transfer Rate (ITR) obtainable with the dry cap with that previously established for the full cap for an existing paradigm using the same subjects. Classification results are summarized in Table 2.6. *Feedback - Gel Cap* (top) reports feedback data from an earlier study [13]. The first line shows the bit/min information transfer rate of 1D cursor control averaged over 8 sessions consisting of 25 trials each. The second line gives the average time per trials and the peak performing session result. *Feedback - Dry Cap* (middle) as above. Note that here 4 sessions of 100 trials each were evaluated. Also the peak performance was computed as the best 25 consecutive trials. The lower part (bottom) of the table summarizes the relative loss in performance of the respective setups for the subjects. A negative sign indicates lower performance of the dry electrode cap. *% of MVC* stands for the power of feedback trials, as compared to the maximum voluntary contraction (MVC). *EMG-fb* stands for the EMG activity in the activity in the actual feedback trials, as compared to the preparatory phase of each feedback trial, *EMG-pre*.

The locations of the 6 channels used were determined with the aid of a sensitivity analysis on full cap data similar to [41]. After a CSP matrix W is calculated, the row with the lowest sum of absolute values is labeled as the least-significant channel in terms of classification. After elimination of this channel from further analysis, the entire CSP/LDA classification procedure can be re-run. By performing channel elimination iteratively, we can approximate the expected error for any *best* $m < n$ channels and derive a relative ranking of channel relevance (see Figure 2.7).

While subject experience and proper instruction can alleviate the confounding role of EMG and EOG by encouraging performance in which no such activity can be detected (2 of the subjects had no detectable artifact) in most subjects, artifacts are unavoidable as they are involuntary in nature. The results in Table 2.6 (lowest part) show that classification based on EOG/EMG is either close to chance level, or much less accurate than the classification based on EEG. Furthermore note that in trials in which EOG or EMG analysis erred in classification, EEG can still be consistently classified with the same accuracy as in other trials.

With only 6 dry electrodes approximately placed above the motor cortex (Figure 2.2 d), the information transmission rate achieved a peak of 36.5 bits/min (on par with any EEG-BCI performance reported) and is on average 30.8% slower than previous experiments with 64 wet electrode caps on the same subjects.

Despite its simplicity the CSP algorithm and extensions thereof [35, 79] remains among the highest consistent performers among the many EEG-BCI analysis techniques developed and attempted [16]. For general scientific interest, a BCI algorithm needs to do more than simply show a high ITR. Critical is the identification

Subjects	al	zg	ay	zk	aw	Average
Feedback – Gel Cap						
1D (bit/min)	24.4	13.0	22.6	8.8	5.9	14.9
correct (%)	98.0	98.0	95.0	86.8	80.5	91.7
time/trial (s)	2.1	3.9	1.9	3.0	2.9	2.8
peak (bit/min)	35.4	19.6	31.5	23.4	11.0	24.2
Feedback – Dry Cap						
1D (bit/min)	17.6	3.4	14.1	7.9	5.0	9.6
correct (%)	91.8	79.2	94.8	84.5	83.8	86.8
time/trial (s)	2.0	4.7	3.1	2.9	4.4	3.4
peak (bit/min)	36.5	14.0	25.0	23.1	16.8	23.1
Percentage difference Gel Cap – Dry Cap						
1D (%)	−27.8	−63.4	−37.6	−10.2	−15.2	−30.8
correct (%)	−6.3	−19.1	−0.2	−2.6	3.9	−4.9
time/trial (%)	4.7	−18.1	−38.7	−4.0	−34.1	−18.0
peak (%)	3.0	−28.4	−20.6	−1.3	34.5	−2.6
Feedback Classification Accuracy EEG-EOG-EMG						
EEG (%)	91.8	79.2	94.8	84.5	83.8	86.8
EMG (%)	72.3	47.5	52.2	61.1	85.8	63.8
EEG (% on EMG-)	90.4	78.7	94.3	83.5	89.9	87.4
EOG (%)	72.8	49.0	55.1	58.5	80.6	63.2
EEG (% on EOG-)	91.2	76.4	95.5	85.1	88.4	87.3
EMG (% of MVC)	2.7	1.2	1.7	1.3	0.7	1.5
EMG-fb (% of EMG-pre)	107.9	102.5	98.1	103.0	109.4	104.2

Figure 2.6: Results of feedback sessions for dry vs. full cap.

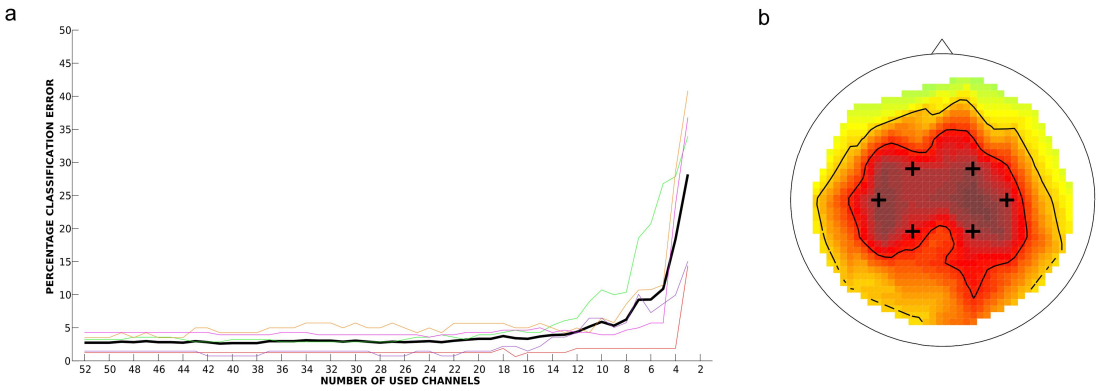

Figure 2.7: Relationship of ITR to number of electrodes and position: a) Predicted error rates vs. number of channels for different subjects (colored lines) and average (black line). b) electrode importance ranking averaged across subjects, plus dry cap electrode placement.

and description of the physiological origin of signal that provides for discrimination. It would be useful to perform *EEG source localization*, i.e. a spatial de-mixing of the signal which provides for electrical dipole locations back-calculated from the recorded signal. Using algorithms designed for this particular purpose, it has been shown that motor imagery based BCI does indeed localize to the motor cortex [137]. Although source localization from only 6 channels of recording cannot be done without an unacceptable loss in accuracy, we had full-cap data from the same paradigm at our disposal.

Interestingly, the CSP algorithm was originally conceived to be a signal-driven source localization technique which can locate known dipole sources [69]. As such, the primary CSP patterns of the full-cap data for left- and right- classes do indeed show highest sensitivity around the contra-lateral motor cortical areas (compare Figure 2.2 and Figure 2.7) as would be expected from basic motor neurophysiology. Further evidence is gained by simply asking the question: if we only had m electrodes available, where should they be placed in order to maximize classification? We performed a sensitivity study where the electrode that least contributed to the CSP-based classification was iteratively removed from the analysis. Results are shown in Figure 2.7. The *best* placement for electrodes varies from subject to subject but is fundamentally fronto-parietal and bilateral (i.e. above the motor cortical areas). Note also that for at least one subject the expected 6-channel performance is low, as was confirmed in the dry cap experiment. Since potentials propagate perpendicularly from the folded cortical surface, varying anatomy and cranial electrical properties among subjects means that one cannot just place electrodes 'above the motor cortex' and expect maximal performance. However, our study does show

that such a simplifying strategy works surprisingly well, based on a ranking of electrode location relevance (see Figure 2.7) averaged across subjects. Individualized electrode placement will likely improve performance, but not without considerable cost, however. Further technical development of the electrode design - and specialized research - may also be necessary in order for the recording pins design to improve in such a manner that they bypass all hair-types and make consistent contact with the scalp. The subjects tested were not chosen with any such criteria in mind and good results were obtained from 5 out of the first 7 people tested.

EEG analysis, whether it is classification or localization, can be compromised by EOG and EMG even if these are produced involuntarily. However, arm muscle activation or bodily movement must be considerably large in order to affect EEG [33, 20]. In our experiments, no movement is visible and measured hand EMG magnitudes averaged 1.5% of maximum voluntary contraction (MVC). Note that this is not necessarily phasic activity but mostly tonic co-contraction. EMG levels during cue presentation (i.e. movement imagination) are from -1.9% to 9.5% greater than EMG levels during the brief rest period between trials. A look at the last rows of Table 2.6 shows that EMG classification accuracy correlates with the magnitude of this difference (on the order of 0.15% of MVC) rather than the overall EMG magnitudes. Being based on overall differences so slight, EMG affords significantly poorer classification than EEG.

EOG represents mainly ocular muscle activity but can also partially reflect facial, tongue and jaw muscle activity. As EOG electrodes are closer to the scalp than EMG electrodes, their activity, even if moderate, is more likely to represent an artifact in EEG. The EMG/EOG classifiers operated on feedback trial data and not calibration trial data, may have contained other types of eye movement patterns due to the absence of visual target presentation.

Prior analysis of artifact influence in BCI experiments has shown that the type of movement can be determined earlier and more accurately in EEG than in EMG/EOG [14]. That EEG, in this study, still indicates mental states in trials and subjects in which artifact, whether EMG or EOG, cannot discriminate the mental class further reinforces the idea that the classifier responds mainly to cortical activity patterns, in a physiologically expected location and frequency range.

2.4 Bristle sensors

In 2011 a study was published by Grozea at al. [55], where a novel dry electrode is introduced and coined *bristle-sensors*. These electrodes consist of metal-coated polymer bristles, as can be seen in Figure 2.4. As already stated, good physical contact between electrodes and the scalp leads to low impedances and is of paramount im-

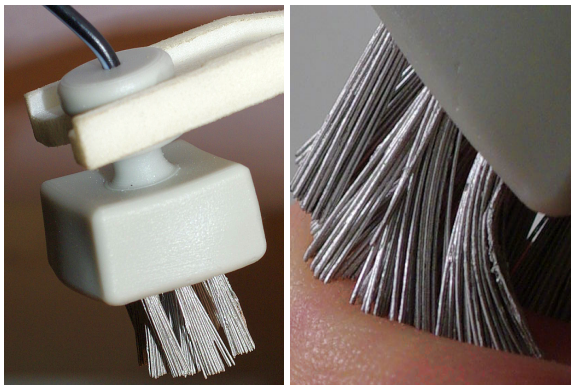

Figure 2.8: On the left: bristle sensor prototype. On the right: Flexibility of the bristles. The figure is reproduced from Grozea et al. [55].

portance for the quality of the EEG signal during acquisition. However, at the same time user discomfort needs to be kept at a minimum. Due to the flexible nature of the bristle-sensors, the pressure they exert is distributed uniformly and therefore the reliability of the contact is also increased.
Classical wet electrodes were measured in close proximity to the dry electrodes. A typical sample of time domain simultaneous recordings can be seen on the left part of Figure 2.9. Furthermore a number of standard EEG paradigms were tested to investigate the signal quality: The α-rhythm of the occipital cortex was recorded and compared for *eyes open* and *eyes closed* conditions. Its grand average of four subjects is depicted on the right panel of Figure 2.9. A standard auditory oddball paradigm was performed. The N100 as well as P300 components were stimulus aligned, and baseline corrected. One-sample t-tests revealed highly significant p-values in the range of 10^{-12}-10^{-2} for the individual subjects.
Finally a small survey was conducted among the participants of the study and most of them agreed that the bristle-sensors are more comfortable to wear, as compared to the pin-based electrodes introduced earlier.

2.5 Conclusions

The implications of dry electrode technology are significant, both in terms of practicability of non-invasive BCI for the severely disabled and in terms of a robust, affordable brain imaging technique for long-term neuroscience experiments (some sessions lasted over 5 hours). Clinical applications may include daily EEG monitoring for epilepsy or narcolepsy. Regarding healthy subjects, dry-electrode BCI opens a more practical outlook for Human-Machine Interaction, for monitoring alertness,

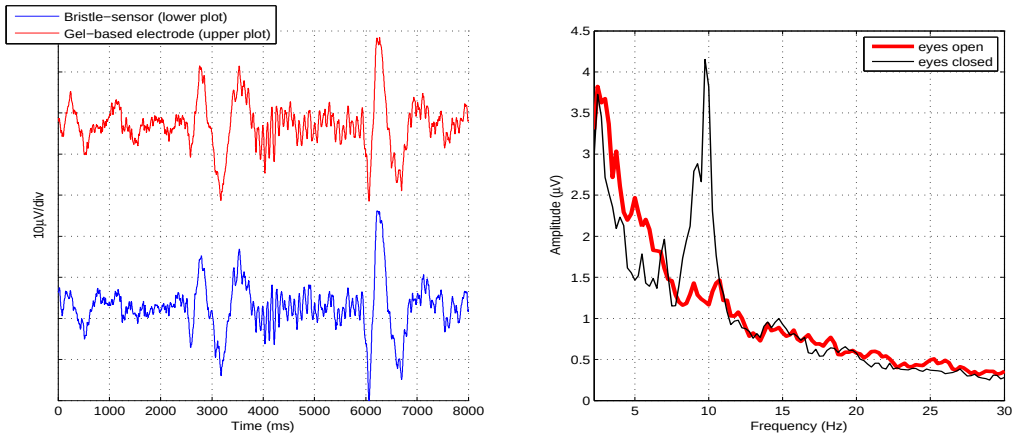

Figure 2.9: Left: Signal quality assessed by direct comparison with simultaneously recorded signal with gel-based electrodes: Sample time domain signal accompanied by signal from a gel-based electrode on a neighbor location, after bandpass filtering 1 to 45 Hz; Alpha rhythm is visible from $t = 4000ms$ to $t = 6000ms$. Right: Spectra of the EEG signal recorded with the prototype for the eyes open/eyes closed conditions, averaged over all subjects with available data. It shows a peak at 10Hz for the eyes closed condition. The figure is reproduced from Grozea et al. [55].

emotion or mental workload.

Here the attempt was made of maximizing the practical value of BCI from the fewest number of recording channels possible. The scientific implications of this approach are that by careful analysis and electrode placement effective functional imaging of the awake, active brain can be achieved non-invasively and in a fairly simple, cost and time-effective manner. Dry electrodes may be sparsely placed elsewhere on the scalp as to focus on other cortical areas that are not motor-related.

The state of current EEG-BCI research makes use of electrophysiological phenomena that contribute to accurate discrimination among mental states in single trials. Miniaturization of EEG equipment as well as the wearability and convenience of novel EEG systems will be a vital factor in determining whether EEG-based related BCI technology will be accepted by the wider community and thus gain widespread use. Future research will focus on further improvements of EEG sensor and data analysis technology and strive towards simple devices that learn to adapt to a user or patient and allow communication even in highly noisy and non stationary real world scenarios.

CHAPTER 3

Ensemble Methods for BCI

Classical BCI-systems relied on subject-training or *operant conditioning* [8],[39]. As discussed before, lately machine learning methods have been introduced for BCI and greatly helped in reducing subject-training [11, 91]. Finding accurate subject-dependent temporal and spatial filters is of paramount importance for achieving high information transfer rates in ERD related BCI systems. Very recently the reuse of old CSP-patterns of expert BCI subjects has been realized [72] and thus feedback sessions without a preceding calibration session could be started. However, for naïve users there is still the need of a calibration session to estimate parameters for spatial filters, temporal filters and classifiers.
In this Chapter we will explore how ensemble learning can assist in estimating suitable classifiers for BCI. The following Chapter is split into two parts. Part one *Ensemble Methods for subject-dependent BCI* (Section 3.2) shows how ensemble methods can help in the estimation of temporal filters for subject-dependent classifiers. The second part *Ensemble Methods for subject-independent BCI* (Section 3.3) shows that it is possible and feasible to use an ensemble methods based approach to obtain a *subject-independent* classifier by formulating an optimization problem that can be solved by various regression and classification methods. We show that ℓ_1-regularized regression and ℓ_1-regularized linear mixed effects models (Section 3.3) are a good choice to fulfill this task.

However before we take a closer look at the two main parts of this Chapter, the two large datasets, which are exploited for this endeavor are introduced.

3.1 Available Data and Experiments

We consider two different sets of BCI data and through both datasets, different aspects of our approach will become apparent.

The first (dataset A) consists of 83 BCI experiments (sessions) from 83 individual subjects, where each session consists of 150 trials. This results in a total of 12450 trials. Our second dataset (dataset B) consists of 90 sessions from 44 subjects. The number of trials of a single session varies from 60 to 600 trials. Table 3.1 gives the exact details of how the trials are distributed within sessions (experiments) and sub-

number of datasets/subject	1	2	3	5	8	9	13
occurance	32	5	3	2	1	1	1
percentage [%]	39.0	11.0	11.6	12.5	6.8	8.0	11.1

Table 3.1: The first row gives the numbers of experiments that exist for a single subject, while the second row shows how often this occurs. Third row shows percentage of trials in that category.

jects. In other words, our first dataset can be considered to be *balanced* in the number of *trials per subjects* and *sessions per subject*. Our second dataset is *unbalanced* in this sense.

As one may expect, the balanced dataset makes it easier to build a zero-training classifier, since not only we do not need to correct for the uneven number of trials per subject but also because we have a larger base of subjects. That enables us to obtain a 'clean' model. However, the unbalanced dataset enables us to examine how individual sessions of the same subject affect the estimation of our model and leads to a more thorough understanding of the underlying processes.

Each trial consists of one of two predefined movement imaginations, being left and right hand, i.e. data was chosen such that it contains only on these two classes, although originally three classes were cued during the calibration session, being left hand (L), right hand (R) and foot (F). 45 EEG channels, which are in accordance with the 10-20 system, were identified to be common in all sessions considered. The data were recorded while subjects were immobile, seated on a comfortable chair with arm rests. The cues for performing a movement imagination were given by visual stimuli, and occurred every 4.5-6 seconds in random order. Each trial was referenced by a 3 second long time-window starting at 500 msec after the presentation of the cue. Individual experiments consisted of three different training paradigms. The first two training paradigms are visual cues in form of a letter or an arrow, respectively. In the third training paradigm the subject was instructed to follow a moving target on the screen. Within this target the edges lit up to indicate the type of movement imagination required. The experimental procedure was designed to closely follow [13]. Electromyogram (EMG) on both forearms and the foot were recorded as well as electrooculogram (EOG) to ensure there were no real movements of the arms and that the movements of the eyes were not correlated to the required mental tasks.

3.2 Ensemble Methods for subject-dependent BCI

Various ways of choosing the temporal filters have previously been proposed. It could be set globally, for example within the α- or β-band. However, if one examined the calibration data of individual subjects, one would find that the frequency ranges at which the most significant differences occur, vary from subject to subject and that finding the exact frequency ranges would enhance overall classification rates significantly (see also Figure 3.2). To this end automatic heuristics have been developed recently to optimize this task (see Algorithm 1) [11, 91].

In the following approach, we employ a multi-classifier system (MCS), based on a predefined filter-bank of temporal filters and apply it to dataset B, that consists of 90 sessions from 44 subjects. As previously stated it comprises 2-class experiments consisting of left and right hand movement imaginations. Our results indicate that our novel approach is a superior alternative to existing methods since it is inherently immune to overfitting and achieves a highly competitive performance [42].

3.2.1 Methods

3.2.1.1 Selection of a Frequency Band

Before introducing the novel, ensemble based approach, we would like to briefly review a popular heuristic, which has proven to be very useful in detecting the most discriminant frequency range for given subjects, if enough calibration data is at hand [18]. Its pseudo code is given by Algorithm 1. However, a less formal description is given here:

1. Use Laplacian or bipolar channels, from motor cortex related electrodes
2. For each trial, channel and frequency in the range from 7 to 35 Hz, calculate the log-bandpower
3. Calculate the correlation coefficient between the log-bandpowers and their true labels
4. Find the frequency with the highest correlation coefficient and broaden the band step-wise in both directions, until the next frequency bin is smaller than 5% of the peak

Note, that the algorithm works best if only few channels are used. A good choice is, e.g., to choose $C = \{c_1, c_2, c_3\}$ with c_i being one from each motor-related areas of the *left hand*, *right hand* and *foot* with $\max \sqrt{\sum_f (\text{score}_c(f))^2}$ [18].

Require: Let $X_{(c,i)}$ denote trial i at channel c with label y_i and let C denote the set of channels.

$\mathrm{dB}_c(f,i) \leftarrow$ log band-power of $X_{(c,i)}$ at frequency f (f from 5 to 35Hz)

$\mathrm{score}_c(f) \leftarrow \mathrm{corrcoef}\big(\mathrm{dB}_c(f,i),\, y_i\big)_i$

$f_{\max} \leftarrow \mathrm{argmax}_f \sum_{c\in C} \mathrm{score}_c(f)$

$$\mathrm{score}^*_c(f) \leftarrow \begin{cases} \mathrm{score}_c(f) & \text{if } \mathrm{score}_c(f_{\max}) > 0 \\ -\mathrm{score}_c(f) & \text{otherwise} \end{cases}$$

$\mathrm{fscore}(f) \leftarrow \sum_{c\in C} \mathrm{score}^*_c(f)$

$f^*_{\max} \leftarrow \mathrm{argmax}_f\, \mathrm{fscore}(f)$

$f_0 \leftarrow f^*_{\max};\ f_1 \leftarrow f^*_{\max}$

while $\mathrm{fscore}(f_0 - 1) \geq \mathrm{fscore}(f^*_{\max}) * 0.05$ **do**

    $f_0 \leftarrow f_0 - 1$

while $\mathrm{fscore}(f_1 + 1) \geq \mathrm{fscore}(f^*_{\max}) * 0.05$ **do**

    $f_1 \leftarrow f_1 + 1$

return frequency band $[f_0, f_1]$

Algorithm 1: Selection of a discriminative frequency band, reproduced from [18]

Filter bank Two idle rhythms of the postcentral somatosensory and precentral motor cortex, namely the μ-rhythm (9-14 Hz) and beta band (16-22 Hz) are found in healthy adults. Preparation of movements or mere imaginations of those can lead to suppression of the idle rhythms contralaterally [100]. On an individual subject level, the μ and β-rhythms have different modulation frequency ranges as well as differing (de-)synchronization strengths. Machine learning techniques have been shown to be a viable approach in finding optimal subject-dependent temporal filters. Here we a present a different approach, where we generate a filter-bank, consisting of 9 different band-pass filters (see Figure 3.1). The temporal filters are designed by using prior neurophysiological knowledge and experience from BCI experiments: While most subjects show optimal performance with μ-rhythm temporal filters, some show optimal performance with β-band filters. The choice of the individual filters in our filter-bank approach reflects these considerations. For each temporal filter we calculate a spatial filter and process the data in parallel as can be seen in Figure 3.2.

3.2.1.2 Setup of the ensemble

3.2.1.3 Validation

Each dataset is split into two chronological halves. A chronological split of the data is proposed, since this represents a setup that is very similar in nature to an actual experiment, where first a training session is obtained and then used to optimize

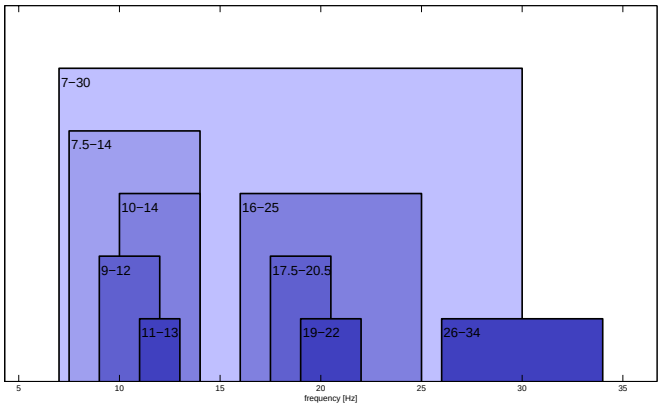

Figure 3.1: Frequency ranges of all temporal filters, used in the ensemble.

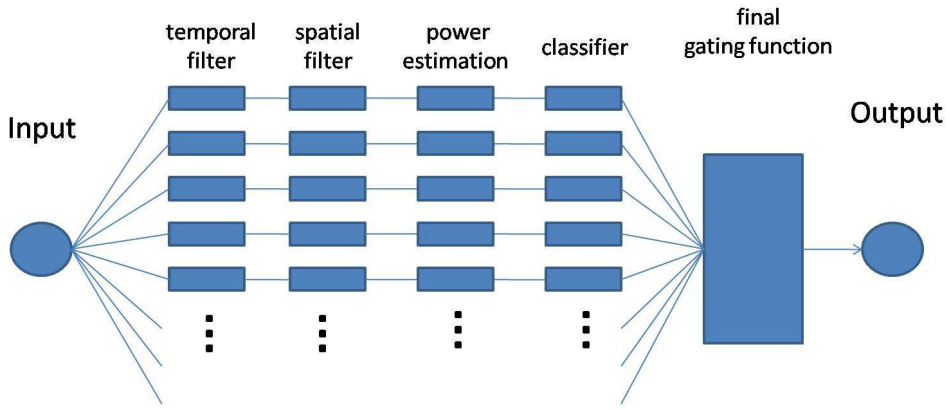

Figure 3.2: The movement imagination data of a given subject is processed by predefined temporal filters in parallel, and subsequently filtered by a spatial filter, which was obtained by training data of that subject) and finally classified, once the spectral power has been estimated. The classifier outputs are then combined via a gating function.

frequency [Hz]	test loss [AOC*100]	best performance [%]
7.5 – 14	**12.4**	16.5
11 – 13	19.6	9.9
10 – 14	12.8	**30.8**
9 – 12	18.6	11.0
19 – 22	42.9	1.1
16 – 22	31.8	6.6
26 – 34	46.4	2.2
17.5 – 20.5	41.4	2.2
7 – 30	14.8	19.8

Table 3.2: Summary of the performance of each temporal filter, we chose to include for the ensemble. *test loss* gives the cross-validated median classification loss over all subjects. *best performance* gives the percentage of datasets for which the given frequency band performed best. Note that for seemingly unsuitable filters, some datasets score their best validation loss.

subject-dependent filters to optimize the BCI performance. In other words, the algorithm was trained on the first half and validated on the second. The outputs for a single trial are given as $X \in \mathbb{R}^{d \times t}$, where d is the number of temporal filters and t the number of trials. The ensemble mean $\hat{y}_m = \frac{1}{d}\sum_{j=1}^{d} X$ performs surprisingly well for many ensemble problems in general [106] as well as for the problem we considered here.

3.2.2 Results

Each LDA output for a given trial indicates how far the feature is from the hyperplane and can thus be interpreted as how confident a classifier is. In this sense the weighting of the individual classifiers is already optimal. It is therefore not surprising that the ensemble mean yields the best results, as can be seen from Table 3.3. As can also be seen from Table 3.3, the heuristic performs very well for good subjects, while for subjects, where the bandpower differences are not so well detectable, a broadband CSP performs favorably. Errors are given as area over the curve (AOC) of the receiver operating characteristic (ROC) [143].

3.2.3 Discussion and Conclusions

The principal aim of this work is to make classifier tuning as automatic and fast as possible. In this sense the ensemble method obviates the need of any parameter estimation in the domain of temporal filters.

	$csp_{[7\text{-}30]}$	csp_{auto}	$emean$	$emax$	$emaj$	$emed$
25%-tile	7.2	4.1	**3.6**	6.8	24.1	5.1
median	14.8	15.5	**11.2**	17.3	43.8	11.3
75%-tile	31.7	36.7	**30.7**	32.7	64.9	31.4

Table 3.3: Results for two baselines and four ways to combine the outputs of the ensemble members. Errors are given as [AOC*100].

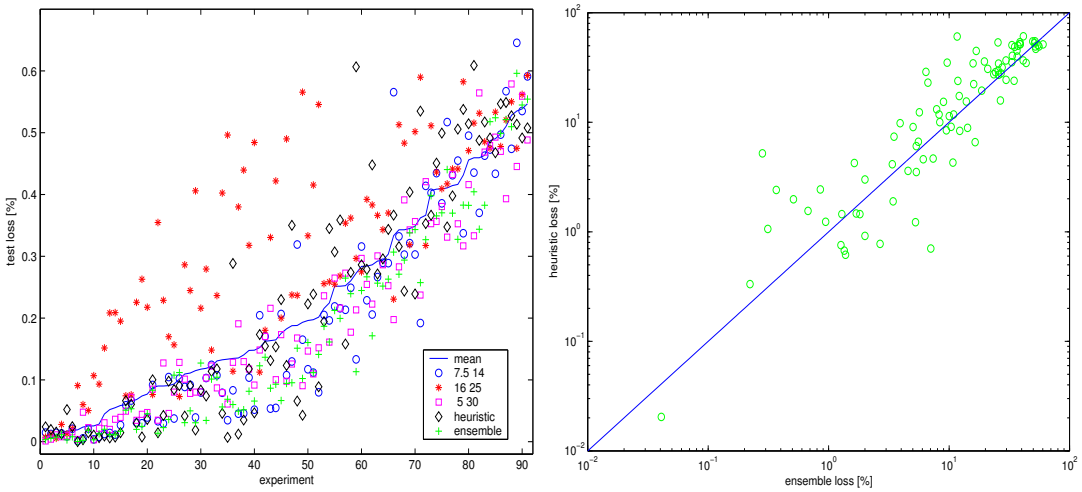

Figure 3.3: The panel on the left shows the resulting loss of 4 different frequency bands, data is sorted by the mean performance of all bands. The panel on the right shows the test loss for each individual experiment for the best ensemble method (mean), versus the classical procedure, with the automatic heuristic.

The results of the ensemble of temporal filters show that that for small numbers of training trials, or for subjects, where the detection of the correct frequency band is difficult, it is possible to improve classification accuracy, since heuristics can fail. By using the ensemble we exploit prior information from neurophysiology and BCI classifier calibration experience and let the ensemble of classifiers decide which band scores the highest confidence at minimal computational cost. While the proposed combination of classifier outputs can be realized in simple and effective manner, it is also less prone to overfitting.

It would be unrealistic to claim that the data presented here can be seen as an unbiased sample of society, as only successful BCI subjects are likely to participate in more than one experiment. However, since most of the BCI community is interested in well performing subjects, the results presented here should be of interest. Furthermore, when possible we look at individual subject performance as well as experiment performance, as to reduce this bias as much as possible.

It remains to be seen, whether by this method, the resulting architecture is more robust to nonstationarities, which may occur over long feedback sessions. Also in the future this could be easily tested by applying the presented method to datasets where non-stationarities are known to exist or by putting the method into practice within a feedback environment.

3.3 Ensemble Methods for subject-independent BCI

A time consuming step in the preparation of a BCI system is the required individualized adaptation to the BCI user, which involves approximately 30 minutes of calibration for assessing a subject's brain signature. Here we aim to also remove this calibration procedure from BCI setup time by means of machine learning. In particular, we harvest a large database of EEG BCI motor imagination recordings (83 subjects, dataset A) for constructing a library of subject-specific spatio-temporal filters and derive a subject independent BCI classifier. Our offline results indicate that BCI-naïve users could start real-time BCI use with no prior calibration at only a very moderate performance loss.

3.3.1 Introduction of ensemble methods for zero training

Modern BCI systems require the recording of a brief calibration session during which a subject conceives a fixed number of brain states, say, movement imagination and after which the subject-specific spatio-temporal filters (e.g. [18]) are inferred along with individualized classifiers [35]. Recently, first steps to transfer a BCI user's filters and classifiers between sessions was studied [72] and a further online-study con-

firmed that indeed such transfer is possible without significant performance loss [74]. In the following sections we will go one step further in this spirit and propose a *subject-independent zero-training BCI* that enables both experienced and novice BCI subjects to use BCI immediately without calibration.

Our offline study applies a number of state-of-the-art learning methods (e.g. SVM, Lasso etc.) in order to optimally construct such one-size-fits-all classifiers from a vast number of redundant features, here a large filter bank available from 83 BCI users. The use of sparsifying techniques specifically tell us what are the interesting aspects in EEG that are predictive to future BCI users. As expected, we find that a distribution of different μ-band features in combination with a number of characteristic common spatial patterns (CSPs) is highly predictive for all users. What is found as the outcome of a machine learning experiment can also be viewed as a compact quantitative description of the characteristic variability between individuals in the large subject group. Note that it is not the best subjects that characterize the variance necessary for a subject-independent algorithm, rather the spread over existing physiology is to be represented concisely. Clearly, our procedure may also be of use apart from BCI in other scientific fields, where complex characteristic features need to be homogenized into one overall inference model.

In the following we present the ensemble learning algorithm, consisting of the procedure for building the filters, the classifiers as well as the gating function, where we apply various machine learning methods. Interestingly we are able to successfully classify trials of novel subjects with zero training suffering only a small loss in performance. Finally we put our results into perspective.

3.3.2 Generation of the Ensemble

The ensemble consists of a large redundant set of subject-dependent common spatial pattern filters (CSP cf. [18]) and their matching classifiers (LDA). Each dataset is first preprocessed by 18 predefined temporal filters (i.e. band-pass filters) in parallel (see upper panel of Figure 3.4). A corresponding spatial filter and linear classifier is obtained for every dataset and temporal filter. Each resulting CSP-LDA couple can be interpreted as a potential basis function.

To give an example, let us consider our *balanced dataset*, which we introduced in Section 3.1. The design matrix X and targets y for the regression are generated as follows: Each trial of each subject is first processed by 18 predefined band-pass filters, CSPs and then linearly classified. Since we have 83 subjects with 18 classifiers each, the total number of features is $18 \cdot 83 = 1494 \Rightarrow \beta \in \mathbb{R}^{1494}$. Each of the 83 subjects performed 150 trials, therefore we have $150 \cdot 83 = 12450$ data points. The data matrix X and the targets y have thus the dimensionalities $X \in \mathbb{R}^{12450 \times 1494}$ and $y \in \mathbb{R}^{12450}$. Note that contrary to the usual use case of ℓ_1-regularization, our regres-

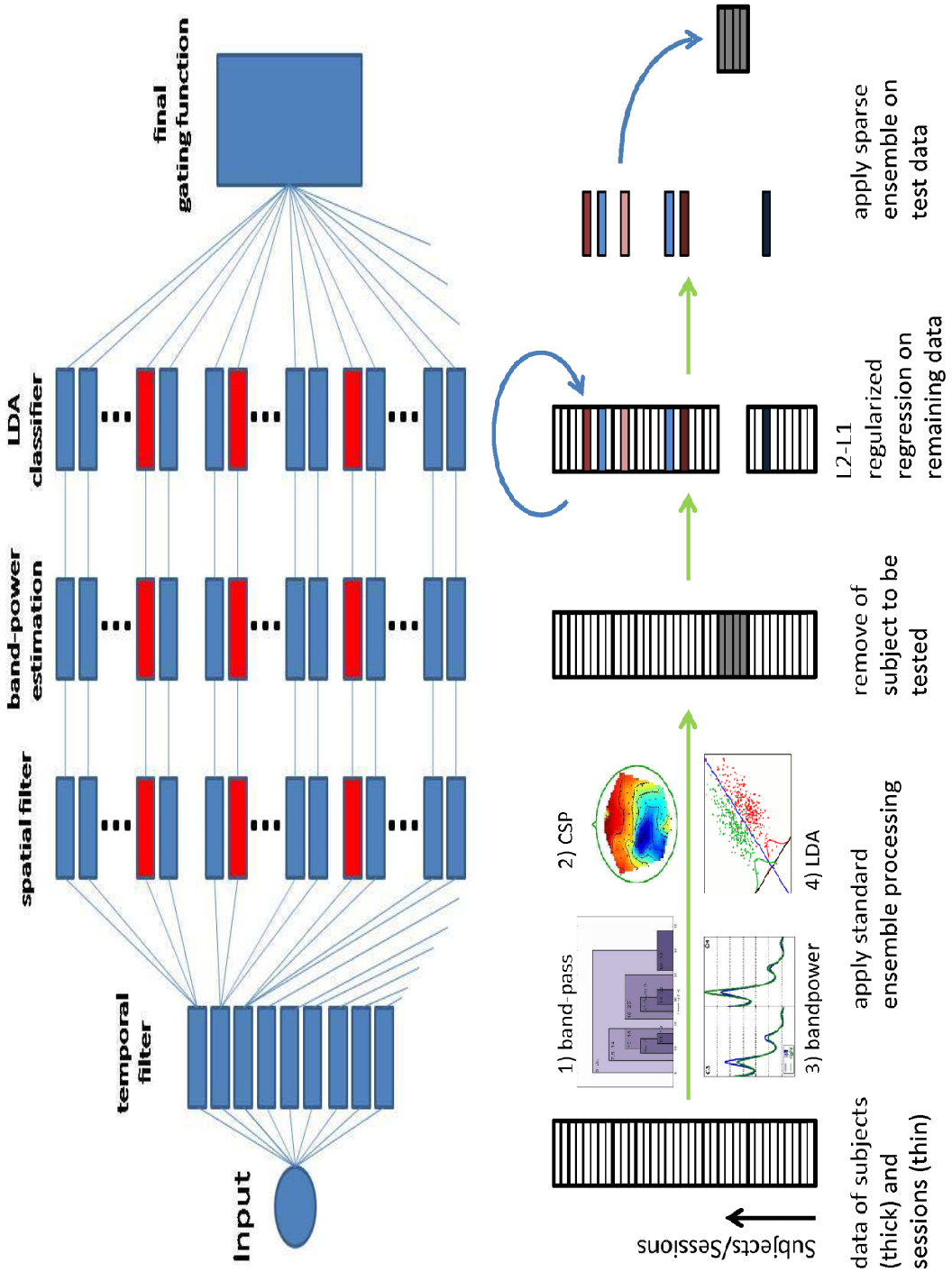

Figure 3.4: 2 Flowcharts of the ensemble method. The red patches in the top panel illustrate the inactive nodes of the ensemble after sparsification.

sion problem is not ill-posed, i.e., in our case, $n > p$.

Finding an appropriate weighting for the classifier outputs of these basis functions is of major importance for the accurate prediction. We employed different forms of regression and classification in order to find an optimal weighting for predicting the movement imagination data of unseen subjects[11, 14]. This processing was done by leave-one-subject-out cross-validation, i.e. the session of a particular subject was removed, the algorithm trained on the remaining trials (of the other subjects) and then applied to this subject's data (see lower panel of Figure 3.4).

3.3.3 Temporal Filters

We identified 18 neurophysiologically relevant temporal filters (see left part of Figure 3.7, of which 12 lie within the μ-band, 3 in the β-band, two in between μ- and β-band and one broadband $7-30$Hz. In all following performance related tables we used the percentage of misclassified trials, or 0-1 loss.

3.3.4 Final gating function

The final gating function combines the outputs of the individual ensemble members to a single one. This can be realized in many ways. For a number of ensemble methods the mean has proven to be a surprisingly good choice [106]. As a baseline for our ensemble we simply averaged all outputs of our individual classifiers. This result is given as *mean* in Table 3.5.

Classification and Regression We employ various classification methods such as k Nearest Neighbor (kNN), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and a Linear Programming Machine (LPM), all of which already introduced in Section 1.3.2.
Furthermore, we also performed classic least squares regression (*LSR* in Table 3.5), as well as quadratic regression with ℓ_1 regularization. For the dataset we consider it can be expressed as

$$\underset{\beta_{ij}^{(k)}}{\operatorname{argmin}} \sum_{x\in X\setminus X_k} \left(h_k(x)-y(x)\right)^2 + \alpha \sqrt{\sum_{i=1}^{B}\sum_{j\in S\setminus S_k}\sum_{x\in X\setminus X_k} c_{ij}(x)^2}\left(\sum_{i=1}^{B}\sum_{j\in S\setminus S_k}|\beta_{ij}^{(k)}|+|b|\right) \tag{3.1}$$

$$h_k(x) = \sum_{i=1}^{B}\sum_{j\in S\setminus S_k}\beta_{ij}^{(k)} c_{ij}(x) - b \quad ,$$

where $c_{ij}(x) \in [-\infty;\infty]$ is the continuous classifier output, before thresholding,

obtained from the session j by applying the bandpass filter i, B is the number of frequency bands, S the complete set of sessions, X the complete data set, S_k the set of sessions of subject k, X_k the dataset for subject k, $y(x)$ is the class label of trial x and β_{ij}^k in equation (3.2) are the weights given to the LDA outputs. The hyperparameter α in equation (3.1) was varied on a logarithmic scale and multiplied by a dataset scaling factor which accounted for fluctuations in voting population distribution and size for each subject. The dataset scaling factor is computed using $c_{ij}(x)$, for all $x \in X \setminus X_k$.

For computational efficiency reasons the hyperparameter was tuned on a small random subset of subjects whose labels are to be predicted from data obtained from other subjects such that the resulting test/train error ratio was minimal, which in turn affected the choice (leave in/out) of classifiers among the 83x18 candidates. The ℓ_1 regularized regression with this choice of α was then applied to all subjects, with results (in terms of feature sparsification) shown in Figure 3.5.

The exemplary CSP patterns shown in the lower part of the Figure exhibit neurophysiologically meaningful activation in motorcortical areas. The most predictive subjects show smooth monopolar patterns, while subjects with a higher self-prediction loss slowly move from bipolar to rather ragged maps. From the point of view of approximation even the latter make sense for capturing the overall ensemble variance.

The implementation of the regressions were performed using `CVX`, a package for specifying and solving convex programs [53]. We coupled an ℓ_2 loss with an ℓ_1 penalty term on a linear voting scheme ensemble.

3.3.5 Validation

The subject-specific CSP-based classification methods with automatically, subject-dependent tuned temporal filters (termed reference methods) are validated by an 8-fold cross-validation, splitting the data chronologically. The chronological splitting for cross-validation is a common practice in EEG classification, since the non-stationarity of the data is thus preserved [35].

To validate the quality of the ensemble learning we employed a leave-one-subject out cross-validation (LOSO-CV) procedure, i.e. for predicting the labels of a particular subject we only use data from other subjects.

3.3.6 Results

The performances of the various ensemble methods as well as a number of baselines are presented in Table 3.5. As a reference method, performances of subject-

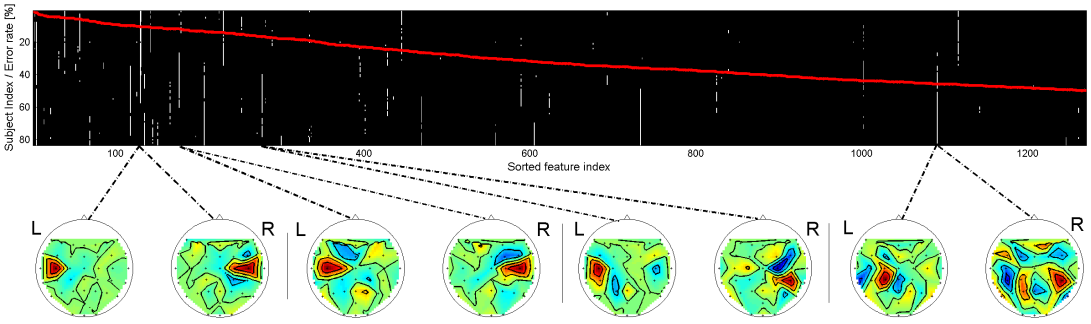

Figure 3.5: Feature selection during cross-validation: white dashes mark the features kept after regularization for the prediction of the data of each subject. The numbers on the vertical axis represent the subject index as well as the Error Rate (%). The red line depicts the baseline error of individual subjects (classical auto-band CSP). Features as well as baseline errors are sorted by the error magnitude of the self-prediction. Note that some of the features are useful in predicting the data of most other subjects, while some are rarely or never used.

specific CSP-based classification with heuristically tuned frequency bands [18] are presented and termed *self*. Furthermore, we considered much simpler (zero-training) methods as a control. *Lap* stands for the power difference in two Laplace filtered channels (C3 vs. C4) and simple band-power (named *BP*) stands for the power difference of the same two channels without any spatial filtering. For these simple zero-training methods we chose a broad-band filter of $7-30$ Hz, since it is the least restrictive and scored one of the best performances on a subject level (for a comparison, please refer to Figure 3.2.

The bias b in equation (3.2) can be tuned broadly for all sessions or corrected individually by session, and implemented for online experiments in multiple ways [74, 114, 73]. In our case we chose to adapt b without label information, but operating under the assumption that class frequency is balanced. We therefore simply subtracted the mean over all trials of a given session. Table 3.4 shows a comparison of the various classification schemes. We evaluate the performance on a given percentage of the training data in order to observe information gain as a function of datapoints. Clearly the two best ML techniques are on par with subject-dependent CSP classifiers and outperform the simple zero-training methods (not shown in Table 3.4 but in Table 3.5) by far. While SVM scores the best median loss over all subjects (see Table 3.4), ℓ_1-regularized regression scored better results for well performing BCI subjects (Figure 3.6 column 1, row 3). In Figure 3.6 and Table 3.5 we furthermore show the results of the ℓ_1-regularized regression and SVM versus the auto-band reference method (zero-training versus subject-dependent training) as well as vs. the simple zero-training methods Laplace and band-power.

	classification				regression	
% of data	kNN	LDA	LPM	SVM	LSR	LSR-ℓ_1
10	31.3	45.3	37.3	31.3	46.0	30.7
20	32.0	40.0	38.0	**28.7**	42.0	31.3
30	32.7	38.7	37.3	33.1	38.0	30.0
40	32.7	36.0	37.9	31.3	36.7	**29.3**

Table 3.4: Main results of various machine learning algorithms.

approach	machine learning							classical		
	zero training									self
method	mean	kNN	LDA	LPM	SVM	LSR	ℓ_1	Lap	BP	CSP
# <25%	31	30	18	14	29	19	**36**	24	11	**39**
25%-tile	17.3	17.3	27.3	26.7	18.7	26.0	16.0	22.0	31.3	11.9
median	30.7	31.3	36.0	37.3	**28.7**	36.7	**29.3**	34.7	38.7	**25.9**
75%-tile	41.3	42.0	43.3	44.0	41.3	44.0	40.7	45.3	45.3	41.4

Table 3.5: Comparing ML results to various baselines.

Figure 3.7 shows all individual temporal filters used to generate the ensemble, where their color codes for the frequency they were used to predict labels of previously unseen data. As to be expected mostly μ-band related temporal filters were selected. Contrary to what one may expect, features that generalize well to other subjects' data do not exclusively come from BCI subjects with low self-prediction errors (see white dashes in Figure 3.5), in fact there are *some* features of weak performing subjects that are necessary to capture all variance of the ensemble. However there is a strong correlation between subjects with a low self-prediction loss and the generalizability of their features to predicting other subjects, as can be seen on the right part of Figure 3.7.

3.3.6.1 Focusing on a particular subject

In order to give an intuition of how the ensemble works in detail we will focus on a particular subject. We chose to use the subject with the lowest reference method cross-validation error (10%). Given the non-linearity in the band-power estimation (see Figure 3.4) it is impossible to picture the resulting ensemble spatial filter exactly. However, by averaging the chosen CSP filters with the weightings, obtained by the ensemble and multiplying them by their LDA classifier weight, we get an approximation:

$$P_{ENS} = \sum_{i=1}^{B} \sum_{j \in S \setminus S_k} w_{ij} W_{ij} C_{ij} \tag{3.2}$$

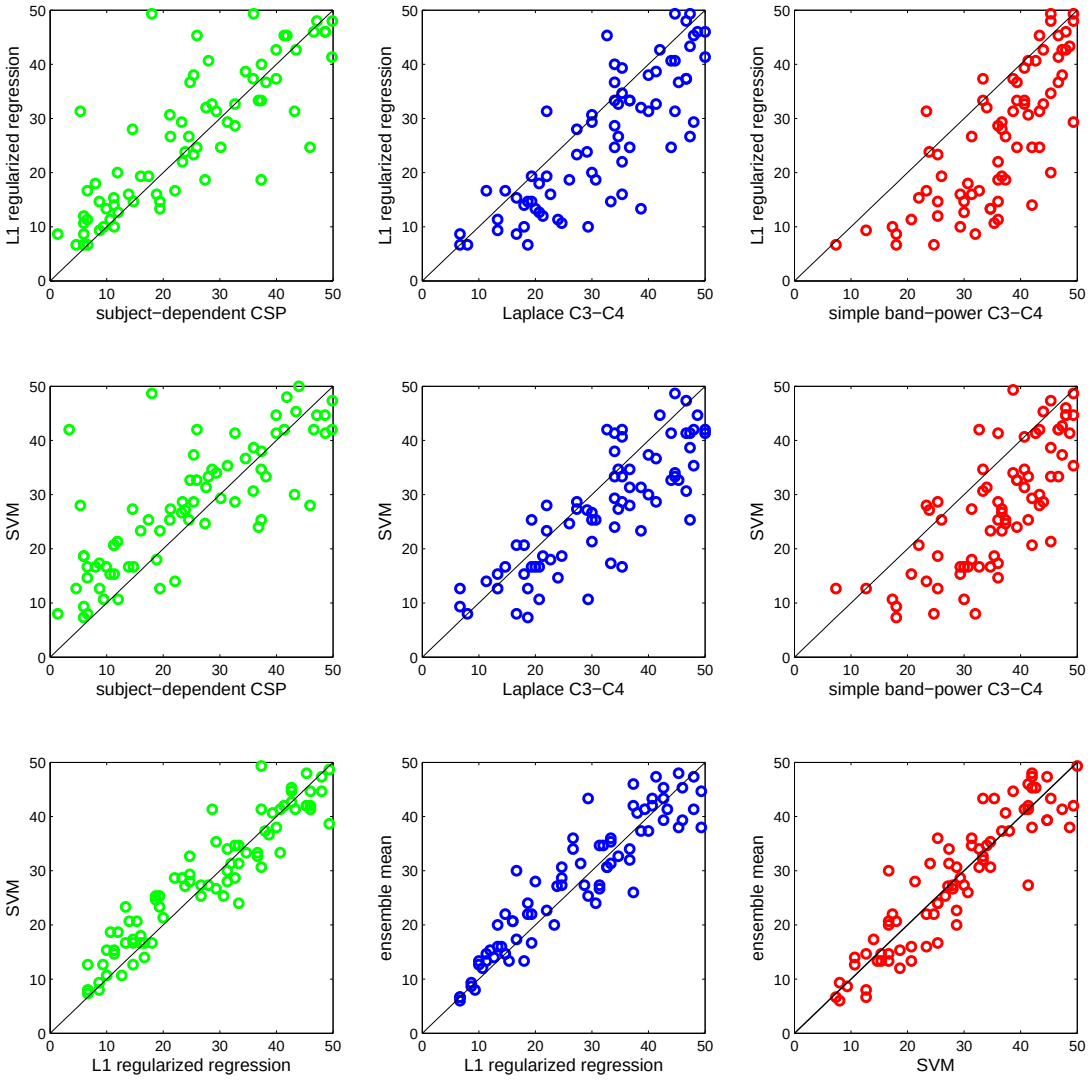

Figure 3.6: Compares the two best-scoring machine learning methods ℓ_1-regularized regression and Support Vector Machine to subject-dependent CSP and other simple zero-training approaches. The axes show the classification loss in percent.

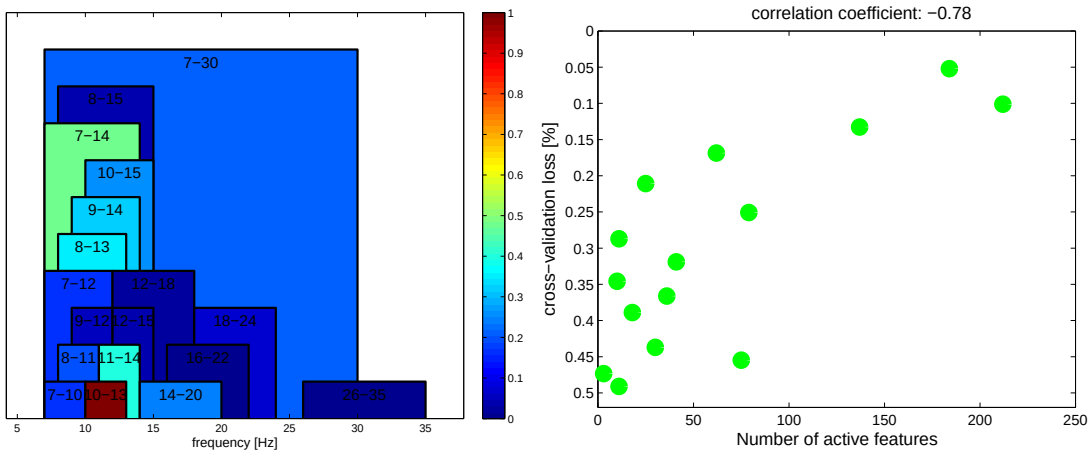

Figure 3.7: On the left: The used frequency ranges of the temporal filters and in color-code their contribution to the final ℓ_1-regularized regression classification (the scale is normalized from 0 to 1). Clearly μ-band temporal filters between $10-13$Hz are most predictive. On the right: Number of features used vs. self-predicted cross–validation. A high self-prediction can be seen to yield a large number of features that are predictable for the whole ensemble.

where w_{ij} is the weight matrix, resulting from the ℓ_1 regularized regression, given in equations (3.1) and (3.2), W_{ij} the CSP filter, corresponding to temporal filter i and subject j and C_{ij} the LDA weights (B in Figure 3.8). For the case of classical auto-band CSP this simply reduces to $P_{CSP} = WC$ (A in Figure 3.8).

Another way to exemplify the ensemble performance is to refer to a transfer function. By injecting a sinusoid with a frequency within the corresponding bandpass filter into a given channel and processing it by the four CSP filters, estimating the bandpower of the resulting signal and finally combining the four outputs by the LDA classifier, we obtain a response for the particular channel, where the sinusoid was injected. Repeating this procedure for each channel results in a response matrix. This procedure can be applied for a single CSP/LDA pair, however we may also repeat the given method for as many times as features were chosen for a given subject by the ensemble and hence obtain an accurate description of how the ensemble processes the given EEG data. The resulting response matrices are displayed in panel C of Figure 3.8. While the subject-specific pattern (classical) looks less focused and more diverse the general pattern matches the one obtained by the ensemble. A third way of visualizing how the ensemble works: we show the primary projections of the CSP filters that were given the 6 highest weights by the ensemble on the left panel (F) and the distribution of all weights in panel D. The spatial positions of highest channel weightings differ slightly for each of the CSP filters given, however the maxima of the projection matrices are clearly positioned around the

primary motor cortex.

In the upper part of Figure 3.9 the outputs of all basis classifiers are applied to each trial of one subject. The top row (broad) gives the label, the second row (broad) gives the output of the classical auto-band CSP, and each of the following rows (thin) gives the outputs of the individual classifiers of other subjects. The individual classifier outputs are sorted by their correlation coefficient with respect to the class labels. The trials (columns) are sorted by true labels with primary key and by mean ensemble output as a secondary key. The row at the bottom gives the sign of the average ensemble output. The lower left part of Figure 3.9 depicts the covariance matrix of all broad-band classifier outputs. The lower right part shows the covariance matrix of the concatenated classifiers of all 9 temporal filters. The classifiers of both covariance matrices are sorted according to their average correlation with all respective other classifiers.

3.3.7 Conclusion

The offline analysis in the previous sections presents evidence that it is possible to generate a *subject-independent* classifier, which enables expert as well as BCI-naïve users to start a feedback session without the necessity of recording a calibration session in advance. We have taken great care in this work to exclude data from a given subject when predicting his/her performance by using the previously described LOSOCV. In contrast with previous work on ensemble approaches to BCI classification based on simple majority voting and Adaboost [134, 21] that have utilized only a limited dataset, we have profited greatly by a large body of high quality experimental data accumulated over the years. This has enabled us to choose by means of machine learning technology a very sparse set of voting classifiers which performed as well as standard, state-of-the-art subject calibrated methods. ℓ_1 regularized regression in this case performed better than other methods (such as majority voting) which we have also tested.

Note that, interestingly, the chosen features (see Figure 3.5), do not exclusively come from the best performing subjects, in fact some average performer was also selected. However most white dashes are present in the left half, i.e. most subjects with high auto-band reference method performance were selected. Interestingly some subjects with very high BCI performance are not selected at all, while others generalize well in the sense that their model are able to predict other subject's data. No single frequency band dominated classification accuracy – see Figure 3.7. Therefore, the regularization must have selected diverse features. Nevertheless, as can be seen in the upper and lower part of Figure 3.9, there is a high redundancy between the individual classifiers of the ensemble. Our approach of finding a sparse solution reduces the dimensionality of the chosen features significantly. For very able

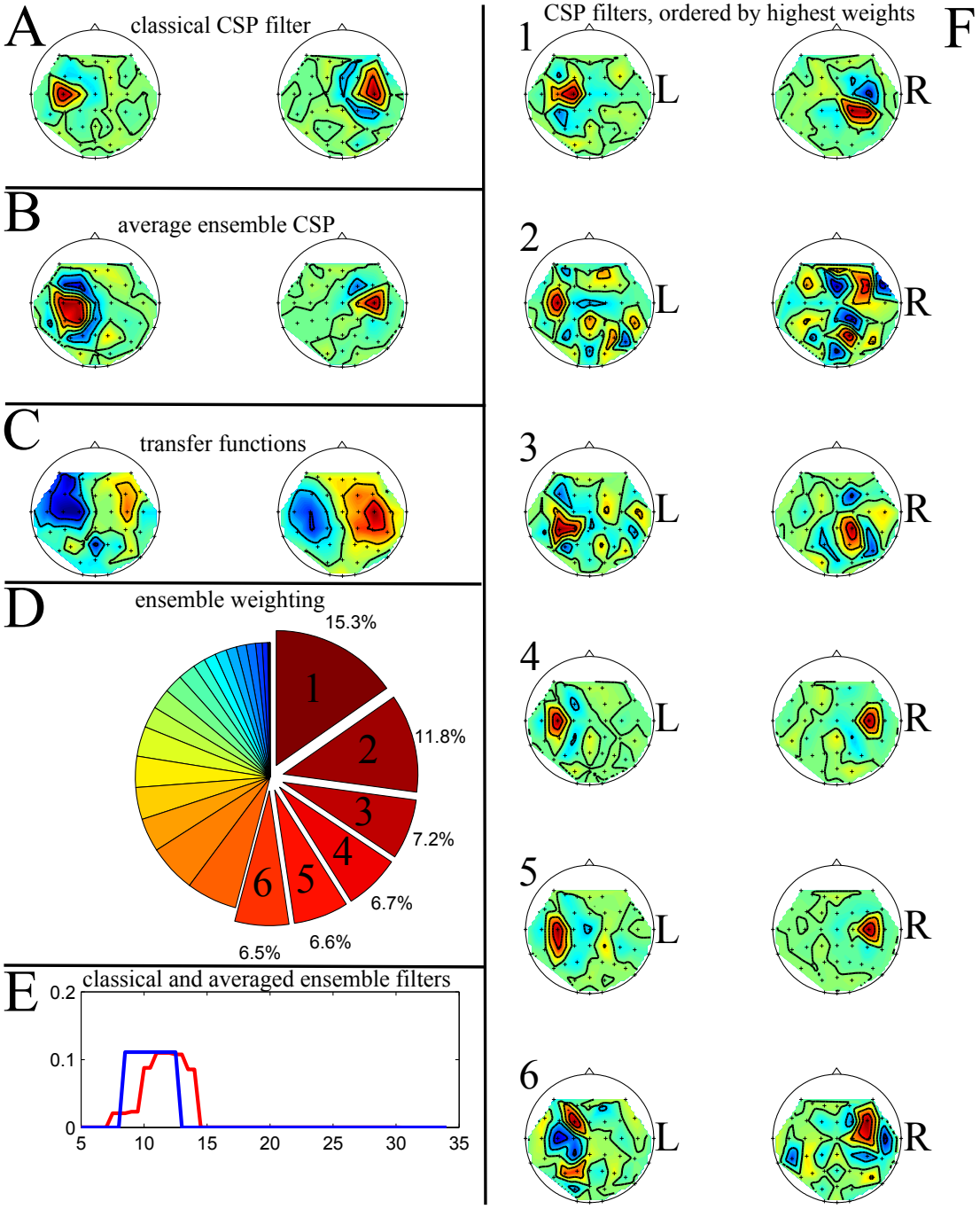

Figure 3.8: A: primary projections for classical auto-band CSP. B: linearly averaged CSP's from the ensemble. C: transfer function for classical auto-band and ensemble CSP's. D: weightings of 28 ensemble members, the six highest components are shown in F. E: linear average ensemble temporal filter (red), heuristic (blue). F: primary projections of the 6 ensemble members that received highest weights.

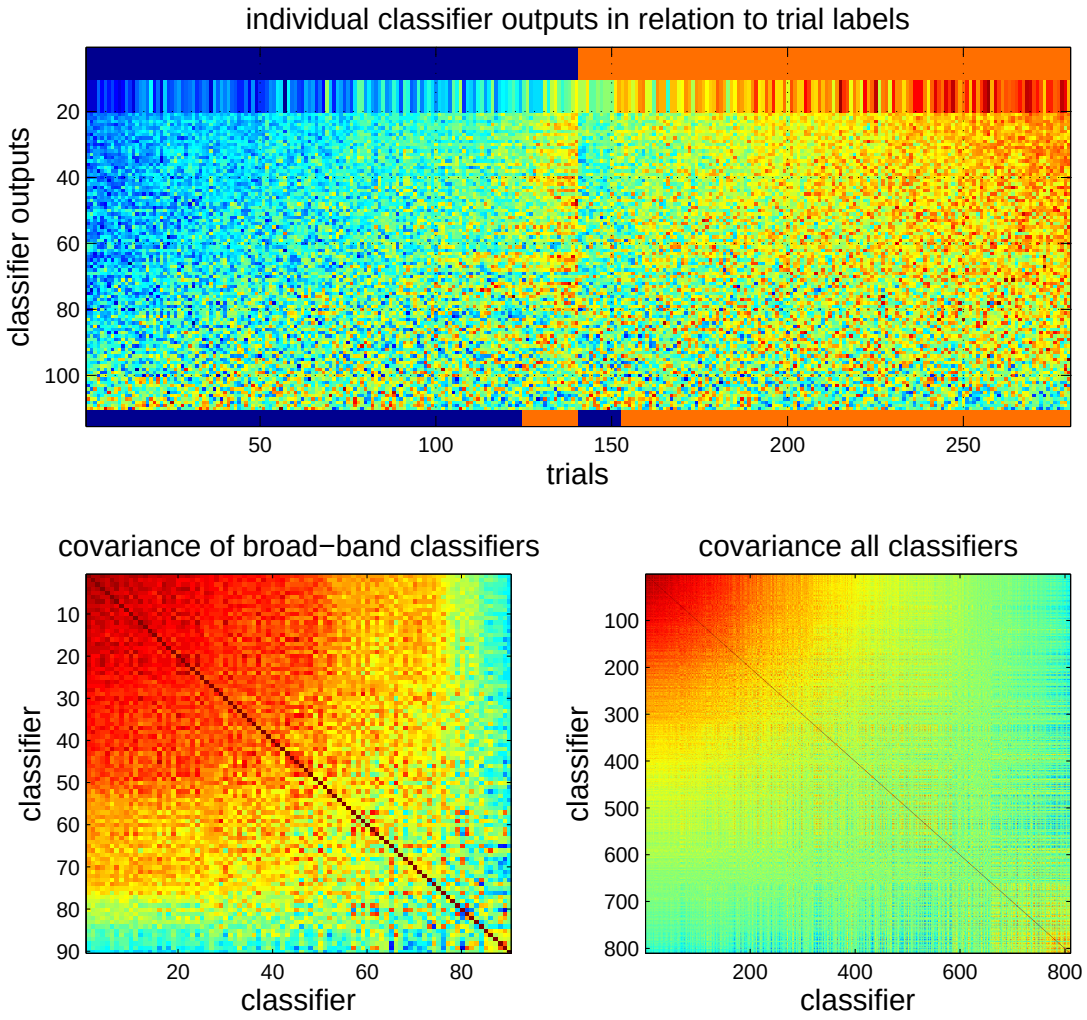

Figure 3.9: Top: Broad-band version of the ensemble outputs for a single subject. The outputs of all basis classifiers are applied to each trial of one subject. The top row (broad) gives the label, the second row (broad) gives the output of the classical auto-band CSP, and each of the following rows (thin) gives the outputs of the individual classifiers of other subjects. The individual classifier outputs are sorted by their correlation coefficient with respect to the class labels. The trials (columns) are sorted by true labels with primary key and by mean ensemble output as a secondary key. The row at the bottom gives the sign of the average ensemble output. Lower left: Covariance matrix of sorted broad-band classifier outputs. Lower right: Covariance matrix of all sorted classifiers outputs.

subjects our zero-training method exhibits a slight performance decrease, which however will not prevent them from performing successfully in BCI.

The sparsification of classifiers, in this case, also leads to potential insight into neurophysiological processes. It identifies relevant cortical locations and frequency bands of neuronal population activity which are in agreement with general neuroscientific knowledge. While this work concentrated on zero training classification and not brain activity interpretation, a much closer look is warranted. Movement imagination detection is not only determined by the cortical representation of the limb whose control is being imagined (in this case the arm) but also by differentially located cortical regions involved in movement planning (frontal), execution (fronto-parietal) and sensory feedback (occipito-parietal). Patterns relevant to BCI detection appear in all these areas and while dominant discriminant frequencies are in the α range, higher frequencies appear in our ensemble, albeit in combination with less focused patterns.

What we have found from our machine learning algorithm can be interpreted as representing the characteristic neurophysiological variation of a large subject group, which in itself is a highly relevant and interesting result. While here we present results of a motor-imagery paradigm, future studies may show that the ensemble approach may also be applied for other paradigms.

3.4 ℓ_1-penalized Linear Mixed-Effects Models for zero-training BCI

When measuring experimental data we typically encounter a certain inbuilt heterogeneity: data may stem from distinct sources that are all additionally exposed to varying measuring conditions. Such so-called group, respectively individual effects need to be modeled separately within a global statistical model. Note that here the data are not independent: a part of the variance may come from the individual experiment, while another may be attributed to a *fixed* effect. Such mixed-effects models [104] are known to be useful whenever there is a grouping structure among the observations, e.g. the clusters are independent but within a cluster the data may have a dependency structure. Note also that mixed-effects models are notoriously hard to estimate in high dimensions, particularly, if only few data points are available.

In the following we will for the first time use a recent ℓ_1-penalized estimation procedure for high-dimensional linear mixed-effects models [112] in order to estimate the mixed effects that are persistent in experimental data from neuroscience. This novel method builds upon Lasso-type procedures [125, 89, 142], assuming that the number of potential fixed effects is large and that the underlying true fixed-effects vector is sparse. The ℓ_1-penalization on the fixed effects is used to achieve

sparsity. The idea of ℓ_1-penalized likelihood approaches in linear mixed-effects models is not novel. The work of [19] and [62] present ℓ_1-penalized methods for linear mixed effects models. While the latter [19, 62] only studied the low-dimensional setting, only [112] have succeeded in investigating the high-dimensional case (i.e. $n \ll p$).

In BCI we encounter high variability both between subjects and within repetitions of an experiment for the same subject. The novel approach splits up the overall inherent variance into a *within-group* and a *between-group variance* and therefore allows us to model the unknown dependencies in a meaningful manner. While this is a conceptual contribution to adapt the mixed effects model for BCI, we also contribute practically: Due to the more precise modeling of the dependency structure we cannot only quantify both sources of variance but also provide an improved ensemble model that is able to serve as a one-size-fits-all BCI classifier – the central ingredient of a so-called zero-training BCI [74, 45, 2]. In other words we can minimize the usually required calibration time for a novel subject – where the learning machine adapts to the new brain (e.g. [11, 12]) – to practically zero.

The following section will introduce the novel statistical model. The BCI setup and data basis were already introduced before (see Section 3.1). Section 3.4.3 will discuss the experimental results.

3.4.1 Statistical Model

We will investigate a so called linear mixed-effects model [104], due to the dependence structure inherent to the two sources of variability: within-subject (dependence) and between-subject (independence). The classical mixed-effects framework has two limiting issues: (1) it cannot deal with high-dimensional data (i.e. the total number of observations is smaller than the number of explanatory variables) and (2) fixed-effects variable selection gets computationally intractable if the number of fixed-effects covariates is very large. By using a LASSO-type concept [125] these limits can be overcome in the present method [112], thus allowing application in the real world as we will see in the next sections.

3.4.1.1 Model Setup

Let $i = 1,\dots,N$ be the number of subjects, $j = 1,\dots,n_i$ the number of observations per subject and $N_T = \sum n_i$ the total number of observations. For each subject we observe an n_i-dimensional response vector y_i. Moreover, let X_i and Z_i be $n_i \times p$ and $n_i \times q$ covariate matrices, where X_i contains the fixed-effects covariates and Z_i the corresponding random-effects covariates. Denote by $\beta \in \mathbb{R}^p$ the p-dimensional fixed-effects vector and by $b_i, i = 1,\dots,N$ the q-dimensional random-effects vec-

tors. Then the linear mixed-effects model can be written as ([104])

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \qquad i = 1,\dots,N \quad , \tag{3.3}$$

where we assume that $i)$ $b_i \sim \mathcal{N}_q(0,\tau^2 I_q)$, $ii)$ $\varepsilon_i \sim \mathcal{N}_{n_i}(0,\sigma^2 I_{n_i})$ and $iii)$ that the errors ε_i are mutually independent of the random effects b_i.

From (3.3) we conclude that

$$y_i \sim \mathcal{N}_{n_i}(X_i\beta, \Lambda_i(\sigma^2,\tau^2)) \quad \text{with} \quad \Lambda_i(\sigma^2,\tau^2) = \sigma^2 I_{n_i} + \tau^2 Z_i Z_i^\top. \tag{3.4}$$

It is important to point out that assumption $i)$ is very restrictive. Nevertheless, it is straightforward to relax this assumption and assume that $b_i \sim \mathcal{N}_q(0,\Psi)$ for a general (or possible structured) covariance matrix Ψ.

To give the reader an intuition of the method, we generated a simple toy example that demonstrates why estimating mixed-effects can help in finding a superior solution that takes possible shifts in the input-space of multiple-subject data into account: The data is generated with the model given in Equation (3.3) and by setting $Z_i = \mathbf{1}_{n_i}$ and $b_i \in \mathbb{R}$ we assume a random-intercept model or one bias per group. The top left panel of Figure 3.10 shows the five groups of input data we generated, each consisting of 40 trials with the following parameters: $\beta_{\text{ORIG}} = 0.5$, $b_{\text{ORIG}} = [-2;-1;0;1;2]$ and a noise level of $\epsilon_{\text{ORIG}} \sim \mathcal{N}(0,0.2)$. While least-square regression (LSR) estimates $\beta_{\text{LSR}} = 0.048$ and $b_{\text{LSR}} = 0.075$, the proposed mixed-effects model is far more accurate and estimates $\beta_{\text{LMM}} = 0.504$ and the individual biases to be $b_{\text{LMM}} = [-1.96;-1.015;-0.014;0.973;2.013]$, as can be seen in the lower part of Figure 3.10. Figure 3.11 depicts a flowchart, which also gives an intuition of when mixed-effects models should be considered.

foreach $(\sigma^2,\tau^2,\lambda)$ **do**
  foreach i **do**
    (Whiten data and labels)
    $\Lambda_i = \sigma^2 I_{n_i} + \tau^2 Z_i Z_i^\top$
    $\bar{X}_i = \Lambda_i^{-1/2} X_i, \quad \bar{y}_i = \Lambda_i^{-1/2} y_i$
  end
  (Fit ℓ_1-penalized least-squares to concatenated data)
  $\hat{\beta} = \text{argmin} \left\| \bar{X}\beta - \bar{y} \right\|_2^2 + 2\lambda \sum_{k=2}^{p} |\beta_k|$
  foreach i **do**
    (Find random effects)
    $\hat{b}_i = [Z_i^\top Z_i + \sigma^2/\tau^2 I_q]^{-1} Z_i^\top (y_i - X_i\hat{\beta})$
  end
end

Algorithm 2: algorithm for fitting the mixed effects model

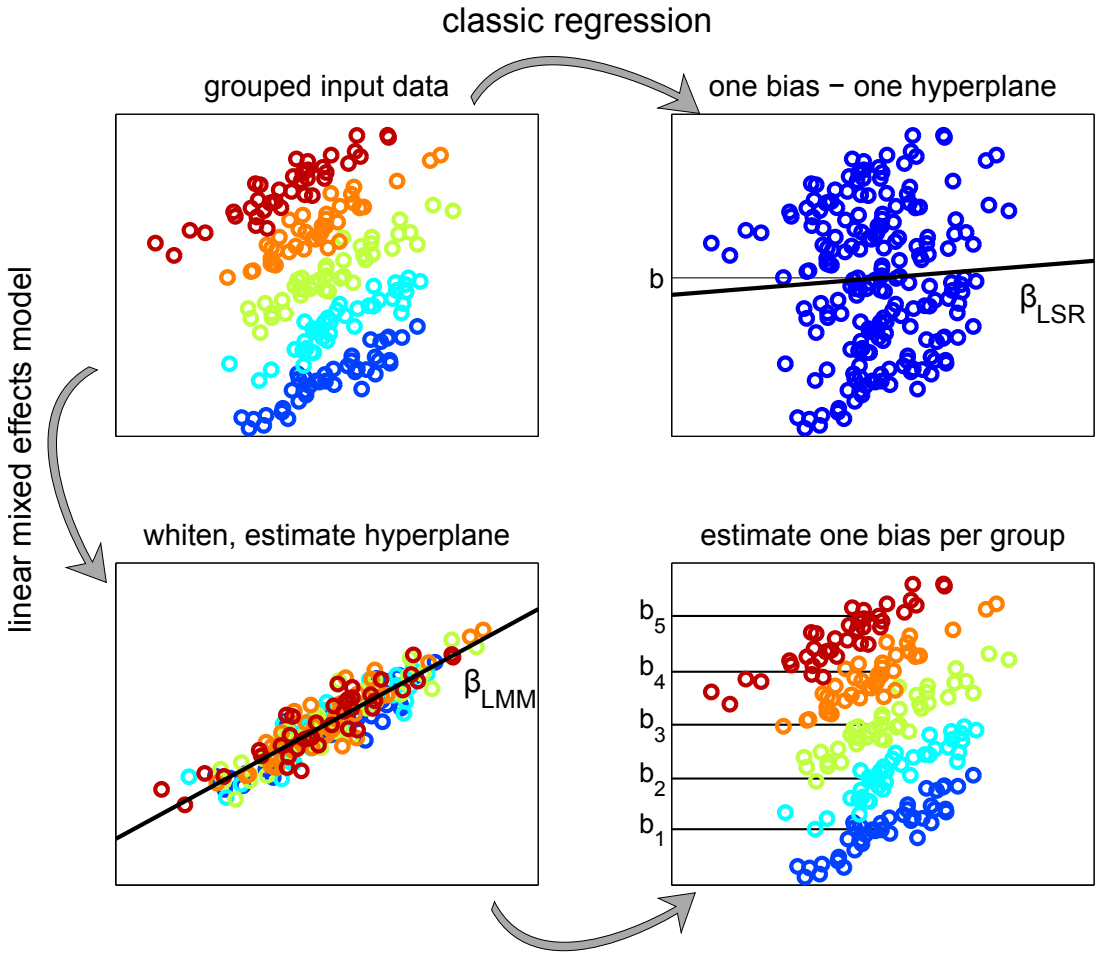

Figure 3.10: Illustration of the fitting procedure for a linear mixed-effects model with $Z = \mathbf{1}_{n_i}$, i.e. a random intercept model: groups have the same slope but different intercepts. The colors distinguish groups. If fitted with a classical regression, the fixed-effect is not recovered correctly. By applying Algorithm 2, the data are first whitened with Λ_i and then the fixed-effect is estimated from the whitened data by linear regression. In as a second step, the random effects are recovered.

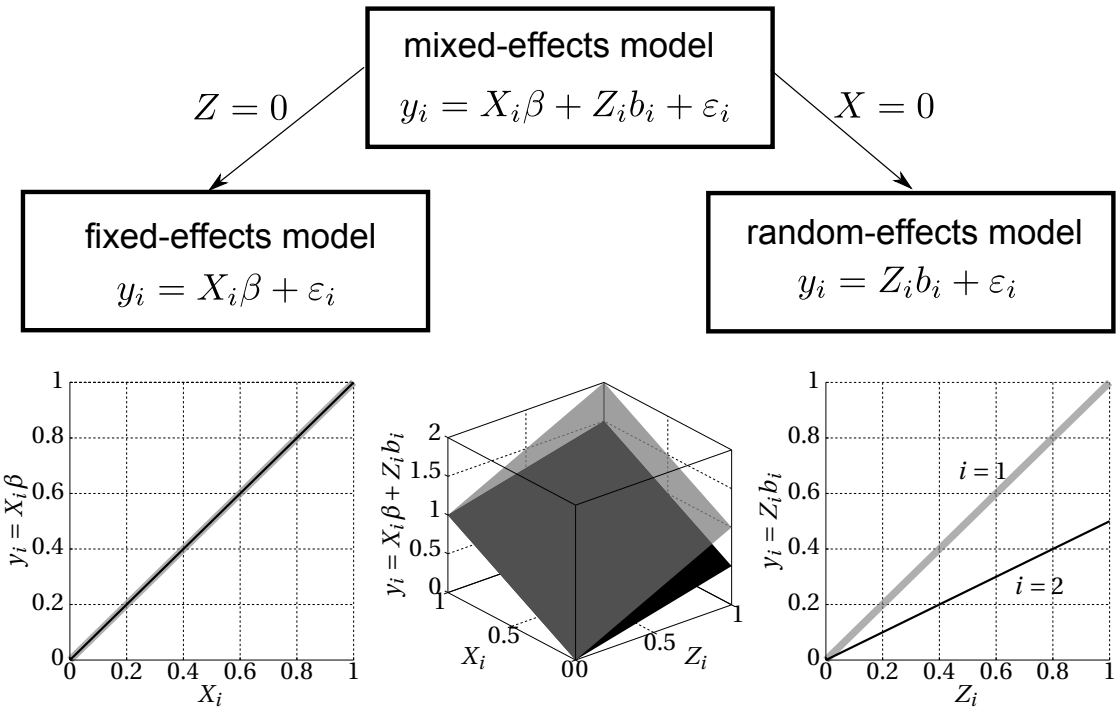

Figure 3.11: Upper part: The flowchart gives an overview of the mixed-effects, random-effects and fixed-effects models. Lower part: Plot of the mixed-effects model $y = X_i\beta + Z_i b_i$ without noise, with $i = \{1,2\}$, $\beta = 1$, $b_1 = 1$, $b_2 = 1/2$. Grey: group 1, black: group 2. The figure in the middle shows the plot for the general case. In the left plot, where $Z_1 = Z_2 = 0$, i.e. with only fixed effects present, the model reduces to a linear function in X_i: the two curves coincide. In the right plot, where $X_i = 0$, i.e. only random effects are present, the groups are decoupled and form two independent linear functions.

3.4.1.2 ℓ_1-penalized Maximum Likelihood Estimator

Since we have to deal a large number of covariates, it is computationally not feasible to employ the standard mixed-effects model variable selection strategies. To remedy this problem, in [112] a Lasso-type approach is proposed by adding an ℓ_1-penalty for the fixed-effects parameter β. This idea induces sparsity in β in the sense that many coefficients $\beta_j, j = 1,\dots,p$ are estimated exactly zero and we can perform simultaneously parameter estimation and variable selection. Consequently, from (3.4) we derive the following objective function

$$S_\lambda(\beta,\sigma^2,\tau^2) := -\frac{1}{2}\sum_{i=1}^{N}\left\{\log|\Lambda_i| + (y_i - X_i\beta)^\top \Lambda_i^{-1}(y_i - X_i\beta)\right\} - \lambda\sum_{k=1}^{p}|\beta_k| \quad , \tag{3.5}$$

where λ is a nonnegative regularization parameter.
Hence, estimating the parameters β, σ^2 and τ^2 is carried out by maximizing $S_\lambda(\beta,\sigma^2,\tau^2)$:

$$\hat{\beta},\hat{\sigma}^2,\hat{\tau}^2 = \underset{\beta,\sigma^2,\tau^2}{\operatorname{argmax}} S_\lambda(\beta,\sigma^2,\tau^2) \quad . \tag{3.6}$$

It is worth noting that $S_\lambda(\beta,\sigma^2,\tau^2)$ is a non-concave function, which implies that we can not apply a convex solver to maximize (3.5).

3.4.1.3 Prediction of the random-effects

The prediction of the random-effects coefficients $b_i, i = 1,\dots,N$ is done by the maximum a posteriori (MAP) principle. Given the parameters β, σ^2 and τ^2, it follows by straightforward calculations that the MAP estimator for $b_i, i = 1,\dots,N$ is given by $b_i = [Z_i^\top Z_i + \sigma^2/\tau^2 I_q]^{-1} Z_i^\top (y_i - X_i\beta)$. Since the true parameters β, σ^2 and τ^2 are not known, we plug in the estimates from (3.6). Hence the random-effects coefficients are estimated by

$$\hat{b}_i = [Z_i^\top Z_i + \hat{\sigma}^2/\hat{\tau}^2 I_q]^{-1} Z_i^\top (y_i - X_i\hat{\beta}). \tag{3.7}$$

3.4.1.4 Model Selection

The optimization problem in (3.6) is applied to a fixed tuning parameter λ. In practice, the solution of (3.6) is calculated on a grid of λ values. The choice of the optimal λ-value is then achieved by minimizing a criterion, i.e. a k-fold cross-validation score or an information criteria. We propose to use the Bayesian Information Criterion (BIC) defined as

$$-2\ell(\hat{\beta},\hat{\sigma}^2,\hat{\tau}^2) + \log N_T \cdot \hat{df}_\lambda \quad , \tag{3.8}$$

where $\hat{df}_\lambda = |\{1 \le j \le p; \hat{\beta}_j \neq 0\}|$ denotes the number of nonzero fixed regression coefficients and $\ell(\hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2)$ denotes the likelihood function following from the model assumptions in (3.3). The BIC works well in the simulation examples presented in [112] and is computationally fast.

3.4.2 Computational Implementation

With τ and σ fixed, the cost function (3.5) is equivalent to an ℓ_1-penalized linear regression after whitening by the covariances Λ_i:

$$\hat{\beta} = \underset{\beta|\tau,\sigma}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \Lambda_i^{-1/2} \left(X_i \beta - y_i \right) \right\|_2^2 + 2\lambda \sum_{k=2}^{p} |\beta_k| \tag{3.9}$$

We solve the resulting convex optimization problem for b with fixed σ and τ using the orthant-wise limited memory quasi-Newton algorithm [3]. As suggested in [112], the optimization is performed over a grid of (σ^2, τ^2) to find the optimum of the considered parameters.

Since, in our case, the labels y_i are binary (i.e, 0 when the left hand was cued and 1 for the right hand), we have also fitted the logistic regression equivalent to the least-squares regression presented in Section 3.4.1.2. Preliminary analysis indicates that a so called random-intercept (i.e. one bias per group) is appropriate for our data, i.e., $Z_i = 1$ and $\beta_i \in \mathbb{R}$:

$$y_{ij} = f\left(x_{ij}^{\mathsf{T}} b + \beta_i \right) + \varepsilon_{ij} \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i \tag{3.10}$$

where $f(x) = 1/\left[1 + \exp(-x)\right]$ is the sigmoid function. We assume that $\beta_i \sim \mathcal{N}(0, \tau^2)$, $\varepsilon_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$ and that ε_i are mutually independent of β_i. We solve the resulting numerical optimization problem using the orthant-wise limited memory quasi-Newton algorithm [3].
In the context of (3.10), σ^2 corresponds to the *within-subject variability* and τ^2 to the *between-subject variability*. By estimating σ^2 and τ^2 we are able to allocate the variability in the data to these to sources.

3.4.3 Results

3.4.3.1 Subject-to-Subject Transfer

As explained in Section 3.1, we use our first balanced dataset to find a zero-training subject-independent classifier. Figure 3.12 shows the results of fitting a least-squares and a logistic regression model, both ℓ_1-regularized, fit to a) a linear model with one

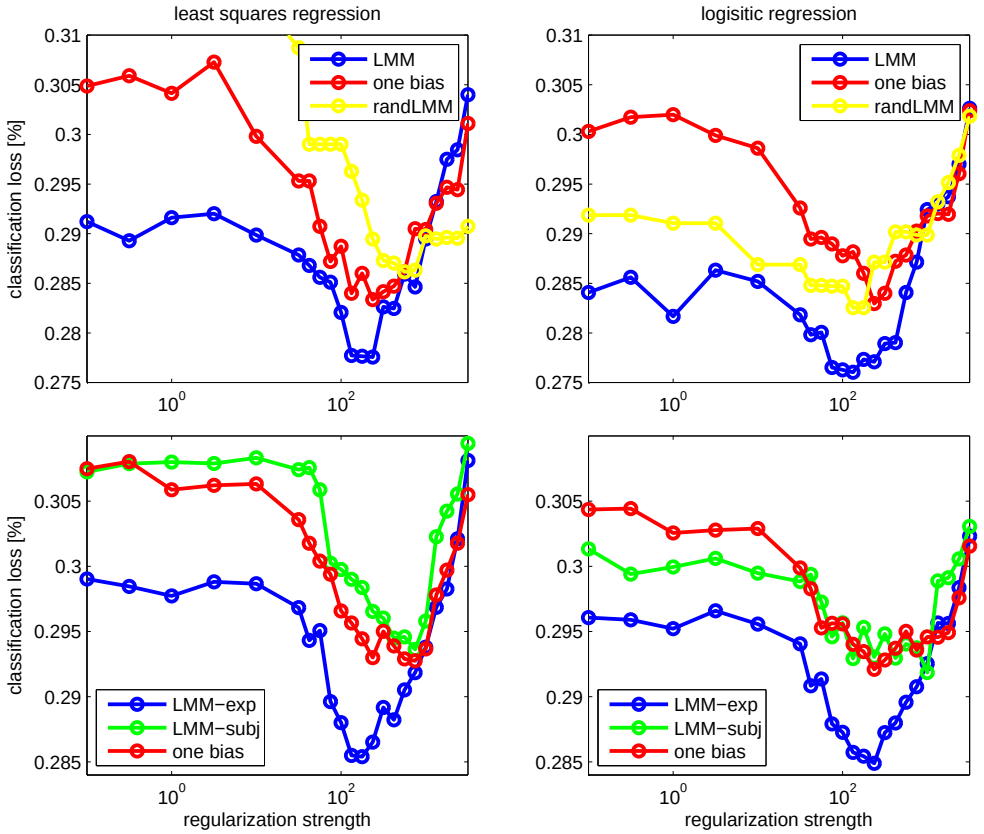

Figure 3.12: The two top figures show the mean classification loss over subjects for the *balanced* dataset as a function of the regularization constant λ. The LMM approach is compared to classical ℓ_1-regularized least squares (left) and logistic (right) regression. The two lower figures show the same results the *unbalanced* dataset. LMM-subj estimates one bias per subject and LMM-exp one bias per experiment (session).

bias and b) a mixed-effects model with one bias per subject. We are able to improve classification by use of the mixed-effects model for both regressions.

To further explore the group effect of the mixed-effects model, we assigned the trials in the *balanced dataset* to random groups. The results can be seen in the yellow lines in the top panels of Figure 3.12 as well as in Table 3.6 (indicated by *rand LMM*). As expected, Table 3.6 shows that the mixed-effects model with randomly assigned groups does not improve the classification, while the meaningfully applied mixed-effects model gets very close to the *self-prediction error*. *Self-prediction error* denotes the average cross-validation error when using the training data of a subject to predict his test data, i.e., performing conventional, subject-dependent BCI. The *self-prediction error* could therefore be interpreted as a lower bound for subject-

	least squares regression	logistic regression
one bias	28.34%	28.30%
rand LMM	28.62%	28.25%
LMM	27.76%	27.60%
self-prediction	27.51%	
Laplace	33.95%	
band-power	37.60%	

Table 3.6: Classification loss of the *balanced dataset* for various methods. In the *one bias* method only one bias is estimated for the whole dataset, in *random LMM* one bias per group is estimated, however members of groups are randomly assigned. In *LMM* one bias per group is estimated.

independent classifier loss.

In Figure 3.13 we compare the performance of our method on the basis of individual subjects with other methods and perform t-tests to examine their statistical significance. The p-values are included within the figure. As the most simple baseline method we used 'Laplace features' by calculating the difference of two motor related channels (namely 'C3' and 'C4') within a time interval of $750-3500$ ms, after broadband ($7-30$ Hz) temporal and Laplacian spatial filtering of the individual channels. This method scored an average loss of 33.95% as can be seen in Table 3.6. In Table 3.6 the results of the various ensemble approaches, as well as the considered baselines are given for the *balanced dataset*.

As can be seen on the left side of Figure 3.13 our novel method performs very favorably. LMM improves classification performance for 89.2% of the subjects considered with high significance and leads to an average loss of 27.6%. Furthermore, we compare with with the previously proposed zero-training procedure [43] (see also Section 3.3.6), which is very similar to the LMM method described here, except that it performs ℓ_1-regularized regression for combining the outputs of the individual classifiers (average loss 28.3%). Also here we achieve a significant improvement. Finally, we compare our method to the subject-dependent, cross-validated classifier loss, derived from the data themselves (average loss 27.51%). A per se unfair comparison. Given that the subject-dependent classifier is not significantly better ($p = 0.93$), we may state that we are on par.

The sparsity of our results becomes apparent from Figure 3.14, where we display the magnitude of weights for each run of the LOSO-CV. For LMM on average 28.9% of all features are active, while for 'one-bias' 33.5% of all features are non-zero. Note that for both methods most of the active features lie within a vertical line, indicating that the feature is also active for most other subjects and can thus be considered

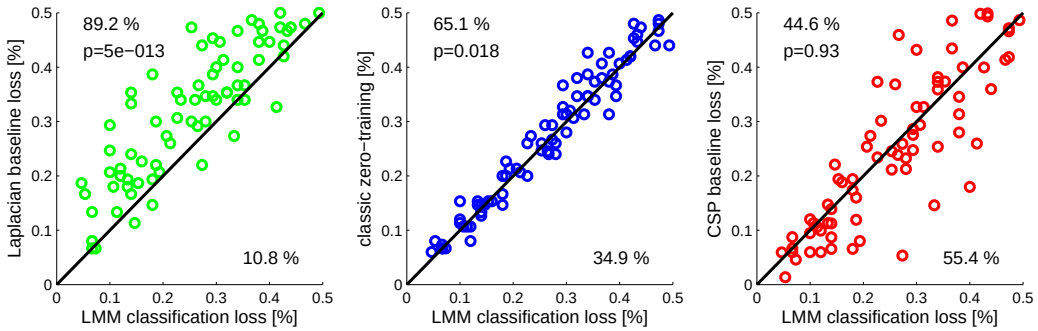

Figure 3.13: Scatter plot, comparing the proposed method with various baselines on a subject specific level.

particularly stable.

As can be seen in Figure 3.15 (left panel) the LMM method needs less features per subject ($N_{\mathrm{LMM}} \approx 310$) as compared to estimating only one bias ($N_{\ell_1} \approx 500$). Besides from selecting less features in total, the LMM chose a higher fraction of features with low self-prediction errors. This is shown in the middle panel, where we display the cumulative sum of features, sorted by increasing self-prediction accuracy.

To visualize differences between weight vectors resulting from the LOSO-CV procedure, the right panel displays these vectors, projected to two dimensions. The matrix of Euclidean distances between all pairs of weights was embedded into a 2×83-dimensional space and projected onto the resulting point cloud's first two principal axes for visualization. The mixed effects model absorbs more of the variability into its bias terms and thus results in more consistent weight vector estimates.

3.4.3.2 Session-to-Session Transfer

To investigate how the results of the method can be understood in terms of individual subjects and their (possibly multiple) sessions, we validated the method in two ways. First we allow each experiment to have an individual bias. In the second approach, we allow only one bias per subject, i.e. multiple experiments/sessions from the same subject will be grouped. The results are shown in the right panel of Figure 3.16. For both validation approaches the logistic regression captures a higher between-group variability as compared to linear regression and can thus be seen as the more appropriate method, as is also apparent from the lowest cross-validation loss (see Figure 3.12). Furthermore and more interestingly, we see sub-

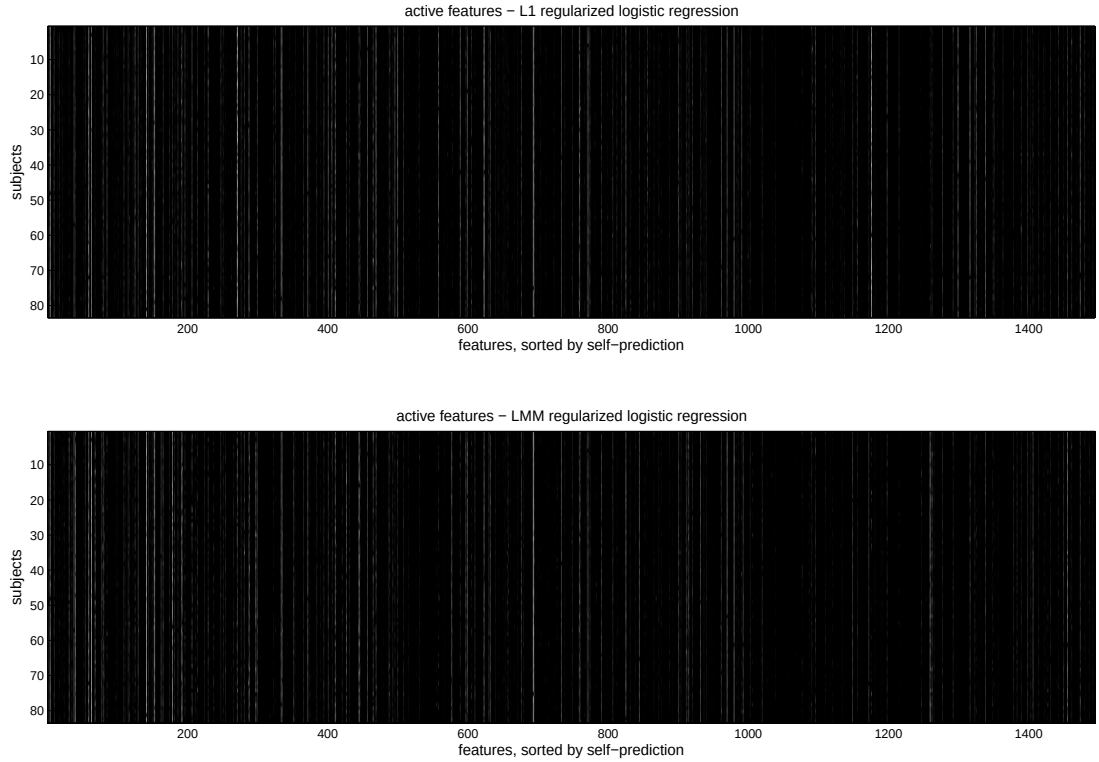

Figure 3.14: Both plots show the selected features in white, while inactive features are black. The x-axis represents all possible features, sorted by their cross-validated 'self-prediction'. The y-axis represents each subjects resulting weight vector.

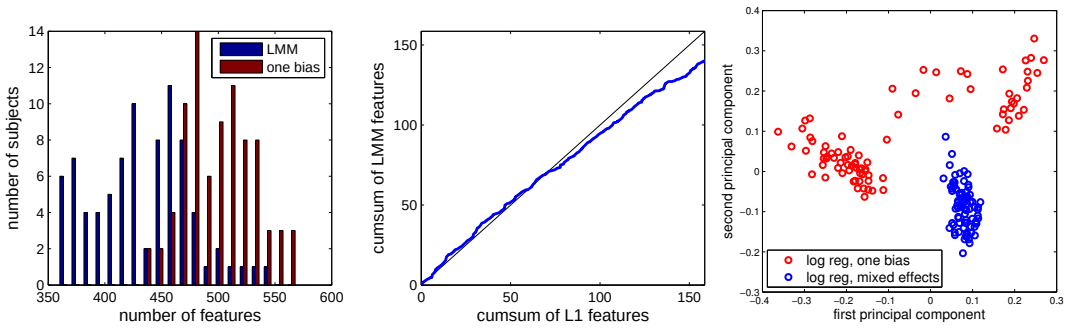

Figure 3.15: Left: histogram of the number of selected features for all subjects. Middle: cumulative sum of features, sorted by 'self prediction'. LMM rather chooses features, that had a good 'self prediction', and needs less features in total. Right: Variability between classifier weights b of the two models for each of the $N = 2 \times 83$ LOSO-training runs using the best regularization strength.

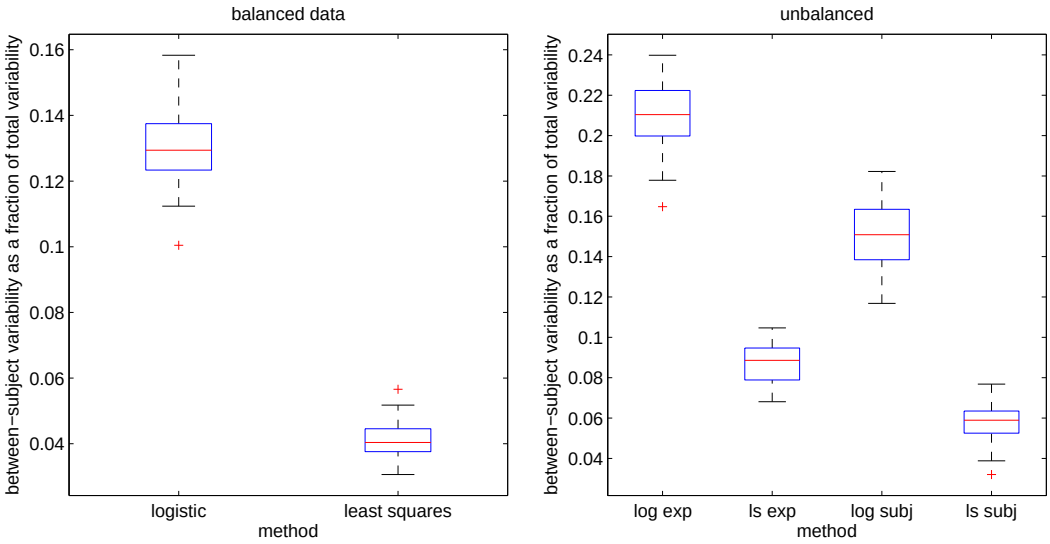

Figure 3.16: Both figures show the magnitude of between-subject variability as a fraction of total variability. On the left: Results for the first *balanced* dataset. One the right: Results for the *unbalanced* dataset. *log* stands for logistic regression, *ls* for least squares, *exp* for one bias per experiment and *subj* for one bias per subject.

stantially higher between-group-variability if we allow biases for each experiment. This result does not only confirm knowledge from previous publications, that the transfer of classifiers from sessions to sessions required a bias correction [74], but also underlines the validity of our approach in the sense that we are able to capture a meaningful part of the variability which would otherwise be ignored as noise.

3.4.4 Relation of baseline misclassification to σ^2 and τ^2

Using standard methods for ERD-related BCI decoding [18], we obtain a mean classification loss for each subject within our *balanced* dataset, based on the cross-validation of band-pass and spatially filtered features. In Figure 3.17 we examine the relationship between this *baseline loss* and the *within-subject variability* σ.^{2} and *between-subject variability* τ.^{2}. The *baseline loss* and σ^2 have a strong positive correlation, with high significance. This makes intuitive sense: a dataset that is well classifiable should also exhibit low variance of its residuals. We furthermore examine the relation of τ^2 and σ^2 and find a strong positive relation.

Interestingly we do not find a significant relation between the baseline loss and τ^2. In other words it is not possible to draw conclusions about the quality of a subject's data by the variance of its assigned biases.

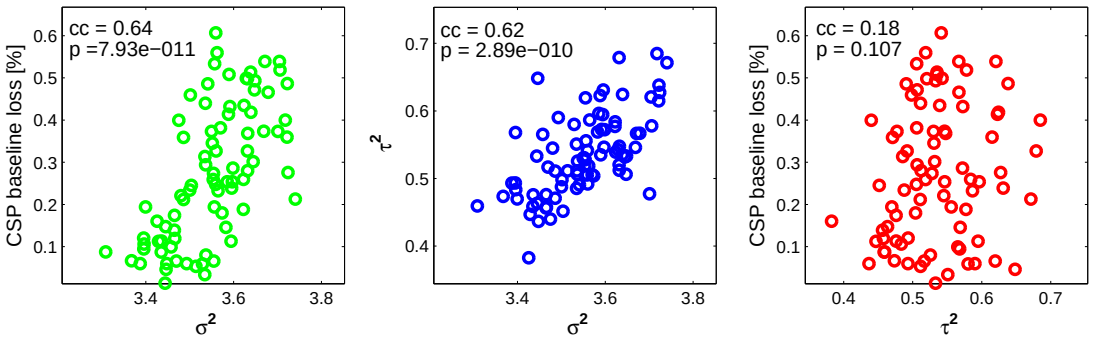

Figure 3.17: The three scatterplots show relations between *within-subject variability* $\sigma.^2$, *between-subject variability* τ^2 and the baseline cross-validation misclassification for every subject. *cc* stands for correlation coefficient and *p* stands for paired t-test significance.

3.4.5 Effective spatial filters and distances thereof

To estimate the similarity of effective spatial filters, we use a transfer function as described in [45]: By injecting a sinusoid into a given channel and processing it by the spatial filter, estimating the bandpower and applying the classifier, we obtain a response for one particular channel. Repeating this procedure for each channel results in a response matrix that can be easily visualized. We define a distance measure for each individual subject between her original CSP filter and those estimated via 'LMM' and 'one bias' methods. The measure we use is the angle between their vectorized response matrices (see [74]).

For four subjects the resulting response matrices, based on the original CSP pattern, are shown on the top row of the left part of Figure 3.18. To obtain a response matrix for the ensemble approaches, we calculate the weighted sum of responses, determined by β (see middle and lower parts of Figure 3.18).

In the right part of Figure 3.18 the resulting distances between 'LMM' or 'one bias' and the original CSP based response function are plotted against all subjects with self-prediction loss of less than x. As one would expect both distances increase on average, as more subjects with higher self-prediction loss are added to the analysis. It shows that the linear mixed-effects model is consistently closer, irrespective of the subject's self-prediction error.

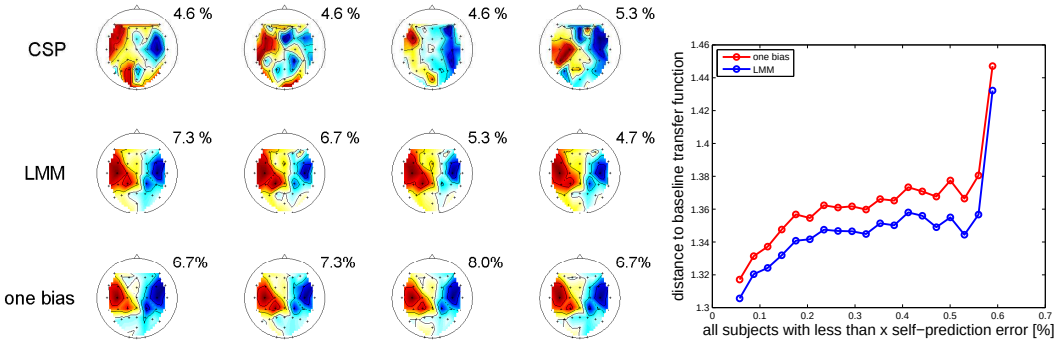

Figure 3.18: Left part: Response matrices of the four best subjects for 'original CSP', 'LMM' and 'one bias'. Classification loss is given as percentage numbers. Right part: Response distances of 'LMM' and 'one bias' versus self-prediction error [%].

3.4.6 Discussion and Conclusions

When analyzing experimental data, it is of generic importance to quantify variation both across the ensemble of acquired data and within repetitions of measurements. Distinguishing and modeling such mixed effects is of high interest e.g. in medicine, biology, physics and the neurosciences.

In this Chapter we have applied a recent sparse modeling approach from statistics [112] based on a so-called ℓ_1-penalized linear mixed-effects model and proposed its first time use for a large BCI data set, leading to a novel BCI zero-training model (see also [74, 45]). In this manner we could efficiently model the different dependencies and variabilities between and within subjects. Note that the novel statistical model not only gave rise to a better overall prediction – in other words to an *improved zero-training model* – but it furthermore allowed to quantify the differences in variation more transparently and also interpretability. By attributing some of the total variability, in other methods considered as noise, to differences between subjects, we are now able to obtain a solution that is sparser and at the same time superior in prediction accuracy. Not only features with high prediction performance are preferably chosen, but also responses of the novel ensemble are more similar to its original counterpart.

Furthermore, we would like to note that while more complex random effects would in principle be conceivable, our random intercepts model was not just chosen by intuition but from our experience with BCI: When performing an experiment with the same subject on two subsequent days, on the second day the classifier can often be reused without much retraining, only the bias needs to be adjusted [74, 114].

We have developed a statistical framework that can be applied to a large number of scientific experiments from a large number of domains, where inter-dependencies of input space exist and have shown that our approach leads to more robust feature selection and is superior in its classification accuracy. Future research will study on-line adaptation of penalized linear mixed-effects models in the context of medical diagnosis and may well find its way into a broader scientific context.

CHAPTER 4

Multimodal NIRS and EEG measurements for BCI

4.1 Combined NIRS-EEG measurements enhance Brain Computer Interface performance

Since its precursors in the early 70's [127] BCI technology has developed many variants and employed a large number of neuroimaging methods (please see Section 1 for further details and references). Combinational approaches for EEG features from multiple domains [34], such as movement related potentials (MRPs) and event-related desynchronizations (ERD), as well as combinations of EEG and peripheral parameters like electromyography [78] have been shown to increase the robustness of the classification.

These positive findings for combined approaches have motivated us for an evaluation of a simultaneous EEG and NIRS setup which preserves the advantages of both non-invasive techniques namely low costs, portability and easiness to handle. NIRS measures the concentration changes of oxygenated and deoxygenated hemoglobin ([HbO] and [HbR]) in the superficial layers of the human cortex. While concentration of [HbO] is expected to increase after focal activation of the cortex due to higher blood flow, [HbR] is washed out and decreases [75, 131, 84]. Thereby, it measures a comparable effect to the blood oxygenation level dependent (BOLD) contrast in functional magnetic resonance imaging (fMRI), since also here the washout of [HbO] is the major constituent [65].

The idea of using NIRS as an optical BCI has been introduced by Coyle et al. in 2004 [32]. Since then a number of groups followed the direction of using NIRS as a basis for optical BCI [31, 115, 140, 4, 64, 85], by either examining the resulting signals for motor imagery or classifying the NIRS signals directly. A recent publication used NIRS as a 'brain switch' and combined it with an EEG-based SSVEP for the operation of an orthosis [103]. However, to our knowledge our study is the first report of simultaneous EEG and NIRS measurements for SMR-based Brain-Computer-Interfacing. In general a multi-modal approach can have a number of benefits: As every neuroimaging method suffers from its particular limitations (EEG from spatial resolution, while NIRS or fMRI from the sluggishness of the underlying vascular response limiting its temporal resolution), it now becomes possible to

partly overcome these by focusing on their individual strengths [49, 7]. Furthermore and maybe more importantly, the information gained from these various sources complement each other to some degree [6, 92]. Due to this reasoning it becomes apparent, why simultaneous NIRS and EEG measurements are widely used in order to research language processing [135, 38, 124, 111, 54] and the visual cortex [97, 59]. A recent study that examines the somato-motoric activity following median nerve stimulation [122] shows the reliability of simultaneous measurements of NIRS and EEG in the motor and somatosensory domain, which proves to be valid for SMR-based BCI as well.

By extracting relevant NIRS features to support and complement high-speed EEG-based BCI and thus forming a *hybrid* BCI [101], we exploit the responsiveness of EEG (i.e. high ITR) as well as enhance and robustify overall BCI performance by using information from the vascular response, which are not contained within the EEG. Moreover, we evaluate the time delay and spatial information content of the hemodynamic response during a SMR-based BCI paradigm.

The following section introduces the setup and design of our study, as well as the statistical tools we applied for the analysis of the acquired data. In Sections 4.5 and 4.6 we present the experimental results and Section 4.7 concludes the work by discussing our findings and puts them into perspective with future work.

4.2 Participants and Experimental Design

Fourteen healthy, right-handed volunteers (aged 20 to 30) participated in the study, which lasted approximately four hours. The experiment was approved by the local ethics committee (Charité University Medicine, Berlin, Germany), and performed in accordance with the policy of the Declaration of Helsinki. The subjects were seated in a comfortable chair with armrests and were instructed to relax their arms. The experiment consists of 2 blocks of motor execution by means of hand gripping (24 trials per block per condition) and 2 blocks of real-time EEG-based, visual feedback controlled motor imagery (50 trials per block per condition). For all blocks the first 2 s of each trial began with a black fixation cross, that appeared at the center of the screen. Then, as a visual cue, an arrow appeared pointing to the left or right. For the case of motor imagery, the fixation cross started moving for 4 s, according to the classifier output. After 4 s the cross disappeared and the screen remained blank for 10.5 ± 1.5 s. The online processing was based on the concept of coadaptive calibration [129] and is described in detail in Section 4.4. For the case of executed movements the fixation cross remained fixed and the subjects were instructed to open and close their hands with an approximate frequency of 1 Hz. Also here after 4 s the cross disappeared and the screen remained blank for 10.5 ± 1.5 s.

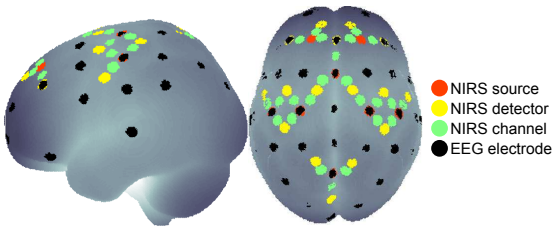

Figure 4.1: Locations of EEG electrodes; sources, detectors and actual measurement channels of NIRS. Note that electrodes and optodes might share a location.

4.3 Data Acquisition

During both tasks simultaneous measurements of EEG and NIRS were performed. The NIRS-System (NIRScout 8-16, NIRx Medizintechnik GmbH, Germany) was equipped with 24 optical fibers (8 sources with wavelengths of 850 nm and 760 nm, 16 detectors convolving to 24 measurement channels). Frontal, motor and parietal areas of the head were covered as shown in Figure 4.1. The sampling frequency was $f_{\mathrm{NIRS}} = 6.25$ Hz. EEG, electrooculogram (EOG) and electromyogram (EMG) were recorded with a multichannel EEG amplifier (BrainAmp by Brain Products, Munich, Germany) using 37 Ag/AgCl electrodes, 2 bipolar EMG, 2 bipolar EOG (vertical as well as horizontal EOG), sampled at $f_{\mathrm{EEG}} = 1$ kHz and downsampled to 100 Hz. NIRS probes and EEG electrodes were integrated in a standard EEG cap (extended 10-20 system with a possibility of 256 electrodes) with inter-optode distances between 2 and 3 cm. The optical probes are constructed, such that they fit into the ring of standard electrodes. This enables us to situate the NIRS channel positions according to the standard 10-20 system, as can be seen in Figure 4.1.

4.4 Data Analysis

Based on a recent development coined *Co-adaptive Calibration* the user was given instantaneous EEG-based BCI feedback for the two blocks of motor imagery [129]. During the first block of 100 trials a subject-independent classifier, depending on band power estimates of Laplacian filtered, motor-related EEG channels, was used. For the second block subject-dependent spatial and temporal filters were estimated from the data of the first block and combined with some subject-independent features, namely band power of Laplacian filtered, motor-related EEG electrodes, to form the classifier for the second block. During the online feedback features were calculated every 40 ms with a sliding window of 750 ms.

The analysis of NIRS data was performed offline. Concentration changes of

hemoglobin were calculated according to the modified Lambert-Beer law on the NIRS data (differential path length factor of 5.98 (higher wavelength: 830 nm) and 7.15 (lower wavelength: 760 nm), extinction coefficients for [HbO] 2.5264/1.4866 (higher/lower wavelength) and [HbR] 1.7986/3.8437 (higher/lower wavelength), and an inter-optode-distance of 3 cm). This procedure converts attenuation changes measured by the NIRS system into concentration changes of oxygenated [HbO] and deoxygenated [HbR] hemoglobin [28, 66]. NIRS data was low-pass filtered at 0.2 Hz using a one-directional filter method, namely a 3^{rd} order Butterworth-filter. A baseline interval was defined from -2 s to 0 s before stimulus onset, and its mean was subtracted from each trial. To examine how well the NIRS data classifies the given tasks we analyzed the time courses with the help of a moving window (width 1 s, step size 500 ms) that we apply from 6 s, prior to stimulus onset and up to 15 s after stimulus onset. Time courses of [HbO] and [HbR] were averaged over the time length of the moving window width, resulting in average concentration changes for each of the 24 channels. These time-averaged concentration changes were then used as features for a linear discriminant analysis (LDA). Validation was performed by a cross-validation with an 8-fold chronological split. Previous studies have shown that a chronological split maintains non-stationarities of the data and thus represents a relatively conservative measure [80]. We used the time interval of the global peak classification accuracy and performed paired t-tests to test whether classification of motor imagery shows a significantly earlier peak accuracy as compared to executed movements and in which chromophore accuracy was higher. Trials of the two measured blocks per condition were combined.

Offline EEG decoding was performed as follows: for both paradigms (real movements and motor imagery) the two blocks were combined. Subject-dependent band-pass filter coefficients were estimated by means of an established procedure (a heuristic, based on r^2-values) [18]. The selected band-pass filter coefficients for *executed movements* were mostly in the α-band (5 of 14 subjects) and in the β-band (7 of 14 subjects). For a small proportion of subjects (2 of 14) a broad-band filter was selected. For the case of *motor imagery* the discriminant information was highest in the α-band (10 of 14 subjects), followed by the β-band (3 of 14 subjects). Only for one subject a broad-band filter was chosen.

A spatial filter, in form of Common Spatial Patterns [50, 69, 109, 18] was estimated and an LDA classifier computed. The previously mentioned parameters for subject-dependent temporal filters, spatial filters and linear classifier were estimated solely on the training set of each cross-validation step [80]. The cross-validation followed the same principle as mentioned for the NIRS signals. For the time course of classification accuracy the same moving window was applied as for the NIRS data. Furthermore to establish a single measure of classification accuracy for each subject and paradigm, the time interval was chosen to be $[750-3500]$ ms

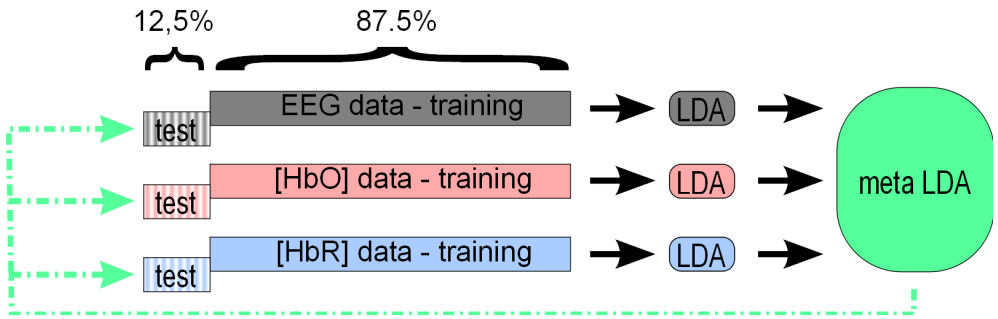

Figure 4.2: Flowchart of the first step of the cross-validation procedure: The EEG and NIRS data is split into 7/8 training data and 1/8 test data. First an individual LDA classifier is computed for EEG, [HbO] and [HbR]. Then a meta-classifier is estimated for optimally combining the three LDA outputs. All LDA classifiers are then applied to the test set (dotted green line) and a test loss is computed. The procedure is repeated for 8 chronological splits.

after stimulus onset for all subjects.

To examine the possible benefits of combining both signal domains, classification results were calculated for EEG and NIRS separately, but also in combination by estimating a meta-classifier. After estimation of the three individual classifiers (one for the EEG induced band power changes and one each for the evoked deflection from baseline [HbO] and [HbR]) and their performance, we explore a number of possible combinations (such as EEG, [HbO] or EEG, [HbO], [HbR] etc.).

Our selection of NIRS features for the combination with EEG was based on the global peak cross-validation accuracy for each individual subject. As a meta-classifier we used an LDA. The LDA weights are re-estimated within each cross-validation step in order to avoid a bias in the estimation of the generalization error [80]. The general procedure can be seen in Figure 4.2. To graphically investigate the potential improvement of a combination of NIRS and EEG measurements as compared to a BCI, solely dependent on EEG, we show scatter plots comparing EEG classification accuracy and the improvement for EEG in combination with each NIRS chromophore as well as both chromophores.

To gain topographical maps of significant features, and thereby show the physiological validity of our approach, we calculated point-biserial correlation coefficients [123]. The point-biserial correlation coefficient is a special case of the Pearson product-moment correlation coefficient and measures the association of a binary random variable and a continuous random variable and was introduced in Section 1.3.1.4.

Mutual information is an information theoretic measure, which estimates the

information that two random variables share. It can be expressed in terms of conditional entropies of random variables X and Y:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{4.1}$$

The conditional entropy $H(X|Y)$ quantifies the remaining entropy of X, after the value of Y is known. If $H(X|Y) = H(X)$, then $I(X;Y) = 0$: the variables are independent. On the other hand, if X and Y are identical, then $H(X|Y) = 0$ and hence $I(X;Y) = H(X)$. $I(X;Y)$ is symmetric and its values are in the range of 0 and 1: $I(X;Y) = I(Y;X) \in [0;1]$ [86]. To examine the degree of independence between the NIRS and EEG-based classifier outputs, we restrict their outputs to values 0 and 1 and estimate their mutual information.

To further investigate, whether mostly the same trials are classified wrongly by EEG and by NIRS, we form two groups of trials: one group consists only of trials, where EEG classification was correct, while in the other group only misclassified trials are included. By comparing the NIRS classification of each of these groups to the mean classification of both groups, we can examine to which extent the NIRS classification results resemble those of the EEG.

4.5 Physiological reliability of NIRS features

Our first aim is to show the physiological reliability of NIRS feature classification both in time and location. We performed single trial classification of left vs. right motor execution (and imagery) with a moving time window after stimulus onset. Classification accuracies for each subject over time can be seen in Figure 4.3 for EEG (top row) and both chromophores of NIRS (middle: [HbO], bottom: [HbR]). The left column shows motor imagery and the right column executed movements. A classification accuracy of 100% means that the two conditions are perfectly separable, while a classification accuracy of 50% represents random guessing when considering a binary classification task.

Average EEG classification peaks at $\langle t_{\text{eeg}}^{\text{real}} \rangle = 1680 \pm 1014$ ms for executed movements and at $\langle t_{\text{eeg}}^{\text{imag}} \rangle = 1430 \pm 707$ ms for motor imagery. Peak classification times of [HbO] are at $\langle t_{\text{hbo}}^{\text{real}} \rangle = 7430 \pm 2201$ ms and at $\langle t_{\text{hbo}}^{\text{imag}} \rangle = 6501 \pm 1579$ ms and of [HbR] at $\langle t_{\text{hbr}}^{\text{real}} \rangle = 6966 \pm 2484$ ms and $\langle t_{\text{hbr}}^{\text{imag}} \rangle = 6109 \pm 1339$ ms for executed movements and motor imagery, respectively. EEG features are thus earlier classifiable as compared to [HbO] and [HbR] for executed movements ($p < 10^{-6}$ and $p < 10^{-5}$) and for motor imagery ($p < 10^{-6}$ and $p < 10^{-6}$).

Average EEG classification accuracy for executed movements (90.8%) is higher than that of [HbO] (71.1%) and [HbR] (73.3%). Paired t-tests between EEG and the

two NIRS chromophores yields highly significant results ($p < 10^{-3}$ and $p < 0.01$). While also for motor imagery EEG scores higher average classification rates (EEG: 78.2%, [HbO]: 71.7%, [HbR]: 65.0%), here not both p-values are significant ($p = 0.09$ and $p < 0.05$). For motor imagery [HbO] shows a significantly higher classification accuracy, as compared to [HbR] ($p < 0.01$).

The topology of significant EEG and NIRS features can be seen in Figures 4.4 and 4.5. Here $log(p)$ significances of executed and motor imagery are shown, respectively. The time-dependent scalp plots show grand-averages over all subjects, based on the point-biserial correlation coefficient r_{pb}, as described above. The colorbar scales on the right side indicate the significance levels of the individual imaging methods. Note that the *width* of the scale illustrates the maximum level of significance. Red colors denote higher values of the *left* class, while blue colors indicate higher values within the *right* class. As can be seen for both paradigms EEG as well as NIRS chromophores show highly significant patterns in motor-related cortical areas. Note that for EEG (top rows of Figures 4.4 and 4.5) we observe event-related desynchronization (ERD) which is followed by a event-related synchronization (ERS), a previously described physiological effect for EEG oscillations in the alpha and beta band [100].

Interestingly, we find higher significance levels of [HbR] in both paradigms, as compared to the classification results, where [HbO] yielded higher accuracies for motor imagery. A second interesting point to note is the inverted polarity of [HbO] for motor imagery. This effect can also be seen in the averaged time courses of NIRS data shown in Figure 4.6. [HbO] has the expected shape of a hemodynamic response function in the motor execution task, although it ascends in both hemispheres but decreases in the imagery condition. [HbR] shows the expected time courses for both tasks (imagery/executed) and both conditions (left/right).

4.6 Enhancing EEG-BCI performance by NIRS features

While the examination of the NIRS classification itself provided information about the quality and spatial specificity of the NIRS features, a second aim was to actually combine NIRS and EEG features to form a hybrid-BCI. As stated in Section 4.4 a meta classifier was derived for combining the individual signals. Table 4.1 shows classification accuracies for EEG, [HbR] and [HbO] and their combinations for both tasks. Furthermore we show scatter plots, where the EEG performance is plotted against possible combinations (see Figure 4.7). Dots above the green line indicate that a subject's performance is increased by the combination of the NIRS chromophore(s) as compared to using only EEG. The percentage within the figure indicates the percent of subjects, for whom the combination leads to equal or improved

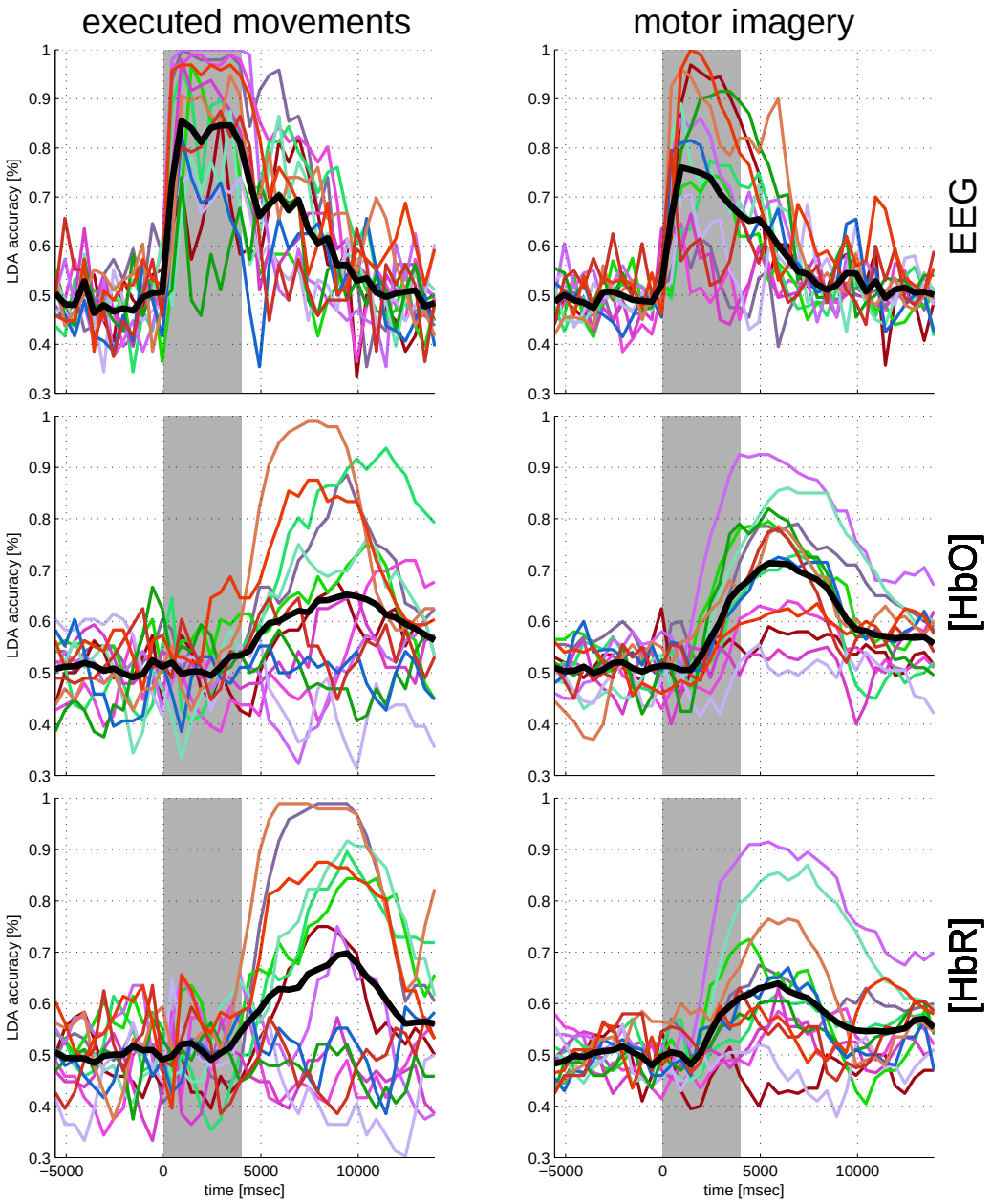

Figure 4.3: EEG and NIRS classification accuracy [%] (LDA) for a 1 s moving time window (top: EEG, middle: [HbO], bottom: [HbR], left: motor execution, right: motor imagery). The x-axis denotes the center of the moving window. Colored lines show the accuracy for the single subjects while the black line is the average over subjects. The grey bar indicates the time interval of cue presentation.

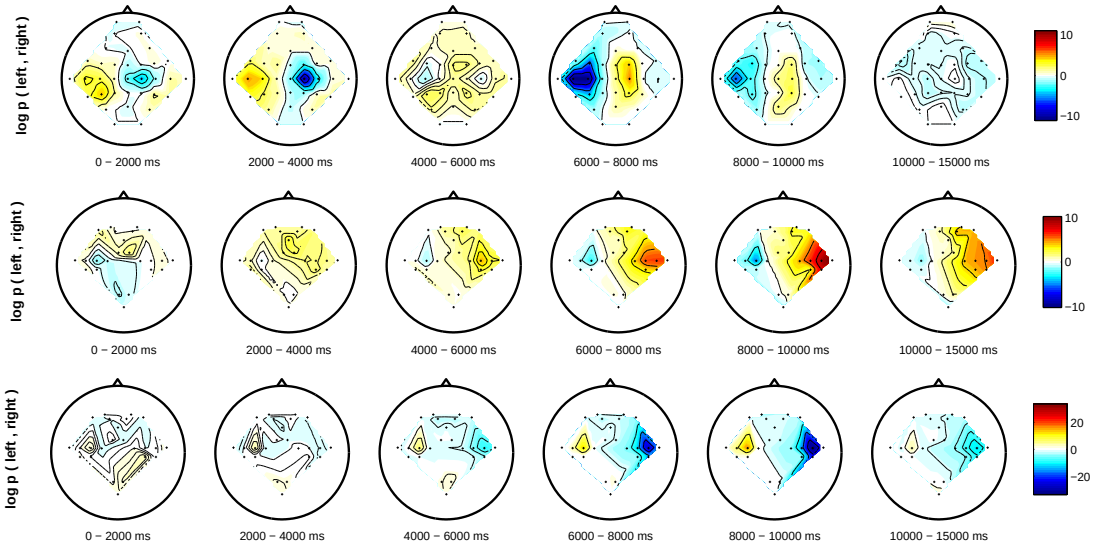

Figure 4.4: Scalp evolution of grand-average log p values for motor execution in EEG and NIRS over all subjects (top: EEG, middle: [HbO], bottom: [HbR]). Red colors denote higher values of the *left* class, while blue colors indicate higher values within the *right* class. Note that the width of the color-scale on the right indicates the level of significance.

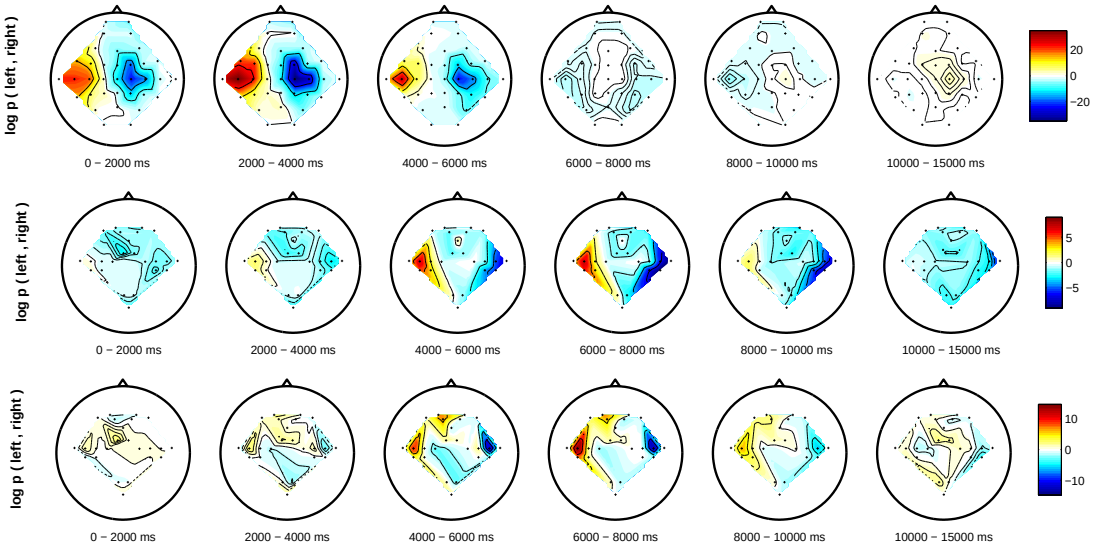

Figure 4.5: Scalp evolution of grand-average log p values for motor imagery in EEG and NIRS over all subjects (top: EEG, middle: [HbO], bottom: [HbR]). Red colors denote higher values of the *left* class, while blue colors indicate higher values within the *right* class. Note that the width of the color-scale on the right indicates the level of significance.

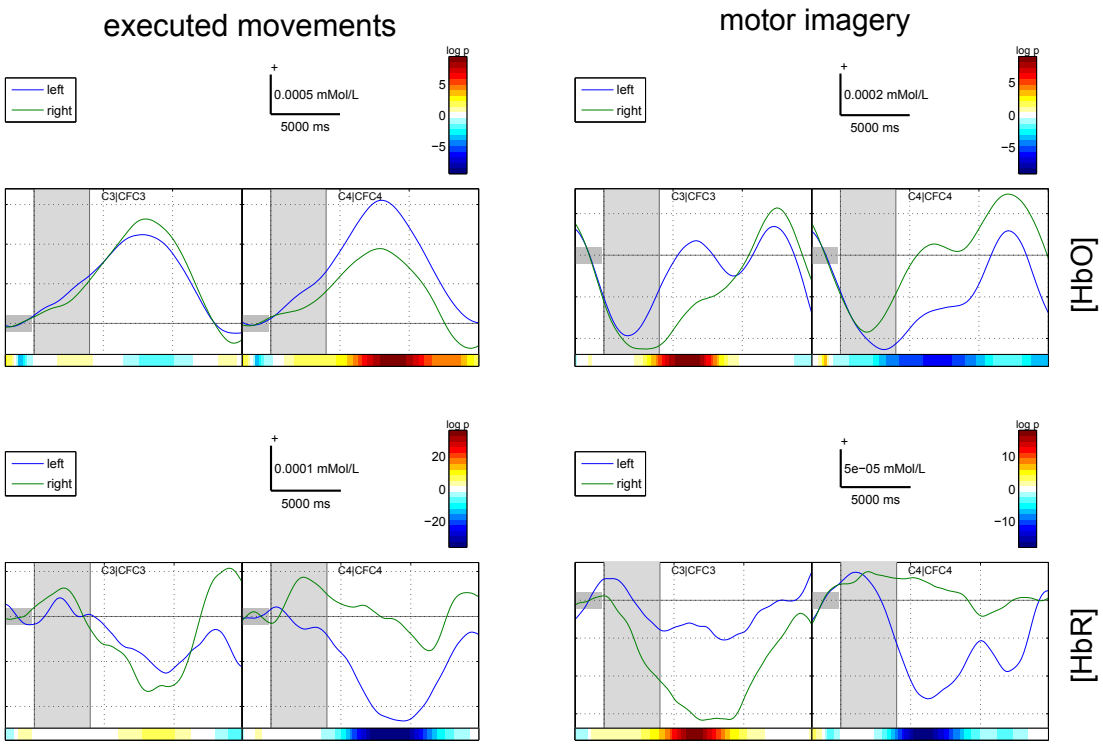

Figure 4.6: Group-average time courses for the two NIRS channels (namely *C3|CFC3* and *C4|CFC4*) with highest discriminability for both conditions (*left* and *right*) and chromophores (*[HbO]* and *[HbR]*). *Executed movement* timecourses are shown on the left panels, while *motor imagery* timecourses on the right. Top panels depict [HbO] and bottom panels [HbR]. The small grey patch before the first vertical line indicates the baseline, which was set from −2 s to 0 s. The second, larger grey patch indicates the time period of cue presentation (0 s to 4 s).

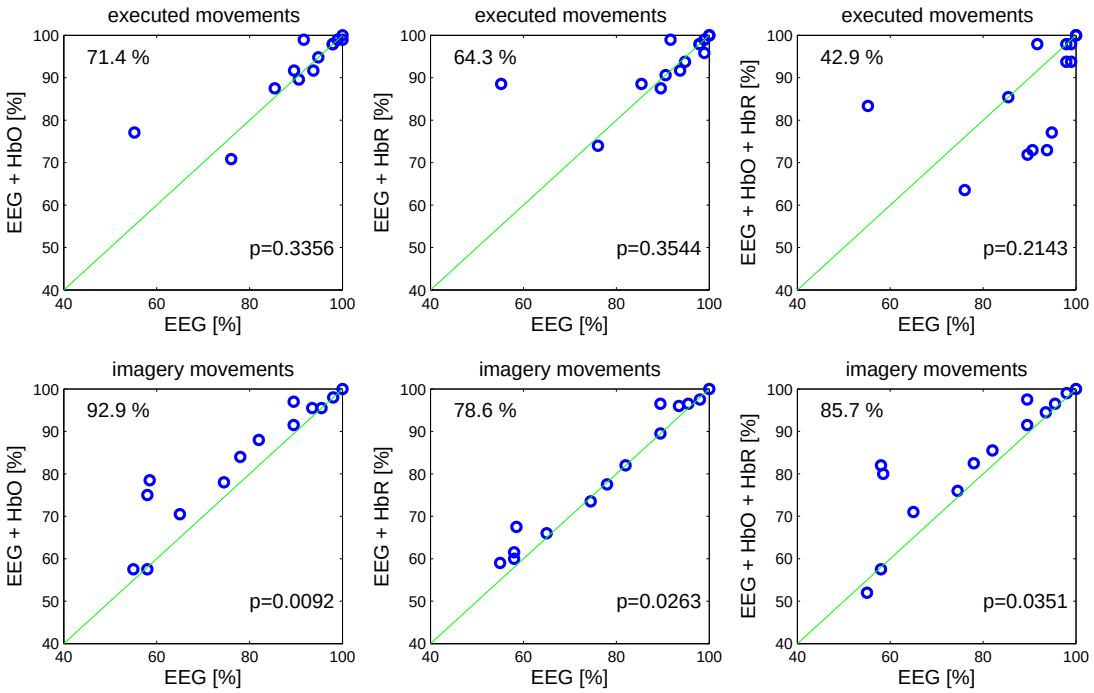

Figure 4.7: Scatter plot comparing classification accuracies and significance values of various combinations of NIRS and EEG for real and motor imagery. The x-axis depicts the EEG classification accuracy. The y-axes depict the classification accuracy of the combinations: EEG + [HbO], EEG + [HbR] and EEG + [HbO] + [HbR] (from left to right).

decoding, and the p-value indicates the significance of the improvement.

While the results in Table 4.1 indicate that combinations of EEG and NIRS are beneficial for average decoding success for both paradigms, only combinations for motor imagery score (highly) significant improvements. When comparing EEG with combined EEG/[HbO] for motor imagery, there was an average 5% classification accuracy increase across all subjects. This increase is highly significant ($p < 0.01$) and the combination scores higher or equal classification rates for 13 out of 14 subjects. Interestingly, two subjects (VPeaa and VPeam) with very bad performance in EEG-BCI were much better classifiable when EEG/NIRS was used (with rates of 81% and 80.5%, respectively). The two other subjects with very low EEG performance, namely VPeac and VPeal, did not show further improvements.

Figure 4.8 shows the relation of the classification performance of the individual measurement methods (EEG, [HbO] and [HbR]) in relation to their mutual information content (I(EEG; [HbO]) and I(EEG; [HbR]). The left column shows these results

VP	executed movements						motor imagery					
	NIRS		EEG	EEG +			NIRS		EEG	EEG +		
	[HbO]	[HbR]		[HbO]	[HbR]	[HbO+R]	[HbO]	[HbR]		[HbO]	[HbR]	[HbO+R]
VPeaa	84.4	99.0	100.0	99.0	100.0	100.0	77.5	65.0	58.5	78.0	70.0	81.0
VPeab	64.6	75.0	85.4	87.5	89.6	85.4	61.5	50.5	98.0	98.5	98.0	98.5
VPeac	69.8	56.2	99.0	99.0	99.0	96.9	62.0	57.5	65.0	69.5	65.0	72.0
VPead	91.7	79.2	99.0	100.0	96.9	93.8	72.0	58.5	74.5	80.0	75.0	77.5
VPeae	69.8	85.4	97.9	97.9	97.9	94.8	80.0	70.0	82.0	86.5	84.0	85.0
VPeaf	59.4	52.1	94.8	93.8	93.8	82.3	57.0	59.0	58.0	57.5	57.5	57.0
VPeag	58.3	69.8	100.0	100.0	100.0	100.0	91.5	90.0	89.5	97.0	96.0	97.0
VPeah	77.1	90.6	55.2	77.1	90.6	85.4	85.0	85.0	95.5	95.5	96.0	95.0
VPeai	52.1	49.0	76.0	68.8	69.8	63.5	52.5	54.5	55.0	56.5	58.5	48.5
VPeaj	60.4	63.5	90.6	91.7	92.7	77.1	81.5	58.5	89.5	93.0	89.0	92.5
VPeak	59.4	59.4	93.8	94.8	91.7	72.9	70.0	65.5	78.0	82.5	79.0	84.5
VPeal	99.0	99.0	91.7	99.0	99.0	99.0	75.0	74.5	93.5	95.0	96.0	94.0
VPeam	65.6	60.4	89.6	89.6	86.5	75.0	76.0	57.0	58.0	75.5	64.0	80.5
VPean	84.4	87.5	97.9	97.9	97.9	97.9	62.0	64.5	100.0	100.0	100.0	100.0
mean	71.1	73.3	90.8	92.6	93.2	87.4	71.7	65.0	78.2	83.2**	80.6*	83.1*

Table 4.1: Individual LDA classification accuracies for features of both NIRS chromophores ([HbO] and [HbR]) and EEG, and their combinations with a meta-classifier. * marks significant ($p < 0.05$), ** highly significant ($p < 0.01$) improvements for the individual combinations of EEG and NIRS features versus plain EEG decoding. p-values are based on paired t-tests.

for executed movements, while the left column shows the results for the motor imagery. Generally speaking the mutual information content rises with higher classification accuracy for all considered methods. If for a given subject method X scores a low classification accuracy, one would expect the conditional entropy $H(X|Y)$ to be of similar magnitude as $H(X)$ and therefore the mutual information content is very low. On the other hand if both methods score very high classification accuracies, $H(X|Y)$ will be low, leading to a high mutual information content.

However, for some subjects we see that, while the classification accuracy of a given method is high, we observe a low mutual information content. This can be interpreted in two ways. Either the other classification method does not work well (and its output is thus very different) or their information content is *complementary*. The average mutual information over all subjects for *executed movements* are given as: $I(\text{EEG};[\text{HbO}]) = 0.125 \pm 0.177$ bit and $I(\text{EEG};[\text{HbR}]) = 0.194 \pm 0.277$ bit. For *motor imagery* $I(\text{EEG};[\text{HbO}]) = 0.096 \pm 0.127$ bit and $I(\text{EEG};[\text{HbR}]) = 0.067 \pm 0.110$ bit.

The left part of Figure 4.9 shows the relation of [HbO] classification performance to [HbO] classification performance of trials that were correctly classified by EEG (*HbO(EEG+)*) and to [HbO] classification performance of trials that were misclassified by EEG (*HbO(EEG-)*). The right part shows the same analysis, comparing EEG classification accuracy to *EEG(HbO+)* and to *EEG(HbO-)*. As can be seen for both plots most points lie close to the angle bisector, only a few blue marks appear below the diagonal. However, these are caused by very small subgroups (the size of the squares encode the number of trials). This means that [HbO] and EEG generally misclassify different trials. If they did not and for example *HbO(EEG+)* would classify more accurately as compared to *HbO(EEG-)*, green dots would generally be substantially higher than blue ones. However, since this is not the case we conclude that the classifier outputs, coming from the two signals are independent to some degree. While we do not explicitly show the results here, results are similar for [HbR].

To illustrate the spatial distribution of significant NIRS features with respect to EEG classification accuracy, we formed two groups of trials: The first group consists of all trials, which were correctly classified by EEG *(EEG+)*, while the second group consists of all trials, which were erroneously classified by EEG *(EEG-)*. For these two groups we calculate the grand average significances of left vs. right hand movement trials. The resulting scalp maps can be seen in Figure 4.10. Two subjects, namely *VPeab* and *VPean*, had to be removed from this analysis since their EEG classification was so accurate (98% and 100%, respectively) that not enough *EEG*-trials were present of both classes (namely left and right hand imagination). In Figure 4.10 the *[HbO]* chromophore is illustrated. This chromophore showed higher classification accuracy for imagined movements over all subjects (the detailed results are given in Table 4.1). As one would expect the level of significance is higher

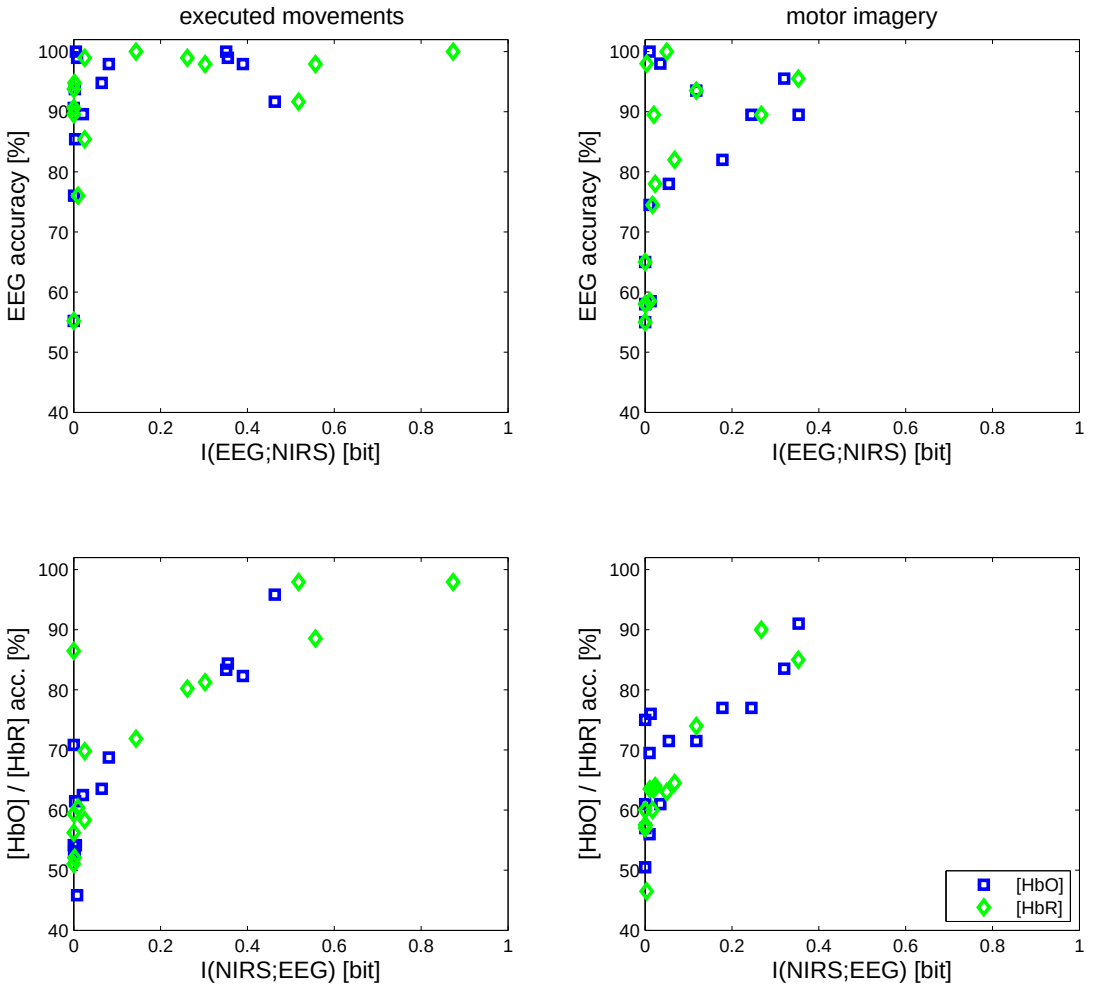

Figure 4.8: Mutual information of EEG and NIRS classifier outputs (x-axes) are compared with their respective classification performances (y-axes). Squares and diamonds represent the results of single subjects (blue - [HbO] ; green - [HbR]). The left column shows *executed movements*, the right column *motor imagery*. top rows: EEG classification accuracy vs. the mutual informations of I(EEG; [HbO]) and I(EEG; [HbR]). bottom rows: classification accuracies of [HbO] and [HbR] vs. their respective mutual information with EEG.

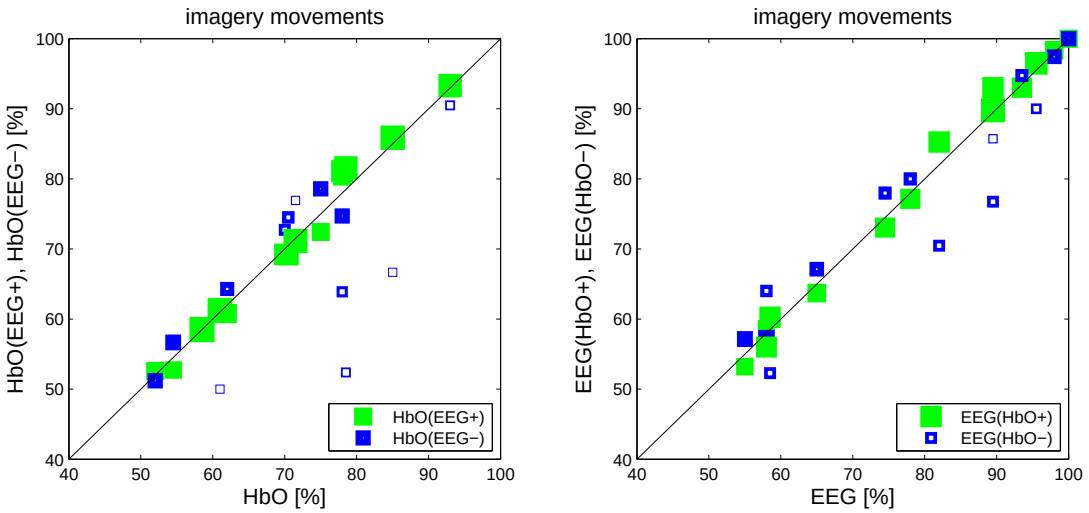

Figure 4.9: Left: Scatter plot comparing [HbO] classification accuracy of all trials to [HbO] classification accuracy, whose EEG classification was correct (green dots) or incorrect (blue dots). Right: comparing EEG classification accuracy of all trials to EEG classification accuracy of trials, where [HbO] was correct/incorrect. The sizes of squares encodes the number of trials.

within the *EEG+* group, as compared to the *EEG-* group. However, we would like to point out that some channels still exhibit highly significant p-values ($p < 10^{-4}$) for the *EEG-* group. Furthermore, and most importantly, we see that the spatial organization shows highest activations within expectable regions of the motor-related cortical areas, very similar to the *EEG+* trials.

4.7 Discussion and Conclusions

Recently BCIs that solely rely on NIRS have been realized [126, 1]. However, when looking at plain NIRS classification rates it becomes apparent that NIRS cannot be seen as a viable alternative to EEG-based BCIs on its own. However, in a combination with EEG we find that NIRS is capable of enhancing event-related desynchronization (ERD)-based BCI performance significantly. Not only does it increase performance for most subjects, but it also allows meaningful classification rates for those who would otherwise not be able to operate a solely EEG-based BCI.

Given that the typical behavior of hemoglobin oxygenation during brain activation consists of an increase in [HbO] approximately mirrored by a decrease of [HbR] [83, 117], for motor imagery (Figure 4.6) only [HbR] clearly showed the typi-

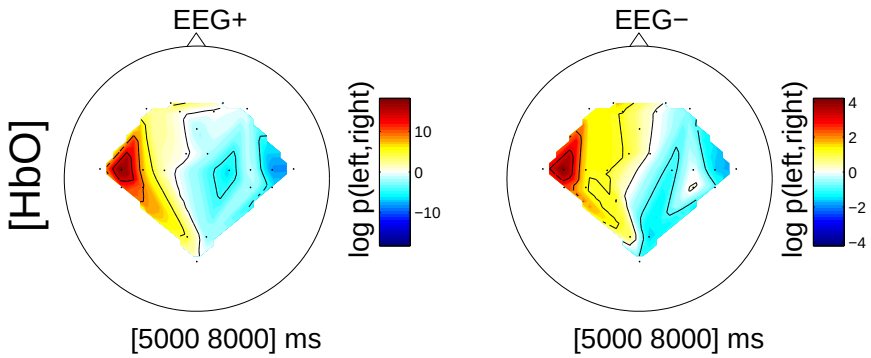

Figure 4.10: Left: Grand average significance of NIRS features for all *correct* EEG trials (EEG+). Right: Grand average significance of NIRS features for all *incorrect* EEG trials (EEG-).

cal behavior. For [HbO] there seems to be an initial drop followed by a subsequent rise. While we have no simple explanation for this finding, following are some considerations which may be relevant: The overall amplitudes during motor imagery are much smaller than during motor movements (note the different scaling) in line with previous fMRI experiments [58]. Therefore, spontaneous fluctuations of [HbO] and [HbR] may appear much more dominant to the point that they can obscure some small simulation-related changes. Since spontaneous fluctuations are much stronger for [HbO] than for [HbR] this may be part of the observed discrepancy. Furthermore, in NIRS with its poorer spatial resolution as compared to fMRI activated and non-activated or deactivated brain areas may be within the sample volume and such partial volume effects may further "dilute" the effect of stimulation. Given that during motor imagery premotor cortex may be activated and primary motor cortex not [58] which is different from motor movements, it seems possible that such partial volume effects have occurred during motor imagery. Finally, as seen in Figure 4.6, [HbO] is rapidly changing during the "baseline" period, i.e. the average of this time period may not serve as an optimal definition of baseline for [HbO] making the quantitative interpretations referring to this baseline difficult. A last consideration refers to potential extracerebral contributions which are stronger for the HbO signal than the HbR signal and which may be related to such systemic factors as e.g. blood pressure. Further research is needed to clarify this point. We are currently preparing a similar study with EEG-feedback controlled SMR-BCI during simultaneous fMRI recordings. Therewith, we will be able to relate BOLD fMRI findings to the EEG and (indirectly) the NIRS recordings. Furthermore, we hope that a simultaneous NIRS-fMRI study with measures of systemic variables such as blood pressure and breathing will give us further evidence of the origin of this effect.

An obvious concern that arises from the addition of NIRS to EEG-based BCI feedback is the long time delay of the hemodynamic response. While we show that classification accuracy increases substantially by employing NIRS, one may rightly argue that information transfer rates, which measure *information per unit time*, could suffer from the inferior temporal responsiveness of such a combination. To this end we would like to offer the following arguments. Firstly, for subjects (and patients) which are not able to operate a BCI, solely based on EEG, this combination presents a viable alternative. Secondly, one could imagine a feedback scenario, where a *secondary* NIRS-derived classifier is only turned on in particular trials, when the *'primary'* EEG-based classification is likely to fail.

In terms of information content, we show that the mutual information of both methods rises with their individual classification accuracy. However, there are also a few examples, where this relationship does not hold true and the mutual information of EEG and NIRS classifier outputs is very small, as compared to their individual accuracies (see Figure 4.8). To further examine these cases we offer an additional analysis, which is given in Figure 4.9. As can be seen here the individual methods mostly misclassify different trials. In combination with the fact, that increased classification accuracy does arise by combining the classifier outputs meaningfully, we interpret these findings such that the individual methods complement each other in terms of information content.

In our study we validated the NIRS data as well as its combinations with EEG in an offline fashion, but our methods could also be applied to a real time experiment. In addition, a large number of potential extensions are possible in order to make the combined system faster to set-up. The current (wet) EEG channels could be replaced by dry electrodes [107, 113, 55] and a zero-training classifier in the spirit of [46, 45, 43] could be established for NIRS. A further interesting aspect would be to study non-stationarities during an experiment [114] and techniques for compensating it [120, 132] also for the present multi-modal BCI setup.

CHAPTER 5

Conlusions and Outlook

In this work many of the shortfalls of state-of-the-art BCI are addressed, such as high hardware-related preparation costs, the need for subject-specific calibration sets and instability of BCI performance across subjects. For most of these given problems novel solutions are introduced, implemented and tested.

The most elementary of EEG-BCI challenges for healthy users is not – at first glance – a computational one. Standard EEG practice involves the tedious application of conductive gel on EEG electrodes in order to provide for accurate measurements of the micro-volt level scalp potentials that constitute EEG signals. Without *dry-cap* technology the proper set-up of BCI sessions in, say, a home environment, is too tedious, messy and therefore impractical. The computational challenges which we have addressed are optimal placement of the reduced number of electrodes and robustness of BCI algorithms to the smaller set of recording sites. With only 6 uni-polar electrodes one can achieve about 70% of full gel cap BCI performance at sites above the motor cortex. The feasibility of the patented dry electrode technology, as has been presented in this thesis, has already lead to a startup company, where an advanced version thereof is being developed as a product.

Our ensemble framework is able to estimate *subject-independent* classifiers for BCI. As seen, these *subject-independent* classifiers are on par with their *subject-dependent* counterparts in terms of classification performance. The difference being that *subject-dependent* classifiers require a calibration dataset of the user. Thus, the proposed approach allows both experienced and novice BCI subjects to engage in BCI feedback sessions immediately without prior calibration. In addition, we show that the ensemble framework, in combination with the appropriate machine learning algorithm, is able to represent the characteristic neurophysiological variation of a large subject group. While we presented results of a motor-imagery paradigm, the given approach may be transferred to a broad range of other experimental designs.

However, EEG-based BCI remains inoperable for some users. To this end we proposed a multi-modal neuroimaging approach, based on NIRS and EEG. It shows that in combination with EEG, NIRS is capable of enhancing ERD-based BCI performance significantly over all subjects. In addition it also allows meaningful classification rates for subjects who would otherwise not be able to operate a solely

EEG-based BCI. Finally, our findings also show that the individual methods complement each other in terms of information content. While in Chapter 4 a multimodal approach for ERD-based BCIs is presented, there is per se no reason why a NIRS-EEG combination would not also be beneficial in other BCI paradigms, such as event-related potential (ERP)-based BCIs or steady-state visual evoked potential (SSVEP)-based BCIs, among others. Future studies will show if these type of combinations will also lead to beneficial results there. It is therefore highly likely that multi-modal approaches, such as ours will become more frequent in the future.

Given the still low information transfer rates of typical non-invasive BCI systems, it is clear that a BCI system will not replace common communication paradigms, such as keyboard and mouse or even telephony of video conferencing in the near future. However the possible applications lie within the domains of patient communication as well as within the gaming industry. Other possible applications, which have recently been discussed and published are rehabilitation purposes after stroke, mental work load monitoring (also in industrial environments) or early breaking detection in an automobile environment, among others.

For the future we anticipate a BCI scenario in which users purchase an affordable computer peripheral which is simply placed on the head and requires no gel. Novel users will not need to undergo a calibration procedure to interact with the BCI system in a game environment, to control a robot, a wheelchair or otherwise. At repeated use, parameters from previous sessions are recalled and re-calibration is rarely necessary. Such a system, capable of an average performance of about >20 bits/min, is achievable within the next few years.

REFERENCES

[1] A. F. Abdelnour and T. Huppert. Real-time imaging of human brain function by near-infrared spectroscopy using an adaptive general linear model. *Neuroimage*, 46:133–143, May 2009.

[2] M. Alamgir, M. Grosse-Wentrup, and Y. Altun. Multitask learning for Brain-Computer Interfaces. In Y. W. Teh and M. Titterington, editors, *Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, pages 17–24, 2010.

[3] G. Andrew and J. Gao. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40, Corvalis, Oregon, 2007. ACM.

[4] G. Bauernfeind, R. Leeb, S. C. Wriessnegger, and G. Pfurtscheller. Development, set-up and first results for a one-channel near-infrared spectroscopy system. *Biomed Tech (Berl)*, 53:36–43, 2008.

[5] H. Berger. Über das Elektroenkephalogramm des Menschen. *Arch. Psychiat. Nervkr.*, 99:555–574, 1933.

[6] F. Bießmann, F. Meinecke, A. Gretton, A. Rauch, G. Rainer, N. Logothetis, and K.-R. Müller. Temporal kernel CCA and its application in multimodal neuronal data analysis. *Machine Learning*, 79:5–27, 2010.

[7] F. Bießmann, S. Plis, F. Meinecke, T. Eichele, and K.-R. Müller. Analysis of Multimodal Neuroimaging Data. *IEEE Reviews in Biomedical Engineering*, submitted, 2011.

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

[9] N. Birbaumer, T. Hinterberger, A. Kübler, and N. Neumann. The thought-translation device (TTD): neurobehavioral mechanisms and clinical outcome. *IEEE Trans Neural Syst Rehabil Eng*, 11(2):120–123, Jun 2003.

[10] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[11] B. Blankertz, G. Curio, and K.-R. Müller. Classifying Single Trial EEG: Towards Brain Computer Interfacing. In T. G. Diettrich, S. Becker, and Z. Ghahramani,

editors, *Advances in Neural Inf. Proc. Systems*, volume 14, pages 157–164. MIT Press, 2002.

[12] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio. The non-invasive Berlin Brain-Computer Interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

[13] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, V. Kunzmann, F. Losch, and G. Curio. The Berlin Brain-Computer Interface: EEG-based communication without subject training. *IEEE Trans Neural Syst Rehabil Eng*, 14:147–152, 2006.

[14] B. Blankertz, G. Dornhege, S. Lemm, M. Krauledat, G. Curio, and K.-R. Müller. The Berlin Brain-Computer Interface: Machine learning based detection of user specific brain states. *Journal of Universal Computer Science*, 12:2006, 2006.

[15] B. Blankertz, M. Kawanabe, R. Tomioka, F. Hohlefeld, V. Nikulin, and K.-R. Müller. Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In *Advances in Neural Inf. Proc. Systems (NIPS 08)*, volume 20, 2008.

[16] B. Blankertz, K.-R. Müller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlögl, G. Pfurtscheller, J. del R. Millán, M. Schröder, and N. Birbaumer. The BCI competition. III: Validating alternative approaches to actual BCI problems. *IEEE Trans Neural Syst Rehabil Eng*, 14(2):153–159, Jun 2006.

[17] B. Blankertz, C. Sannelli, S. Halder, E.M. Hammer, A. Kübler, K.R. Müller, G. Curio, and T. Dickhaus. Neurophysiological predictor of SMR-based BCI performance. *NeuroImage*, 51(4):1303–1309, 2010.

[18] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Proc Magazine*, 25(1):41–56, 2008.

[19] H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection of fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077, 2010.

[20] M. H. Bonnet and D. L. Arand. Impact of activity and arousal upon spectral EEG parameters. *Physiol. Behav.*, 74:291–298, Oct 2001.

[21] R. Boostani and M. H. Moradi. A new approach in the BCI research based on fractal dimension as feature and adaboost as classifier. *J. Neural Eng.*, 1:212–217, 2004.

[22] Peter Brunner, Anthony L. Ritaccio, Joseph F. Emrich, Horst Bischof, and Gerwin Schalk. Rapid communication with a "P300" matrix speller using electrocorticographic signals (ECoG). *Front Neuroscience*, 5:5, Feb 2011.

[23] S. Butterworth. On the theory of filter amplifiers. *Experimental Wireless and the Wireless Engineer*, 7:536–541, 1930.

[24] A. Buttfield, P. W. Ferrez, and J. Millan. Towards a robust BCI: error potentials and online learning. *IEEE Trans Neural Syst Rehabil Eng*, 14:164–168, Jun 2006.

[25] R. Canton. The electric currents of the brain. *British Medical Journal*, 2:278, 1875.

[26] Britton Chance, Qingming Luo, Shoko Nioka, David C. Alsop, and John A. Detre. Optical investigations of physiology. a study of intrinsic and extrinsic biomedical contrast. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1354):707–716, 1997.

[27] M. Cheng, X. Gao, S. Gao, and D. Xu. Design and implementation of a brain-computer interface with high transfer rates. *IEEE Trans Biomed Eng*, 49:1181–1186, Oct 2002.

[28] M. Cope, D. T. Delpy, E. O. Reynolds, S. Wray, J. Wyatt, and P. van der Zee. Methods of quantitating cerebral near infrared spectroscopy data. *Adv. Exp. Med. Biol.*, 222:183–189, 1988.

[29] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 10.1007/BF00994018.

[30] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, January 1967.

[31] S. Coyle, T. Ward, and C. Markham. Brain-computer interface using a simplified functional near-infrared spectroscopy system. *J Neural Eng*, 4:219–226, Sep 2007.

[32] S. Coyle, T. Ward, C. Markham, and G. McDarby. On the suitability of near-infrared (NIR) systems for next-generation brain-computer interfaces. *Physiol Meas*, 25:815–822, Aug 2004.

[33] F. Lopes da Silva. *Dynamics of EEGs as signals of neuronal populations: Models and theoretical considerations*. Williams & Wilkins, London, fourth edition, 1999.

[34] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Trans Biomed Eng*, 51:993–1002, Jun 2004.

[35] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors. *Towards Brain-Computer Interfacing*. MIT Press, 2007.

[36] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161, Cambridge, MA, 1997. MIT Press.

[37] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition, November 2001.

[38] A. C. Ehlis, T. M. Ringel, M. M. Plichta, M. M. Richter, M. J. Herrmann, and A. J. Fallgatter. Cortical correlates of auditory sensory gating: a simultaneous near-infrared spectroscopy event-related potential study. *Neuroscience*, 159:1032–1043, Mar 2009.

[39] T. Elbert, B. Rockstroh, W. Lutzenberger, and N. Birbaumer. Biofeedback of slow cortial potentials. I. *Electroencephalogr. Clin. Neurophysiol.*, 48:293–301, 1980.

[40] C. Falco, F. Sebastiano, L. Cacciola, F. Orabona, R. Ponticelli, P. Stirpe, and G. Di Gennaro. Scalp electrode placement by EC2 adhesive paste in long-term video-EEG monitoring. *Clin Neurophysiol*, 116:1771–1773, Aug 2005.

[41] J. Farquhar, N. J. Hill, T. N. Lal, and B. Schölkopf. Regularised CSP for sensor selection in BCI. In G. R. Müller-Putz, C. Brunner, R. Leeb, R. Scherer, A. Schlögl, S. Wriessnegger, and G. Pfurtscheller, editors, *3rd International Brain-Computer Interface Workshop and Training Course 2006*, pages 14–15, Graz, Austria, 09 2006. Verlag der Technischen Universität Graz.

[42] S. Fazli, C. Grozea, M. Danóczy, B. Blankertz, K.-R. Müller, and F. Popescu. Ensembles of temporal filters enhance classification performance for erd-based bci systems. In *4th International Brain-Computer Interface Workshop and Training Course 2006*. Verlag der Technischen Universität Graz, 2008.

[43] S. Fazli, C. Grozea, M. Danoczy, B. Blankertz, F. Popescu, and K.-R. Muller. Subject independent EEG-based BCI decoding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 513–521. MIT Press, 2009.

[44] S. Fazli, J. Mehnert, G. Curio, A. Villringer, K.-R. Müller, J. Steinbrink, and B. Blankertz. Enhanced performance by a hybrid NIRS-EEG Brain Computer Interface. *NeuroImage*, pages 519–529, January 2012.

[45] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea. Subject independent mental state classification in single trials. *Neural Networks*, 22:1305–1315, 2009.

[46] Siamac Fazli, Márton Danóczy, Jürg Schelldorfer, and Klaus-Robert Müller. ℓ_1-penalized linear mixed-effects models for high dimensional data with application to BCI. *NeuroImage*, 56(4):2100 – 2108, 2011.

[47] E. A. Felton, J. A. Wilson, J. C. Williams, and P. C. Garell. Electrocorticographically controlled brain-computer interfaces using motor and sensory imagery in patients with temporary subdural electrode implants. Report of four cases. *J. Neurosurg.*, 106:495–500, Mar 2007.

[48] D. Flotzinger, G. Pfurtscheller, C. Neuper, W. Mohl, and H. Berger. Classification of non-averaged EEG data by learning vector quantization and the impact of signal processing. In *Proceedings of the 15th Annual International Conference of the IEEE*, pages 263–264, 1993.

[49] Karl J. Friston. Modalities, modes, and models in functional neuroimaging. *Science*, 326(5951):399–403, 2009.

[50] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, Boston, 2nd edition edition, 1990.

[51] A.S. Gevins, D. Durousseau, and J. Libove. Dry electrode brain wave recording system, Oct 1990. US Patent 4,967,038.

[52] D. Goldman. The clinical usage of the 'average' reference electrode in monopolar recording. *Electroenceph. clin. Neurophysiol.*, 2:209, 1950.

[53] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming (web page and software). *http://stanford.edu/ boyd/cvx*, 2008.

[54] T. Grossmann, R. Oberecker, S. P. Koch, and A. D. Friederici. The developmental origins of voice processing in the human brain. *Neuron*, 65:852–858, Mar 2010.

[55] C. Grozea, C.D. Voinescu, and S. Fazli. Bristle-sensors - Low-cost Flexible Passive Dry EEG Electrodes for Neurofeedback and BCI applications. *J. Neural Eng.*, 8:025008, 2011.

[56] S. Harmeling, G. Dornhege, D. Tax, F. Meinecke, and K.-R. Müller. From outliers to prototypes: ordering data. *Neurocomputing*, 69(13-15):1608–1618, 2006.

[57] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction (Springer Series in Statistics)*. Springer New York, 2nd edition, 2001.

[58] D. Hermes, M.J. Vansteensel, A.M. Albers, M.G. Bleichner, M.R. Benedictus, C. Mendez Orellana, E.J. Aarnoutse, and N.F. Ramsey. Functional MRI-based identification of brain areas involved in motor imagery for implantable brain-computer interfaces. *J Neural Eng*, 8:025007, Apr 2011.

[59] M. J. Herrmann, T. Huter, M. M. Plichta, A. C. Ehlis, G. W. Alpers, A. Muhlberger, and A. J. Fallgatter. Enhancement of activity of the primary visual cortex during processing of emotional stimuli as measured with event-related functional near-infrared spectroscopy and event-related potentials. *Hum Brain Mapp*, 29:28–35, Jan 2008.

[60] D. Hill and R. Keynes. Opacity changes in stimulated nerve. *The Journal of Physiology*, 108:278–281, 1949.

[61] Harold Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232, 1953.

[62] J.G. Ibrahim, H. Zhu, R.I. Garcia, and R. Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503, 2011.

[63] F. F. Jobsis. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323):1264–1267, 1977.

[64] S. Kanoh, Y. M. Murayama, K. Miyamoto, T. Yoshinobu, and R. Kawashima. A NIRS-based brain-computer interface system during motor imagery: system development and online feedback training. *Conf Proc IEEE Eng Med Biol Soc*, 2009:594–597, 2009.

[65] A. Kleinschmidt, H. Obrig, M. Requardt, K. D. Merboldt, U. Dirnagl, A. Villringer, and J. Frahm. Simultaneous recording of cerebral blood oxygenation changes during human brain activation by magnetic resonance imaging and near-infrared spectroscopy. *J. Cereb. Blood Flow Metab.*, 16:817–826, Sep 1996.

[66] L. Kocsis, P. Herman, and A. Eke. The modified Beer-Lambert law revisited. *Phys Med Biol*, 51:N91–98, Mar 2006.

[67] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint Conference on artificial intelligence*, volume 14, pages 1137–1145, 1995.

[68] Z. J. Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalogr Clin Neurophysiol*, 79:440–447, Dec 1991.

[69] Z. J. Koles and A. C. K. Soong. EEG source localization: implementing the spatio-temporal decomposition approach. *Electroencephalogr. Clin Neurophysiol*, 107:343–352, 1998.

[70] M. Krauledat. *Analysis of Nonstationarities in EEG signals for improving Brain-Computer Interface performance*. PhD thesis, Technische Universität Berlin, Fakultät IV – Elektrotechnik und Informatik, 2008.

[71] M. Krauledat, G. Dornhege, B. Blankertz, and K.-R. Müller. Robustifying EEG data analysis by removing outliers. *Chaos and Complexity Letters*, 2:259–274, 2007.

[72] M. Krauledat, M. Schröder, B. Blankertz, and K.-R. Müller. Reducing calibration time for brain-computer interfaces: A clustering approach. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Inf. Proc. Systems (NIPS 07)*, volume 19, pages 753–760, 2007.

[73] M. Krauledat, P. Shenoy, B. Blankertz, R. P. N. Rao, and K.-R. Müller. Adaptation in CSP-based BCI systems. In *Toward Brain-Computer Interfacing*, pages 305–309. MIT Press, 2007.

[74] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller. Towards zero training for brain-computer interfacing. *PLoS ONE*, 3:e2967, 2008.

[75] K. K. Kwong, J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, and R. Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. U.S.A.*, 89:5675–5679, Jun 1992.

[76] Mikhail A. Lebedev and Miguel A.L. Nicolelis. Brain-machine interfaces: past, present and future. *Trends in Neurosciences*, 29(9):536 – 546, 2006.

[77] J. H. Lee, J. Ryu, F. A. Jolesz, Z. H. Cho, and S. S. Yoo. Brain-machine interface via real-time fMRI: preliminary study on thought-controlled robotic arm. *Neurosci. Lett.*, 450:1–6, Jan 2009.

[78] R. Leeb, H. Sagha, R. Chavarriaga, and J. Del R Millan. Multimodal Fusion of Muscle and Brain Signals for a Hybrid-BCI. *Conf Proc IEEE Eng Med Biol Soc*, 1:4343–4346, 2010.

[79] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Trans Biomed Eng*, 52:1541–1548, Sep 2005.

[80] S. Lemm, T. Dickhaus, B. Blankertz, and K.-R. Müller. Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387 – 399, 2011.

[81] M. Sande Lemos and B. J. Fisch. The weighted average reference montage. *Electroenceph. clin. Neurophysiol.*, 1991:361–370, 1990.

[82] E. C. Leuthardt, K. J. Miller, G. Schalk, R. P. Rao, and J. G. Ojemann. Electrocorticography-based brain computer interface–the Seattle experience. *IEEE Trans Neural Syst Rehabil Eng*, 14:194–198, Jun 2006.

[83] U. Lindauer, G. Royl, C. Leithner, M. Kühl, L. Gold, J. Gethmann, M. Kohl-Bareis, A. Villringer, and U. Dirnagl. No evidence for early decrease in blood oxygenation in rat whisker cortex in response to functional activation. *Neuroimage*, 13:988–1001, Jun 2001.

[84] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412:150–157, Jul 2001.

[85] S. Luu and T. Chau. Decoding subjective preference from single-trial near-infrared spectroscopy signals. *J Neural Eng*, 6:016003, Feb 2009.

[86] David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

[87] A. Maki, Y. Yamashita, Y. Ito, and H. Koizumi. Spatial and temporal analysis of human motor activity using noninvasive NIR topography. *Med. Phys.*, 22, 1995.

[88] Tadayuki Matsuo, Kazuhiro Iinuma, and Masayoshi Esashi. A barium-titanate-ceramics capacitive-type eeg electrode. *Biomedical Engineering, IEEE Transactions on*, BME-20(4):299 –300, July 1973.

[89] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society*, 70:53–71, 2008.

[90] K. J. Miller, G. Schalk, E. E. Fetz, M. den Nijs, J. G. Ojemann, and R. P. Rao. Cortical activity during motor execution, motor imagery, and imagery-based online feedback. *Proc Natl Acad Sci U S A*, 107:4430–4435, Mar 2010.

[91] K.-R. Müller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz. Machine learning techiques for brain-computer interfaces. *Biomed Tech*, 49(1):11–22, 2004.

[92] Y. Murayama, F. Bießmann, F. C. Meinecke, K.-R. Muller, M. Augath, A. Oeltermann, and N. K. Logothetis. Relationship between neural and hemodynamic signals during spontaneous activity studied with temporal kernel CCA. *Magn Reson Imaging*, 28(8):1095 – 1103, Jan 2010.

[93] M. A. Nicolelis. Actions from thoughts. *Nature*, 409:403–407, Jan 2001.

[94] P. L. Nunez. Methods to estimate spatial properties of dynamic cortical source activity. In G. Pfurtscheller and F. H. Lopes da Silva, editors, *Functional Brain Imaging*, pages 3–9. Hans Huber, Toronto, 1988.

[95] P.L. Nunez. Toward a quantitative description of large-scale neocortical dynamic function and EEG. *Behavioral and Brain Sciences*, 23(03):371–398, 2000.

[96] P.L. Nunez and R. Srinivasan. *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA, 2006.

[97] H. Obrig, H. Israel, M. Kohl-Bareis, K. Uludag, R. Wenzel, B. Muller, G. Arnold, and A. Villringer. Habituation of the visually evoked potential and its vascular response: implications for neurovascular coupling in the healthy adult. *Neuroimage*, 17:1–18, Sep 2002.

[98] H. Obrig and A. Villringer. Beyond the visible–imaging the human brain with light. *J. Cereb. Blood Flow Metab.*, 23:1–18, Jan 2003.

[99] L. Parra, C. Alvino, A. Tang, B. Pearlmutter, N. Yeung, A. Osman, and P. Sajda. Linear spatial integration for single-trial detection in encephalography. *Neuroimage*, 17:223–230, Sep 2002.

[100] G. Pfurtscheller. Event-related synchronization (ers): an electrophysiological correlate of cortical areas at rest. *Electroencephalography and Clinical Neurophysiology*, 83(1):62 – 69, 1992.

[101] G. Pfurtscheller, B. Z. Allison, C. Brunner, G. Bauernfeind, T. Solis-Escalante, R. Scherer, T. O. Zander, G. Mueller-Putz, C. Neuper, and N. Birbaumer. The Hybrid BCI. *Front Neurosci*, 4:42, 2010.

[102] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *Neuroimage*, 31(1):153–159, May 2006.

[103] G. Pfurtscheller, T. Solis-Escalante, R. Ortner, P. Linortner, and G.R. Muller-Putz. Self-paced operation of an SSVEP-based orthosis with and without an imagery-based 'brain switch:' a feasibility study towards a hybrid BCI. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 18(4):409–414, Aug. 2010.

[104] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-Plus*. Springer, New York, 2000.

[105] T. Pistohl, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring. Prediction of arm movement trajectories from ECoG-recordings in humans. *J Neurosci Methods*, 167:105–114, Jan 2008.

[106] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[107] F. Popescu, S. Fazli, Y. Badower, B. Blankertz, and K.-R. Müller. Single trial classification of motor imagination using 6 dry EEG electrodes. *PLoS ONE*, 2(7):e637, 2007.

[108] F. Popescu, S. Fazli, Y. Badower, and K.-R. Müller. Dry electrode cap for electro-encephalography. *WO/2008/067839, PCT/EP2006/011843*, 2008.

[109] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans Rehabil Eng*, 8(4):441–446, 2000.

[110] N. F. Ramsey, M. P. van de Heuvel, K. H. Kho, and F. S. Leijten. Towards human BCI applications based on cognitive brain systems: an investigation of neural signals recorded from the dorsolateral prefrontal cortex. *IEEE Trans Neural Syst Rehabil Eng*, 14:214–217, Jun 2006.

[111] S. Rossi, I. B. Jurgenson, A. Hanulikova, S. Telkemeyer, I. Wartenburger, and H. Obrig. Implicit Processing of Phonotactic Cues: Evidence from Electrophysiological and Vascular Responses. *J Cogn Neurosci*, Jul 2010.

[112] J. Schelldorfer, P. Bühlmann, and S. Van de Geer. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

[113] Eric W. Sellers, Peter Turner, William A. Sarnacki, Tobin Mcmanus, Theresa M. Vaughan, and Robert Matthews. A novel dry electrode for brain-computer interface. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part II*, pages 623–631, Berlin, Heidelberg, 2009. Springer-Verlag.

[114] P. Shenoy, M. Krauledat, B. Blankertz, R. P. N. Rao, and K.-R. Müller. Towards adaptive classification for BCI. *Journal of Neural Engineering*, 3(1):R13–R23, 2006.

[115] R. Sitaram, H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu, and N. Birbaumer. Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain-computer interface. *Neuroimage*, 34:1416–1427, Feb 2007.

[116] B. Sorger, B. Dahmen, J. Reithler, O. Gosseries, A. Maudoux, S. Laureys, and R. Goebel. Another kind of 'BOLD Response': answering multiple-choice questions via online decoded single-trial brain signals. *Prog. Brain Res.*, 177:275–292, 2009.

[117] J. Steinbrink, A. Villringer, F. Kempf, D. Haux, S. Boden, and H Obrig. Illuminating the BOLD signal: combined fMRI-fNIRS studies. *Magnetic Resonance Imaging*, 24:495–505, 2006.

[118] G. Strangman, J.P. Culver, J.H. Thompson, and D.A. Boas. A quantitative comparison of simultaneous bold fmri and nirs recordings during functional brain activation. *NeuroImage*, 17(2):719–731, 2002.

[119] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[120] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:985–1005, December 2007.

[121] B.A. Taheri, R.T. Knight, and R.L. Smith. A dry electrode for EEG recording. *Electroencephalography and clinical neurophysiology*, 90(5):376–383, 1994.

[122] M. Takeuchi, E. Hori, K. Takamoto, A. H. Tran, K. Satoru, A. Ishikawa, T. Ono, S. Endo, and H. Nishijo. Brain cortical mapping by simultaneous recording of functional near infrared spectroscopy and electroencephalograms from the

whole brain during right median nerve stimulation. *Brain Topogr*, 22:197–214, Nov 2009.

[123] Robert F. Tate. Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of Mathematical Statistics*, 25(3):603–607, Sep. 1954.

[124] S. Telkemeyer, S. Rossi, S. P. Koch, T. Nierhaus, J. Steinbrink, D. Poeppel, H. Obrig, and I. Wartenburger. Sensitivity of newborn auditory cortex to the temporal structure of sounds. *J. Neurosci.*, 29:14726–14733, Nov 2009.

[125] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[126] T. Tsubone, T. Muroga, and Y. Wada. Application to robot control using brain function measurement by near-infrared spectroscopy. *Conf Proc IEEE Eng Med Biol Soc*, 2007:5342–5345, 2007.

[127] J.J. Vidal. Toward direct brain-computer communication. *Annu. Rev. Biophys.*, 2:157–180, 1973.

[128] J.J. Vidal. Real-time detection of brain events in EEG. *IEEE Proc*, 65:633–664, 1977.

[129] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz. Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural Computation*, 23(3):791–816, 2011.

[130] A. Villringer and B. Chance. Non-invasive optical spectroscopy and imaging of human brain function. *Trends Neurosci.*, 20:435–442, Oct 1997.

[131] A. Villringer, J. Planck, C. Hock, L. Schleinkofer, and U. Dirnagl. Near infrared spectroscopy (NIRS): A new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience Letters*, 154(1-2):101 – 104, 1993.

[132] Paul von Bünau, Frank C. Meinecke, Franz C. Király, and Klaus-Robert Müller. Finding stationary subspaces in multivariate time series. *Phys. Rev. Lett.*, 103(21):214101, Nov 2009.

[133] S. Waldert, H. Preissl, E. Demandt, C. Braun, N. Birbaumer, A. Aertsen, and C. Mehring. Hand movement direction decoded from MEG and EEG. *J Neurosci*, 28:1000–1008, Jan 2008.

[134] S. Wang, Z. Lin, and C. Zhang. Network boosting for BCI applications. *Book Series Lecture Notes in Computer Science*, 3735:386–388, 2005.

[135] I. Wartenburger, J. Steinbrink, S. Telkemeyer, M. Friedrich, A. D. Friederici, and H. Obrig. The processing of prosody: Evidence of interhemispheric specialization at the age of four. *Neuroimage*, 34:416–425, Jan 2007.

[136] N. Weiskopf, R. Veit, M. Erb, K. Mathiak, W. Grodd, R. Goebel, and N. Birbaumer. Physiological self-regulation of regional brain activity using real-time functional magnetic resonance imaging (fMRI): methodology and exemplary data. *Neuroimage*, 19:577–586, Jul 2003.

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

[138] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clin Neurophysiol*, 113:767–791, Jun 2002.

[139] S. Wray, M. Cope, D.T. Delpy, J.S. Wyatt, and E.O.R. Reynolds. Characterization of the near infrared absorption spectra of cytochrome aa3 and haemoglobin for the non-invasive monitoring of cerebral oxygenation. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 933(1):184–192, 1988.

[140] S. C. Wriessnegger, J. Kurzmann, and C. Neuper. Spatio-temporal differences in brain oxygenation between movement execution and imagery: a multi-channel near-infrared spectroscopy study. *Int J Psychophysiol*, 67:54–63, Jan 2008.

[141] S. S. Yoo, T. Fairneny, N. K. Chen, S. E. Choo, L. P. Panych, H. Park, S. Y. Lee, and F. A. Jolesz. Brain-computer interface using fMRI: spatial navigation by thoughts. *Neuroreport*, 15:1591–1595, Jul 2004.

[142] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, 68:49–67, 2006.

[143] M.H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561, 1993.