Analysis of Nonstationarities in EEG Signals for Improving Brain-Computer Interface Performance

vorgelegt von
Diplom-Mathematiker
Matthias Krauledat

Von der
Fakultät IV, Elektrotechnik und Informatik,
der Technischen Universität Berlin

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Reinhold Orglmeister
Berichter: Prof. Dr. Klaus-Robert Müller
Berichter: Prof. Dr. Stefan Jähnichen

Tag der wissenschaftlichen Aussprache: 13. 3. 2008

Berlin 2008
D 83

Acknowledgments

I would like to thank Prof. Dr. Klaus-Robert Müller for his inspiring ideas and for his infectious enthusiasm for research. He gave me the opportunity to work in this amazing research field; this work would not have been possible without him. Special thanks go to the BBCI project leader, Dr. Benjamin Blankertz, for his constant support. Under his guidance, I learned a lot about thorough data analysis and scientific reasoning. I also thank all the members of the Intelligent Data Analysis (IDA) group for creating such a welcoming and friendly working environment. I enjoyed to jointly elaborate new ideas, and I had many fruitful discussions. In particular, I would like to thank my friend Dr. Guido Dornhege with whom it was fun to work, and whose ideas and energy were always inspiring. Likewise, thanks to Dr. Michael Tangermann for our joint work. I would further like to thank our partners Prof. Dr. Gabriel Curio, Dr. Florian Losch and Dr. Volker Kunzmann from the Charité Berlin, and Pradeep Shenoy from the University of Washington, for all our successful collaborations.

I thank Prof. Dr. Stefan Jähnichen and Prof. Dr. José Millán, who kindly agreed to review my thesis.

I am indebted to all the proofreaders, Dr. Benjamin Blankertz, Christine Carl, Dr. Guido Dornhege, Dr. Motoaki Kawanabe, Frank Meinecke, Verena Scharstein, Dr. Michael Tangermann, Dr. Carmen Vidaurre and Thorsten Zander, who with their suggestions helped to improve the quality of this thesis. Thanks to Dr. Anton Schwaighofer for his LaTeX“telephone hotline”.

This work was supported by grants of the *Deutsche Forschungsgemeinschaft* (DFG), FOR 375/B1, and of the *Bundesministerium für Bildung und Forschung* (BMBF), FKZ 01IBE01A/B and FKZ 01IBB02A/B. A large part of the work has been performed at Fraunhofer FIRST, and some of it at the University of Potsdam.

Finally, my sincere gratitude goes to my family and to Verena, for constant backup and encouragement.

Abstract

Brain-Computer Interface (BCI) research aims at the automatic translation of neural commands into control signals. These can then be used to control applications such as text input programs, electrical wheelchairs or neuroprostheses. A BCI system can, e.g., serve as a communication option for severely disabled patients or as an additional man-machine interaction channel for healthy users. In the classical "operant conditioning" approach, subjects had to undergo weeks or months of training to adjust their brain signals to the use of the system. The Berlin Brain-Computer Interface project (BBCI) has developed an Electroencephalogram-(EEG-)based system which overcomes the need for operant conditioning with advanced machine learning methods. By adapting classifiers to the highly subject-specific brain signals, even subjects with no prior experience in BCI can achieve high information transfer rates from their first session.

However, after an initial calibration, the brain signals are rarely so stationary that the first classifier can be reused in the next experimental session. Even if the classifier was fitted to the subject on data from the same day, we sometimes encountered long periods of low performances. These drawbacks can clearly impede the continuous use of the system, which is particularly important for disabled people.

The reason for this flaw is the nonstationarity in the EEG data. Due to changes in the characteristic properties of the data, classification can often be corrupted.

In this work, I will present a new framework for nonstationary data analysis, which encompasses methods for the quantification and visualization of nonstationary processes. The analysis of data acquired in BCI experiments will be used to exemplify the power of the methods. In particular, I show some neurophysiological evidence for the sources of the nonstationarity. Once the underlying reasons for the nonstationarity are known, classification can be adaptively enhanced; I will present some surprisingly simple methods. Finally, I will construct classifiers that are largely robust against the changes from one experimental session to the next. This novel type of classifiers can be applied without initial calibration and has the potential to drastically improve the applicability of BCI devices for daily use.

While the BCI scenario was used as a testbed for the framework, it can be applied to a wide range of problems. Nonstationarity can occur in any field of machine learning, whenever the measured systems under observation change their properties over time.

Zusammenfassung

Ein Brain-Computer Interface (BCI, “Gehirn-Computer-Schnittstelle”) ist ein System, das neuronale Kommandos in Steuersignale umsetzt. Diese können genutzt werden, um Anwendungen wie Texteingabeprogramme, elektrische Rollstühle oder Neuroprothesen zu steuern. Ein BCI kann beispielsweise Schwerstbehinderten zur Kommunikation verhelfen, oder auch gesunden Benutzern einen zusätzlichen Kanal zur Mensch-Maschine-Interaktion bieten. Im klassischen Ansatz, der “operanten Konditionierung”, mussten Benutzer in einem wochen- bis monatelangen Training ihre Gehirnstrommuster an die Wirkungsweise des Systems anpassen. Das Berliner Brain-Computer Interface (BBCI) hingegen hat ein auf dem Elektroenzephalogramm (EEG) basierendes System entwickelt, das durch den Einsatz von neuartigen Methoden des maschinellen Lernens keine Konditionierung mehr benötigt. Hierbei passen sich Klassifikatoren automatisch an die Daten an, die zwischen den Benutzern oft stark variieren. So können selbst Benutzer, die zum ersten Mal mit einem BCI arbeiten, hohe Informationstransferraten erzielen.

Nach der anfänglichen Kalibrierung sind die Gehirnströme jedoch selten so stationär, dass der Klassifikator der ersten Sitzung auch später erfolgreich angewandt werden kann. Selbst bei Klassifikatoren vom gleichen Tag können manchmal längere Abschnitte auftreten, in denen die Übertragungsraten sehr niedrig sind. Durch diese Probleme wird der permanente Gebrauch des Systems beeinträchtigt, der gerade für Behinderte besonders wichtig ist.

Der Grund dafür ist die Nicht-Stationarität in EEG-Signalen. Durch Veränderung der charakteristischen Eigenschaften der Daten wird die Klassifikation oft in Mitleidenschaft gezogen.

In dieser Dissertation werde ich eine Theorie für die Analyse nicht-stationärer Daten entwickeln, die Methoden für die Quantifizierung und Visualisierung nicht-stationärer Prozesse beinhaltet. Anhand der Analyse von Daten aus BCI-Experimenten werde ich die Effizienz dieser Methoden veranschaulichen. Insbesondere werde ich neurophysiologische Anhaltspunkte für Quellen der Nicht-Stationarität aufzeigen. Sind die Prozesse bekannt, die der Nicht-Stationarität zugrunde liegen, kann man die Klassifikation durch Adaptation verbessern. Hierzu werde ich einige erstaunlich einfache Methoden entwickeln. Abschliessend werde ich Klassifikatoren konstruieren, die gegenüber Veränderungen von einer experimentellen Sitzung zur nächsten weitgehend robust sind. Diese neuartige Kategorie von Klassifikatoren kann ohne anfängliche Kalibrierung angewandt werden und hat daher das Potential, die tägliche Benutzbarkeit von BCI-Systemen zu ermöglichen.

Obwohl ausschliesslich BCI-Daten zur Auswertung herangezogen wurden, können die Methoden auf eine Vielzahl von Problemen angewandt werden. Nicht-Stationarität kann in jedem Bereich des maschinellen Lernens auftreten, sobald sich die Eigenschaften der beobachteten Systeme zeitabhängig verändern.

Contents

1. Introduction

> "There is nothing so stable as change"
> (Bob Dylan)

Nonstationarity is an ubiquitous problem in signal processing and machine learning, when dynamical systems change their properties over time. It appears in application fields where the estimation of the state of a process relies on "real-world" data, typically acquired by (multiple) sensors. In some prominent research areas such as automatic processing of audio and video data and problems like speech recognition, image recognition and gesture detection, nonstationarity has been recognized as an important challenge. In all these fields, the application of automatized algorithms at different points in time has to be performed with special care. Problems can particularly arise if the algorithms rely on calibration data or the estimation of parameters on small fractions of the available data.

This thesis will address this problem with a new framework for nonstationarity in machine learning applications; it encompasses methods for the analysis and visualization and quantification of data. In particular, I will add a new perspective to data analysis, by regarding the parameters of machine learning algorithms as input for my methods. This perspective allows for abandoning the simplistic view of machine learning methods as "black box" systems; the methods carry valuable information – incorporated in their parameters – about the problems they are applied to.

Brain-Computer Interface research aims at the automatic translation of neural commands into control signals. These signals can then be used to control applications such as text input programs, wheelchairs or neuroprostheses. A BCI system can, e.g., serve as a communication option for severely disabled patients or as an additional man-machine interaction channel for healthy users. In the classical "operant conditioning" approach, subjects had to undergo weeks or months of training to adjust their brain signals to the use of the system. The Berlin Brain-Computer Interface project (BBCI), of which I am a member, has developed an Electroencephalogram-(EEG-)based system which overcomes the need for operant conditioning with advanced machine learning methods. By adapting classifiers to the highly subject-specific brain signals, even subjects with no prior experience in BCI can achieve high information transfer rates from their first session.

However, after an initial calibration, the brain signals are rarely so stationary that the first classifier can be reused in the next experimental session. Even if the classifier was fitted to the subject on data from the same day, we sometimes encountered long periods of low performances. These drawbacks can clearly impede the continuous use of the system, which is particularly important for disabled people. This makes Brain-Computer Interface research a particularly difficult and inspiring application area with respect to nonstationarity. The reasons for the changes in the dynamics are among the following:

- The physical properties of the sensors (such as electrodes and amplification units) change over time. This can be due to drying of the conductive electrode gel (which depends on the room temperature and the gel consistency) or electromagnetic fields of nearby power lines.

- Neurophysiological conditions (e.g., awakeness), even in a single subject, can display a large variability. This can also affect mental strategies for the communication with the devices.
- Psychological parameters, such as attention, task involvement and motivation, are also variable over time.
- Finally, BCI always requires two interacting “systems”, namely the user and the computer, whose internal states depend on each other.

This list already gives an impression of the various disciplines involved in BCI research. More specifically, some preliminary knowledge in each of them is required for the successful construction of BCI devices:

- **Metrology:** It is a highly challenging task for, e.g., electrical engineers to develop new devices for the measurement of brain activity. For a suitable implementation of a BCI, the preparation times as well as weight and size of the devices have to be reduced, while providing convenient use and high resolution in frequency and space.
- **Neurophysiology:** For the localization of neuronal processes and the development of paradigms, neurophysiological experience is required. In another perspective, BCI devices can serve to get insight into specific brain functions.
- **Psychology:** This research field is required for developing models for the interaction of the user with a machine.
- **Computer Science** with subdisciplines:
  - **Signal Processing:** Data of brain activity, such as EEG data, are high-dimensional time series with a low signal-to-noise ratio. Therefore, advanced signal processing has to be applied to reveal the relevant part of the signals.
  - **Machine Learning:** Brain signals are highly subject-specific and display a large variability. Therefore, if the application should be adapted to the user’s brain signals, it is required to adjust specific settings by automated machine learning methods.
  - **Software Engineering:** For online BCI experiments, it is necessary to bring together realtime data acquisition, data analysis and the display for the user (typically graphical feedback). The developed applications have to be safe and comprehensible for the user and need to incorporate all the specifications from the above-mentioned research fields.

1.1. Outline of this work

My work in the BBCI project, as well as the work presented in this thesis, has been mainly in the domain of computer science. In a highly interdisciplinary research area such as BCI research, there are some prerequisites to understanding and communicating ideas. Therefore, I will introduce some basic facts about measurement techniques, neurophysiology and paradigms in Section 2.1. I will then summarize some methods of signal processing and machine learning in Section 2.2, and will report on the state of the art in the Berlin BCI (BBCI) in Section 2.3. Then, I will introduce a notion of nonstationarity in chapter 3, as required for modeling machine learning processes.

In chapters 4–6, I will describe the main achievements of my work[1].

- **Analysis of Nonstationary Signals:** I developed methods for the quantification and visualization of the changes of these signals over time. Applying these methods to BCI data, I discovered that a commonly encountered source of nonstationarity is the influence of a particular frequency modulation in the visual cortex. I will present these findings and show how much the features of the EEG are affected by the change. I will also show how much the classification performance is impaired by the nonstationarity (see chapters 4 and 5).

- **Adaptive Classification:** If the reasons for the nonstationarity are known, it is possible to find remedies. One option is the development of adaptive classifiers. For this purpose, it is crucial to assess the appropriate update frequency and the amount of data required for the first update, since an adaptation with too few samples can lead to a bad estimation of the adaptation parameters, and consequently to a degraded classifier. Considering that for many subjects the standard classifiers already work quite well, I will also analyze how much of the classifier's structure should be preserved. As a result, I found a surprisingly simple, but effective adaptation method for the course of a single session (see Chapter 5).

- **Classifiers for Session-to-Session transfer:** The changes in the brain signals from one session to the next can also be regarded as nonstationary. I developed a framework in which the classifiers across sessions can be compared. This comparison led to a new method for training a classifier which works without lengthy calibration measurement. This reduces the preparation time for BCI experiments drastically. In Chapter 6, I will show the feasibility of this method with online BCI feedback experiments.

[1] Parts of this thesis are based on work published in Krauledat et al. [68] (chapter 4), Shenoy et al. [126] (Section 5.1), Krauledat et al. [71] (Section 5.2) and Krauledat et al. [70] (section 6.2).

2. Basic BCI Ingredients

2.1. Brain-Computer Interface (BCI)

Brain-Computer Interfaces are systems which allow for direct control of, e.g., a computer application or a neuroprosthesis, solely by human intentions as reflected by suitable brain signals ([150]). The usual motor output pathways and peripheral nerves are bypassed in this approach. It is this independence which makes the development of a BCI such an important and realistic choice for the construction of prostheses for severely handicapped people, such as patients suffering from tetraplegia or late stages of Amyotrophic Lateral Sclerosis (ALS). These patients can have a strongly reduced communication ability due to their physical condition; in the case of ALS, it can even lead to a state where they are "completely locked in", which means that they have no control over any muscle. For these groups of patients, a BCI can also be a useful option as a communication device. Besides the obvious use for the severely disabled, other applications such as the development of games or the constant monitoring of attentional states in working environments with high risks are conceivable.

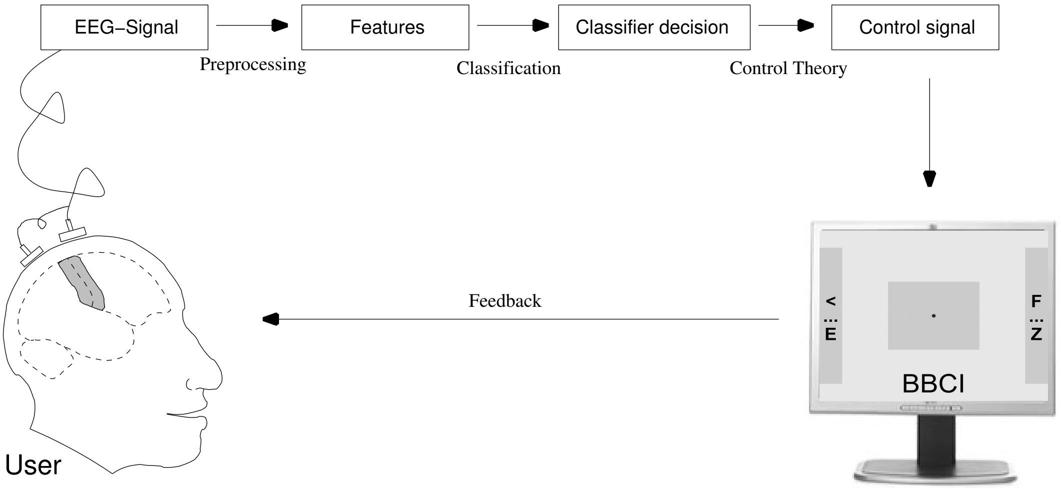


Figure 2.1.: In this figure, the classical feedback-loop of a Brain-Computer Interface is depicted: the user's brain signals are acquired, and by different methods of feature extraction, classification and control logic, a feedback is generated and presented to the user.

The development of such a device is a highly interdisciplinary research topic, which brings together scientists from many different fields, such as psychology, neurophysiology, physics, engineering, mathematics and computer science. Therefore, this work will, although mainly focused on the data analysis and computer science part of BCI, always try to point into the related topics and fields.

In this first chapter, I will report on different options for BCIs in terms of measurement techniques for neural activity. Specifically for the EEG measurement, I will describe some of the most prominent neural correlates of brain functions for the analysis of EEG data, and then some of the methods for the extraction and classification of the corresponding features. Finally, I will report on the methods commonly used in the Berlin Brain-Computer Interface (BBCI) project, which I will analyze throughout this work.

2.1.1. Measurement Techniques

This section summarizes some of the most common brain-imaging methods. I will discuss the usefulness in a BCI setting for each of the methods below, according to important parameters like the degree of invasivity, the portability of the system, but also the signal-to-noise ratio, time-frequency– and the spatial resolution.

Microelectrodes

Microelectrodes can be bundles of wire electrodes or silicon-based electrode arrays, arranged in a dense grid. The size of these electrodes depends on the type as well as on the material used; the width usually ranges from 5 to around 100 μm. After inserting them deep in the tissue of the cortex, they can be used to record action potentials from single neurons as well as signals from larger neuron populations. Microelectrodes are also used for the electrical stimulation of specific brain regions. Due to the high risk of infections and tissue damage, BCI research with these electrodes was restricted to animals like monkeys, where features like the firing rates of neurons can be translated into control signals. In recent experiments, see [57], it was shown that this type of feedback can also be performed by a human user who is willing to undergo surgery for the implantation of the electrodes. Due to the small recording sensors, the integration of these devices into a portable BCI system is quite realistic. Some groups reported high information transfer rates and online control of motor prostheses with this recording technique (see [26, 104, 32, 138]). However, the longterm signal stability still remains an issue to be resolved (see e.g. [131, 78, 43]), since the movement of the electrode relative to the cortex, as well as the scarring process in the tissue deteriorate the signal quality over time. This drawback makes the insertion of microelectrodes impossible for the longterm use.

Electrocorticography (ECoG)

Electrocorticography (ECoG) signals are acquired with a set of electrodes placed directly on the brain. The electrode grid, a flexible foil or strip with imprinted electrodes, is usually located subdurally (i.e., under the dura mater, a thick membrane inside the skull), but an epidural (i.e., above the dura mater) location is also possible. With this technique it is not possible to record activity of single neurons, but compared to extracranial EEG recordings (see below), the signal is less attenuated and less exposed to spatial smearing by the skull and tissue layers. Therefore, the signal has a higher signal-to-noise ratio and higher spatial and temporal resolution. The influence of muscle artifacts, which is quite high for ordinary EEG, has also been reported to be reduced. Since ECoG is often used for finding the locus of epileptic seizures, some BCI research has been conducted with patients who had implanted electrode grids while they were under medical surveillance (see e.g. [81, 75]).

Positron Emission Tomography (PET)

For the preparation of Positron Emission Tomography scans, radioactive tracer isotopes with a short decay period are incorporated into metabolically active molecules (e.g., sugar). If these molecules are injected into the blood flow, the decay can be measured especially at positions where high metabolic activity is performed ([115]). In neurophysiological research, this measurement can be performed to determine regions of high neuronal activity. A long list of drawbacks includes the long time lag due to the metabolic and hemodynamic response as well as the risks connected to the dose of ionizing radiation. Also the scanner needed to detect the isotope decay is quite big, such that it can not be part of a portable system for BCI use.

functional Magnetic Response Imaging (fMRI)

Changes in the blood flow and in the blood oxygenation in the brain are related to the neural activity, since nerve cells consume oxygen in an active state. fMRI recordings are performed by applying a strong magnetic resonance pulse and measuring the response of the atoms in the body. Since oxygenated hemoglobin, the oxygen carrier protein in the blood, has different magnetic properties than deoxygenated hemoglobin, it is possible to determine a Blood Oxygenation Level Dependent (BOLD) change. While the spatial resolution of this method is very high, the temporal resolution is low due to the hemodynamic time lag. Significant BOLD changes can only be encountered after some seconds of neural activity. The signal is an indirect measure for brain activity, since it does not measure the communication between cells, but rather the energy household of cell populations.

Moreover, fMRI devices are large and stationary, due to the parts that generate the magnetic field. Lightweight and portable devices can not be constructed in a straightforward way. Despite these limitations, fMRI signals have recently been used for BCI research, e.g. [54, 147, 103, 128].

Near-Infrared Spectroscopy (NIRS)

Similar to the functioning principle of fMRI, the near-infrared spectroscopy relies on the physical differences between oxygenated and deoxygenated hemoglobin. The differences can be found in a modified light absorption in the near-infrared light frequency band. Therefore, neural activity can be measured according to the hemodynamic response. The sensors are typically placed on the head, accompanied by infrared light emitting sources aimed at the scalp surface. The light in this band penetrates the scalp to a sufficient depth as to allow for acquiring the vascular activity of the cerebral cortex. Being bound to a similarly low temporal resolution as the fMRI, the NIRS setup can be made sufficiently small to construct a portable system for BCI use (see [79, 129]).

Magnetoencephalography (MEG)

MEG recordings measure the magnetic field which is induced by the ionic currents flowing in the dendrites of neurons in the brain. Note that due to the orthogonality of magnetic field and electrical current, only large neuron populations with dendrites oriented tangentially to the scalp surface can be recorded in the MEG. With its high spatial and temporal resolution, it is a very promising recording technique which has already been used in proof-of-concept BCI experiments ([3, 75, 87]). Unfortunately, the possible applications are very limited due to the size of the recording device and to the necessity of a shielded room, since even small electrical devices induce noise which superimposes the signal of interest.

Electroencephalography (EEG)

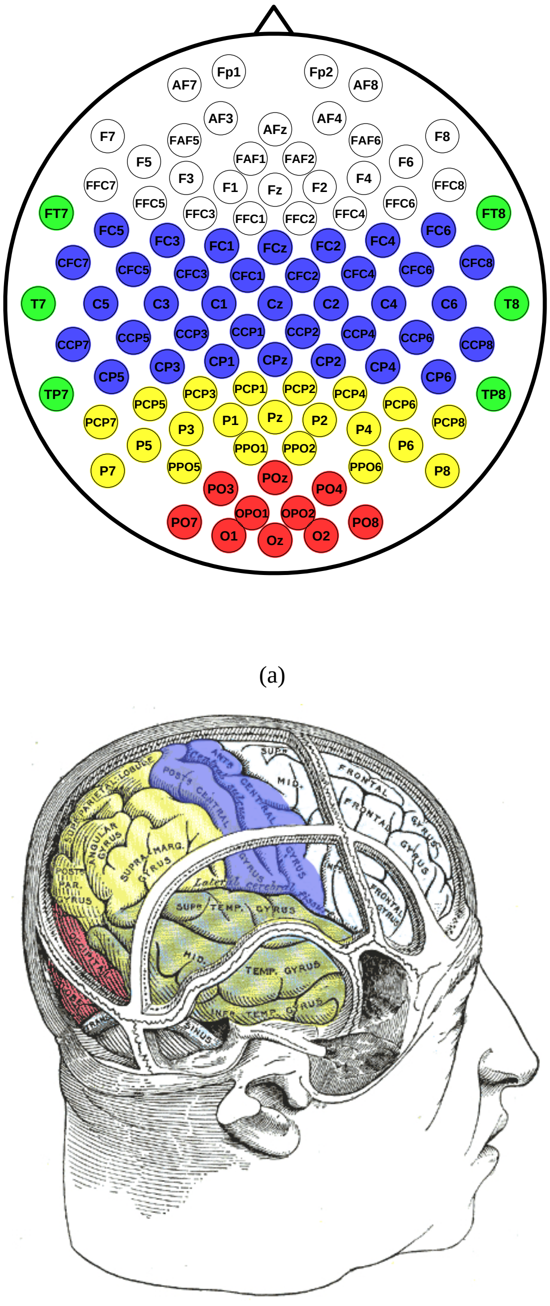
The recording technique that I will be focusing on for the rest of this thesis, the EEG, is a non-invasive, small sized recording device. When the impedance of electrodes on the scalp surface is lowered sufficiently by applying conductive gel, the EEG signals, which rely on the ionic current of neural activity, can be acquired at high spatial and temporal accuracy. The signal of interest is a modulation of the electrical scalp potential at a particular electrode position (see figure 2.2) with respect to one or more reference electrodes. If a large neuronal population is orthogonally oriented with respect to the scalp surface, its induced potentials are large enough to be acquired at the electrodes outside the head. In the EEG setup of the BBCI, the reference electrode is attached to the nose, to prevent muscle activity to deteriorate the signal.

Although most BCI research with human subjects is conducted with the EEG ([6, 150, 113, 11]), the applicability suffers from the long preparation time for the application of the electrodes on the head. Currently, a lot of effort is put into the development of "dry electrodes" to overcome this restriction (see e.g. [117]). The interpretability of the EEG signals as neuronal activity of particular regions of the brain is restricted by the spatial smearing induced by the layers of tissue, skull and hair which separate the sensors from the cortex. Furthermore, EEG recordings are often distorted by noise from various sources. This topic will be discussed in detail in Section 4.1.

2.1.2. Neural Features of the EEG

This section is intended as an overview of the features of scalp EEG that are most frequently used for BCI purposes. [41] provides a broader and more complete review.

Many of the features described here are event-related potentials (ERP), a class of neural signals that large populations of neurons emit phase-locked to some event. Classical ERP phenomena are evoked potentials, P300, the error potentials and the Lateralized Readiness Potential (LRP), which I will introduce below. The common method for the analysis of ERPs is to average the time course over many independently recorded trials, locked to the stimulus or response event. By the independence assumption, the signal-to-noise ratio improves as the number of trials is increased, such that the underlying ERP is visible with sufficiently many recorded trials. Note that it is a crucial requirement for BCI research that the signals are classified in single trial analysis; therefore methods are needed which classify the signals without the necessity to average over many recorded trials beforehand. See section 4.1 for



(a)

(b)

Figure 2.2.: Part (a) of this figure shows the arrangement of EEG electrodes according to an extension of the so-called “10-20” system, [60]. In this figure, just as in all similar plots that follow, the head is projected as viewed from above, the small triangle on top of the circle marks the nose. The colors are encoding the scalp region these electrodes are assigned to: frontal (white), central (blue), parietal (yellow), occipital (red) and temporal (green) electrodes. Typical reference electrodes are attached to the nasion or to mastoids. Alternatively, scalp electrodes such as Cz or Fz can be used for referencing. – Part (b) of this figure shows the major lobes of the brain, as viewed from the temporal perspective, inside the head. The color coding has been synchronized with the electrode montage. Adapted from [51].

methods for robustifying the averaging process, such that fewer trials are required.

Other correlates of neural activity are oscillatory features such as the Event-Related (De-) Synchronization. For most of this work, I will be focusing on the ERD/ERS complex, which will be explained below.

Steady State Visually Evoked potentials (SSVEP)

Attended visual stimuli, presented at a fixed frequency between 6 Hz and 24 Hz, elicit a rhythm in the posterior visual areas with the same fundamental frequency; [55] even reports significant rhythm modulations up to a frequency of 100 Hz. The evoked response is relatively stable and easy to detect in many subjects and has therefore often been used for BCI purposes, see e.g. [88, 101, 99]. The basic setup of these interfaces consists of several targets that are highlighted at a different frequency. By measuring the brain response and comparing the elicited frequency to the stimulus frequencies, it can then be decided which target the subject is focusing on. Note that for this visual attention, it is necessary that the users have control over their eye movements and are not otherwise visually impaired. For some patients, e.g. those suffering from an advanced stage of ALS, this requirement is not met.

Recently, a tactile variant, the so-called Steady-State Somatosensory Evoked Potential (SSSEP), has been explored in [100]. It has been shown that the responses in the EEG to attention shifts of healthy subjects to either of their index fingers which are under constant tactile stimulation, can be classified at accuracies between 70 and 80%.

P300

The P300 component of the EEG is a positive potential that occurs in the context of the "oddball paradigm", where a series of standard stimuli is randomly interleaved with non-standard stimuli, termed "deviants", see [135]. After the presentation of each deviant, the large positive potential, which occurs with a relative latency of approx. 300 ms to the stimulus, is called P300 (or P3). This positive component, which is not present in the standard trials, is predominantly found in parietal electrodes. Amplitude and shape of this component are known to be influenced by various factors, such as the deviant-to-standard ratio, the presentation modality (e.g. visual, auditorial or tactile), attendance and task relevance.

The first use of the P300 in the BCI context has been demonstrated in [45] and [31], where a matrix with 6 rows and 6 columns contained all 26 letters of the alphabet and the 10 digits. While the subject was instructed to attend to a single letter, the rows and columns were highlighted randomly. The correct letter was decoded by averaging over the rows and columns separately and by selecting the row/column pair to which the subject responded with the largest P300 component.

A variant of this approach, relying on auditory stimuli, has been presented in [49], where proof-of-concept experiments were reported.

Although P300 speller feedback systems are studied extensively by many researchers (e.g. [74, 86, 125]), it can still a quite exhausting experience for the users, since the concentration on the flickering symbols (or is often reported as annoying.

Auditory Evoked Potentials

In [56], the time-locked response to auditory stimuli was used for classification. Concurring sequences of auditory stimuli were presented separately to both ears of the subject. The shift of the subjects' attentional focus on either of the two sequences changed the neuroelectrical responses to the stimuli such that binary BCI decisions were possible.

Error Potentials

During the evaluation of the correctness of an event, the so-called error potentials modulate in amplitude between an "error" event and a "correct" event. The event-related response can be divided in two different responses, namely a slow negative potential wave (termed error negativity) and the following positive potential (error positivity), see [44]. While the negative wave is present in both correct and wrong trials and only changes the amplitude, the error positivity can only be seen in error trials and is out of these two the more discriminating feature.

While it is hard to imagine a paradigm where a Brain-Computer Interface is solely controlled with error potential features, some research has been conducted on using it as an add-on to existing BCI systems, see e.g. [46, 110, 15, 120]. If a choice has been taken by the user, the BCI can perform a check for the error negativity and repeat the last decision in case of a positive outcome of this check. If the last choice was erroneous, this repetition gives the user the option to select the correct choice. On the other hand, a "false positive" error potential detection can lead to a much longer decision process, which is obviously not desirable. [34] shows some considerations on the required error correction accuracy, which should be reached as a minimum for successful error correction.

Slow Cortical Potentials

In [6], a brain-computer interface for paralyzed patients was demonstrated; two subjects suffering from advanced amyotrophic lateral sclerosis (ALS) were trained over the period of 4 years to voluntarily produce a slow negative shift of the scalp EEG. They could then use this ability to control a spelling device. Although the training was first intended to produce a negative shift according to [5], it was found that a positive variation was more reliable and more responsive to training with imagery strategies. The signals that were used in this series of experiments were termed "slow cortical potentials" (SCP).

Lateralized Readiness Potentials (LRP)

According to the model known as homunculus, for each part of the human body there is a respective region in the motor and somatosensory area of the neocortex. The 'mapping' from the body to the respective brain areas preserves topography, i.e., neighboring parts of the body are represented in neighboring parts of the cortex. While the region of the feet is at the center of the vertex, the left hand is represented lateralized on the right hemisphere and the right hand on the left hemisphere. In the preparation of hand movements, such as keystrokes with the fingers, a slow negative potential is building up in the corresponding regions of the brain. This process is called "Lateralized Readiness Potential" or "Bereitschaftspotential".

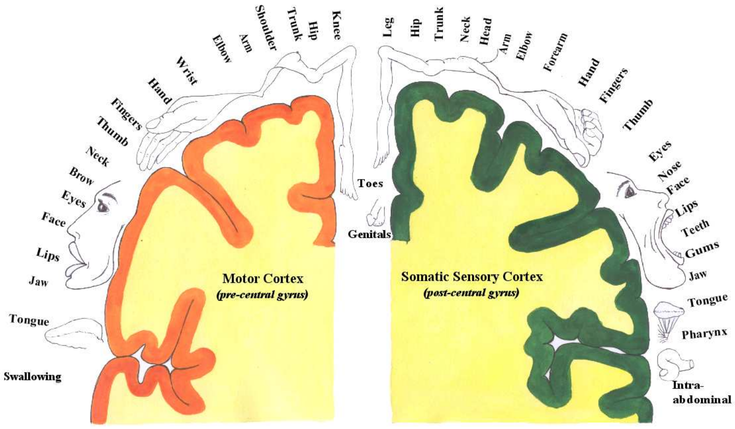


Figure 2.3.: This figure shows two sections of the central area of the brain (shown in blue in figure 2.2), viewed from a dorsal perspective. The motor cortex, shown on the left, is located in the precentral gyrus (i.e., anterior to the central sulcus), while the somatic sensory cortex, shown in the right part of the figure, lies in the post-central gyrus (i.e., posterior to the central sulcus). The size of the body parts displayed on top of the cortex is shown according to the size of the cortex regions that are representing them. This reveals a slightly different topography for motor cortex and somatic sensory cortex. Note that only the left half of the motor and the right half of the somato-sensory cortex are displayed for simplicity. From [72].

The analysis of multi-channel EEG recordings has shown that the involved brain areas contribute to this shift with different intensity ([29, 76]). The focus is in the frontal lobe of the corresponding motor cortex, i.e., contralateral to the performing hand (see figure 2.4). The laterality of an upcoming hand movement can be classified with high accuracy based on the spatial distribution of this EEG signal, more than 100 ms prior to the actual execution of the movement, see [15, 66, 67].

Phase Synchronization

Phase synchronization is a phenomenon that occurs in many natural systems ([116]), and it is also a measure that is used to quantify the interaction between different sources in the brain. There are many different methods how a phase synchronization can be assessed, e.g. by estimating the difference between instantaneous phases

$$\Delta\phi(t) := \phi_1(t) - \phi_2(t),$$

where t is a point in time. The instantaneous phase can be obtained by Hilbert transform or wavelet analysis.

The pitfall in this kind of analysis is the fact that EEG recordings never represent signals of the actual sources of brain activity, but only their superpositions. This can induce high values of synchronization between different electrodes, even if only a single signal is mixed into both electrodes. In [85, 105, 106], various methods were proposed to counter this

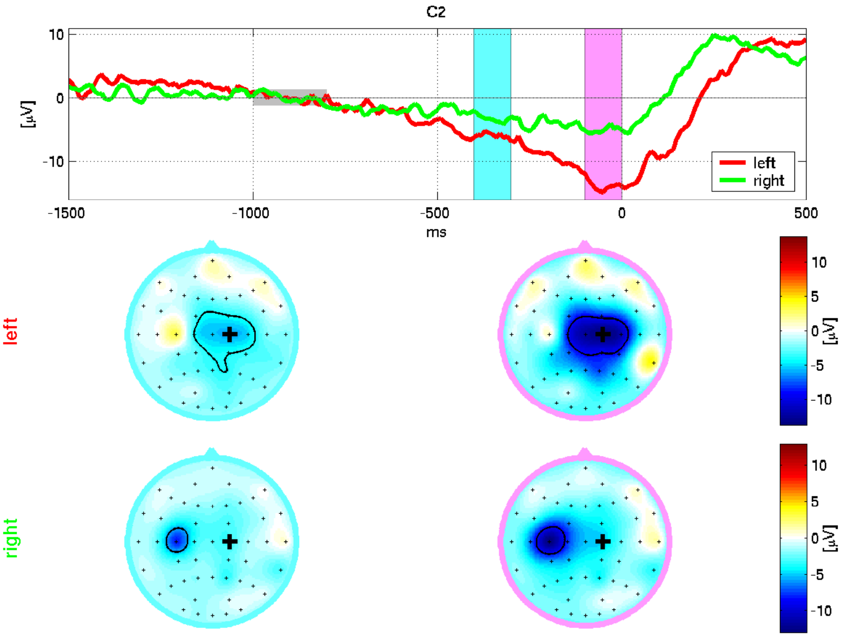


Figure 2.4.: EEG data during the execution of keypresses with left or right little finger reveal an early onset of a slow negativity on central electrodes. While the peak of this process can be found on the left hemisphere for a right hand keypress, it is on the right hemisphere for a left hand keypress. The first row shows the timecourse of the EEG at right-hemisphere electrode C2, averaged over 80 trials per movement type. For the left hand, the curve is clearly below the right hand curve. For each of the marked intervals, scalp topographies are shown separately for left and right hand movement. Electrode C2 is marked with a "+" on these scalp maps.

problem. They involve the unmixing of the sources beforehand by means of Independent Component Analysis (ICA) and then calculating the synchronization index.

In [23], it was demonstrated that even without avoiding the above pitfall, online BCI control can be established. The drawback of this method is the limited interpretability of these results.

Amplitude Modulation of the Sensorimotor Rhythm (SMR)

Some of the event-related changes of the EEG consist either of decreases or increases of the power in given frequency bands. This can be accounted to a decrease or an increase in synchrony of the measured neuronal populations. These phenomena are termed "Event-Related Desynchronization" (ERD) and "Event-Related Synchronization" (ERS) and can be found in EEG and MEG recordings during the execution of a variety of mental states and mental tasks, such as sensory-semantic processing, memory and movement tasks.

Some brain states are characterized by the intensity of specific frequency bands over specific brain areas. A very predominant frequency for the EEG is in the α-band, ranging from approx. 7 Hz to 13 Hz. This frequency band is very strong in the parietal and occipital re-

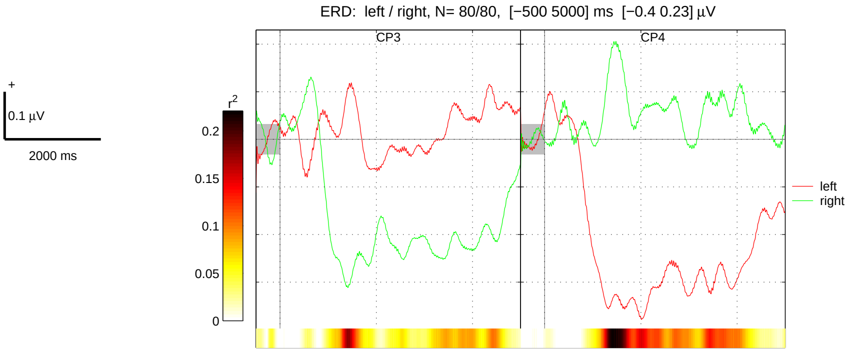


Figure 2.5.: The timecourse of the bandpower from 11–15 Hz, in two selected electrodes over the author's motor region. During imagination of a left hand movement, the bandpower in CP4 (on the right, i.e. contralateral hemisphere) is reduced. During the imagination of a right hand movement, the same holds for electrode CP3 (on the left hemisphere). The bars below denote the discriminability of the curves at every point in time, in terms of r^2-values (see Section 3.1.2). Higher values correspond to a better discriminability.

gion of the cortex, but due to volume conduction, it can also be measured over more frontal electrodes. This rhythm is known to modulate according to visual processing, fatigue and attentional state, see [4, 127].

During executed or imagined hand or foot movements, the μ-rhythm in the corresponding motor area can be observed to be attenuated ([114, 113]). This is a paradigm which can easily be used for BCI purposes, since motor imagery can be performed spontaneously and without previous training. The μ-band is at a similar frequency as the α-band, but the spatial distribution of the ERD of motor tasks is centered at the corresponding motor cortices. As an example for ERD, figure 2.5 shows the author's bandpower over two selected EEG electrodes (CP3 and CP4), during imagination of left and right hand movements. After bandpass filtering the data between 11–15 Hz, a sliding window of 200 ms length was used to estimate the power in these two electrodes. The β-band, at frequencies from 15–30 Hz, is known to undergo similar (de-)synchronization effects.

γ-band (30–80 Hz) oscillations as well as higher frequencies have also been reported to encode information about intended movements; even above this frequency range, information is encoded. [50] presented motor-related amplitude modulations at frequencies up to 200 Hz, which were termed "Very High Frequency Oscillations" (VHFO). In a study on 12 healthy subjects, the laterality of upcoming hand movements could be predicted at high accuracy.

In this work, I will focus on the modulation of frequencies in μ- and β-band ([4, 127, 112]).

2.2. Signal Processing and Machine Learning

For the classification of brain signals, there are two important steps: signal processing (which corresponds to the "feature extraction" process in machine learning terms) and the application of a classifier. For both steps a large variety of options is used by different BCI groups; this is due to the fact that the neural signals of interest (see Section 2.1.2) exhibit such a diversity. This, in turn, entails a large variability of the signal properties. On the BCI Meeting 2005, many researchers tried in a joint effort to find a taxonomy of all methods used for BCI ([82]), but even this list was by no means exhaustive and could only provide a selection of the methods used by some of the workshop participants. Another detailed, but still not complete, list of methods is given in [41]. In this section, I will briefly introduce some methods most of which will be applied later in this work.

2.2.1. Feature Extraction

Feature extraction is a process which is intended to reduce the dimensionality and likewise the complexity of a dataset to a few dimensions with the largest information content. For the application in BCI frameworks, it is an important prerequisite that the process is computationally efficient, robust against noise influences, and only relies on data samples from the past.

While some of the feature extraction methods are generally applicable in the EEG context (such as the frequency filters) or were derived from much different fields of application (such as Independent Component Analysis (ICA)), some of them are specifically taylored to the signals of interest. While, for example, the Common Spatial Patterns (CSP) algorithm was originally introduced as a fairly general method ([48]), that found its way into the BCI research community ([118]), its spatio-temporal extensions ([80, 38, 139, 140]) were developed with the goal to improve the feature extraction process for BCI applications.

Frequency Filters

In some cases it is advisable to reduce the frequency content of the EEG signal to some frequency band of interest; this can be indicated if neurophysiological models suggest that the signal is mainly located at a specific frequency.

Since, for example, the ERD/ERS-complex (see Section 2.1.2) can be found predominantly in the μ- and β-band, it is advisable to apply a frequency filter with this particular bandpass to the signals before extracting bandpower features. With most of the time-frequency representations, such as Fast Fourier Transformation (FFT) or Wavelet Transformations, it is even possible to use the frequency coefficients directly as estimates for the frequency content.

Digital Filters A digital Infinite Impulse Response (IIR) filter consists of two finite sequences $a \in \mathbb{R}^{n_a}$ and $b \in \mathbb{R}^{n_b}$, which are chosen according to specific filter design criteria (see [108]). By convolution with these two sequences, the signal x is filtered to the signal y as follows:

$$\begin{aligned} a(1)y(t) &= \sum_{i=1}^{n_b} b(i)x(t-i-1) \\ &\quad -\sum_{i=2}^{n_a} a(i)y(t-i+1) \end{aligned}$$

for all t.

A special case of IIR filters, a Finite Impulse Response Filter (FIR) is obtained by choosing $n_a = 1$ and $a(1) = 1$, which makes the second term in the above equation vanish. Note that these filters introduce a time delay into the signal.

FFT-based Filters

The Fast Fourier Transformation (FFT) is a mapping of the signal from the time domain to its frequency domain representation ([108]). A filter can be obtained with this mapping by selecting the frequency bins of interest and applying the Inverse Fast Fourier Transformation (IFFT). Since both FFT and IFFT are linear methods, they can be implemented in a computationally efficient way.

Wavelet-based Filters

As a further method of translating signals into their frequency representation, wavelets [30], orthonormal bases of finite time series with a specific frequency content, can be applied. By scaling and translating a prototypical “mother wavelet”, the resulting “daughter wavelets” can approximate the signal efficiently. Again, by restriction of this representation on wavelets within a specific frequency range, the signal can be filtered.

Spatial Filters

If $X \in \mathbb{R}^{T \times C}$ is the matrix representation of EEG data, where T is the number of samples in time and C is the number of channels, a spatial filter for X is any $w \in \mathbb{R}^{C}$. The spatially filtered signal $S \in \mathbb{R}^{T \times 1}$ is then defined by

$$S = X \cdot w.$$

Since every EEG electrode only measures a superposition of signals derived from various sources in the brain, it is a difficult task to find the signal that originates at a specific scalp location. Spatial Filters are tools for the extraction of specific sources, but they can also be used to alleviate the influence of non-cerebral signals such as eye blinks or head movements.

For most neurophysiological analyses, predefined filters which target specific brain regions are defined, e.g. Bipolar filters, Laplace filters and the Common Average Reference method, which itself can be understood as a spatial filter. Although applying the very same filter to different datasets makes the resulting findings more comparable, it does not account for the individual differences between the recordings. Another approach are data-derived filters, obtained from methods like PCA, ICA or CSP. All these methods reflect certain properties of the EEG and the optimal parameters can therefore again be regarded as features of the data. Note that although the inter-subject comparability within the feature space is not granted, the beauty of these methods lies in the duality of the filters: on the one hand, a filter is computed which can be used to derive a signal from a particular source in the brain, on the other hand, a “pattern” that corresponds to the spatial distribution of the same source on the head.

This view is derived from the general framework in which all these methods can be formulated: the measured signal, $X \in \mathbb{R}^{T \times C}$, is a mixture of other (source) signals, $S \in \mathbb{R}^{T \times C}$,

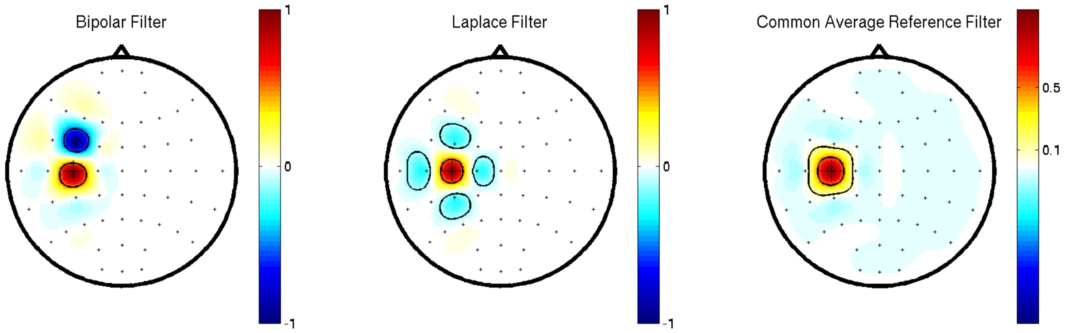


Figure 2.6.: Three commonly used spatial filters in EEG research: the bipolar filter subtracts the signals of two electrodes (in this case C3 and FC3) and the Laplace Filter subtracts the surrounding neighbor electrodes. For the Common Average Reference, the average signal of all electrodes is subtracted from every single electrode. In all three pictures, electrode C3 is marked with a black cross.

with an invertible mixing matrix $A \in \mathbb{R}^{C \times C}$.

$$\begin{aligned} X &= S \cdot A \quad \text{and} \\ S &= X \cdot W, \text{ with } W = A^{-1}. \end{aligned}$$

In the context of EEG analysis, this means the following: the first equation implies that row number i of the "mixing matrix" (i.e., pattern i) denotes the influence of source number i on each electrode. The second equation shows that column number i of the "de-mixing matrix" A^{-1} (i.e., filter number i) denotes the factor with which each electrode must be scaled in order to receive the source signal in column i of S. Both the filters and the patterns can be displayed with their spatial distribution on the scalp.

Bipolar Filter A very simple method of spatial filtering is the differential signal between two (usually neighboring) electrodes. The signals from very distant sources are superimposed over both electrodes with approximately the same intensity. The subtraction can then minimize the influence of these other sources. The signal is then simply calculated as follows:

$$s^{\text{BIP}} := s_{i_1} - s_{i_2}.$$

This corresponds to a filter with the coefficients

$$w_j^{\text{BIP}} = \begin{cases} 1, & j = i_1 \\ -1, & j = i_2 \\ 0, & \text{otherwise.} \end{cases}$$

Bipolar measurements are usually not regarded as an actual filter, since they don't require to apply more than two electrodes. But even this minimal setup is often used for EEG analysis in the BCI context, see e.g. [144]. Figure 2.6 shows a typical Bipolar filter between electrode C3 and FC3.

Laplace Filter Again with the idea of removing signal content which does not originate from near the recording electrode, a Laplace filter subtracts the signal of surrounding electrodes. More specifically, if s_i is the signal recorded at electrode i, and if $s_{i_1}, \dots, s_{i_n}$ are the n electrodes from a neighborhood of electrode i, then

$$s^{\mathrm{LAP}} := s_i - \frac{1}{n} \sum_{j=1}^{n} s_{i_j}$$

is the laplace-filtered signal at electrode i. The parameter n depends on the electrode montage used for recording. The filter has the following coefficients:

$$w_j^{\mathrm{LAP}} = \begin{cases} 1, & j = i \\ -\frac{1}{n}, & j \in \{i_1, \dots, i_n\} \\ 0, & \text{otherwise.} \end{cases}$$

Common Average Reference Although the Common Average Reference is a re-referencing method rather than a filter method, it can still be formulated as a spatial filter. For each electrode, the mean signal over all electrodes is subtracted, i.e.,

$$s_i^{\mathrm{CAR}} := s_i - \frac{1}{C} \sum_{j=1}^{C} s_j$$

is the CAR-signal at electrode i. This corresponds to the following filter:

$$w_j^{\mathrm{CAR}} = \begin{cases} 1 - \frac{1}{C}, & j = i \\ -\frac{1}{C}, & \text{otherwise.} \end{cases}$$

This method can be applied if, for example, the reference electrode introduces some noise into the data. Since it often subtracts very distant channels, some of the spatial resolution of the signals is lost after this transformation.

Principal Component Analysis (PCA) The k principal components of a set of data points $x_1, \dots, x_n \in \mathbb{R}^m$ are the solutions $\hat{y}_1, \dots, \hat{y}_k \in \mathbb{R}^m$ of the optimization problem

$$\min_{y_1, \dots, y_k, a} \sum_{i=1}^{n} ||x_i - (\mu + \sum_{j=1}^{k} a_{i,j} y_j)||_2,$$

where μ is the empirical mean of the data. In other words, PCA components span the k-dimensional affine subspace of $\mathbb{R}^m$ that describes the data with minimal error.

A simple calculation shows (see [42]) that the principal components can be computed as the eigenvectors of the scatter matrix $\Sigma := \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^\top$ corresponding to the k largest eigenvalues. The principal components correspond to the directions with the largest variance. This method is therefore often used for dimensionality reduction. This sort of analysis is useful for the analysis and quantification of unlabeled data, but it does not necessarily reflect the directions with the best discriminability.

PCA has been extended to its non-linear version “kernel PCA” (kPCA), [98]. This algorithm can describe the data in a higher-dimensional space and is therefore not guaranteed to reduce the dimensionality of the data.

Independent Component Analysis (ICA) Instead of representing the data in a least-squares sense, ICA tries to find directions in the data which are most independent from each other. This goal of ICA can be understood in the framework of Blind Source Separation (BSS) as in the beginning of this section:

Suppose the measured signal $X \in \mathbb{R}^{T\times C}$ with T samples and C dimensions is actually a linear mixture of sources, i.e.,

$$X = S\cdot A,$$

where $A \in \mathbb{R}^{C\times C}$ is an unknown mixing matrix and $S \in \mathbb{R}^{T\times C}$ is the source signal.

In order to recover the source signal, a further requirement, namely the statistical independence of the sources, is necessary. If we now try to find a de-mixing matrix W such that the de-mixed signals $S = X\cdot W$ are spatially as independent as possible, the original source signals can be recovered.

Depending on the specific assumptions on the underlying sources, many different approaches have been proposed ([25, 59, 2, 153]). In this work, I will apply FastICA (see [59]) for the identification of outlier trials (section 4.1) and IBICA (see [83, 84]) for finding inlier points in a set of parameters (section 5.1).

Common Spatial Patterns The Common Spatial Pattern (CSP) algorithm is very useful in calculating spatial filters for detecting ERD/ERS effects ([63]) and can be applied to ERD-based BCIs, see [118]. It has been extended to multi-class problems in [36], and a robustified version has been proposed for making it invariant to influences by other signals, such as changes in the visual α-bandpower.

Given two distributions in a high-dimensional space, the (supervised) CSP algorithm finds directions (i.e., spatial filters) that maximize variance for one class and simultaneously minimize variance for the other class. After having band-pass filtered the EEG signals to the rhythms of interest, high variance reflects a strong rhythm and low variance a weak (or attenuated) rhythm. Let us take the example of discriminating left hand vs. right hand imagery. The filtered signal corresponding to the desynchronization of the left hand motor cortex is characterized by a strong motor rhythm during imagination of right hand movements (left hand is in idle state), and by an attenuated motor rhythm during left hand imagination. This criterion is exactly what the CSP algorithm optimizes: maximizing variance for the class of right hand trials and at the same time minimizing variance for left hand trials. Furthermore the CSP algorithm calculates the dual filter that will focus on the area of the right hand. It will even calculate several filters for both optimizations by considering the remaining orthogonal subspaces.

Let Σ_i be the covariance matrix of the trial-concatenated matrix of dimension [concatenated time-points × channels] belonging to the respective class $i \in \{1,2\}$. The CSP analysis consists of calculating a matrix W and diagonal matrix D with elements in $[0,1]$ such that

$$W^\top\Sigma_1 W = D \qquad \text{and} \qquad W^\top\Sigma_2 W = I - D.$$

This can be solved as a generalized eigenvalue problem. The projection that is given by the i-th column of matrix W has a relative variance of d_i (i-th element of D) for trials of class 1 and relative variance $1-d_i$ for trials of class 2. If d_i is near 1, the filter given by the i-th column of W maximizes variance for class 1, and since $1-d_i$ is near 0, the same column

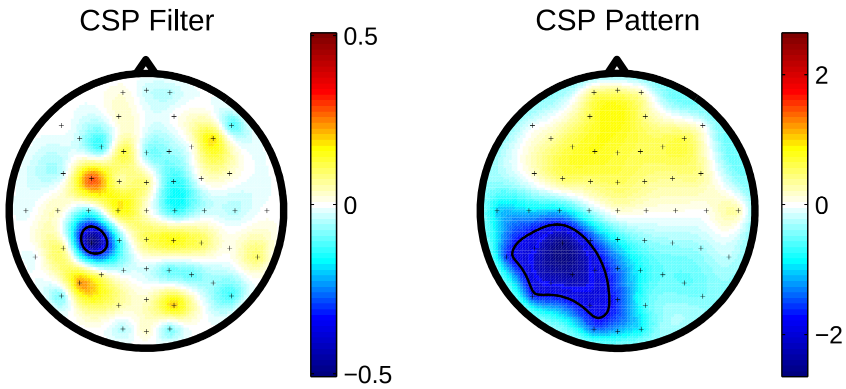


Figure 2.7.: The author's CSP filter and pattern for the detection of ERD during imagination of a right hand movement. The filter (the left part of this figure) is the first row of the de-mixing matrix W, the pattern (the right part of the figure) is the first column of the mixing matrix W^{-1}. In contrast to the filters displayed in figure 2.6, both CSP filter and pattern can be checked for neurophysiological validity. In this case, they exhibit a strong focus on the central area of the left hemisphere, which is the motor cortex associated to the right hand.

minimizes variance for class 2. Typically one would obtain projections corresponding to the three highest eigenvalues d_i, i.e., CSP filters for class 1, and projections corresponding to the three lowest eigenvalues, i.e., CSP filters for class 2. Figure 2.7 shows the filter and the dual pattern corresponding to the minimization of the bandpower in the μ-band from 11 Hz to 15 Hz, for the imagined movement of the right hand. EEG data were recorded during the imagination of left hand and right hand movement imagery in 70 trials per class. Both filter and pattern are focussed over the left motor area, i.e. contralateral to the performing hand. Figure 2.8 shows the spectra for the signal projected by this filter, along with the spectra of the filters shown in Fig. 2.6. While a desynchronization can be recognized for all filtered signals as a difference between the red and green graph, the CSP projection clearly optimizes this difference. For this comparison, see also the argument in [41].

A very concise tutorial on CSP is given in [22].

Spatio-temporal Filters Although the classification of bandpower estimates on the spatially filtered data by means of the CSP algorithm is very effective, the problem of the correct choice of the temporal (i.e., frequency) filter remains. It is not actually a problem which can be optimized independently from the spatial filter, since for different frequencies, different spatial filters are optimal and vice versa. Therefore, a simultaneous optimization of both filters is highly desirable.

In [80], a new method (termed "Common Spatio-Spectral Patterns" (CSSP)) was introduced to optimize both filters by simply performing the CSP calculation on the signal, concatenated with a time-delayed version of itself. The resulting filters can be split into frequency filter and spatial filter. Depending on the time delay, this method can significantly increase the classification accuracy as compared to the usual CSP approach.

Another method, "Common Sparse Spectral Spatial Patterns" (CSSSP) (cf. [38]), expands

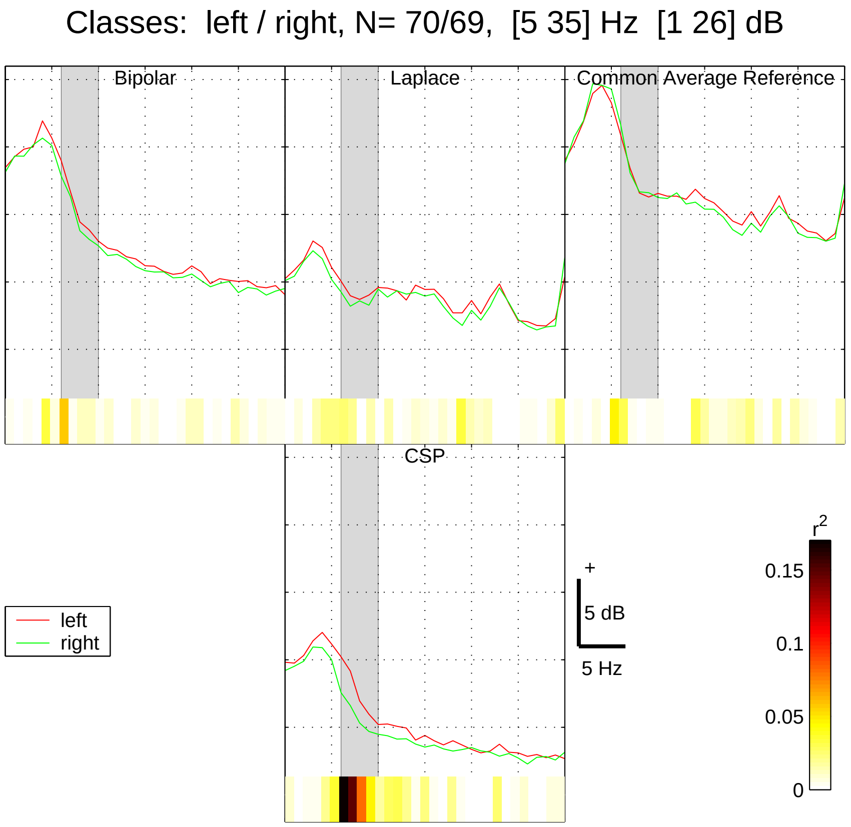


Figure 2.8.: For the filters shown in Figure 2.6 and 2.7, these are the spectra of the projected signals. The spectra are given for left (red) and right hand imagination (green) separately. The gray shaded area denotes the frequency band the CSP filters were calculated on. Below the spectra, the color intensity denotes the separability of the frequency bins, in terms of the bi-serial correlation coefficient (see Section 3). The best class separability is achieved by the CSP filter.

the range of possible frequency filters, by explicit simultaneous optimization of both the parameters of the digital FIR filter and the spatial filter. The performance is similar to the performance of CSSP.

There are many other approaches to the joint optimization of spatial and spectral filters, e.g. [139, 140], where the optimization is performed in the spectral domain.

2.2.2. Classification

According to [42], a classifier on a given feature space $\mathscr{X} \subset \mathbb{R}^n$ can be defined as a set

$$\mathscr{C} := \{g_i : \mathscr{X} \longrightarrow \mathbb{R} | i = 1, \dots, C\},$$

where C is the number of classes. The classifier assigns the feature $x \in \mathscr{X}$ to a class c if

$$g_c(x) > g_i(x) \text{ for all } i \neq c.$$

Intuitively speaking, this characterizes a decision function $f_{\mathscr{C}} : \mathscr{X} \longrightarrow \{1,\dots,C\}$, which assigns a class label to each point in the feature space, by defining

$$f_{\mathscr{C}}(x) := \begin{cases} \text{argmax}_{i\in\{1,\dots,C\}} g_i(x), & \text{if this maximum exists} \\ 0 & \text{otherwise.} \end{cases}$$

A classifier partitions the feature space into decision regions $\mathscr{R}_1,\dots,\mathscr{R}_C$, which consist of all the points of the feature space that the classifier assigns to the respective class label. If the functions g_i constituting the classifier are sufficiently simple (e.g. continuous), it can be interesting to analyze the decision boundaries, i.e. the set $\mathscr{X} \setminus (\bigcup_{i=1}^{C} \mathscr{R}_i)$. It corresponds to the points where the largest g_i have the same function value.

In this work, I will focus on a simple case where the classifier only compares $C = 2$ classes (such a classifier is also called “dichotomizer” or “binary classifier”). In the case of a binary classifier and $\mathscr{C} := \{g_1, g_{-1}\}$, the decision function can be reduced to the form

$$f_{\mathscr{C}}(x) := \text{sgn}(g_1(x) - g_{-1}(x)).$$

It is common practice to inspect the “graded” classifier output (which is the function value of $(g_1(x) - g_{-1}(x))$ before applying the sgn-function) as well as the classifier decision.

I will also only consider classifiers whose classification function and the corresponding decision boundaries will be linear (these classifiers are called “linear”). Under some assumptions on the distributions of the underlying classes in the feature space, namely known normal distributions with equal covariances, a linear classifier is the optimal choice in the sense that it minimizes the probability for misclassification (“Bayes-optimal”).

The decision for linear classifiers is not simple, but one of the most important arguments is the small number of parameters which have to be estimated on the training data. While the extension to richer function classes can enhance the training accuracy, there is always a considerable risk of overfitting: with a sufficiently large function class to choose the classifier from, any finite amount of training data can be classified perfectly, but the generalization ability of the classifier is not always guaranteed. Therefore, I will restrict myself to the case where most of the power of the classification process is actually performed in the feature extraction: if the data in the feature space are linearly separable, they can be easily classified. I will only present some methods that I will use later throughout this work. Also note that other methods, such as regularization, will not be applied here. For a more detailed discussion of linear and non-linear methods, see [93].

Linear Discriminant Analysis (LDA)

If $X \in \mathbb{R}^n$ and $Y \in \{1,-1\}$ are random variables ($n \in \mathbb{N}$) with $X|(Y = i) \sim \text{N}(\mu_i, \Sigma)$ for some $\mu_i \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$ ($i \in \{1,-1\}$), and if the class priors are equal
(i.e., $P(Y = 1) = P(Y = -1)$), then the decision function

$$f(x) = (\mu_1 - \mu_{-1})^\top \Sigma^{-1} x - 0.5(\mu_1 - \mu_{-1})^\top \Sigma^{-1} (\mu_1 + \mu_{-1})$$

is the Bayes-optimal classifier for this problem. Since in the general case, μ_1, μ_{-1} and Σ are not known, they can be estimated by the class means

$$\hat{\mu}_i := \frac{1}{m_i} \sum_{j\in\{k|y_k=i\}} x_j,$$

where m_i is the number of samples from class i ($i \in \{1, -1\}$), and the averaged classwise scatter matrix

$$\begin{aligned}\hat{\Sigma} &:= 0.5(\hat{\Sigma}_1 + \hat{\Sigma}_{-1}) \\ &= 0.5(\frac{1}{m_1} \sum_{j \in \{k | y_k = 1\}} (x_j - \hat{\mu}_1)(x_j - \hat{\mu}_1)^\top \\ &\quad + \frac{1}{m_{-1}} \sum_{j \in \{k | y_k = -1\}} (x_j - \hat{\mu}_{-1})(x_j - \hat{\mu}_{-1})^\top).\end{aligned}$$

This classifier can easily be extended to the case where the class priors are not equal. A further extension to the case of unequal class covariance matrices makes the decision boundary $\{x \in \mathbb{R}^n | f(x) = 0\}$ non-linear. The method is then called Quadratic Discriminant Analysis (QDA).

Least Squares Regression (LSR)

Using linear regression on the labels, another classifier can be introduced. In regression problems, a relation between data $x_i \in \mathbb{R}_n$ and function values $y_i \in \mathbb{R}^m$ (for $i \in \{1, \ldots, N\}$, for some $n, m, N \in \mathbb{N}$) is described by choosing one function out of a function class which minimizes the (squared) error between its function values and the target values y_i.

In order to find the linear classifier whose classification values are as close as possible to the labels, we can set $m = 1$ and choose $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ such that they minimize

$$\begin{aligned}\sum_{i=1}^N (w^\top x_i + b - y_i)^2 &= \sum_{i=1}^N (\begin{pmatrix} w \\ b \end{pmatrix}^\top \begin{pmatrix} x_i \\ 1 \end{pmatrix} - y_i)^2 \\ &= || \begin{pmatrix} w \\ b \end{pmatrix}^\top \begin{pmatrix} x \\ 1 \ldots 1 \end{pmatrix} - y^\top ||_2^2.\end{aligned}$$

The last term is minimized by setting

$$\begin{pmatrix} w \\ b \end{pmatrix} := \begin{pmatrix} x \\ 1 \ldots 1 \end{pmatrix}^{+\top} y,$$

where "$^+$" denotes a pseudo-inverse operator. Note that this corresponds to LDA with a modified scaling.

A nice property of the LSR classifier is that applying it to the class means μ_1 and μ_{-1} yields 1 and -1, respectively. This behavior is desirable if, like in the BCI context, the classification values should be finally translated into control signals. This control can be improved if the expected function values of the classifier for "typical" input values like the class means is known beforehand.

Support Vector Machines (SVM)

The Support Vector Machine, introduced in [142], is based on the idea of separating the training data $x_i \in \mathbb{R}^n$ with labels $y_i \in \{-1, 1\}$ ($i \in \{1, \ldots, N\}$ for some $N \in \mathbb{N}$) by means of a linear hyperplane, such that the minimal distance of each point from the hyperplane, the so-called “margin”, is maximized. In other words, the weight vector $w \in \mathbb{R}^n$ and the offset $b \in \mathbb{R}$ can be determined by the optimization problem

$$\begin{array}{rc} \operatorname{argmin}_{b\in\mathbb{R},w\in\mathbb{R}^n} & ||w||_2^2 \\ \text{s.t.} & y_i(w^\top x_i + b) \geq 1, (i = 1, \dots, N). \end{array}$$

The x_i for which the constraint is fulfilled with equality are called “Support Vectors” of the hyperplane, since they determine the location and angle of the hyperplane and, geometrically interpreted, “support” the outer borderline of the margin. This reduces the complexity of the problem, even in high dimensions, to a few support vectors.

Since the data need not necessarily to be separable, such a hyperplane does not always exist. For this case, the optimization criterion can be relaxed by introducing slack variables $\xi_i \in \mathbb{R}$ ($i = 1, \dots, N$) and a regularization parameter $C > 0$ in the following way:

$$\begin{array}{rc} \operatorname{argmin}_{b\in\mathbb{R},w\in\mathbb{R}^n,\xi\in\mathbb{R}^N} & ||w||_2^2 + C\sum_{i=1}^N \xi_i^2 \\ \text{s.t.} & y_i(w^\top x_i + b) \geq 1 - \xi_i^2, (i = 1, \dots, N). \end{array}$$

The regularization parameter C controls the tradeoff between two objectives: a smaller C will result in a larger margin around the hyperplane, but might result in a higher error on the training data. Larger Cs decrease the training error, but possibly reduce the generalization error by enlarging the margin.

Support Vector Machines can easily be extended to non-linear cases. This and a more detailed overview of Support Vector Machines can be found in [24, 98, 123].

2.3. The Berlin Brain-Computer Interface (BBCI)

For the presentation of the BBCI, I will first give a very general overview of the past and ongoing projects. Then I will report the most commonly used methods for training a classifier on data from mental imagery, and how the resulting classifiers can be applied to drive feedback applications such as the 1-dimensional control of a computer cursor.

2.3.1. Overview and History

In the year 2000, the Berlin Brain-Computer Interface project was initiated as a cooperation between the Fraunhofer Institute FIRST and the Department of Neurology of the Charité Berlin. Recently, the Technische Universität Berlin also became involved in this ongoing research process.

Before the BBCI entered the field, the majority of BCI research was performed by long training periods for the users of BCIs (e.g. [150, 6]). This training can take months or even years, until a reasonable communication performance can be established. Guided by the motto “Let the machines learn!”, the focus of the BBCI is to shift the main burden of learning away from the user onto the analyzing and classifying computer. This can be done by combining knowledge from different ends of this interdisciplinary field, namely neurophysiology and machine learning techniques.

The BBCI has covered a wide range of BCI paradigms; it has for example demonstrated how upcoming keypress movements with left or right index finger could be classified with a high accuracy, even 120 ms before the actual movement was performed ([8, 67, 66]). Further works have shown how error potentials can be usefully integrated into BCI applications, [15, 21].

Much work in the BBCI has been performed on the ERD/ERS paradigm in motor imagery data (e.g. [12, 9]). We could show that it is possible to provide BCI feedback to completely untrained subjects, after a short calibration period of approx. 20–30 minutes. The focus of my work in this project is to even reduce this short amount of calibration to the absolute minimum, by thorough analysis of the feedback experiments and the behavior of the classifiers throughout the experiment. I will in the following mainly exemplify my methods on ERD features and the used classifiers.

An important ingredient to a Brain-Computer interface is a rich feedback application. Therefore, we developed various games and text input devices and demonstrated how they can successfully be operated. These applications include a BCI-controlled Cursor (e.g. [12, 9], see also chapter 5.1), text input devices like a binary speller ([71]) and the "Hex-O-Spell" interface ([13]). Among the implemented games are "Brain-pong", a variant to the 1970s arcade game "PONG", for one or two players, and more recently the control of a real-world Pinball machine. Since this list is by no means a full report of the possible applications, I refer to [65, 96] for a more complete overview.

The basis for this success is the application of machine learning on high-dimensional EEG-data. The BBCI has shaped this concept in the BCI community by organizing BCI classification competitions ([20, 19, 119]), where the participating researchers all over the world could benchmark their own algorithms on data from BCI experiments. For better comparison and to avoid overfitting, the results were only released at the end of the competition, when the labels of the test set were published.

In the following sections, I will give a short overview of the standard procedures applied for BCI motor imagery feedback sessions in the BBCI. Most of the experiments reported here followed this procedure, until I introduced a new method (see Section 5.1).

2.3.2. Measurement

All the experiments conducted for this work have been performed with non-invasive scalp EEG. For each subject, brain activity was recorded by means of 64–128 Ag/AgCl electrodes, attached to an EEG cap. The data were mostly recorded simultaneously with surface EMG (electromyogram) of the right foot and both forearms, as well as EOG (electrooculogram). This was exclusively to make sure that the subjects performed no real limb or eye movements correlated with the mental tasks that could directly (artifacts) or indirectly (re-afferent feedback from muscles and joint receptors) be reflected in the EEG and thus be detected by the classifier, which operated on the EEG signals only. Amplifiers and recording software from the company "Brain Products GmbH" were used, and the data were recorded at a rate of 1000 Hz.

2.3.3. Calibration

The subjects were sitting in front of a computer screen, with the hands in a relaxed position on armrests. Every 5.5 (± 0.25) seconds one of three different visual stimuli (see Fig. 2.9 for an example) indicated for 3.5 seconds which mental task the subject should accomplish

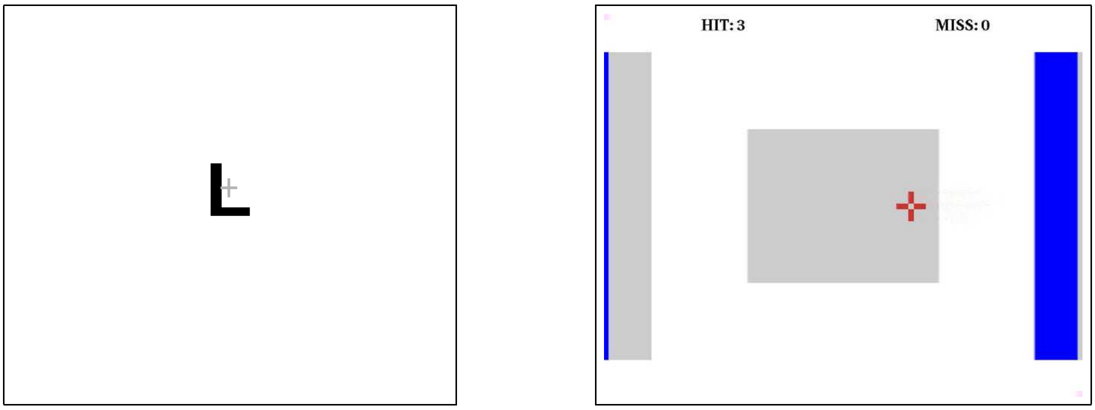


Figure 2.9.: The left part of the figure shows the computer screen during the calibration measurement. For a duration of 3.5 seconds, a letter indicates the mental task the subjects have to fulfill. In the right part of the figure, a “Cursor control”-feedback is depicted. See text for details.

during that period. The investigated mental tasks were imagined movements of the left hand (*L*), the right hand (*R*), and the right foot (*F*). Between 70 and 200 repetitions for each class were recorded. In this work I investigate only binary classifications, but the same classification setup can be used in the multi-class case, [36, 37].

2.3.4. Feature Extraction and Classification

After the calibration measurement, a classifier was trained on the two best discriminable classes. There are several parameters in this feature extraction and classification procedure that can be specifically chosen for each subject to obtain optimal results. In the online experiments this is done semiautomatically by combining machine learning, expert knowledge and visual inspection of some characteristic curves such as spectra and ERD curves, see [10], so the following parameters can be slightly adjusted by the experimenters.

After choosing all channels except the EOG and EMG and a few outmost channels of the cap, a causal band-pass filter from 7–30 Hz is applied, which encompasses both the μ- and the β-rhythm. The data we extract are from the windows 750–3500 ms after the presented visual stimulus, since in this period discriminative brain patterns are present in most subjects. Afterwards we apply the CSP algorithm (see Section 2.2.1) to the data. This decreases the number of channels by suitable linear spatial filters which are learned on the training trials. We typically use 3 filters per class, which leads to 6 remaining channels, chosen by the magnitude of the corresponding eigenvalues and by visual inspection; a more refined method of automatic selection of the best channels is presented in [22]. We then calculate the logarithm of the variances of for these channels. The resulting feature vectors are a measure of the amplitude in the specified frequency band.

After the presented preprocessing usually between 70 and 200 six-dimensional feature vectors for each class remain. Since the data have in most cases a Gaussian distribution, we apply a linear classifier such as LDA or LSR.

2.3.5. Feedback

In the BCI context, a “feedback” is the output of the system to the input it receives when measuring the neuronal activity. Feedback applications can have different output modalities (such as visual, tactile or auditory) and can differ in the timing of the return, the appearance of the stimuli and many other properties of the presentation. As typical examples for feedback applications, I will explain the setup for some variants of the “Cursor control” feedback (see Fig. 2.9), where the subjects can control the horizontal position of a cursor on the screen.

Cursor Control Feedback

During the feedback period, the EEG data were acquired from the recording computer and classified (almost) in realtime. Due to recording and digitizing restrictions, the data are acquired every 40 ms, and then the last, e.g., 1000 ms of EEG data are taken into account for the classification.

The data are spatially filtered with the pre-computed CSP filters. Then the bandpower in these signals is estimated by applying the frequency bandpass-filter and calculating the logarithm of the variance. The resulting features are then fed into the classifier.

In the “Cursor control”-feedback, two rectangular targets are placed at the left and right side of the screen. At the beginning of each trial, one of the targets is highlighted and the subject attempts to navigate the cursor into the target, using the two imagined movement types. The graded output from the classifier is then used to move the cursor either in a position-controlled, or in a rate-controlled manner. This means that the scaled classifier output is either used to move the cursor by a small amount to the chosen direction or is mapped directly to a horizontal position on the screen. Each “trial” lasts until the subject hits one of the two targets, and as a result the trials are of variable length. A block of (typically 25–100) feedback trials, not interrupted by a break, is called a “feedback run”.

Mental Typewriter Feedback

There are various ways in which a one-dimensional continuous output of a BCI can be used to enter text (e.g. [13, 91, 6, 152, 107]). The basis for the mental typewriter in this example is a continuous movement of the cursor in the horizontal direction. A “rate controlled” scenario was used, i.e., at the beginning of each trial, the cursor is placed in a deactivated mode in the middle of the screen. Every 40 ms, the current classifier output is added to the position of the cursor, thus moving it left or right. The feedback enables the subjects to type letter by letter on the basis of binary choices. The alphabet is divided into two contiguous sets of letters with approximately equal probability of occurence in the german language. The first and last letter of each division appear in a rectangle on the left and right end of the computer screen, see Fig. 2.10. By moving the cursor into one of the targets, the subjects can choose the set of letters containing the one they wish to type. The chosen set is then divided into smaller sets, until a single letter is selected. For correction purposes, one further symbol (<), for deleting one letter, is added to the alphabet. In case of failing to hit the correct letter, the subject can then try to select this delete-symbol to erase the erroneous letter. Note that after an error of only one binary choice, it is impossible for the subject to return to the node of the

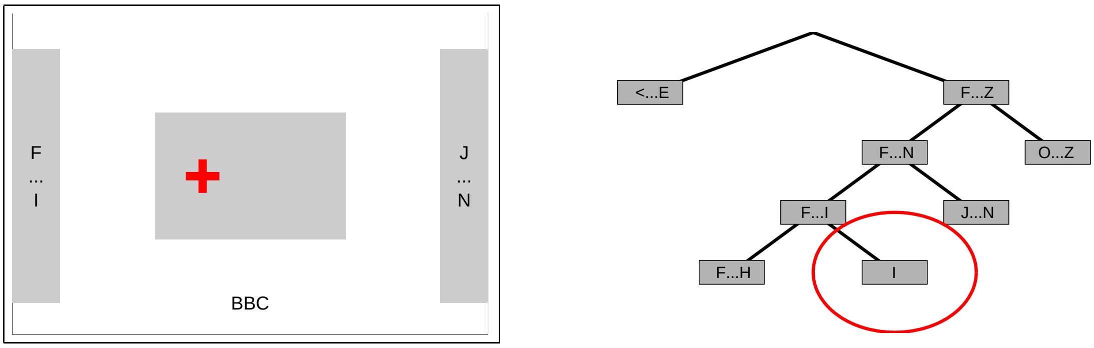


Figure 2.10.: The figure on the left is a screenshot of the feedback that was given to the subjects. The position of the cross is controlled by the classification output of the current EEG signal. By moving the cross into the right or left box, the respective set of letters is selected. For completing the acronym “BBCI”, the subject would try to select the left box, since the letter “I” is associated to it. A unique series of decisions (right–left–left–right) leads to the selection of this letter; this corresponds to the binary decision tree shown in the right figure.

decision tree containing the correct letter. Thus, a wrong letter will be selected regardless of the next decisions. In our studies, subjects often used this period to relax or stretch. This period of the experiment, however, should be excluded from any offline analysis schemes, since it does not contain useful information about the intended task.

Fixed Duration Cursor Control

For specific analysis of the feedback data, it can be problematic that the trials of the conventional Cursor Control Feedback application can have significantly different length. A modification of this setup can facilitate the analysis: instead of ending the trial when the cursor hits any of the two targets, the “fixed duration cursor control feedback” lets the subject control the cursor for a predetermined amount of time (typically 3.5 seconds). Just like in the regular case, the graded classifier output is used to control the cursor in horizontal direction in a rate-controlled fashion. After 3.5 seconds, the cursor is fixed again and the outcome of the trial is determined by the horizontal position of the cursor. If the cursor is on the correct side of the screen, the trial is counted as “hit”, and as “missed” otherwise. The target box is then colored according to the trial outcome in green (for a successful trial) or red (in the other case). After a short intertrial break of 1 second the next target is presented.

Feedback of Results

Another variant of the “Cursor Control” feedback concerns the visibility of the cursor. In our studies, we frequently encountered subjects who were distracted by the constant feedback given to them in form of the horizontal position of the cursor. This led to the development of a paradigm where the only difference to the standard scenario is that the cursor is no longer visible. Subjects only receive feedback at the end of each trial, by the color of the previously ordered target: “green” for success and “red” for failure.

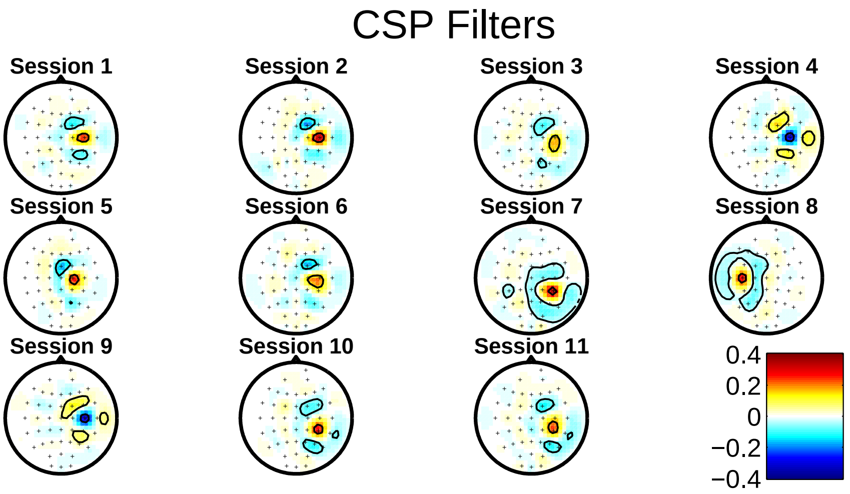


Figure 2.11.: The author's optimal CSP filters across sessions. For each session, the CSP filters were calculated on the calibration measurement, for the discrimination of left hand and right foot imagery. Only the filter corresponding to the lowest eigenvalue (i.e., for the minimal bandpower during left hand imagery) is displayed. The focus is mostly on electrode C4 (sessions 1, 2, 4, 6 and 9) or CP4 (sessions 3, 10, 11) or on surrounding electrodes (session 5). In session 8, it is even on the ipsilateral (left) hemisphere.

This feedback type can also be used to force different levels of visual attention, since it can be quite fatiguing to focus on a screen with almost no change of the visual scene (see Section6.1).

2.3.6. Problems in this Approach

Although this approach works well for a large number of untrained subjects (see [11]), there were still some issues to be resolved:

- After training the classifier, the control could often only be established after adding a fixed real value to the classification output ("bias term"). Since the classifier worked well on the data from the calibration measurement, it was unclear why this was necessary. Also, the need for this manual adjustment is an unpleasant detail in the otherwise highly individualized and fine-tuned system.

- During the presentation of the feedback, there were sometimes periods when the subjects completely lost the ability to control the BCI system. It was an open issue both how to re-adapt the classifier parameters and how to do it online and in realtime.

- For longterm BCI users such as severely disabled people, a daily calibration period would be tiresome and annoying. The straight-forward approach to re-use the first classifier ever set up for a subject will clearly fail, as figure 2.11 shows. The CSP filters for a single subject display a large variability, such that it is not evident how a particularly robust filter can be found from the training data of previous experiments.

I will address all of these topics in the upcoming chapters, and I will show how some of them can be solved by the aid of advanced machine learning techniques.

3. Introduction to Nonstationarity

In the BCI context, we usually have labeled time series with a large variability. The sources for the variability include the following categories:

- Measurement artifacts and measurement noise
- Physiological artifacts
- Influences of other, not task-related neurophysiological processes
- Changes in the feedback setup (stimulus modality, stimulus appearance,...)
- Changes in psychological parameters

Since so many different factors are contributing to the variability of the brain signals, it is hard to quantify and describe the nature of their influence. Some of them only increase the noise level, but are relatively stable over time, such as the 50 Hz-noise induced by the alternating current of european power lines, which can be assumed not to undergo a large variation over the course of a BCI session.

If the performance of BCI classifiers changes over time, this is often referred to as "nonstationarity". This term is not limited to the application to BCI, and in the literature, many definitions and concepts, often tailored to the specific field of application, have been proposed (e.g. [62, 111, 102, 109, 121, 136, 133]). In the BCI field, it is of particular interest to find remedies against nonstationary behaviour of classifiers, to maintain the ability of the user to control the system. In this chapter, I will first go one step back and introduce the concept of nonstationarity (see the following definitions) and then discuss a variety of methods that can be applied for the characterization and the quantification of nonstationary time series (Sections 3.1, 3.2 and 3.3). These methods can be applied to get a deeper understanding of the underlying processes that are inducing the nonstationarity.

Definition Let $\mathbb{P} = (\Omega, \mathscr{F}, \mathrm{P})$ be a probability space, $n \in \mathbb{N}$ and $I \subset \mathbb{R}$.

A set of the form

$$S = \{F_t | t \in I\},$$

where each F_t is a random variable over $\mathbb{P}$ with values in $\mathbb{R}^n$ is called *Stochastic Process* with state space $\mathbb{R}^n$.

Mathematical properties and methods for stochastical processes and the concept of random variables can be found in [64] and [33]. In the following, I will also refer to stochastic processes as multivariate time series. This view puts more emphasis on the time course, but it should nevertheless be clear that the chosen probability space is of importance for the following definitions.

Definition Let $(X_t)_{t \in I}$ be a multivariate time series, where $I \subset \mathbb{R}$ is an index set. Then (X_t) is called *(strict-sense-) stationary time series*, if the probability distribution does not change over time, i.e.,

$$P_{X_{t_i}} = P_{X_{t_j}}$$

for all $t_i, t_j \in I$.

A time series is called *nonstationary*, if it is not stationary.

In classification problems on time series, we are usually given a time series together with a series of labels. For this setting, the above definition is not yet appropriate, since the labels will have to be modeled in the probability distribution. If the labeled data are regarded as a time series $(X_t, Y_t)_{t \in I}$ on some index set $I \subset \mathbb{R}$, where the labels Y_t are also random variables, the definition of stationarity would entail that the joint distribution of the labels and the data is unaltered over the whole index set. For the purpose of the investigations presented here, the main focus of attention will only be the evaluation of the conditional probability distribution $P_{X_t|Y_t}$, not on the entire joint distribution P_{X_t,Y_t}.

Therefore, a modification of the above definition for the special case of time series with labels will make this explicit:

Definition Let $(X_t, Y_t)_{t \in I}$ be a labeled multivariate time series, where $I \subset \mathbb{R}$ is an index set and $Y_t \in \mathscr{C} \subset \mathbb{R}$ for all t. Then (X_t, Y_t) is called *stationary labeled time series*, if the probability distribution for each class does not change over time, i.e.,

$$P_{X_{t_i}|Y_{t_i}} = P_{X_{t_j}|Y_{t_j}}$$

for all $t_i, t_j \in I$.

This implies: $(X_t, Y_t)_{t \in I}$ is stationary iff for all classes $c \in \mathscr{C}$ the time series $(X_t)_{t \in \{s \in I | y_s = c\}}$ is stationary.

A labeled time series is called *nonstationary*, if it is not stationary.

Now the question arises how it can be shown that a labeled time series is nonstationary. According to the definition, it is only required to find two points in time where the distributions are different. On the other hand, there is always the problem of a sufficiently accurate estimation of the probability density at a given point in time. If a stationary time series is generated from a normal distribution with a large covariance matrix (e.g., caused by measurement noise), it is not trivial to decide from the data whether the time series comes from the same distribution. In order to identify a nonstationary process, it is important that tests for the change of underlying parameters can be done at a reasonable significance level.

If it is safe to assume a parametric model for the distribution of the time series, it is sufficient to demonstrate that the parameters of the model are changing over time. For the case of a multivariate normal distribution, this corresponds to investigating the mean and covariance of the data and how they change over time[1].

The usual setup of the BBCI uses bandpower features for the classification. This requires the measurement of EEG over a time window of 100–1000 ms. In order to let the different

[1] The concept of stationarity, which only requires the first and second order moments to not vary over time, is commonly refered to as *wide-sense stationarity*. Note that "(strict-sense) stationarity" implies "wide-sense stationarity", and therefore "wide-sense nonstationarity" implies "(strict-sense) nonstationarity". The investigation that I am conducting are, in this notion, testing for wide-sense nonstationarity.

samples be independent, a property that is required for the correct estimation of mean and covariance of a normal distribution, I have to ensure that the windows for the bandpower estimation do not overlap. In the following, I will use the data of different trials – either in the calibration or in the feedback setup as described in Section 2.3.3 and 2.3.5 – as samples, i.e., $t \in I$ is the number of a particular trial and X_t is some feature derived from the EEG recording of this trial.

3.1. Probability Distribution Comparison

According to the definition of nonstationarity in labeled time series, the check for nonstationarity involves the comparison of estimates of the distribution of two classes at two given points in time. There are various methods to compare the distribution of two random variables, I will introduce the most prominent ones and show how they are related.

3.1.1. Kullback-Leibler Divergence

Definition The Kullback-Leibler Divergence (sometimes refered to as “Kullback-Leibler Distance”, although this is mathematically not quite accurate, as the considerations below will demonstrate) of the probability distributions P and Q with respective probability densities p and q is defined by

$$\mathrm{KL}(P,Q) := \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

For two n-dimensional random variables X_1, X_2 with $X_1 \sim \mathrm{N}(\mu_1, \Sigma_1)$ and $X_2 \sim \mathrm{N}(\mu_2, \Sigma_2)$, this amounts to

$$\begin{aligned}
\mathrm{KL}(P_{X_1}, P_{X_2}) &= -\frac{1}{2}\big[\log(|\Sigma_1\Sigma_2^{-1}|) + E(X_1-\mu_1)^t\Sigma_1^{-1}(X_1-\mu_1) \\
&\qquad\qquad - E(X_1-\mu_2)^t\Sigma_2^{-1}(X_1-\mu_2)\big] \\
&= -\frac{1}{2}\big[\log(|\Sigma_1\Sigma_2^{-1}|) + \mathrm{trace}(E(X_1-\mu_1)(X_1-\mu_1)^t\Sigma_1^{-1}) \\
&\quad -\mathrm{trace}(E(X_1-\mu_1)(X_1-\mu_1)^t\Sigma_2^{-1}) - (\mu_2-\mu_1)^t\Sigma_2^{-1}(\mu_2-\mu_1)\big] \\
&= -\frac{1}{2}\big[\log(|\Sigma_1\Sigma_2^{-1}|) + \mathrm{trace}(I - \Sigma_1\Sigma_2^{-1}) - (\mu_2-\mu_1)^t\Sigma_2^{-1}(\mu_2-\mu_1)\big],
\end{aligned}$$

where I denotes the n-dimensional identity matrix.

Note that the Kullback-Leibler Divergence is non-negative, i.e., $\mathrm{KL}(P,Q) \geq 0$ for all P,Q. The equality holds if and only if $P = Q$. However, the Kullback-Leibler Divergence does not define a metric in the mathematical sense, because it is not symmetric. It is therefore sometimes used in a symmetric version by defining

$$\mathrm{KL}_{\mathrm{sym}}(P,Q) := \mathrm{KL}(P,Q) + \mathrm{KL}(Q,P).$$

A simple example shows that the Kullback-Leibler Divergence does also not satisfy the triangle inequality. Suppose we have three Bernoulli Distributions $P_1 = B(d)$, $P_2 = B(0.5)$

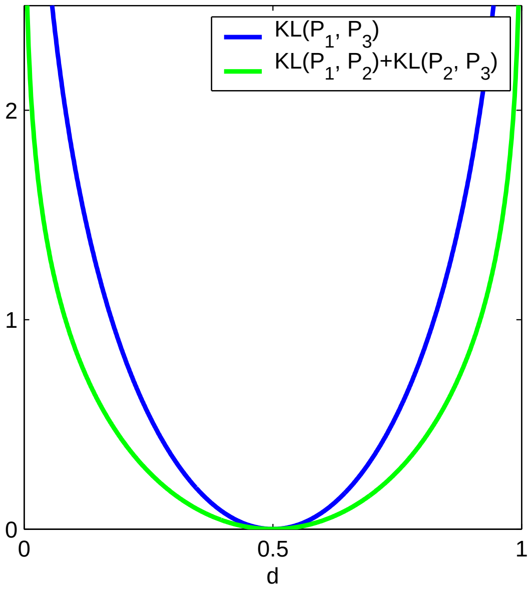


Figure 3.1.: This figure shows the effect of varying the parameter of three Bernoulli distributions on the associated Kullback-Leibler divergence values. The triangle inequality is only satisfied for a single choice of parameters.

and $P_3 = B(1-d)$ for some $d \in]0,1[$. The corresponding Kullback-Leibler divergence terms $\mathrm{KL}(P_1, P_3)$ and $\mathrm{KL}(P_1, P_2) + \mathrm{KL}(P_2, P_3)$ are shown in Fig. 3.1.1. The triangle inequality only holds (trivially) for $d = 0.5$; this can be shown with straightforward calculus. Note that the symmetric version of the Kullback-Leibler Divergence still does not fix the triangle inequality. The KL divergence is a very general tool, such that it can be used to express some information-theoretic concepts, as shown below.

Shannon Entropy

If X is a discrete random variable with probability mass function $p(x_i) = p_i(i = 1, \ldots, n)$, the Shannon Entropy of X is defined as

$$H(X) := -\sum_{i=1}^{n} p_i \log(p_i).$$

In information theory, the Shannon entropy is a measure for the uncertainty associated to the transmission of an information.

The Shannon entropy can be expressed with the KL divergence in the following way:

$$H(X) = \log(n) - \mathrm{KL}(P_X, P_U),$$

where U is a uniformly distributed variable. In other words, the less information is contained in X (i.e., the closer X is to a uniform distribution), the larger the associated Shannon entropy.

In BCI research, the Shannon entropy is often used to evaluate the performance of a particular setup, see e.g. [34, 73]. The bitrate is the expected number of bits that can be ("almost surely") transferred over a particular channel in a specific amount of time.

Mutual Information

If X and Y are random variables with probability density functions $p(x)$ and $p(y)$ and joint density function $p(x,y)$, the mutual information of X and Y is defined as

$$I(X,Y) := \int_X \int_Y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy.$$

In information theory, the mutual information is a measure of the dependence between X and Y, with $I(X,Y) = 0$ iff X and Y are independent.

Expressed with the Kullback-Leibler divergence,

$$I(X,Y) = \mathrm{KL}(P_{X,Y}, P_X P_Y),$$

which means that it denotes the distance between the joint distribution and the product of the two distributions. From this expression, the above property is directly deducted.

3.1.2. Bi-serial Correlation Coefficient (r-value)

The (point-)bi-serial correlation coefficient r measures how much information one feature dimension (of the data $x \in \mathbb{R}^d$) provides about the labels. For each dimension i of x, it is computed in the following way:

$$r_i = \frac{(\mu_1 - \mu_2)}{\sigma} \sqrt{\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)}},$$

where μ_j is the class-specific empirical mean of dimension i of x, σ the sample standard deviation of dimension i of x , and n_j denotes the number of samples for class $j \in \{1,2\}$.

This value describes the separability of the data in one dimension by scaling the difference of the empirical means with the inverse of the sample standard deviation. It is often used in the squared version, where high r^2-values correspond to high discriminability of the respective feature dimension. The signed r^2-value ($\mathrm{sgn}(r) \cdot (r^2)$) additionally preserves the information which class has the higher mean.

3.1.3. Area Under the Curve (AUC)

The Area Under the Curve (AUC) is a feature of the Receiver Operating Characteristic (ROC) curve.

If the discrimination threshold of a binary classifier is varied, the ROC curve is a graphical plot of the sensitivity of the classifier ("True Positive Rate") against (1-specificity) ("False Positive Rate").

Let X be a real-valued random variable, modeling the graded output of a given classifier. Suppose the values of X can be interpreted as detector for the event $Y = 1$, where $Y \in \{1, -1\}$ is a random variable. Then the quantities mentioned above are defined in the following way for a given discrimination threshold x^*:

True Positive Rate: $P(X > x^* | Y = 1)$
False Positive Rate: $P(X > x^* | Y = -1)$.

The ROC curve is then drawn by varying the decision threshold in the interval $[-\infty, +\infty]$. If the classifier discriminates well between the two classes, the ROC-curve is far from the diagonal line (which corresponds to no separability at all), therefore the area below the curve (calculated as the integral from 0 to 1) gives an impression of the discrimination ability of the classifier. AUC values can range from 0 to 1, where 1 means perfect separability. Note that all values below 0.5 can mean that the detection problem should be formulated with a reversed sign.

The AUC value has the advantage that it works just as well for discrimination problems with different class priors. Also note that for a given classifier, the discrimination is assessed independently from any additive bias term.

3.1.4. Classification Error

While both the r^2-values and the AUC-values can be applied independently from a given classifier, the classification error just denotes the percentage of errors that some classifier committed on a test data set. This measure obviously does not give a general impression on the separability of the test data, since the classifier might just be sub-optimal. Yet, this value can be used to assess a change in the feature distributions from one point in time to the other: by adjusting the classifier parameters on a "training" data set and applying it to "test" data, the performance on the test data is high if training and test data were drawn from similar distributions.

In usual machine learning applications, a robust prediction for the classification error on unseen data is often computed by training the classifier on all data except for a single point and then applying it to this point. By repeating this procedure for each point in the dataset, the number of errors can be counted and be divided by the total number of points. This fraction is then called the "(leave-one-out-) cross validation error". In chapter 5, the classification error with varying training and test sets will be used to check the stability of the features in a given feature space.

The classification error, although highly relevant for the analysis of BCI performance, is a very complex measure, which does not necessarily give insights on the underlying differences between distributions. However, it is sensitive to the change of the distributions, if this shift is relevant to the discriminability of the data. This is exemplified in the next section.

3.2. Pairwise Probability Density Comparison

If the performance of a classification-based Brain-Computer Interface does not meet the performance as predicted from the training data, some relevant change must have occured. In other words, the distributions of the data in parts of the feedback session do not resemble the distribution on the calibration data. Regardless of the neurophysiological and psychological factors, this change can easily be documented in the feature space by just inspecting the relation of class means and covariances to the classification boundary. Fig. 3.2 gives some examples for schematic differences between training and test set. It is important to note that these examples can not only be applied to the particular setting with calibration and feedback data, but they can also serve as a comparison between different parts of the feedback period, or between different sessions. In Section 3.3, I will further comment on why this might be useful in the BCI context.

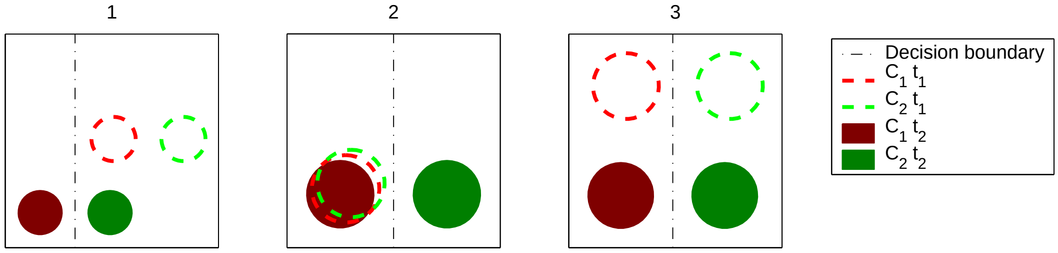


Figure 3.2.: Some changes that can occur in the binary classification of labeled time series. The solid shapes depict the classwise standard deviation of the training data (at time t_1) around the class means. For the test data (time t_2), the standard deviations are given in dashed lines. The optimal decision boundary corresponding to a linear classifier, is depicted with the dashed line.

For the sake of simplicity, the examples given here are restricted to two dimensions, but it is obvious that the same argument holds for any finite-dimensional feature space. The examples of Fig. 3.2 contain in detail:

1 While the covariance does not change for either of the classes, the class means are shifted considerably. Although the separability is unaltered at time t_2, the decision boundary is not useful for the discrimination of the classes, such that the classification error will be at chance level.

2 Only the class mean of class 2 is shifted at time t_2. The discriminability is drastically reduced, such that the classification error will be at chance level.

3 Again, both class means are shifted by the same amount. In contrast to the first example, the shift has occured along the decision boundary, such that the separability of the classes is not changed and the classification error at time t_2 corresponds to the error at time t_1.

Comparing the first two examples, it turns out that the classification error for one particular classifier can not reflect the overall separability of the data. In order to improve the classification rate by adjusting the classifier to the new data, it would be useful to know if the new class distributions can be discriminated at all, and which actions must be taken accordingly. Moreover, as the third example shows, the classification error does not necessarily reflect all the changes that are relevant to the data. This is not exactly a problem, since the information transfer rate of the interface is not affected by a change of this sort, but one should keep in mind that a stable classification performance does not necessarily mean that the class distributions are stable over the whole time. It is, on the contrary, highly interesting to observe what kind of change a particular classifier can be invariant against.

These examples show that the classification error alone can never provide sufficient insight into the changes within a given feature space. It is therefore necessary to include other means of quantifying the degree of alteration of the involved classes, such as binary comparisons between some of the four associated class distributions. Some of the possible comparisons are illustrated in Fig. 3.3.

I The first distance of interest is the between-class distance at time t_1. If the classification problem has equal class priors and if equal distributions are assumed (as for LDA

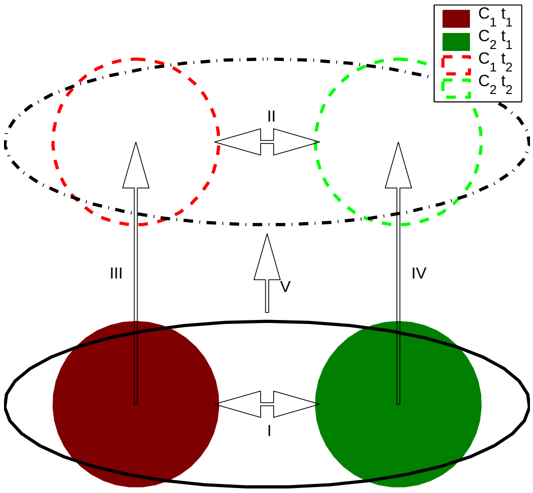


Figure 3.3.: This figure shows some of the binary comparisons of distributions involved in a labeled time series. The distributions of the two classes C_1 and C_2 are depicted with red and green circles, respectively. At time point t_1, the class distributions are shown with solid circles, and at time point t_2, with dashed circles. The overall distribution at time t_1 and t_2 is shown in a solid and a dashed black line, respectively.

classification), a symmetric measure should be used.

II The between-class distance at time t_2.

III The shift of class C_1 from t_1 to t_2. This distance is not required to be symmetric, since we often have a comparison of a “ground truth” distribution (e.g., from a calibration measurement) that a new distribution (e.g., from a feedback experiment) is supposed to be compared against.

IV The shift of class C_2 from t_1 to t_2.

V The shift of the overall distributions from time t_1 to t_2. Note that this unit can be estimated without the knowledge of the class labels, if equal class priors are assumed.

Since there are many other possible combinations that compare two different class distributions, I will exemplify the power and the shortcomings of these binary comparisons. Some examples are given in Fig. 3.4.

1 Both classes are shifted by the same amount. Although the common distribution (V) and the classwise distributions (III and IV) will change considerably, the class separability (I and II) is not affected.

2 In this example, the classes are only flipped. In contrast to the first example, the overall distribution of the samples (V) will not notice this change.

3 The class separability decreases drastically due to larger classwise covariances. If the overall distribution is estimated only by assessing the mean and the pooled covariance, this change will again go unnoticed by measure V.

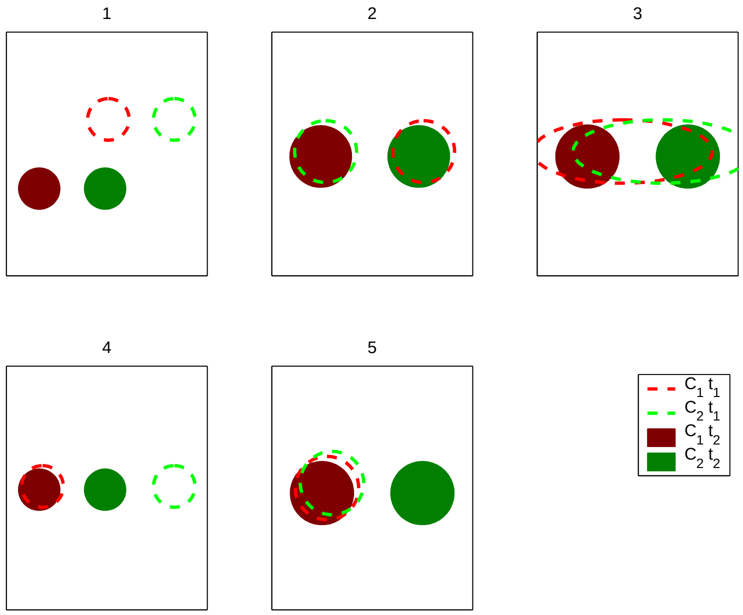


Figure 3.4.: This figure shows some examples of distribution changes that can occur in labeled time series. The presented distance measures I-V (see Fig. 3.3) respond quite differently to these changes.

4 The class separability is increased in this example (I vs. II), but only the second class distribution is changing. This change can be observed by checking the distance measures II, IV and V.

5 Although measures III, IV and V are affected in a similar way as in the previous example, the class separability (I vs. II) is drastically reduced. These examples show that the class separability always needs to be regarded in addition to the overall changes of the distributions.

In Chapters 5 and 6, the comparisons will be performed mainly for one-dimensional distributions: if the bandpower features are calculated for each scalp electrode separately, the distance between the distributions generates a scalp topography which can then be interpreted neurophysiologically. A first example is given in figure 3.5.

3.3. Possible Choices of Time Windows

In BCI research, the time series we are inspecting originate from online feedback experiments or from the preceding calibration measurements. If we regard the whole experiment as a time series, there is only one single instance of this time series for every subject. This means that it is impossible to assess the data distribution at a given point in time, since this would require multiple repetitions of the same time series with the same nonstationary behavior. For the goal of this work, namely to improve the classification performance of BCI systems, it is inevitable that the necessary actions all rely on the current session and can be performed without repetitions, i.e., in an online fashion.

Therefore, I will restrict myself in this work to the comparison of distinct time windows taken from the same time series. In a completely stationary time series, any two chosen windows in time contain samples which are drawn from the same distribution. If we further assume Gaussian distributions and conditional independence of the samples (which is the case if we only take one sample per trial of the recording), we can estimate the sufficient statistics (sample mean and sample covariance) of the distribution within each window separately. This estimate converges to the underlying mean and covariance, if the window sizes are sufficiently large. Using the estimates for both windows separately and applying the distance measures introduced in Section 3.1, it can then be decided how similar the distributions are. If they differ significantly, this proves that the time series is not stationary.

In the following chapters, I will apply a variety of different measures to particular choices of time windows to assess their degree of nonstationarity. These choices include

1. Comparison of calibration measurement vs. feedback measurement (see sections 5.1 and 5.2)
2. Comparison of entire sessions (see sections 6.2 and 6.3)
3. Within-session comparison (see sections 5.1, 5.2 and 6.1)

To illustrate the variability over these different time periods, the changes of the discriminability of the author's brain signals are depicted as scalp topographies in Fig. 3.5. For these figures, the discriminability of calibration (panel (a) and (c)) and feedback data (panel (a) and (b)) has been analyzed. Each panel exhibits a considerable variability of the region with maximal discriminability as well as of the magnitude of the r^2-values.

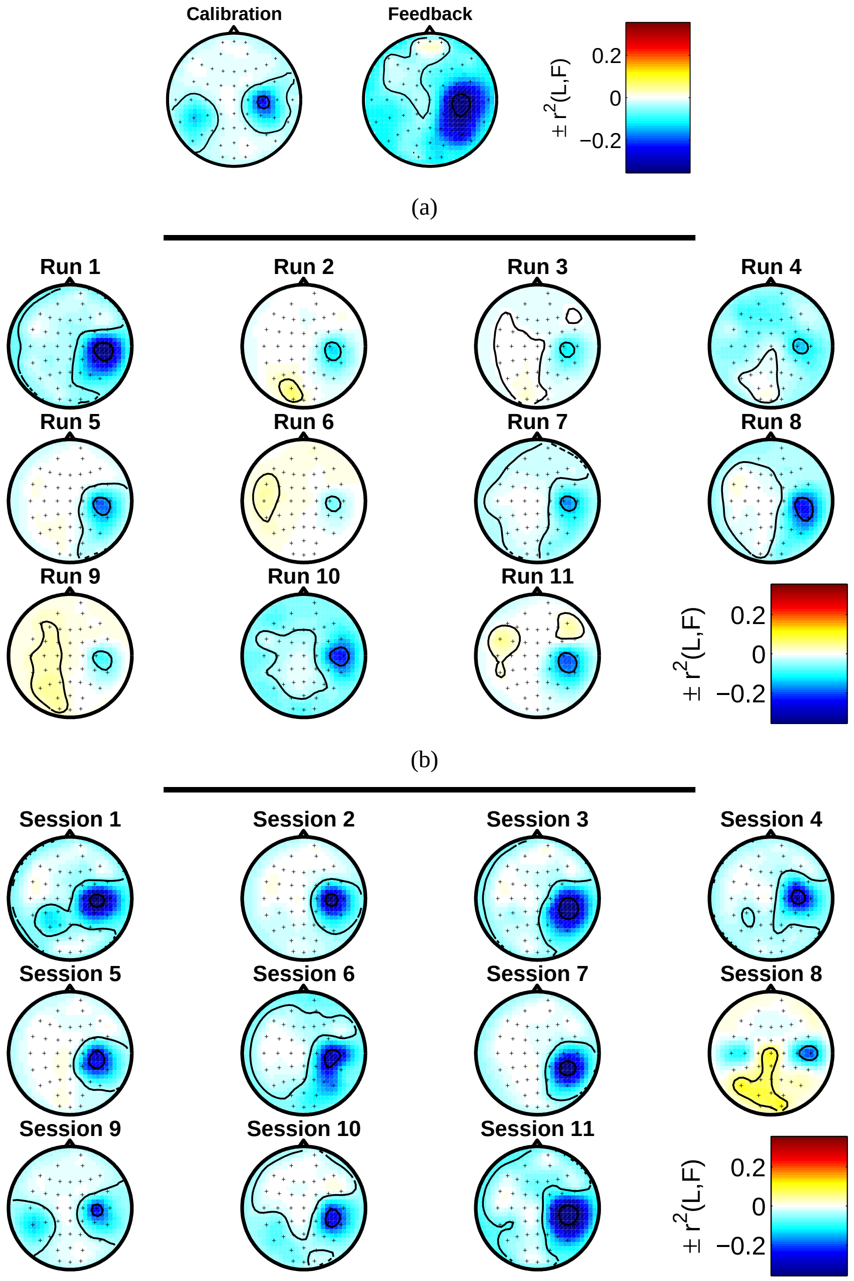
Note that important sources of variability in neurophysiological data, such as inter-subject differences, are not covered in this work. Although the methods presented here can clearly be applied to that scenario, it would be beyond the scope. For more details on inter-subject variability, see e.g. [124].

Among the most important comparisons are the time windows from a single session. For the BCI-context, stationarity over this period would also imply a stable performance of a static classifier over the whole session. Unfortunately, this stationary case is rarely observed. It is nevertheless important to identify the reasons for nonstationarity within single sessions and to use this knowledge to design remedies.

3.4. Adaptation

A frequently encountered problem in using EEG-based Brain-Computer Interfaces is that the performance decreases when going from offline training sessions to online operation of the BCI. One could suspect this to be caused by bad model selection strategies which could in principle choose overly complex classification models that overfit the EEG data. Yet I will show in the following chapters that the nonstationarities in the EEG statistics can actually account for this failure. If the subject's brain processes during feedback cause the distributions to wander astray on a sometimes very local timescale, counter measures have to be applied which alleviate the effect of the nonstationarity.

Various approaches were suggested to cope with nonstationary behavior of EEG signals. In the BCI context, the large variety of methods that are used for classification also enable



(a)

(b)

(c)

Figure 3.5.: This figure shows scalp topographies of the discriminability of the author's brain activity during the imagination of left hand and right foot movement. For each scalp map, EEG data in each electrode has been bandpass filtered to the range of 10–25 Hz, and the log bandpower was calculated by calculating the logarithm of the variance in each trial, 1000–3000 ms after the presentation of the stimulus. Finally, the signed r^2-value has been calculated for each electrode in order to find regions with the largest between-class differences.

the experimenter to adapt different parameters of the algorithm in the run of the session; moreover, the success of the adaptation algorithm used might depend on the chosen BCI scenario.

In [151], a visual BCI feedback was described in which the user was able to control a computer cursor in two dimensions, trying to hit one of eight possible targets. The classification algorithm used two distinct bandpower features acquired from a small subset of 64 scalp electrodes. Several scaling factors were used to translate these features into positions on the screen, four of which were successively adapted to the individual user during the session.

Similarly, [89] demonstrated that in a classification of four classes, the estimation of mean and covariance matrix for each of the classes can be iteratively updated in a simulated online scenario; based on these parameters, the predicted online performance for these subjects improved considerably. In this case, several channels from centroparietal scalp regions were used for the feature extraction.

In another offline study, this finding was backed by [145]; here, the parameters of a quadratic classifier (QDA) were adapted after each trial of a cursor-movement task. After a careful update parameter selection, the resulting classification was superior to the static classifier that was used from the start.

In each of these studies, the used method of adaptivity differs slightly and it is hard to transfer these results to other classification approaches, since the underlying changes in the models might differ. In [90], a broader selection of adaptive systems is contrasted, encompassing

1. Bias and LDA adaptation in a CSP-based BCI system
2. Discontinuous and continuous LDA adaptation on bandpower features and Adaptive Autoregressive Parameter (AAR) features
3. Stochastic Meta Descent for a multiclass Statistical Gaussian classifier on bandpower features.

I will present the first of these adaptation strategies in sections 5.1 and 5.2, where I will analyze its performance in an offline evaluation, but also discuss the feasibility of this approach in online experiments.

As a more general question in the context of the evaluation of adaptive methods, it is always important to consider the problem of the choice of an appropriate length of the adaptation time window, which has an influence on the adaptation rate. If this window is chosen too large, the classifier responds very slowly to ongoing changes, whereas a short time window can result in poor estimation of the classifier parameters and therefore in a suboptimal classifier. This problem will be discussed in detail in Section 5.2.

4. Nonstationarity – or just Outliers?

It is often hard to determine the degree of nonstationarity that the brain signals are affected by, since the correct estimation of the underlying distribution is often prevented by a large variability due to outliers. As a simple example, Fig. 4.1 shows how the estimation of class-specific parameters, such as class mean and class covariance, can fail in the presence of outlier samples.

In this chapter, a general concept of outliers will be introduced. Based on this concept, methods will be developed to alleviate their effect on the data. If this treatment alone will remove any detrimental influence of the large variability of the data, the underlying data can further be assumed to be stationary. Yet, I will demonstrate that outlier reduction can not be the only answer to this problem, because the data still prove to be inherently nonstationary after outlier elimination.

4.1. The Outlier Concept

Biomedical signals such as EEG are typically contaminated by measurement artifacts, outliers and non-standard noise sources. I will propose to use techniques from robust statistics and machine learning to reduce the influence of such distortions. Two showcase application scenarios are studied: (a) Lateralized Readiness Potential (LRP) analysis, where I can show that a robust treatment of the EEG allows to reduce the necessary number of trials for averaging and the detrimental influence of e.g. ocular artifacts and (b) single trial classification in the context of Brain-Computer Interfacing, where outlier removal procedures can strongly enhance the classification performance.

4.1.1. Introduction

Identifying outlier points in a dataset can enhance our understanding of the data. By removing outliers, it is possible to improve the estimation of intrinsic properties such as mean or covariance matrix, and to analyze the data in single trial analysis. Various definitions of the outlier concept have been suggested, e.g. [1, 58, 52, 7, 122, 137, 77]. I will in the following introduce some model assumptions about the EEG data, and by outliers simply refer to those points not fulfilling these assumptions. I will show how this concept can be used to robustify the analysis of motor-related EEG data.

Typically EEG signals are distorted by artifacts and noise. If the few training samples that are measured within the 'calibration' time are contaminated by such artifacts, a suboptimal or even highly distorted classifier can be the consequence [93]. Since simple classifiers like Linear Discriminant Analysis (LDA), Regularized Discriminant Analysis (RDA) or Quadratic Discriminant Analysis (QDA) assume Gaussian distributions of the classes in feature space, every deviation from this assumption can result in poor performance of the

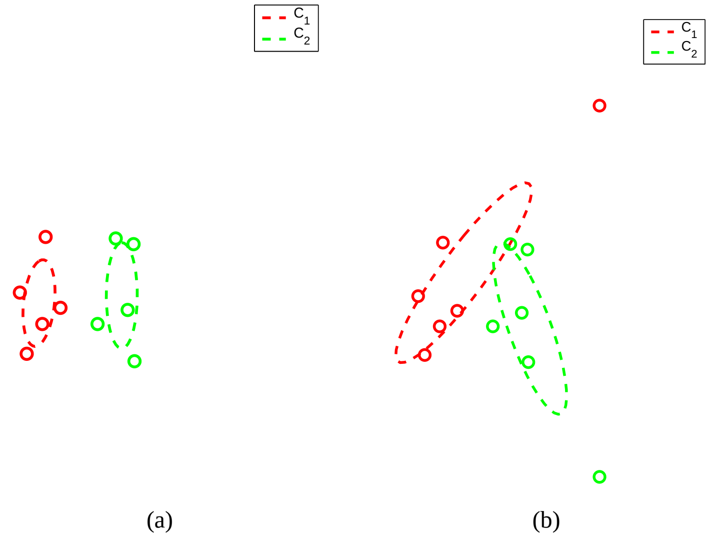


Figure 4.1.: Part (a) of this figure shows five randomly drawn samples of class C_1 and five samples of class C_2. The ellipsoid shapes denote the sample covariances around the sample means. The classes are easily separable. Part (b) shows the influence of outliers: by just adding a single point per class, which is no typical example for the class distribution, the parameter estimation is corrupted.

discrimination method. I will show that outliers can transform the data to a non-gaussian distribution. Therefore it is important to strive for robust machine learning and signal processing methods that are as immune as possible against such distortions.

4.1.2. Robustification Approaches for EEG analysis

The literature points out various methods of how to identify outliers [1, 58, 52, 7, 122, 137, 77]. In Section 4.2, I will use the delta-method ([53], see Section A for a short introduction) to identify outliers. This method does not rely on the estimation of parameters such as mean or covariance matrix of the data in feature space, but rather uses the relative distances of each data point to its k nearest neighbors. In Section 4.3, I will use the Mahalanobis distance [1, 130], which requires to estimate both mean and covariance matrix of the data sample to find points with the largest deviance from the class mean. Points with high distances to all others are really different from the usual data ensemble and should therefore not be considered representative. Furthermore a decision has to be made on how many trials should be removed based on the outlierness curve. Tests to automatize the cut point in this curve did not result in significant changes. Thus, for the purpose of this work I present only results where the 10 %-worst trials were removed.

Apart from the general issue of choosing an outlier detection method, it is also an inherent problem of EEG data that the dimensions of the feature space may have different qualities: usually, data points are given with a certain number of repetitions (trials), and they contain channel informations and the temporal evolution of the signal. A natural approach is to specifically use this information to find outliers within a certain dimension, i.e., removing

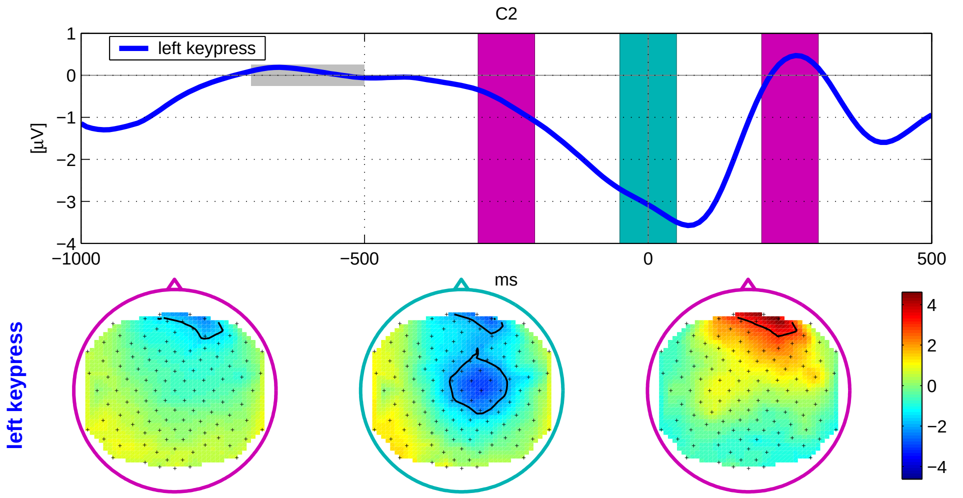


Figure 4.2.: This figure shows the Lateralized Readiness Potential during a finger movement for one subject. The timecourse for electrode C2, averaged over more than 500 trials, is shown above; the spatial distribution corresponding to the timepoints in the gray shaded areas is visible from the three scalp plots below.

channels with an increased noise level (due to high impedances at the specific electrode) or removing trials which are contaminated by artifacts from muscular or ocular activity. These approaches will be explained in detail in Section 4.3.

4.2. Outliers in LRP Features

This section will serve as an introduction to the nature of outliers in neurophysiological data. I will demonstrate exemplarily how outliers can disrupt the estimation of the distribution of certain features of the EEG, which suggests that removing those outlier trials can lead to a more robust estimation of the original LRP signal.

4.2.1. Experimental Setup

EEG data were acquired in 34 experiments from 17 different subjects. Brain activity was recorded from the scalp with multi-channel EEG amplifiers using 32–128 channels, at a sampling rate of 1000 Hz. The subjects pressed buttons of a keyboard with their index fingers in a (selfpaced) rhythm of approximately 0.5 Hz, in a selfchosen, random order. Each experiment consisted of 500–1000 repetitions of these movements (“trials”). The data were then stored for training classifiers for online BCI feedback experiments. In the course of these experiments, a cross-shaped cursor was presented to the subjects on the screen, indicating the estimated laterality of the keypress. The results obtained during training and feedback experiments are presented in previous publications, [15, 66, 67]. I will now use the same feature extraction as it was applied for classification purposes in order to demonstrate qualitative differences between in- and outlier trials.

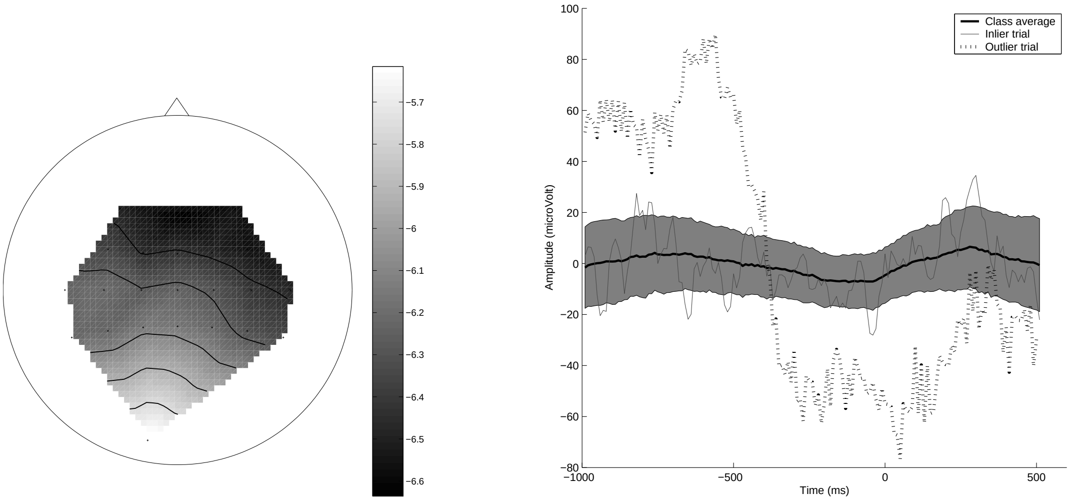


Figure 4.3.: In the left part of this figure, the differences between outlier and inlier trials are presented in terms of the Wilcoxon ranking score, averaged over data from 17 subjects (see text for details). The right part shows the EEG signal of one subject at electrode C2, averaged over more than 500 trials of repeated left index finger keypresses. One trial that has been identified as an outlier trial and a typical inlier trial are shown in the same plot. The gray area depicts the standard deviation of the inlier trials.

Feature Extraction

First, up to 20 central channels are selected that cover the areas corresponding to the motor cortices of the fingers. The data are then bandpass-filtered to 0.8–3 Hz, and the last 150 ms preceding the keypress are subsampled to 20 Hz, such that only three samples per channel remain. The samples are then concatenated over all channels. These steps are explained in detail in [15].

Outlier identification

According to the delta-score (see Chapter A; a more detailed version is given in [53]) obtained by each trial, those 10% of the trials with the highest scores are labeled as outliers. Figure 4.3 shows the difference in the power between outlier- and inlier-trials in terms of the w-scores w_{ch} of the average bandpower fv_{ch} in the frequency band from 0.8 to 5 Hz. The w-score is used in the Wilcoxon test for the comparison of two random samples for equal distribution. It is computed in the following way:

$$w_{\text{ch}} = \frac{R_{\text{ch,in}} - \frac{n_{\text{in}}(n_{\text{in}}+n_{\text{out}}+1)}{2}}{\sqrt{\frac{n_{\text{in}}n_{\text{out}}(n_{\text{in}}+n_{\text{out}}+1)}{12}}},$$

where $n_{\text{in}}, n_{\text{out}}$ are the respective numbers of in- and outliers, and

$$R_{\text{ch,in}} = \sum_{i=1}^{n_{\text{in}}} R(\text{fv}_{\text{ch},i})$$

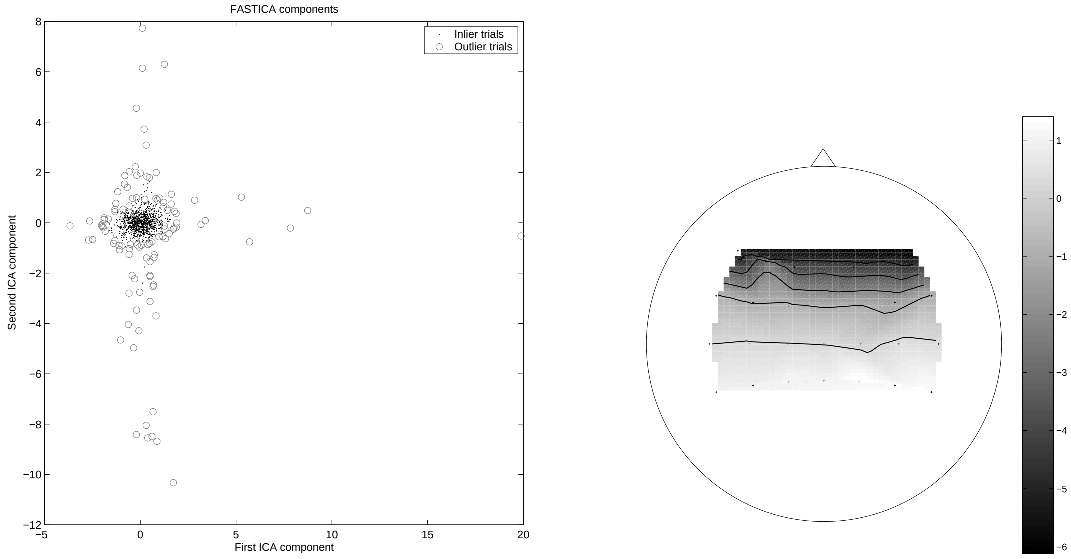


Figure 4.4.: The left part of this figure shows a scatterplot of two normalizations of linear projections of the feature space in one subject. The cross-shape of this plot reveals a non-gaussian structure of the data. The gray circles mark the trials which are identified as outliers. In the right plot, one of the corresponding projection matrices is shown. The spatial distribution suggests that the distribution of this projection is caused by vertical eye movements (such as eyeblinks).

is the sum of the ranks of all inlier trials in the combined sample of in- and outlier trials. A low *w*-value indicates that the variance of the outlier trials in this channel is higher than the variance of the inlier trials. The figure shows the spatial distribution of these differences after averaging over all subjects. Since the *w*-values of all channels are negative, the trials that have been identified by the outlier method have higher variances in this frequency band. By the spatial distribution, it is also apparent that this variance is caused by eye movements, since the influence of eye movements is maximal in the electrodes near the eyes and falls off with increasing distance, see e.g. [28]. In the right part of figure 4.3, the timecourse of the trials with lowest and highest delta-score (i.e., of an in- and an outlier) at electrode C2 are shown for one subject. This also illustrates the high variance of the outlier trials.

Figure 4.4 shows a two-dimensional linear projection of the feature space with the most “non-gaussian” components. These projections are found by applying Independent Component Analysis to the feature space for one subject. It has been shown in [15] that the applied preprocessing converts the data into a feature space where it is safe to assume gaussian distributions for the data. Under this assumption, every projection of the feature space should be normally distributed again, but this figure shows that there is in fact a strong “non-gaussianity” due to the outliers. The gray circles indicate the trials which the delta-method would identify as outliers. After the removal of 10% outlier trials, the projections are no longer significantly different from normal distributions.

Results

In this section I have illustrated that eye movements are a common source of deteriorating influences on the EEG signal when dealing with slow cortical potentials. In our experiments, there is a significant correlation between eye movements and the identification of the trials as outliers. Note that these trials may also be removed from the data ensemble by simple eye artifact-rejection; however, this rejection method assumes that only the eyes are sources of signal deterioration, while outlier detection methods also capture other types of influences, such as muscular activity or movement artifacts.

It has also been shown that outliers in EEG recordings can deteriorate the data in such a way that basic assumptions about the underlying distribution, e.g. gaussianity, are not met and hence a robust estimation of the parameters can not be guaranteed. Removing outlier trials from the recording can help to remove this detrimental effect of the outliers.

4.3. Outliers in Bandpower Features

So far, possible effects of outlier identification and outlier removal have been demonstrated only by their effect on the distribution in the EEG feature space. Now I will quantify this effect by applying the presented methods in a single trial classification context with bandpower features.

In this section I investigate data from 22 EEG experiments with 8 different subjects. All experiments included calibration sessions in which the subjects performed mental motor imagery tasks according to visual stimuli, as explained in Section 2.3.3. The classifiers were CSP-based, see Section 2.2.1, and subject-specifically trained to capture the ERD/ERS-complex connected to the motor imagery task.

There are a number of factors that could degrade the performance of the CSP method: (1) outlier trials where the subject either produces artifacts or does not perform the required mental task, (2) unreliable channels, that are partly noisy due to measurement problems. In this section I investigate two methods that would compensate for (1) in different ways and one method that tries to compensate for (2). The expectation in this study was that robustifying methods could only improve performance in few experiments because we had well controlled EEG measurements on subjects that were highly motivated for the experiments such that they would canonically try to avoid to produce artifacts.

4.3.1. Feature Extraction, Classification and Validation

There are several parameters in this feature extraction procedure that should be specifically chosen for each subject to obtain optimal results. In our online experiments this is done semiautomatically by combining machine learning, expert knowledge and visual inspection of some characteristic curves such as spectra and ERD curves, see [10]. In this comparative offline analysis, absolute performance does not matter, so there is one fixed setup for all subjects.

After choosing all channels except the EOG and EMG and a few outmost channels of the cap, a causal band-pass filter from 7–30 Hz is applied to the data, which encompasses both the μ- and the β-band. The extracted trials are the windows 750–3500 ms after the presented visual stimulus, since in this period discriminative brain patterns are present in most subjects. Afterwards the CSP algorithm (see Section 2.2.1) is applied to the data such that the number of channels is decreased by suitable linear spatial filters which are learned on the training trials. In this example, 3 patterns per class are used, which leads to 6 remaining channels. As a measure of the amplitude in the specified frequency band I calculate the logarithm of the variances of the remaining channels as feature vectors. LDA classifiers were used for classification.

To explore the performance of an algorithm, a 10×10-fold cross-validation is applied to the feature vectors. This means that the data set is randomly split into ten equal parts, each of which is used once as a test set while training is done on the other 90 percent. This procedure is repeated ten times to get 100 test errors.

Since the CSP algorithm and other techniques presented later on exploit label information, these techniques have to be used only on the training set within the cross-validation procedure. Otherwise the cross-validation error could underestimate the generalization error.

To maintain comparability between algorithms, the chosen divisions into training and test sets are stored, such that all algorithms are applied to the same divisions.

4.3.2. Outlier Removal

Channel Removal

Instead of calculating the covariances, the evaluation of the correlation coefficients gives the opportunity to estimate the certainty for each channel. Here I take the difference of the lower bound and upper bound of the 95 % confidence interval for the estimation of the correlation coefficients. Using this as a measure of the goodness resp. badness, unreliable channels can be removed by a simple threshold criterion.

Outlier-trial removal

As a simple and reliable approach I will show here only one way, which performed reasonably well in our studies. For the validation of the presented algorithms outliers were only removed considering the training set, but for the test set all trials without recognizing their outlierness were used. However, the information that a trial is an outlier might also be used in feedback situations, e.g. by freezing the cursor instead of providing the regular feedback. This option would greatly enhance the range of possible application, but as this study is only considering calibration data, I will forgo this option.

The presented outlier removal approach is based on the idea to use the Mahalanobis distance of the variance of each trial and channel as measurement of the outlierness of the trials (cf. [1, 130]).

Robustification by normalization

For the robust estimation of covariance matrices, many different algorithms have been proposed. Other feasible variants include approximating covariances via 1-norm, median absolute deviation (MAD) or using the least informative distribution approach (cf. [58]).

The method I am going to present in this category is to normalize each time point in the filtered EEG signal to have euclidean norm 1 over the channels. With this modified signal, the covariances and the CSPs are calculated and applied to the normalized data with the same processing as before. Different strategies like applying this spatial filter to the original filtered but unnormalized data or normalizing the whole window trialwise result in similar performance. Normalizing the EEG data in this way deletes the absolute amplitude of the signals and retains only the relative amplitudes in their spatial configuration. This is enough information to detect ERD features, and additionally has the effect that outliers have less influence in estimating covariances (of the normalized signals).

4.3.3. Results

As reported in many publications (e.g. [11, 18, 35, 36, 37]), one can see that the usual CSP algorithm often performs quite well. Nevertheless, there are some experiments in which one or more of the robustification approaches can greatly improve classification. Unfortunately, the same new methods can also deteriorate the results in other instances. This means that for the application in BCI feedback experiments, a meta-decision about the robustification method has to be taken, based on the data of the training session for each subject. For the validation of such a procedure on our offline data, two schemes were applied, which used different partitions of each data set.

In the *chron* approach, the data were split into their (chronological) first and second half. On the first half I calculated the cross-validation error for each of the competing algorithms as described in Section 4.3.1. I will call the results here the “expected performance” or “expected error” of the algorithm. Based on the expected performance, the most promising algorithm is chosen for the application on the test data. For this decision, the difference between the expected error of our baseline CSP approach and the expected error of each of the algorithms presented in Section 4.3.2 is calculated. Only if this difference exceeds a certain switching threshold, the alternative algorithm is chosen instead of the CSP approach for the evaluation of the test set. Once the decision is taken for one of the methods, the classifier is trained on this first half and applied to the other half of the data (“test performance”). This evaluation mode closely resembles an actual feedback situation; a fixed classifier is trained using only data from a preceding training session, and is applied to the following feedback data. Note, however, that this evaluation is prone to be affected by nonstationary behaviour of the EEG data, which is often encountered in this type of experiments.

The *nonchron* approach, the second evaluation method, is to a large extent invariant to these local changes in the EEG; here the training set consists of every even trial and the test set of every odd trial, such that slow trends are always present in both training and test data. The evaluation then proceeds as in the *chron* method.

Figure 4.5 compares this test performance gain in different switching thresholds for each of the algorithms and for the best of all of them. Furthermore the percentage of experiments is shown where a switch to a robustification algorithm took place. Obviously, this portion decreases with increasing thresholds, i.e., if we choose a more conservative strategy. On the other hand the mean performance gain increases (i.e., the classification test difference decreases) with increasing threshold, until only few or no false decisions are left. Nevertheless, there are very few experiments where the decision to change was wrong as seen in the figure, but the cases where a change improves the classification accuracy outweigh the others. Between the algorithms no substantial difference is visible, but as their success lies in different experiments, further improvement by combination strategies can be expected.

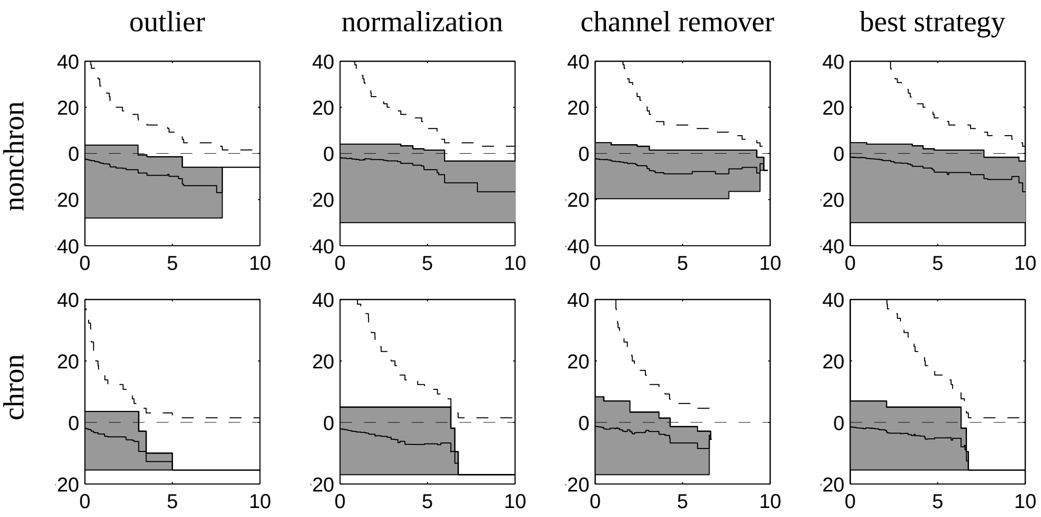


Figure 4.5.: The decision threshold varies between 0 % and 10 % on the *x*-axis. For each experiment, the expected performance of each robustification algorithm is compared to the expected performance of ordinary CSP. If the performance gain exceeds the threshold, the actual test error is evaluated on the test data. Out of all these experiments, where switching to the robustification method seems to be recommendable, the mean of the test error gain on the chosen algorithm against CSP is plotted as a black line. The range of all these values is visualized by the gray shaded area. Below zero, the change to the robustified method was successful: the lower the solid line, the higher the improvement. The dashed line shows the portion of experiments where the robustified method was chosen. In the first three columns each single robustification approach is compared to CSP whereas in the last column the best of all three robustified methods was used respectively. *chron* (for chronological order) denotes an evaluation mode where the expected error is estimated by cross-validation on the first half of the data and the test error is determined on the second half; the *nonchron* mode splits the data into even and odd trials.

In total the *chron* and *nonchron* evaluation strategy lead to similar interpretations. One important difference is that the gray area above the zero line is thinner in the *nonchron* case. That means that in the *chron* evaluation there are several cases in which the result of the chosen robustification method is worse than the baseline CSP result, while in the *nonchron* case there are less severe failures. This gives a clue for the reason of the failure: nonstationarity in the data. If all datapoints were drawn from the same distribution, then *nonchron* and *chron* evaluation should result in similar classification accuracies, but this finding shows that the distributions are undergoing changes throughout the time.

In the end, the figure shows that it can be profitable in some cases to switch to a suitable outlier algorithm for enhancing performance.

4.4. Discussion

EEG data recorded in motor-related tasks are highly challenging to evaluate due to noise, nonstationarity and diverse artifacts, specifically from eye movements or neck and jaw muscles. Thus BCI provides an excellent testbed for testing the quality and applicability of robust machine learning methods (cf. [41, 19]). In this study the effects that outlier trials may have on the distribution of the features were analyzed. It was shown that eye movements are a common source for the outlierness of trials in slow cortical potential data; the result we encountered was a shift of the data towards a non-gaussian distribution, where the removal of outliers may help to restore the model property of gaussianity that is assumed for linear classification. Finally, it was exemplified how outlier removal methods can improve the classification accuracy in the discrimination between different motor actions.

As our BCI system has so far mainly relied on dimension reduction techniques like CSP, this study has explored directions of their robustification against outliers. However in a BCI training protocol it is essential to decide whether to apply one of the robust alternatives or to stick with the conventional baseline algorithm, that obtains better results in some cases. As shown, this meta-decision, if exercised sufficiently conservatively, i.e., only after an expected gain of more than 5 %, can yield significant performance improvements. These encouraging results should nevertheless be carefully put into perspective: (i) no overall best robustification strategy can be observed and (ii) individualized choices need to be made for each subject. Furthermore the more conservative our strategy, the less likely it is to switch and also the less likely it is to have erroneously switched. Part of the reason, why the selected algorithm occasionally performs suboptimal is the intrinsic nonstationarity in a BCI experiment. Obviously BCI users are subject to variations in attention and motivation. Finally, this section has shown that using only outlier reduction techniques can not account for nonstationary behaviour of the data. In order to address this problem, it has first to be investigated how the nonstationarity in the data is generated and to which neurophysiological changes it corresponds. Also, since the classification approaches in this section are quite indirect measures of the changes in the distributions of the data, I will introduce some new approaches in the next section, where the feature space will be more thoroughly investigated.

From the above findings it follows that in order to further improve information transfer rates in BCI, methods have to be found which counter the effects of switching dynamics. Some methods in this spirit will be proposed in the following.

5. Observations in a Fixed Feature Space

5.1. Nonstationarity and Adaptation

Although I have provided evidence that nonstationarity in EEG signals affects the classification accuracy perceptibly, the question remains where the nonstationarity originates and consequently, which psychological or neurophysiological processes are involved. This section is intended to shed some further light on this question. I will present an approach which is based on a fixed feature space, such that only few parameters will have to be estimated.

A systematic quantitative study of data for multiple subjects recorded during offline and online sessions is presented. The methods for analysis of the data and visualization thereof are generally applicable and give a closer insight into the structure of the – global and local – change of the data quality. I will demonstrate the change in distributions of chosen EEG features, and provide evidence of changes both in the transition from offline to online settings, and in the course of a single online session. The former changes turn out to be shifts of the data in feature space, due to the different background activity of the brain during the online feedback task (see Section 5.1.5).

In the second part of the study, adaptive classification techniques for the use in BCIs with CSP-(Common Spatial Patterns) based features are presented, in order to gain quantitative understanding of these changes, and consequently remedial schemes for improving online BCI performance are proposed. When applying adaptive techniques on a variety of datasets collected during online task performance (Section 5.1.7), these results demonstrate that instabilities of the BCI control can be encountered throughout the experiment, but the major detrimental influence on the classification performance is caused by the initial shift from training to the test scenario. Hence, simple techniques that relearn only part of the classifier can account for this change, and significantly improve BCI control.

This study focuses on a feature space that is a low-dimensional projection of 128-channel EEG data computed by the CSP algorithm, see Section 2.2.1. However, the methods of analysis, measurement and visualization, as well as the questions regarding adaptivity addressed in this section are widely applicable and should serve as useful tools in studying nonstationarity in the BCI context.

5.1.1. Experimental Protocol

I will investigate data from a BCI study consisting of experiments with 6 subjects. For one subject no effective separation of brain pattern distributions could be achieved. Thus no feedback sessions were recorded and the data set is left out in this investigation. For the recording of EEG data during motor imagery of left hand, right hand and foot, calibration and feedback experiments were conducted as presented in Section 2.3 with 140 trials for each class.

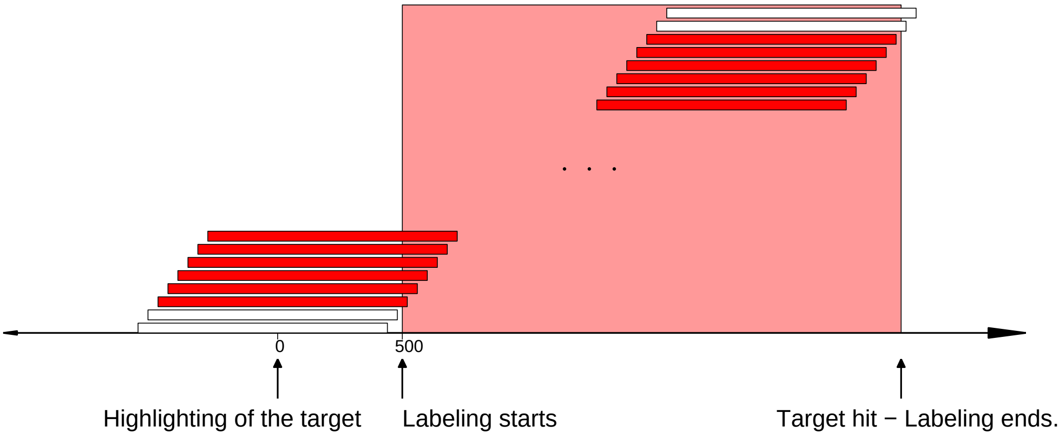


Figure 5.1.: In the feedback run, sliding windows were used for classification. For adaptation and evaluation, the windows (here colored red) between releasing the cursor and the end of the trial are selected. See text for details.

The data were then used to train a classifier for the two best discriminable classes, using the classification scheme presented in Section 2.2. Subsequently, two controlled feedback runs were recorded, in the "Cursor Control" feedback scenario (see Section2.3.5). A 1-second window of data was used to estimate the features, which were classified in overlapping windows every 40 ms, see Fig. 5.1. The continuous output from the classifier was then used to move the cursor either in a position-controlled (i.e., classifier output maps directly to horizontal position on the screen), or in a rate-controlled fashion (i.e., scaled classifier output was used to move the cursor by a small amount in the chosen direction). During each trial, one of the targets was highlighted and the subject was trying to navigate the cursor into the target. At the end of the experimental session, a third run of data was recorded for purposes of studying long-term performance of the trained classifier for 4 of the 5 subjects. This run included the same targets as the feedback session, but no visible cursor ("Feedback of Results", see 2.3.5).

5.1.2. Analyzing Data from Feedback Sessions

Since the online sessions were controlled (i.e., the subject was directed to hit a certain target), I will use this information to label the data collected during an online (or feedback) session. In a realistic BCI application, the labels of ongoing trials may not always be available, and any adaptive schemes we may propose will have to take this into account. For this section, I use the data labels in an offline analysis to provide greater insight into the data.

For labeling the data from a feedback run, I take the signals from the start of each trial until its successful completion, and process the signals in a manner similar to the online scenario; i.e., compute features on overlapping windows of the same size and overlap as used in the online protocol. These data points are labeled according to the appropriate target class. When using the recorded data for testing various classification schemes, I always assign samples coming from one trial either all to the training or all to the test set.

5.1.3. Changes in the Data Distributions

In this section, I will examine the changes in performance of the subjects using a variety of measures and visualizations that help us to characterize the type and degree of changes seen in EEG features used for BCI classification. These findings are also linked to possible neurophysiological changes that may cause these observed changes. The data are visualized in two different manners: (1) by fitting a gaussian distribution[1] on the data over an entire session (or over short term windows), and (2) by examining the optimal separating hyperplane computed using an LDA classifier on the chosen data.

5.1.4. Differences from Calibration to Feedback

Fig. 5.2 shows a comparison between training data collected offline (in the calibration phase), and the test data recorded during a subsequent feedback session. The figure shows, for two subjects, the hyperplanes of the classifiers computed on the training and test data respectively, along with the means and covariances of the data points from each class. For ease of visualization, the data are projected onto two carefully chosen dimensions containing maximal information.

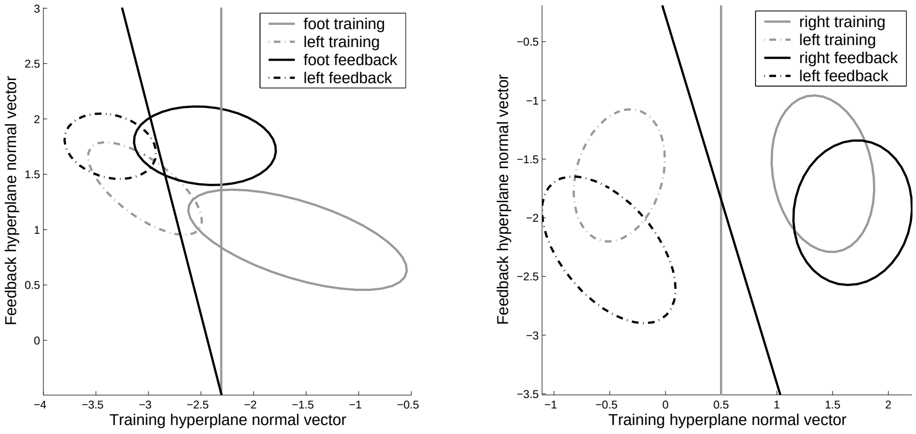
The x-axis shows the projection of the data on normal vector w_{TR} of the original classifier as obtained from the training session. The other dimension is chosen orthogonally to w_{TR}, such that w_{FB} (the normal vector of the optimal separating hyperplane for the feedback data) is contained in this two-dimensional subspace. The black and gray lines denote the intersections of the decision boundaries of the classifiers with the subspace which is shown here. It is a property of this display mode that the relative location of the distributions to the hyperplane can be seen by orthogonal projection, while the angles of the original space are preserved.

It appears in this figure that for subject *av*, the test data distributions look very different from the training data, and in fact, the original classifier would perform quite poorly in the online scenario. This is not always the case, though–for example, in subject *ay*, while the test distributions are different from the training data, the impact of this change on online performance is less severe.

In order to examine this change more closely across all online datasets, we consider the following two possibilities for modifying the training classifier hyperplane: (1) shift the original classifier's hyperplane parallel to itself[2] in order to get the best performance in the online setting, and (2) in addition, rotate the hyperplane to further improve performance on the online data. We call these two methods REBIAS and RETRAIN. Tab. 5.1 summarizes the shift and angle required for optimal performance on each online dataset. The required bias shift alone does not give a quantitative sense of the severity of the problem, and so Tab. 5.1-(a) shows this shift as a fraction of the training data's class mean distance from the training classifier's hyperplane. Note that in some cases, the optimal shift is comparable to the distance of one class mean to the decision boundary. This shows that an adaptation of the bias would be necessary for correct classification. Tab. 5.1-(b) shows the angle between training and test classifiers' hyperplanes on each dataset. In most cases, the angle does

[1] On the plausibility of the assumption of Gaussian distributions in EEG data, see e.g. [15] and also the discussion in [93].

[2] This can be implemented, e.g., by simply adding a *bias* to the classifier output.



(a)

(b)

Figure 5.2.: Changes in the optimal classifier from calibration to feedback: The figure shows, for subjects *av* and *ay*, the optimal hyperplane separating the training data classes (offline), and the test data classes (online). Also shown are the mean and covariance of the respective data distributions. In the case of subject *av* (figure (a)), the original classifier would perform very poorly, whereas for subject *ay*, as indicated in figure (b), the change is less severe.

(a)

Subject	al	aw	av	ay	aa
Shift/Distance	0.11	0.80	0.83	0.07	0.26
	0.12	0.94	0.56	0.09	0.26
	0.01	0.82	0.61	0.04	0.60

(b)

Subject	al	aw	av	ay	aa
Angles (°)	13.2	26.6	15.1	15.1	9.5
	9.7	20.6	28.7	17.7	6.7
	36.2	45.4	4.2	40.5	13.3

Table 5.1.: Measuring the changes in the optimal classifier for offline and online distributions. These are the changes necessary for the classifier to perform optimally on feedback data, for every experiment in this study. Part (a) shows the ratio between the optimal shift for correcting the bias and the distance between class means. Part (b) shows the angle between the old hyperplane (calculated from the offline data) and the optimal hyperplane for the feedback data.

(a)

Subject	al	aw	av	ay	aa
REBIAS/ORIG	0.93	0.79	0.67	1.00	0.97
	0.89	0.74	0.75	0.95	0.93
	1.00	0.75	0.80	0.99	0.82

(b)

Subject	al	aw	av	ay	aa
RETRAIN/REBIAS	0.98	0.99	0.99	0.98	0.98
	0.98	0.99	0.94	0.71	0.98
	0.72	0.87	1.00	0.73	0.97

Table 5.2.: Estimating the expected gain in classification when adapting the separation as calculated from the offline distributions to the online distributions. Any linear decision boundary between two normally distributed random variables misclassifies a certain quantile of both distributions. Here we compared the expected error quantiles for the optimal decision boundary for the training set to the decision boundary for the feedback sessions, when applied to the estimated distributions of the feedback data. Part (a) reflects the gain when only re-adapting the bias, and part (b) shows the improvement when the complete decision boundary is recalculated.

not change substantially. Tab. 5.2 provides a an interpretation of these classifier changes in terms of their impact on classifier performance.

It shows the ratios of estimated error quantiles for the training decision boundary, the bias-adapted decision boundary and the readapted decision boundary. It is evident that the adaptation of the bias results in a significant lower error quantile estimate (which confirms the findings in Tab. 5.1, whereas an additional adaptation of the angle only gives a comparatively small gain.

5.1.5. Explaining the Shift in Data Distributions

Fig. 5.2 and Tab. 5.1 together indicate that the primary difference between the offline and online situation is a *shift* of the data distributions for both classes in feature space, while not significantly changing their orientation. To clarify this aspect, I will display the spatial distributions of the band power on the scalp for the training and feedback situations.

As mentioned in Section 2.2.1, the CSP algorithm is used for feature extraction and the classifiers are trained on these features under the assumption that the spatial distribution of these activation patterns remain fairly stable during feedback.

This assumption can be verified in Fig. 5.3 which displays task specific brain patterns during offline vs. online session for one representative subject. The scalps with red resp. blue circles show band power during left hand resp. right foot motor imagery, calculated from offline (upper row) and online (middle row) session. In the plots of the offline session no systematic difference between the mental states can be seen, since the maps are dominated by a strong parietal α rhythm. Nevertheless the map of r values (see appendix) reveals a dif-

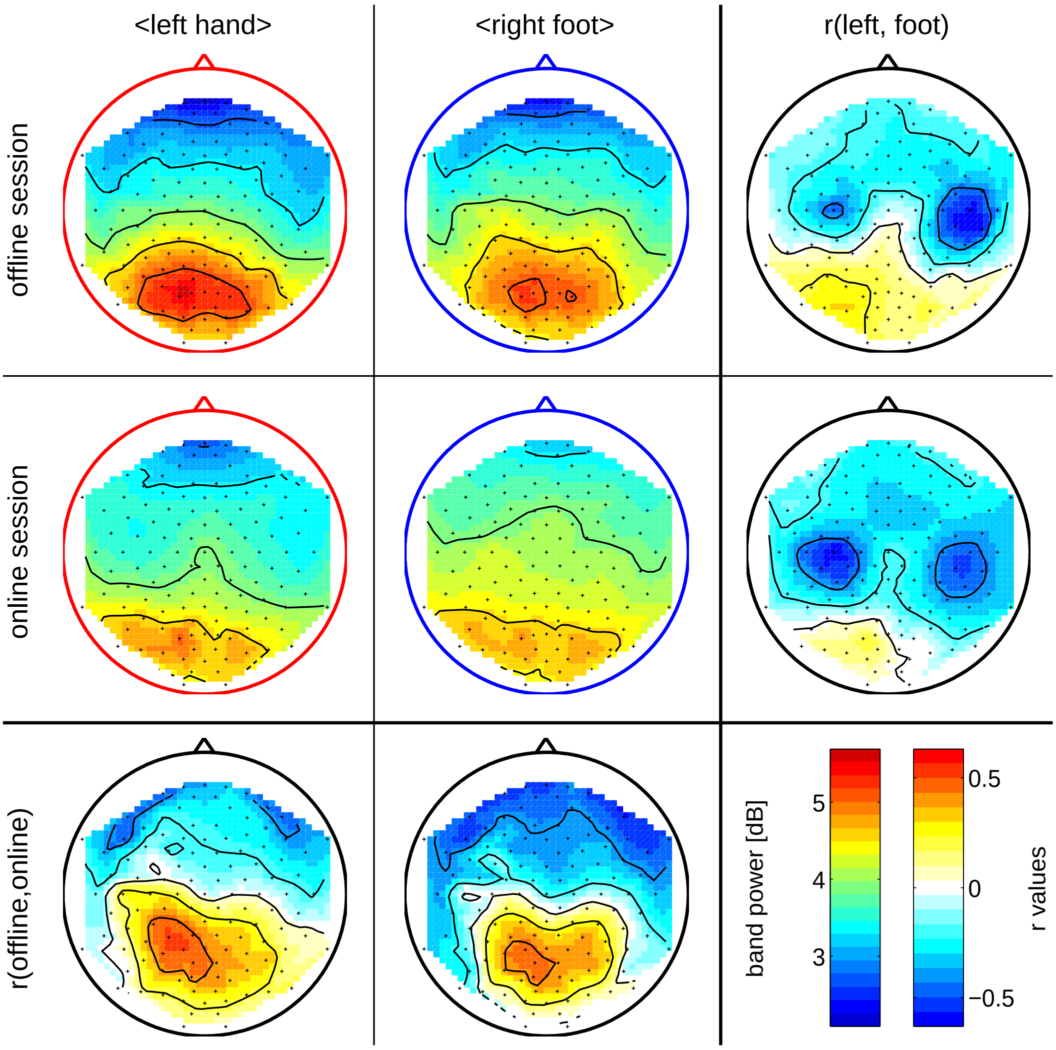


Figure 5.3.: This figure shows the task specific brain patterns and how they differ between offline and online sessions. The upper left 2×2 matrix of scalps displays topographic scalp maps of band power (broad band 7–30 Hz as used for calculating the CSP features in this subject). Maps are calculated from the offline session (upper row) resp. online session (middle row) separately for motor imagery of the left hand (left column) resp. of the right foot (middle column). Maps in the right column show the *r* values of the difference between the tasks, maps in the lower row show *r* values of the difference between offline and online session. While there is a huge and systematic difference between brain activity during offline and online sessions, the significant difference between the tasks stays fairly stable when going from offline to online operation (compare the *r* value maps in the right column).

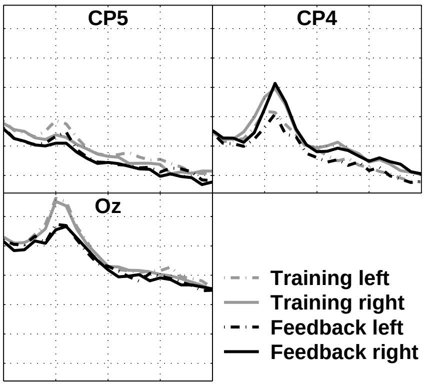


Figure 5.4.: This figure shows the spectra in the frequency range of 5–25 Hz both in training and feedback, for the two classes separately. The amplitudes are in the range of 22–54 dB.

ference focused over sensorimotor cortices. The parietal α rhythm is much less pronounced during the online session (middle row), resulting in a very strong difference between offline and online topographies, see *r* value maps in the lower row. In spite of this strong difference, the relevant difference between the tasks is qualitatively very similar in the offline and online settings (see the *r* value maps in the right column). The topography of the difference between offline and online situation suggests that in the former case a strong parietal α rhythm (idle rhythm of the visual cortex) is present due to the decreased visual input during the calibration measurement, while that rhythm activity is decreased in online operation due to the increased demand for visual processing. The power spectra (see Fig. 5.4) of electrodes in the corresponding regions corroborate this assumption, since there appears to be an increase in the power of the lower alpha band (just below 10 Hz).

Thus there is a difference in *background activity* of the brain in offline and online settings. This difference also strongly influences the CSP features chosen for classification.

5.1.6. Changes in EEG Features During Online Sessions

I will now present the performance of subjects in the course of a single online session. At each point of an online session, I will consider a window for each class containing all data points from the last 10 trials of that class. These data points can be used to get a *local estimate* for the density of each class at that point in time. A gaussian distribution is then fitted to these local windows of data, as well as to the entire online session, to obtain an *overall* density estimate.

Fig. 5.6 shows the Kullback-Leibler-divergence (see Section 3.1) of the local density estimate for each class from the overall density estimate of that class, over time. They are obtained by averaging over the last 10 trials per class and over the whole dataset, respectively. Since these curves alone do not provide information about classifiability of the data, the figure also shows sample visualizations of data from certain time intervals, along with the classifier hyperplane. It turns out that the data distribution for the foot class changes over the course of the experiment, and the KL-divergence curve reflects these changes.

The subject's overall performance was not very good, and the short period of time where

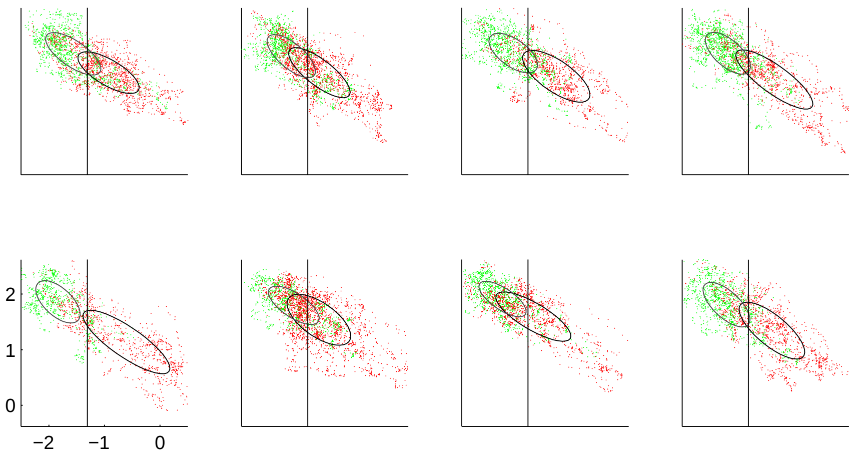


Figure 5.5.: The single plots in this figure represent the development of the feature distributions for subject *av* throughout one feedback experiment, windows representing each run (consisting of 28 trials each). The data are projected on the feature subspace spanned by the optimal hyperplane and the largest PCA component.

the KL-divergence for the foot class is very high corresponds to a period when the subject gained better control over the BCI. This can also be inferred from the corresponding visualizations.

A point to be noted here is that the subject took short breaks at various points during the experiment. Although the data acquired during these intervals were excluded from the analysis, the breaks may potentially influence performance. For example one of the breaks coincided with the end of the phase with good performance—it is likely that upon resuming the experiment the subject was unable to regain the control acquired in the previous phase.

The lower part of figure 5.6 shows local estimates of the distributions of both classes during one feedback session. We first calculated the classifier which is optimal for the feedback session and the largest PCA component w_{PCA} of the features. In this way, the projection shows the dimension with the largest variance. The x-axis shows the projection of the data on normal vector w_{FB} of that hyperplane of the feature space corresponding to the decision boundary of the classifier. The other dimension is chosen orthogonally to w_{FB}, such that w_{PCA} is contained in this two-dimensional subspace. Just like in Fig. 5.2, the projection preserves angles.

For a closer look, Fig. 5.5 shows the data distributions from each uninterrupted run. While the distributions are qualitatively different, it is not clear whether there is a discontinuity at each break. A further study consisting of new long-term experiments has therefore been performed to separate the gradual changes from the sudden changes induced by the breaks. It is presented in Section 5.2. It is, however, clear that the user's performance over a short period of time (about 30 minutes) can show considerable changes.

A new physiological interpretation cannot be given at this point since the patterns encountered in occasional lapses of performance are highly individual; furthermore, the recorded sessions were not sufficiently long to find trends in the EEG that correlate with performance. See sections 5.2 and 6.1 for experiments including longer sessions of BCI usage.

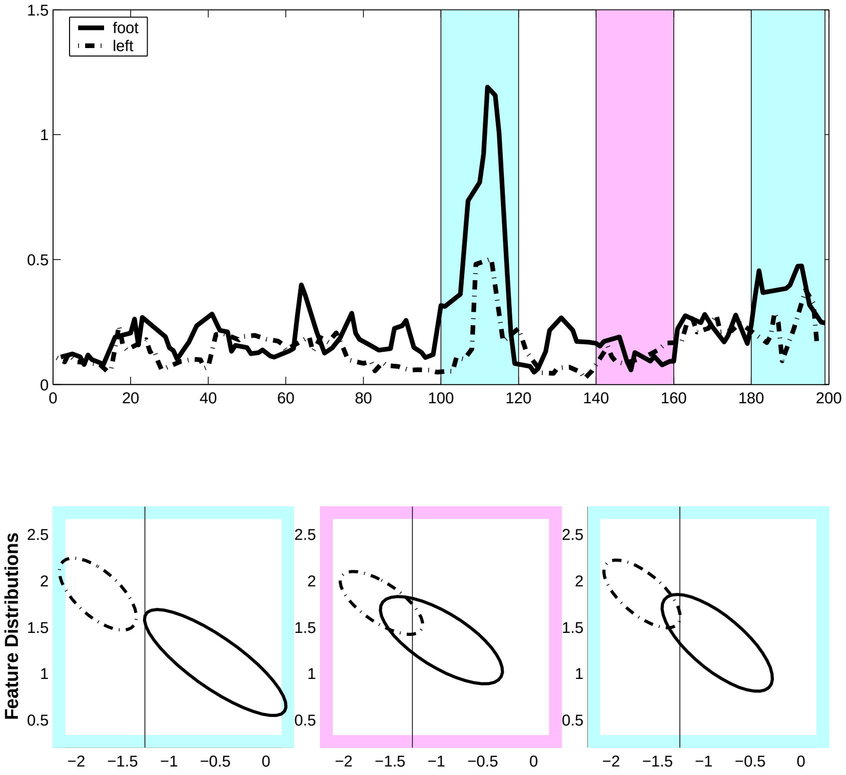


Figure 5.6.: This figure shows the change of the Kullback-Leibler Divergence during the feedback session. The corresponding feature distributions are displayed below for the shaded intervals. The data is projected on the plane spanned by the normal vector of the optimal separating hyperplane for the feedback and the largest PCA-component of the feedback data.

5.1.7. Adaptive Classification

I have shown qualitative and quantitative evidence indicating nonstationarity in the BCI classification problem; however, two questions remain unanswered: (a) What is the impact of this nonstationary behavior on performance in a feedback setting? (b) What remedial measures can we use to address the nonstationary behavior of EEG-related features? In this section, I will propose a range of techniques that aim to quantify the nature and impact of nonstationarity on performance, and thereby suggest adaptive methods for improving online control. Accordingly, I will define and compare a broad range of classifiers and the rationale behind each choice, and subsequently discuss their applicability in an online scenario.

Adaptive methods. The adaptive classification methods investigated are:

ORIG: This is the unmodified classifier trained on data from the offline scenario, and serves as a baseline.

REBIAS: The continuous output of the unmodified classifier, *shifted* by an amount that would minimize the error on the labeled feedback data.

RETRAIN: The features are computed as determined by the offline scenario, but the LDA classifier is retrained to choose the hyperplane that minimizes the error on labeled feedback

data.
RECSP: The offline training data are completely ignored, and the CSP feature selection and classification are trained solely on the feedback data.

The schemes are listed in increasing order of change to the classifier, and correspond to different assumptions on the degree of difference between offline and online data. In addition, there is the option of using *all* the labeled online data up to the current point (temporal), only a window over the immediate past (moving), or only an initial window of data from each session (initial). Each choice corresponds to different assumptions of the volatility of the online classification problem. The adaptation schemes are therefore C-REBIAS[3], C-RETRAIN and C-RECSP, W-REBIAS, W-RETRAIN and W-RECSP, and I-REBIAS, I-RETRAIN and I-RECSP respectively for the three cases considered.

Performance against Non-Adaptive Classifiers. Fig. 5.7-(a) compares the classification error of each adaptive method with the nonadaptive ORIG classifier. The adaptive classifiers were trained on a window of 60 seconds length. That was also the shortest (i.e., first) window of the temporal classifiers.

An inspection of the subplots reveals that the schemes REBIAS and RETRAIN clearly outperform the ORIG classifier, since most of the classification errors on the feedback data decrease. RECSP, on the other hand, does not improve performance. A possible reason for this is the small training sample size, a question which will be revisited in Section 5.2. Further, when examining each row, it can be seen that the I- methods perform better than the W- and C- methods, indicating that the I- methods are more stable than the C- and W- methods.

Also, the I-REBIAS method is comparable to the other algorithms–this is a very useful result because the I-REBIAS method is a lightweight adjustment that only requires a *short initial calibration period*, and is thus relatively nonintrusive. Thus Fig. 5.7-(a) shows that adaptive methods can indeed improve performance, even with simple adaptive schemes.

5.1.8. Performance against Best-Case Baseline

I will now address the central question regarding the online BCI scenario–how nonstationary is the data distribution within the online sessions? For each method, we define an idealized baseline scenario where the method can access the data and labels of both past and future from an online session. We then compare the temporal[4] k-fold crossvalidation error of the method in this baseline scenario to the method trained only on data from the past (as in the previous experiment).

This choice of baseline is aimed at examining whether each method suffers from having “too much” training data, or too little data. For example, if the classification problem were highly nonstationary, the windowed methods can be expected to outperform the baseline, since they can adapt to local changes. If the data are stationary across an online session, then the baseline would be the best possible choice, since it has more training data.

Fig. 5.7-(b) shows the results of this comparison. The following can be inferred from the figure: Firstly, the baseline is better in almost all cases, indicating that the adaptive

[3] C- denotes cumulative, W- denotes fixed window sizes, I- denotes use of only the initial segment of the session.

[4] i.e., the data is divided into k contiguous blocks in order to prevent overfitting

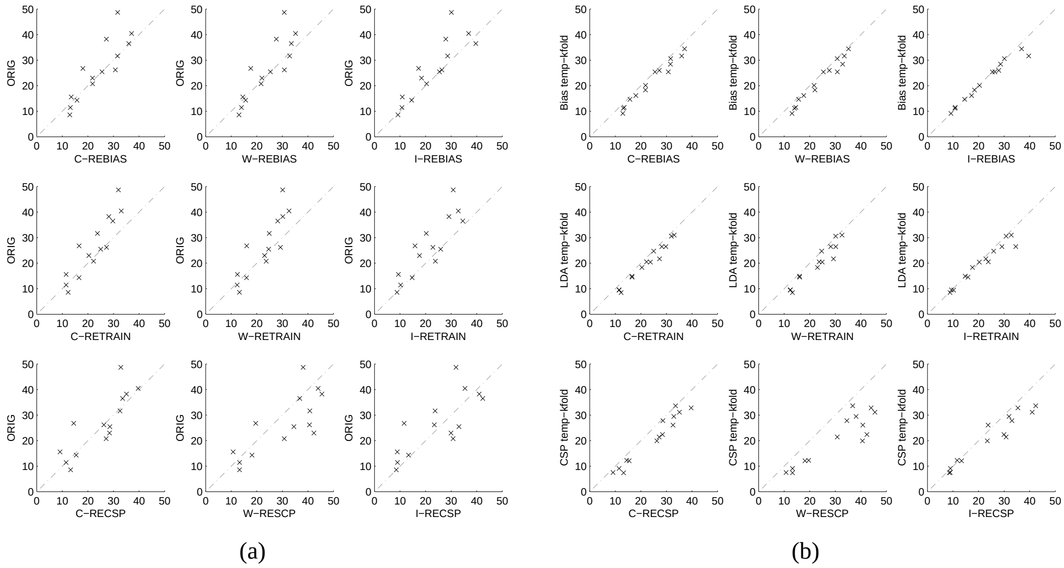


Figure 5.7.: Comparison of various adaptive classification methods on data recorded from online sessions. Each subplot is a scatter plot, with the error rate of a reference method on the *y*-axis and the error rate of the method of investigation on the *x*-axis. The performance of the latter is better for those data points that lie over the diagonal. Error rates are given in %. (a) All the proposed adaptive methods (except RECSP) clearly outperform the unmodified classifer trained on the offline data. (b) The adaptive methods are compared against a theoretical baseline that uses labels of future data points in the online session. See text for more details.

methods have insufficient data. This is especially true for the RECSP algorithms, and is clearly because of the very high-dimensional data they deal with. The REBIAS methods on the other hand do not benefit very much by the addition of more data, and the I-REBIAS error is comparable to the temporal k-fold error on REBIAS.

Note that these results do not necessarily mean that there are no dynamic changes in the data; in fact, in Section 5.1.6 it is shown that the data distributions do in fact move considerably. Instead, these results indicate that within the constraints of the chosen feature space and the adaptive algorithm, more training data will not help. The positive result from this experiment is that the best-performing REBIAS-algorithms, which only rely on an initial window of data, are comparable to the best possible error from the REBIAS algorithm.

5.1.9. Increasing Available Training Data

The choice of feature space is an important factor in the performance of our classification algorithm. Fig. 5.8 shows the error averaged across subjects for each dynamic version of the adaptive algorithms (i.e., the C- and the W- methods), as a function of the data window used for training. The figure confirms that the RECSP methods indeed improve on addition of training data; however, they are still considerably worse than the best-performing algorithm.

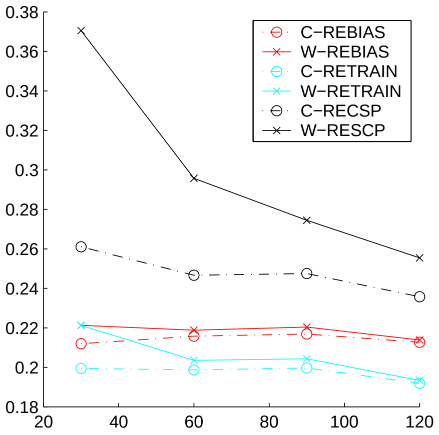


Figure 5.8.: Influence of parameters on the adaptive classification results. This figure shows the average error across all sessions and subjects as a function of the window of data points (in seconds) used for the windowed classification methods. For the C- classifiers, this indicates the size of the first training window.

The experiments were not sufficiently long to examine whether, with sufficient data, the RECSP- algorithms can be competitive. Note, however, that the algorithm is already too heavy-weight in terms of data and computation to be viable as an adaptive algorithm on short time scales. The question how much data the RECSP- algorithm actually needs will be addressed in Section 5.2, where a study with considerably longer experiments will be analyzed.

5.1.10. Discussion

These results show that an important factor affecting online BCI performance are the neurophysiological changes to the mental state of the subjects (as described in Section 5.1.6) between the offline and online settings. These changes do not render the EEG features found on the training data unusable, but require only a slight modification of the classification step. This is mainly attributed to the effectiveness of the applied offline feature selection scheme, and the fact that the basic neurophysiological processes used for control are similar in both scenarios. Our proposed modification can in fact be implemented in practice as a short *calibration phase* in the initial part of a session involving online BCI use.

While changes in performance and feature distributions do occur during online sessions (see Section 5.1.6), the classification results indicate that on average, they do not have a significant effect on performance. It is unclear at this point whether these changes can be affected by a different choice of feature space, or the use of additional features; however, a complete relearning of the feature selection is impractical due to higher computational costs and scarcity of data. Studies of longer-term BCI operation will be presented in Section 5.2 and Chapter 6 to shed further light on the exact nature of the changes during an online setting, and to suggest ways of addressing these changes.

The problem we frequently encountered with our Brain-Computer Interface system is that the performance decreases when going from offline training sessions to online operation of the BCI. One could suspect this to be caused by bad model selection strategies which could in principle choose overly complex classification models that overfit the EEG data. The evidence presented in this section has clearly shown that an alternative reason for failure should also be considered: nonstationarities in the EEG statistics. The subject's brain processes during feedback can cause the distributions to wander astray on a very local timescale. This observation could in principle make simple learning methods rather hopeless and one would have to refer to special statistical modelling that takes into account covariance shifts [133, 134] or even more sophisticated techniques such as transductive inference [143]. However, the successful adaptive methods investigated in this study that are guided by a better understanding of the possible neurophysiological causes of nonstationarity turn out surprisingly simple: a bias adaptation in combination with an offline feature selection scheme significantly increases BCI performance. It was clearly demonstrated that a strong source of nonstationarity stems from the difference between training and feedback session, whereas during the feedback session the statistics seems rather stable on the scale of up to an hour (depending on the subject). So a practical outcome of this study is (1) to correct for the bias between training and feedback session and (2) to furthermore incorporate every half hour one short 2-3 minute controlled feedback session into the neurophysiological paradigm under investigation and retrain or adapt the bias only when changes of the statistics, say due to fatigue, are observed.

5.2. How Much Data Are Required?

In the last section, it was demonstrated that it can be useful to adapt the classifiers in a fixed feature space, where the feature projections are predetermined by some training data. However, the question still remained if it would further improve the BCI performance if the feature space, i.e., the CSP filters, would continuously be adapted to the ongoing signal of a feedback session which is not interrupted by breaks between the runs. Also, we have seen that the adaptation of the feature space, RECSP, shows a suboptimal performance. Up to this point, it is not clear if this is just an effect of the small size of the adaptation window. In order to address these questions, I will examine data recorded from three subjects using a BBCI-based free text spelling experiment in which the labels for the data can be estimated post hoc from the words spelled out by the subjects, and can then be used online to adapt the classifiers used by the BBCI. I will revise some of the adaptive classification schemes from the last section that can use the estimated labels and I will present a comparative performance study of these schemes.

I will show that even in cases where a static classifier already performs quite well, online adaptation of the classifiers does not degrade the classification performance. Only the RECSP-method does still not provide a stable BCI performance if retrained on too short data windows.

5.2.1. Experimental Setup

This section relies on data from 3 subjects, of which one subject was a naive BCI user and the other two subjects had some previous experience. The experiments consisted of two different parts: a calibration measurement and a feedback period. After the calibration measurement, which proceeded as explained in Section 2.3.3, the parameters of the subject-specific translation algorithm were estimated (semi-automatically): selection of two of the three imagery classes and frequency bands showing best discriminability; CSP analysis (see Section 2.2.1) and selection of CSP-filters; calculating a linear separation between

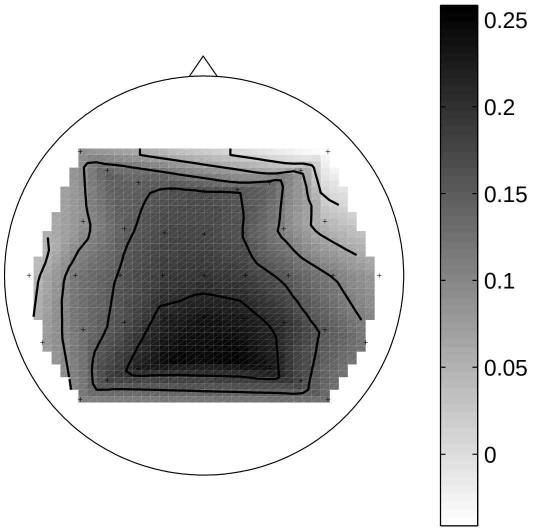


Figure 5.9.: This figure shows the shift in the power of the selected frequency band, in terms of r-values in one subject. Positive values indicate increased bandpower in the selected frequency band in the calibration measurement compared to the feedback session.

bandpower values in the surrogate CSP-channels of the two selected classes by Linear Discriminant Analysis (LDA).

In the feedback session, the classifier output of the ongoing EEG was used to move a cursor horizontally on the screen, as in the “Cursor Control” scenario (see 2.3.5).

5.2.2. Differences from Calibration to Feedback

In many earlier BBCI feedback experiments (as presented in Section 5.1), a strong shift in the features from training to feedback sessions was encountered as the major detrimental influence on the performance of the classifier. Accordingly I introduced an adaptation of the classifier’s bias as a standard tool in our system. To investigate the cause of this shift in data distributions, I compared the brain activity during calibration measurement vs. feedback situation using the bi-serial correlation coefficients r, which was calculated between bandpower values of each channel. The topography of one representative subject shown in Fig. 5.9 suggests that in the former case a strong parietal α rhythm (idle rhythm of the visual cortex) is present due to the decreased visual input during the calibration measurement, while this rhythmic activity is decreased in online operation due to the increased demand for visual processing, which supports the findings from Section 5.1.

5.2.3. Mental Typewriter Feedback

Since the mental engagement with an application is one additional possible source of nonstationarity, the investigation of nonstationarity issues is most interesting during the control of real applications. Therefore I chose a mental typewriter application which was used for free spelling by the subjects. Furthermore this application has the benefit that even in a free operation mode it is possible to assign labels (i.e., subject had intended to move the cursor left or right) to ongoing EEG in an a-posteriori fashion: after the correct completion of a word, one can decide for preceding trials the direction in which the subject was trying to move the cursor. This also applies if the intended word is not known to the experimenter beforehand. A detailed description of this type of feedback is given in 2.3.5.

Labeling Data From Online Feedback

The subjects were instructed to use the mental typewriter interface to write error-free sentences over a period of 30 minutes. After the recording of the data, labels were assigned a posteriori to the binary choices ("trials"), depending on the desired outcome of the letter.

Since the feedback was presented in asynchronous mode (i.e., starting and end point of each trial were not given at a fixed rate by the application, but were based solely on the output of the classifier), the lengths of the trials range from less than one second up to tens of seconds. For this analysis I take only the last 750 ms before the completion of the trial into account.

5.2.4. Adaptation Algorithms

The adaptive classification methods investigated are the same as in Section 5.1:

ORIG: This is the unmodified classifier trained on data from the calibration session, and serves as a baseline.

REBIAS: The continuous output of the unmodified classifier is used, *shifted* by an amount that would minimize the error on the labeled feedback data.

RETRAIN: The features are used as chosen from the offline scenario, but the LDA classifier is re-trained to choose the hyperplane that minimizes the error on labeled feedback data.

RECSP: The offline training data are completely ignored, and CSP feature selection and classification training are performed solely on the feedback data.

In the study previously presented here (Section 5.1), these methods have been shown to have a low computational complexity and a very straightforward applicability in an online scenario. As only the RECSP-method did not improve the classification performance, it is the purpose of this investigation to enquire the reason for this failure. With a much larger amount of training data, I will observe if the adaptation quality can be further increased for this method in particular.

In all adaptive methods a trade-off must be made: taking more training samples for re-training gives more stable estimates, but on the other hand it makes the method less adaptive, i.e., the policy should be to take as little training samples for re-training as possible but enough to allow estimations with reasonable stability. Here the number of training samples necessary for re-training is estimated separately for each method and each subject.

5.2.5. Results

For validation of the proposed classification schemes, I select for each trial from the feedback experiment a preceding window of specified size for re-training. Using the CSP filters and the classifier from the calibration measurement and these new training trials, I update the classifier and apply it to the current test trial – in RECSP training data are essentially

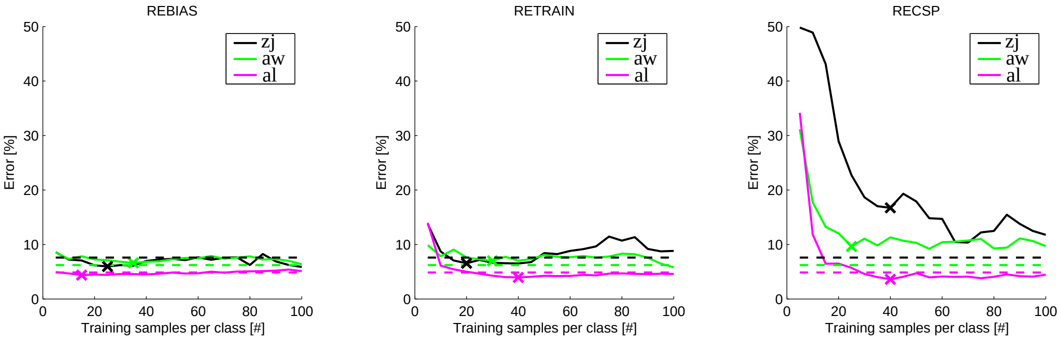


Figure 5.10.: The solid lines show the dependency of each algorithm on the number of training samples. For each subject, a sliding window containing the indicated amount of training samples per class (x-axis) was used for adaptation in the recording of the feedback session, and the resulting classifier was applied to the current sample. The average classification error on the test samples is shown on the y-axis in %, and the position of the optimal adaptation window is marked with a cross. The dashed horizontal lines indicate the respective errors of the ORIG-classifier, applied to all samples of the feedback session.

ignored. Then the predicted laterality is compared with the actual labels. Note that all validations only take into account labels of past trials as it would happen in an online feedback experiment. This procedure corresponds to the W-methods from Section 5.1. Fig. 5.10 shows the influence of the number of training trials on the accuracy of each adaptation method. In all methods under investigation, the error rate decreases with the used amount of training data. The RECSP-method, however, does not produce satisfactory results when used with less than 20 training samples per class. With more samples, the curve stabilizes at a low error rate for one subject, while remaining far above the baseline of ORIG for the other two subjects. Methods REBIAS and RETRAIN perform more stably, producing a reliable estimation with only a few adaptation trials.

Table 5.3 shows the classification errors of all presented adaptation methods, evaluated for a window size that is optimal in the sense that window sizes of up to 10 trials per class more will not decrease the classification error. This window size is also denoted in the table. For subject *al* all suggested adaptation methods show an improvement over the performance of the original classifier, where the gain is increasing with the complexity of the adaptation. However none of these improvements reach the level of significance (using McNemar's test, with a confidence level of $\alpha = 5\%$, see [47] for details). For subject *aw* the opposite effect can be observed. For the last subject REBIAS and RETRAIN again show some improvement while RECSP performs poorly. Taking into account that in this analysis the window size for adaptation was chosen a posteriori to fit optimally to the test (i.e., the evaluation is biased in favor of the adaptive methods), one has to conclude that *in this data* the original classifier can hardly be outperformed by any re-learning method.

	ORIG	REBIAS		RETRAIN		RECSP	
al	4.9	4.4	(15)	3.9	(40)	3.6	(40)
aw	6.2	6.6	(35)	7.0	(30)	**9.7**	(25)
zj	7.6	6.0	(25)	6.6	(20)	**16.7**	(40)
mean	6.2	5.7		5.8		10.0	

Table 5.3.: Validation errors for different adaptation methods, evaluated with a sliding window with an individually chosen number of training trials. The error rates are given in %. The number in brackets denotes the optimal window size (trials per class) for each subject under each method. Only the two numbers printed in bold differ significantly from the ORIG-classifier.

5.2.6. Discussion

This study shows the tradeoff between the various adaptive methods explored. The lightweight adaptive methods such as readjusting bias and angle of the LDA classifier using feedback data can help to improve the performance of the classifier. Here, they do not result in significant increases of the performance. Note that this does by no means indicate that nonstationarities were absent in the *EEG signals*, but it indicates that the BBCI classifier successfully extracted relevant information from sensorimotor areas, while filtering out contributions from sources of nonstationary characteristics like the visual cortex. In fact Fig. 5.9 which shows an enormous difference between the brain activity during calibration measurement and feedback operation was calculated from one of the experiments of this study.

Based on the results presented here one could conjecture that in the idealized case, feature extraction and classification can be successful in extracting a control signal that is not affected by the nonstationarities in the EEG. In fact, classification results on the data investigated in this study could hardly be outperformed by any of the adaptive methods. Nevertheless experience with other data (such as the study presented in Section 5.1) has shown that the change of mental state when turning from the calibration measurement to online operation sometimes needs to be compensated by a lightweight adaptive method such as the manual adaptation of the bias, see [9] or Section 5.1.

In summary, this study has shown that adaptive methods are not generally required for the continuous operation of a BCI. In fact, if robust feature extraction and classification methods are used that manage to eliminate most sources of nonstationarity, the adaptive methods can no longer improve the classification performance. However, even in this case, the straightforward methods of bias and LDA adaptation have shown to have a very stable performance and do, in particular, not compromise the classification performance as compared to the static classifier. These methods can therefore be readily applied in BCI experiments.

Note that all these methods still operate in a fixed feature space, which is not subject to adaptation over the course of the experiment. Due to the nonstationarity of the data (exemplified by Fig. 5.9), one can expect a much larger performance gain if the feature space is also either adapted or robustified against the changes in the data. Unfortunately, the most straightforward method RECSP performs suboptimally, even when adapting on very large time windows. This failure can certainly be accounted to the high dimensionality of the estimated parameters: for the CSP estimation, the covariance matrices of dimension $C \times C$ (C being the number of electrodes) have to be estimated, which is difficult with only a few data points. The next chapter is dedicated to exploring the nonstationarity of these covariance matrices and the associated CSP filters. With more knowledge on the nature of the nonstationarity, it will be easier to find ways to make the feature space robust against these influences.

6. How to Adjust the Feature Space

6.1. A Novel Method for the Quantification of Changes in EEG Data

For the classification of Event-Related Desynchronization (ERD), the estimation of the signal covariance matrices is of central interest. In the calculation of CSP filters and patterns, the classwise calculated sample covariance matrices Σ_1 and Σ_2 are calculated on bandpass-filtered, epoched training data for the classes 1 and 2. The calculation of optimal CSP filters then involves a simultaneous diagonalization of these matrices, as described in Section 2.2.1. In other words, a CSP projection of the EEG can be described as a function of the covariance matrices of the EEG. If the covariance matrices are changing over time, the discriminability of the CSP features is also jeopardized.

In [141], a simple method for the decomposition of these matrices was investigated for adapting the spatial filters across sessions. I will present a slightly different approach, which does not focus on the algebraic properties, but rather on the data distribution of the matrices: if it is possible to describe the change of the covariance matrices, a method can be defined to adapt the spatial filters as well. As a first step towards this goal, I will now present a new view on the covariance matrix space, in order to learn more how the parameters from different sessions are connected.

Note that these matrices are very high-dimensional features of the EEG: if C is the number of electrodes, the matrices have C^2 entries, but due to their symmetry, only the upper triangle matrix (with $\frac{C\cdot(C+1)}{2}$ entries) has to be estimated. For the remainder of this section, I will regard the sample covariance matrices as features of the EEG, and will show that a low-dimensional description for the shift of the covariance matrices is possible for most of the subjects under study. This description in simple terms can be helpful to identify the reasons for the shift and can point to remedies against its influence on the classification performance.

6.1.1. Experimental Setup

The estimation of such a large number of parameters (i.e., quadratic in the number of channels) is only possible with a sufficient number of observations. Therefore, I will report results from a series of experiments with 6 subjects, where 11 BCI feedback runs were conducted per experiment.

The feedback runs were conducted with a “Cursor Control” feedback, with a fixed duration of 3.5 seconds for each trial (see Section2.3.5 for details). Guided by the previous experience with nonstationary bias (see Section 5.1), there were two bias adaptation periods per run. In the beginning, for a period of 20 seconds, a cursor was presented rotating clockwise at constant speed. Based on these 20 seconds of EEG data, the average classifier output was calculated and then the current bias was determined. This method was intended

Subject	*zq*	*ay*	*zp*	*al*	*aw*	*zk*
Classes	LR	LR	LR	FR	LF	LR

Table 6.1.: The classes of mental imagery which the subjects used for the feedback. L and R denote left and right hand, and F denotes foot imagery.

to prevent the control signal from being shifted exclusively to either positive or negative values.

Then, the subject controlled the cursor for 20 trials (10 per class, in random order), and the bias was fine-tuned at the end of this period. With the adjusted bias, the subject controlled the cursor for the following 100 trials (50 per class, in random order). The procedure corresponds to the initial calibration of the bias, as it was found to be a good choice in offline studies, see [126].

In each trial of the feedback, one of the two boxes on either side of the screen was highlighted to indicate a new target. After being fixed in the middle for 750 ms, the cursor was released. The subjects were instructed to now imagine the associated hand or foot movement (see table 6.1), in order to hit the target with the cursor. The classifier output was used to control the cursor in horizontal direction in a rate-controlled fashion. After 3.5 seconds, the cursor was fixed again and the outcome of the trial was determined by the horizontal position of the cursor. If the cursor was on the correct side of the screen, the trial was counted as “hit”, and as “missed” otherwise. The target box was then colored green (for a successful trial) or red (in the other case), and after a break of 1 second, the next target was presented.

Only in runs number 6, 7, 10 and 11, the cursor was not visible to the subjects, such that they only performed their movement imagination and received a feedback of the successfulness of the trial by the color code of the target box at the end of the trial. This type of feedback will be called “feedback of results”. It hast been chosen in order to generate different levels of visual input for the subjects during the experiment. This enables us to supervise the influence of the visual scene on the band power in the visual cortex. More details on the setup of the experiment can be found in Section 6.3.

6.1.2. Methods

After bandpass-filtering the EEG, epochs were extracted in the interval from 500 ms to 4500 ms after the presentation of the target stimulus. The frequency filter was a Butterworth-filter of order 5, in the frequency band of [9 25]Hz.

By considering the class labels for each trial, I calculated the class-wise sample covariances

$$\Sigma_{i,j} = X_{i,j}^{\top} X_{i,j}$$

for class $i \in \{1,2\}$ and run $j \in \{1,\ldots,11\}$, where $X_{i,j}$ is the (#samples) $\times$ (#electrodes)-matrix which results from the concatenation of all trials of class i in run j; by the preceding bandpass-filter, I can assume that $X_{i,j}$ has mean 0 over time. The class-averaged sample covariance matrix is then computed as

$$\Sigma_j := 0.5(\Sigma_{1,j} + \Sigma_{2,j})$$

for run $j \in \{1,\dots,11\}$, if an equal number of trials for both classes has been recorded.

For the comparison of different elements of the vector space $\mathbb{R}^{C\times C}$, an appropriate metric has to be used.

In this example, I therefore used the metric resulting from the Frobenius norm, defined by

$$||A||_F := \sqrt{\sum_{i=1}^{C}\sum_{j=1}^{C}(a_{ij})^2},$$

for all $A \in \mathbb{R}^{C\times C}$. This is equivalent to using the canonical isomorphism

$$\hat{}\colon \mathbb{R}^{C\times C} \to \mathbb{R}^{C^2},$$

which maps a matrix to the concatenation of its columns, and applying the euclidean norm, i.e.,

$$||A||_F = ||\hat{A}||_2.$$

While this metric ignores most of the properties of a matrix, it is nevertheless sensitive to changes such as scaling. The metric regards the matrices $\Sigma_1,\dots,\Sigma_{11}$ as if they were vectors $\hat{\Sigma}_1,\dots,\hat{\Sigma}_{11}$, drawn from a C^2-dimensional normal distribution.

Then, mean and covariance of the C^2-dimensional vectors can be estimated as usual with sample mean and sample covariance. This is depicted in Fig. 6.1, where the solid ellipsoid line denotes the standard deviation of the sample covariance V of $\hat{\Sigma}_1,\dots,\hat{\Sigma}_{11}$ around the sample mean $\hat{\Sigma}_0$. Discriminant theory (see [42]) now tells us that the eigenvector $\hat{\Delta}$ associated to the largest eigenvalue of V is the best direction for a linear approximation of the points $\hat{\Sigma}_j$ ($j = 1,\dots,11$). $\hat{\Delta}$ is called the first principal component of V and can therefore be regarded as the direction of the shift, in the space $\mathbb{R}^{C^2}$.

After calculating $\hat{\Sigma}_0$ and $\hat{\Delta}$, the sample covariance matrices can be approximated by projecting their vectorial representations on the line $L := \{\hat{\Sigma}_0 + r\cdot\hat{\Delta} | r \in \mathbb{R}\}$. In other words, the approximations $\tilde{\Sigma}_j$ are defined as

$$\tilde{\Sigma}_j := \Sigma_0 + r_j \cdot \Delta,$$

where

$$r_j := \frac{(\hat{\Sigma}_j - \hat{\Sigma}_0)^\top \hat{\Delta}}{\hat{\Delta}^\top \hat{\Delta}}$$

for $j = 1,\dots,11$. The r_j can be interpreted as the factor by which the influence of the shift direction is imposed on the EEG data. Fig. 6.2 shows the approximated values and the sizes of the approximation errors for the previous example. In order to assess the quality of the approximation, I will calculate the average error, normalized by the average distance of the point from the mean, i.e.,

$$a := \frac{\frac{1}{11}\sum_{j=1}^{11}||\Sigma_j - \tilde{\Sigma}_j||_F^2}{\frac{1}{11}\sum_{j=1}^{11}||\Sigma_j - \Sigma_0||_F^2},$$

for every subject. Note that the closer this value to 1, the more orthogonal are $\hat{\Sigma}_j - \hat{\Sigma}_0$ and $\hat{\tilde{\Sigma}}_j - \hat{\Sigma}_0$ on average, which suggests a bad approximation.

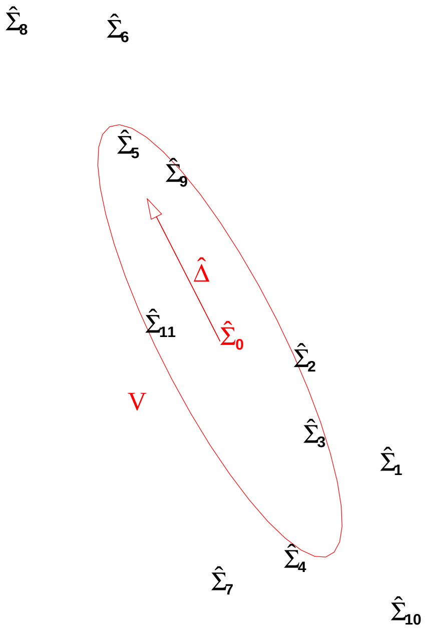


Figure 6.1.: This sketch shows how the sample covariance matrices $\hat{\Sigma}_j$ $(j = 1, \ldots, 11)$ are approximated linearly by principal component analysis (PCA). For this purpose, the sample mean (of the sample covariances), $\hat{\Sigma}_0$, and the eigenvector $\hat{\Delta}$ of the sample covariance matrix (of the sample covariances) V, associated to the largest eigenvalue of V, are estimated. These parameters are depicted in red.

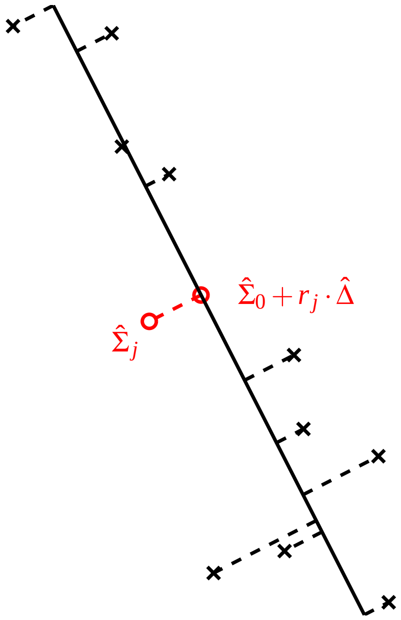


Figure 6.2.: This figure shows the linear approximation of the points in Fig. 6.1. The length of the orthogonal projection of $\hat{\Sigma}_j$ on the line $L = \{\hat{\Sigma}_0 + r \cdot \hat{\Delta} | r \in \mathbb{R}\}$ depicts the approximation error.

Subject	*zq*	*ay*	*zp*	*al*	*aw*	*zk*
a	0.45	0.06	0.08	0.10	0.28	0.03

Table 6.2.: This table shows the average error of the approximation normalized by the average distance to the mean $\hat{\Sigma}_0$. The error is smaller than 0.5 for all subjects, which shows that a considerable part of the $\hat{\Sigma}_j$ is explained by the linear interpolation.

6.1.3. Results

The approximation quality is shown in table 6.2. In all subjects, the value a is below 0.5, which corresponds to an average angle of at most 30° between $(\hat{\Sigma}_j - \hat{\Sigma}_0)$ and $(\hat{\hat{\Sigma}}_j - \hat{\Sigma}_0)$. For subjects *ay, zp, al* and *zk*, the angle is even below 6°, which corresponds to almost perfect linear interpolation. This finding can be supported by visual inspection of the original matrices Σ_j and their approximation counterparts $\tilde{\Sigma}_j$, as in the example in Fig. 6.3. The approximation error is very low, since the typical structure of the covariance matrices is almost completely reproduced by the approximation.

Since it is now possible to identify the shift of the covariance matrix from one run to the next one, I can now try to give an interpretation by analyzing the matrix Δ. The first observation is that it has one large positive eigenvalue, some more positive eigenvalues (approximately 10% of the number of channels), whereas all other eigenvalues are close to 0. Hence, if Δ is regarded as a positive semidefinite matrix, the shift can be interpreted as follows:

Suppose $(X_t)_{t\in I}, (Z_t)_{t\in I}$ are independent time series for some index set I. If $(X_t) \sim \mathrm{N}(0,\Sigma)$ and $(Z_t) \sim \mathrm{N}(0, c\Delta)$ for all t and for some $c \geq 0$, then $(X_t + Z_t) \sim \mathrm{N}(0, \Sigma + c\Delta)$. In other words, Δ can be interpreted as the covariance matrix of another process, independent from the one under observation.[1]

The main source of power of the new process can now be inspected, again by means of principal component analysis. Fig. 6.4 shows the eigenvector δ according to the largest eigenvalue of Δ for every subject. This eigenvector can be interpreted as the source of the main variance of the time series $(Z_t)_{t\in I}$. In all subjects, this δ exhibits a strong focus on parieto-occipital regions of the scalp. This indicates that differences in the α-band activity of the visual cortex are responsible for a shift in the sample covariance matrices from run to run.

For a closer investigation of this conjecture, I will give an analysis of two showcase examples, subjects *ay* and *zk*. These are the subjects with the best approximation performance, which supports the view of Δ as the main difference between runs.

During the experiments, subjects were asked to write down an estimate of their sleepiness, ranging from 1 (awake, not sleepy) to 10 (struggling to keep the eyes open, drowsy) after the completion of each run. Fig. 6.5 plots this “drowsiness index” (on the horizontal axis) against the approximation factor r_j for each run j. The numbers in the plot denote the numbers of the run. Although the drowsiness index was only denoted in discrete steps (i.e., integer numbers), a positive correlation is evident. The closer the subject was to falling

[1] c can be forced to be non-negative in every run, by setting $c_j := r_j - \min_k r_k$ and $\Sigma := \Sigma_0 - \min_k r_k \cdot \Delta$. Then the resulting matrices $\tilde{\Sigma}_j$ are exactly of the form $\Sigma + c_j \cdot \Delta$.

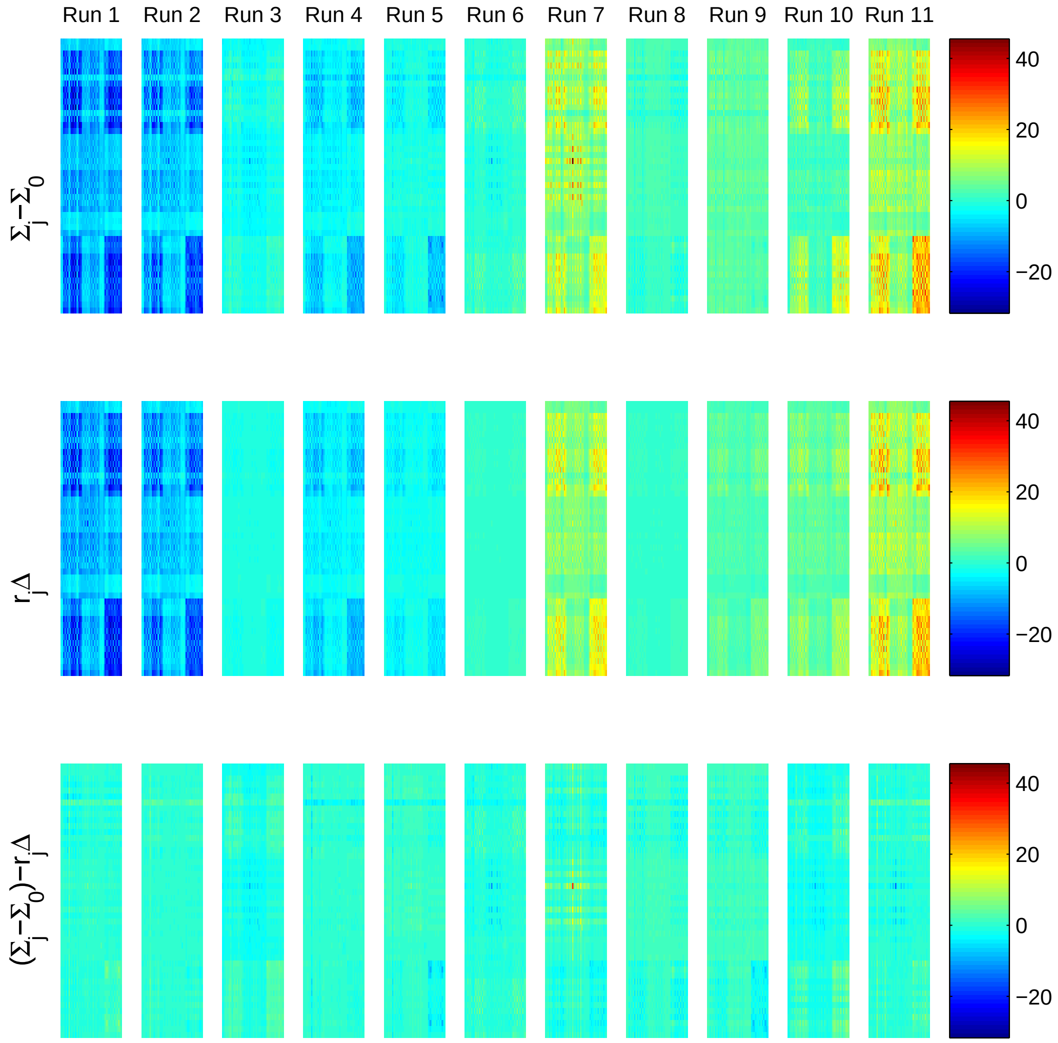


Figure 6.3.: The first row of this figure shows scaled images of the covariance matrices for each run for *ay*. After calculating the mean Σ_0 and the first principal component Δ, these matrices can be approximated by the terms $\tilde{\Sigma}_j = \Sigma_j + r_j\Delta$, as shown in the second row. If the approximation is successful, the remainder (as shown in row 3) is close to 0.

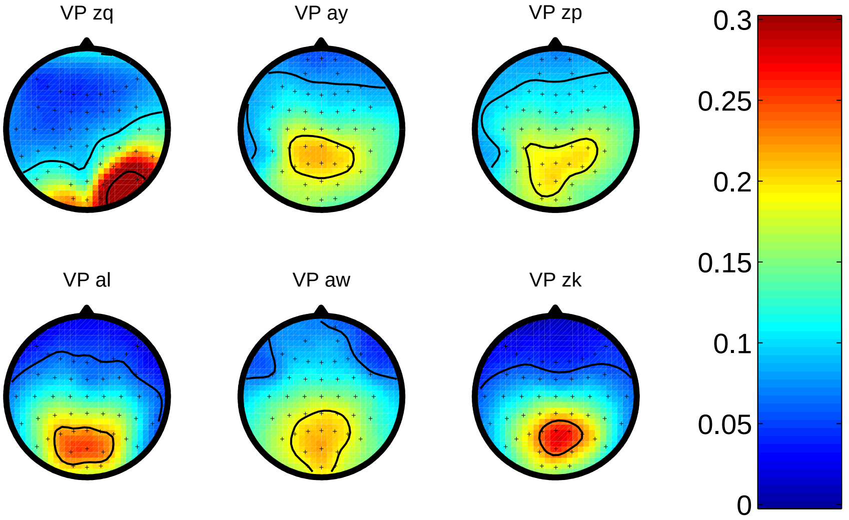


Figure 6.4.: The first principal component δ of the first principal component Δ for all six subjects under study. In all subjects, a strong focus on parieto-occipital regions can be noticed. The absolute scale of the components is irrelevant, since they are normalized. Only the relative distribution and topology are of interest.

asleep in run j, the higher the “covariance shift” index r_j. This finding is similar in five of the six subjects, for whom the correlation between the r_j and the drowsiness index is significant ($p < 0.1$).

For subject *zk*, the shift index is correlated with the modality of displaying the feedback. Fig. 6.6 shows the shift factor r_j for each block j. The runs where the cursor was invisible to the subject (“feedback of results”) are shaded in gray; in these blocks, the shift factor is much higher than in the other blocks. This correlation was only found in subject *zk*.

The presented examples support the interpretation of the covariance shift factor r as the activation strength of the associated principal component δ. The bandpower in the α-band exhibits a large variability from run to run.

6.1.4. Application to Classification Problems

This method can not only be used for analysis of the data, but also for the construction of spatial filters which are robust against the presented trend in the data from run to run and provide a good discriminability between classes. For this purpose, I will come back to the classification problem associated to a labeled time series.

For $i \in \{1,2\}$ and $j \in \{1,\dots,11\}$, let $\Sigma_{i,j}$ denote the sample covariance matrix of all the trials of class i in run j. I have shown that the common class covariance matrix $\Sigma_j = \frac{1}{2}(\Sigma_{1,j} + \Sigma_{2,j})$ for run j can be approximated by

$$\tilde{\Sigma}_j = \Sigma_0 + r_j\Delta.$$

In the light of the previous section, the main contribution of the difference between runs appears to be due to different activation levels of the visual cortex, which is not class-

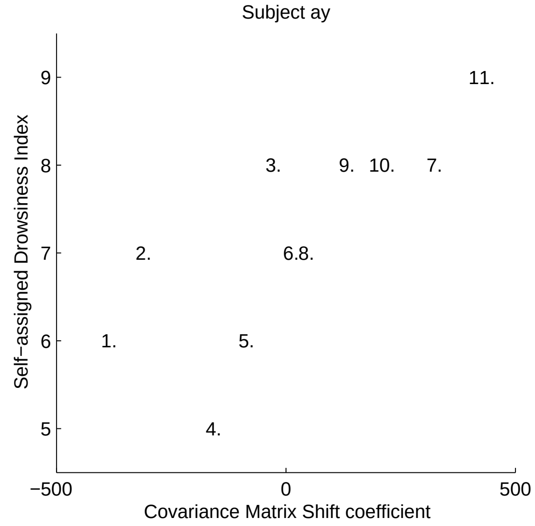


Figure 6.5.: The drowsiness index which the subjects denoted after each run are positively correlated with the covariance matrix shift coefficient. This plot shows the correlation for subject *ay*.

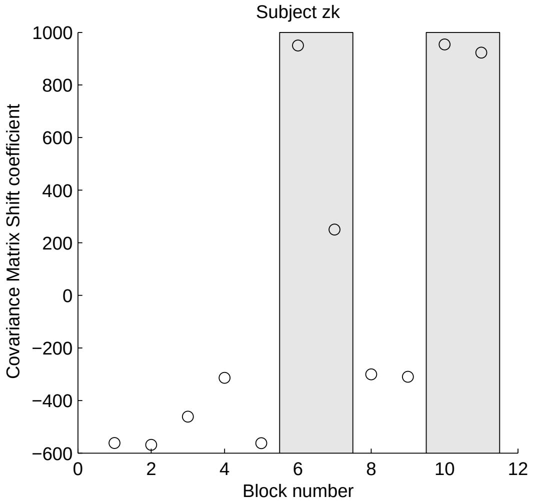


Figure 6.6.: For subject *zk*, this plot shows the covariance shift factor in each run. The gray shaded areas indicate the blocks where only “feedback of results” was given.

dependent; we may therefore assume that one can also approximate $\Sigma_{i,j}$ by

$$\tilde{\Sigma}_{i,j} := \Sigma_{i,0} + r_j \Delta$$

for the same real-valued scalars r_j as defined in the previous section, where $\Sigma_{i,0} := \frac{1}{11}\sum_{j=1}^{11}\Sigma_{i,j}$ is the mean of the class-wise sample covariance matrices for class i.

Note that the Δ is the same principal component as it has been calculated in the previous section, since I only want to consider a common shift for both classes.

With this approximation, it is possible to re-formulate the optimal CSP solution in the following way, as it is derived in [22]:

$$\begin{aligned} w_{CSP} &= \operatorname{argmax}_{w\in\mathbb{R}^C} \frac{w^\top(\Sigma_{1,j}-\Sigma_{2,j})w}{w^\top(\Sigma_{1,j}+\Sigma_{2,j})w} \\ &\approx \operatorname{argmax}_{w\in\mathbb{R}^C} \frac{w^\top(\Sigma_{1,0}+r_j\Delta-\Sigma_{2,0}-r_j\Delta)w}{w^\top(\Sigma_{1,0}+r_j\Delta+\Sigma_{2,0}+r_j\Delta)w} \\ &= \operatorname{argmax}_{w\in\mathbb{R}^C} \frac{w^\top(\Sigma_{1,0}-\Sigma_{2,0})w}{w^\top(\Sigma_{1,0}+\Sigma_{2,0}+2r_j\Delta)w} \\ &= \operatorname{argmax}_{w\in\mathbb{R}^C} \frac{w^\top(\Sigma_{1,0}-\Sigma_{2,0})w}{w^\top(\Sigma_{1,0}+\Sigma_{2,0}+c\Theta)w} \end{aligned}$$

The right hand side (with $c := 2r_j$ and $\Theta := \Delta)^2$ is similar to the formulation of "invariant CSP" (iCSP), see [16], where Θ is the covariance matrix of a process which does not provide discriminative information about the class labels. By adding it to the class-averaged covariance sample matrix $(\Sigma_{1,0}+\Sigma_{2,0})$, the resulting filters are more and more invariant to the process with the covariance matrix Θ, the higher the scalar value c is chosen. This, on the other hand, can make them less responsive to the actual class differences.

This calculation gives a new perspective on how to compute optimal CSP filters for each block: by approximating the class covariances by their estimates $\tilde{\Sigma}_{i,j}$, the calculation results in the iCSP filters which are invariant to the shift defined by the "covariance direction" Δ. The further Σ_j is from the mean covariance Σ_0, the higher the invariance factor $c = 2r_j$.

In this manner, the approximation of class covariances can be used for the calculation of robust classifiers. However, I will not follow this approach, since this method has some shortcomings for the parameter estimation *within sessions*:

1. In order to estimate the direction of the shift, Δ, the recording of several blocks of data is required, possibly under different levels of attention and sleepiness. The number of parameters to estimate is the product of the number of recorded blocks and C^2, where C is the number of electrodes. Therefore, a large number of trials in each block is required.

2. The estimation of the shift factor r is only possible if the direction Δ is known. If transferred to a real-world BCI scenario, this would correspond to an extremely long calibration measurement.

[2] Following a similar argument as in the previous section, it can be assumed without loss of generality that c is non-negative.

Subject	*zq*	*ay*	*zp*	*al*	*aw*	*zk*
a_{Class1}	0.49	0.15	0.52	0.20	0.44	0.07
a_{Class2}	0.63	0.14	0.16	0.21	0.41	0.08

Table 6.3.: This table shows, similar to table 6.2, the error of the approximation for classes 1 and 2 separately. See text for details.

3. The above calculation can only be accurate, if the approximation error

$$a_{\text{Class i}} := \frac{\frac{1}{11}\sum_{j=1}^{11}||\Sigma_{i,j} - \tilde{\Sigma}_{i,j}||_F^2}{\frac{1}{11}\sum_{j=1}^{11}||\Sigma_{i,j} - \Sigma_{i,0}||_F^2}$$

is very small. Table 6.3 shows the approximation quality $a_{\text{Class 1}}$ and $a_{\text{Class 2}}$ for each class and for each subject. If this table is compared to table 6.2, it shows that the magnitude of the approximation error for both classes is considerably higher than the error of the averaged sample covariance matrices, a.

However, for the estimation of parameters *across sessions*, the first two aspects do not restrict the applicability of the method, since each single session provides enough data for the robust estimation of high-dimensional parameters such as channel-wise covariance matrices.

6.1.5. Discussion

At the beginning of this section, some model assumptions were taken which are clearly not valid in a global setting.

For example, it is a common procedure in probability theory that the sample matrices Σ_i are modeled by a Wishart distribution, i.e., $\Sigma_i \sim W_C(\Sigma, 100 \cdot 400)$, where the degree of freedom (here: 40000) is the product of the number of samples in one trial and the number of trials used for estimation, and Σ is the unknown underlying covariance matrix. Unfortunately, this distribution does not give rise to an appropriate metric on $\mathbb{R}^{C\times C}$; due to this lack of direct applicability, I opted for the Frobenius norm.

Furthermore, the sample covariance matrices can not have a Gaussian distribution, since this would also imply that an indefinite or even negative definite matrix could occur with non-zero probability. Yet I have demonstrated that the Σ_j can be *locally* approximated by linear parametrization. These model simplifications resulted in a surprisingly accurate approximation for the sample covariance matrices of most of the subjects under study.

It is surprising that the main direction of change between the different Σ_j is a matrix which is again positive semidefinite except for very small negative eigenvalues. This is not evident, as the following simple example shows:

Suppose

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\Sigma_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then both are positive semidefinite, but the straight line connecting them is characterized by the direction of

$$\Delta := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

which is indefinite.

Since the Δ for the approximation of the covariances of all our subjects is always (nearly) positive semidefinite, the difference can be interpreted by means of an additional neurophysiological component that is only modulated in strength throughout the experiment. For all subjects, the principal source of the component can be localized in the parieto-occipital region of the scalp, and in most cases, the activation index r can be correlated to the level of tiredness that was estimated by the subjects after each run. As shown in Section 5.1, the activation of the visual cortex can have a serious impact on the signals that are used for bandpower feature classification. Here I have presented a completely different approach for the localization of the main source of this activation.

In this section, I have presented a new method for the comparison between the sample covariance matrices of bandpass-filtered EEG signals between different runs. As an example, the data of long experiments (with 11 runs per subject) were presented. The surprising result is that for most subjects the change from session to session can be easily and accurately parametrized by linear interpolation. Both the shift direction Δ and the shift factor r can be related to neurophysiological and psychological parameters, like the sleepiness of a subject or the activity of the parieto-occipital cortex regions. Therefore it is a new and useful tool for neurophysiological data analysis.

The proposed method can readily be used for classification and is closely related to the iCSP method demonstrated in [16]. Since the application for classification problems suffers from some drawbacks, mainly related to the amount of data needed for estimation of spatial filters, I will present a different approach in the next section, where the training data from the same experiment day will be reduced to a minimum.

6.2. Choosing a Robust Feature Space – and Omitting the Calibration

So far, the timescale in which I analyzed nonstationarity was limited to the course of a single session. There also exists a strong variability for a single subject when comparing data from one session to the next. This challenges a stable operation of Brain-Computer Interface (BCI) systems. In our studies, we tried exemplarily to re-use the classifier from a previous session for another online BCI experiment – an attempt which failed due to a significant change of the brain signals. This does not only provide evidence for the nonstationarity *between sessions*, but it leads to a very practical and relevant problem:

To present, the use of machine learning based EEG-BCI systems involves two time-consuming preparational steps at the beginning of every new session. The first one, the montage of an EEG cap, has been largely alleviated by recent advancements (see [117] and the discussion Section 6.3.5 in this chapter). The second step is the recording of calibration data, which I will address with this study. As the signals vary between sessions even for the same user, machine learning based BCI systems rely on the calibration procedure as a requirement for optimal performance (machine training). Even subjects that are experts in the use of machine learning based BCI systems still have to undergo the calibration session of about 20-30 min. From this data their (movement) intentions are so far infered.

Especially for patients with impaired concentration ability, this initial calibration reduces the valuable remaining time for controlling a device or computer software in the so called feedback application phase, but also for healthy users, it can be an annoying procedure.

The present contribution studies to what extent one can *omit* this brief calibration period. In other words, is it possible to successfully transfer information from prior BCI sessions of the same subject that may have taken place days or even weeks ago? While this question is of high practical importance to the BCI field, it has so far only been addressed in [124] in the context of transfering channel selection results from subject to subject. In contrast to this prior approach, I will focus on the more general question of transfering whole classifiers, resp. individualized representations between sessions. Note that EEG patterns typically vary strongly from one session to another, due to different psychological pre-conditions of the subject (see e.g. Fig. 3.5). A subject might for example show different states of fatigue and attention, or use diverse strategies for movement imagination across sessions. A successful session-to-session transfer should thus capture generic 'invariant' discriminative features of the BCI task.

For this I first transform the EEG feature set from each prior session into a 'standard' format (Section 6.2.1) and normalize it. This allows to define a consistent measure that can quantify the distance between representations. I use CSP-based classifiers (see Section 2.2.1) for the discrimination of brain states; note that the line of thought presented here can also be pursued for other feature sets resp. for other classifiers. Once a distance function (Section 6.2.2) is established in CSP filter space, one can cluster existing CSP filters in order to obtain the most salient prototypical CSP-type filters for a subject across sessions. To this end, I apply the IBICA algorithm [83, 84] for computing prototypes by a robust ICA decomposition (see Section 6.2.2). I will show that these new CSP prototypes are physiologically meaningful and furthermore are highly robust representations which are less easily distorted by noise artifacts.

6.2.1. Experimental Setup

The BCI sessions under study were performed with Event-Related (De-)Synchronization (ERD/ERS) phenomena (see Section 2.1.2) in EEG signals related to hand and foot imagery as classes for control. I investigate data from experiments with 6 healthy subjects: *aw* (13 sessions), *al* (8 sessions), *cm* (4 sessions), *zp* (4 sessions), *ay* (5 sessions) and *zq* (4 sessions). These are all the subjects that participated in at least 4 BBCI sessions. Each session started with the recording of calibration data, followed by a machine learning phase and a feedback phase of varying duration. All following retrospective analyses were performed on the calibration data only.

The calibration period for these experiments were performed with the standard setup, see Section2.3.3. The randomized and balanced motor imagery tasks investigated for all subjects except *ay* were left hand (l), right hand (r), and right foot (f). Subject *ay* only performed left- and right hand tasks. Between 120 and 200 trials were performed during the calibration phase of one session for each motor imagery class.

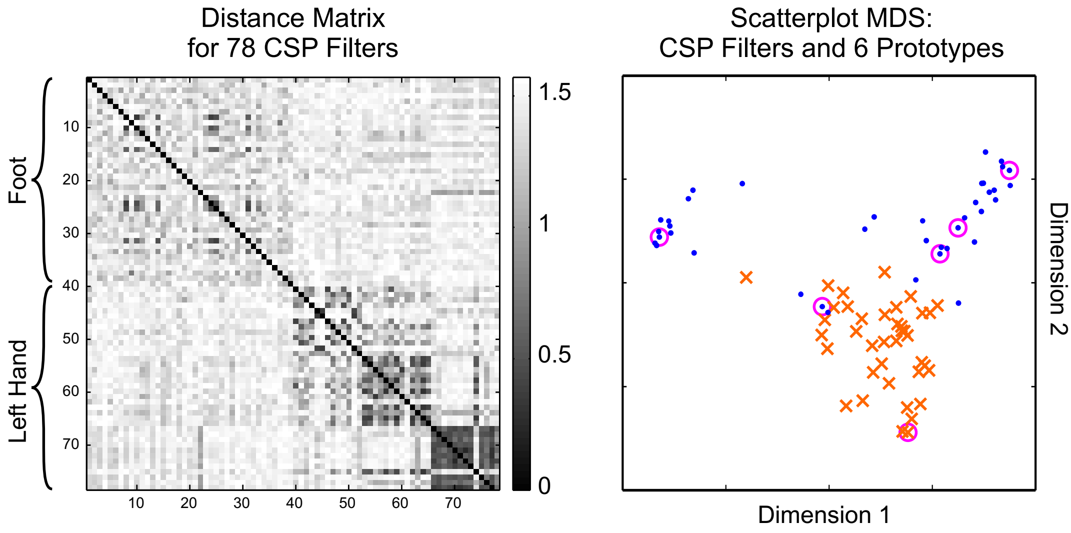


Figure 6.7.: **Left:** Non-euclidean distance matrix for 78 CSP filters of imagined left hand and foot movement. **Right:** Scatterplot of the first vs. second dimension of CSP filters after Multi-Dimensional Scaling (MDS). Filters that minimize the variance for the imagined left hand are plotted as red crosses, foot movement imagery filters are shown as blue dots. Cluster centers detected by IBICA are marked with magenta circles. Both figures show data from *al*.

Data preprocessing and Classification

The time series data of each trial was windowed from 0.5 seconds after cue to 3 seconds after cue. The data of the remaining interval was band pass filtered between either 9 Hz – 25 Hz or 10 Hz – 25 Hz, depending on the signal characteristics of the subject. In any case the chosen spectral interval comprised the subject specific frequency bands that contained motor-related activity.

For each subject a subset of EEG channels was determined that had been recorded for all of the subject's sessions. These subsets typically contained 40 to 45 channels which were densely located (according to the international 10-20 system) over the more central areas of the scalp (see scalp maps in following sections). The EEG channels of each subject were reduced to the determined subset before proceeding with the calculation of Common Spatial Patterns (CSP) for different (subject specific) binary classification tasks.

After projection on the CSP filters, the log-bandpower was estimated by taking the logarithm of the variance over time. Finally, a linear discriminant analysis (LDA) classifier was applied to the best discriminable two-class combination.

6.2.2. A Closer Look at the CSP Parameter Space

The CSP filters are not just randomly drawn points from $\mathbb{R}^C$ (where C is the number of electrodes), but instead represent subject-specific neurophysiological conditions, which suggests that, for a given subject, similar filters should be found across all sessions. I will first define a meaningful notion of similarity in this space and then use this relation to explore the space. It can be expected that the regions with a high density of CSP filters contain examples for filters which are particularly stable across sessions. I will call these regions “clusters”, and I will introduce a method how to sample prototypical filters from the clusters, using a notion of “inlier” points which have a low distance to their nearest neighbors.

Comparison of CSP filters

CSP filters are obtained as solutions of a generalized eigenvalue problem. Since every multiple of an eigenvector is again a solution to the eigenvalue problem every point in the space of CSP filters ($\mathbb{R}^C$) on the line through a CSP filter point and the origin form an equivalence class (except for the origin itself). More precisely, it is sufficient to consider only normalized CSP vectors on the $(C-1)$-dimensional hypersphere. This suggests that the CSP space is inherently non-euclidean. As a more appropriate metric between two points $\mathbf{w}_1$ and $\mathbf{w}_2$ (column vectors of a CSP filter matrix $\mathbf{W}$) in this space, I calculate the angle between the two lines corresponding to these points:

$$m(\mathbf{w},\mathbf{w}_2) = \min\left(\arccos\left(\frac{|\mathbf{w}_1 * \mathbf{w}_2|}{||\mathbf{w}_1|| * ||\mathbf{w}_2||}\right), \pi - \arccos\left(\frac{|\mathbf{w}_1 * \mathbf{w}_2|}{||\mathbf{w}_1|| * ||\mathbf{w}_2||}\right)\right).$$

When applying this measure to a set of CSP filters $(\mathbf{w}_i)_{i\leq n}$, one can generate the distance matrix

$$D = (m(\mathbf{w}_i,\mathbf{w}_j))_{i,j\leq n},$$

which can then be used to find prototypical examples of CSP filters. Fig. 6.7 shows an example of a distance matrix for 78 CSP filters for the discrimination of the variance during imagined left hand movement and foot movement. Based on the left hand signals, three CSP filters showing the lowest eigenvalues were chosen for each of the 13 sessions. The same number of 3×13 filters were chosen for the foot signals. The filters are arranged in groups according to their relative magnitude of the eigenvalues, i.e., filters with the largest eigenvalues are grouped together, then filters with the second largest eigenvalues etc.

The distance matrix in Fig. 6.7 shows a block structure which reveals that the filters of each group have low distances amongst each other as compared to the distances to members of other groups. This is especially true for filters for the minimization of variance in left hand trials.

Finding Clusters in CSP space

The idea to find CSP filters that recur in the processing of different sessions of a single subject is very appealing, since these filters can be re-used for efficient classification of unseen data. As an example of clustered parameters, Fig. 6.8 shows a hierarchical clustering tree (see [42]) of CSP filters of different sessions for subject *al*. Single branches of the tree form distinct clusters, which are also clearly visible in a projection of the first Multi-Dimensional Scaling-Components in Fig. 6.7 (for MDS, see [27]).

Once a suitable distance function is established, it can be used to find regions in the data space consisting of CSP filters, which are more densely sampled than others (‘clusters’). In particular, by identifiying points located in the middle of clusters, it is possible to select them as typical CSP filters.

The proposed metric of Section 6.2.2 coincides with the metric used for Inlier-Based Independent Component Analysis (IBICA, see [83, 84]). This method was originally intended

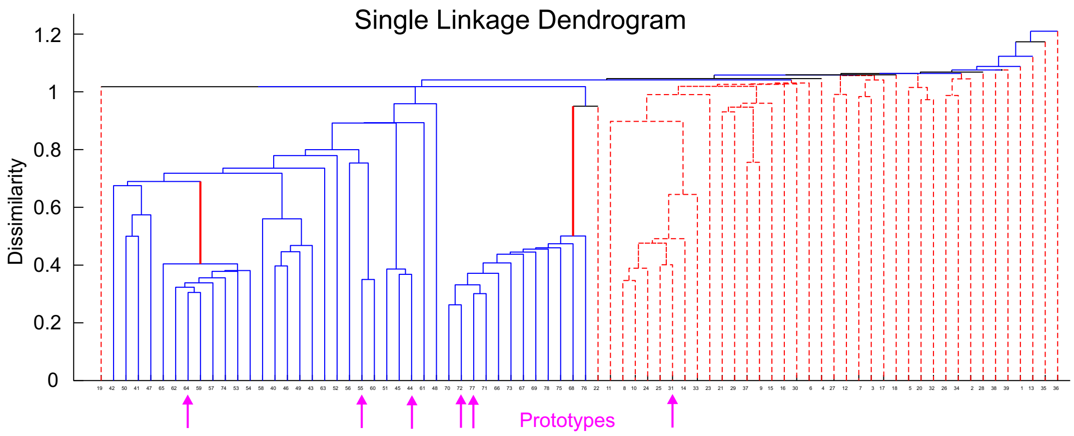


Figure 6.8.: Dendrogram of a hierarchical cluster tree for the CSP filters of left hand movement imagery (dashed red lines) and foot movement imagery (solid blue lines). Cluster centers detected by IBICA are used as CSP prototypes. They are marked with magenta arrows.

to find estimators of the super-Gaussian source signals from a mixture of signals. By projecting the data onto the hypersphere and using the angle distance, it has been demonstrated that the correct source signals can be found even in high-dimensional data. The key ingredient of this method is the robust identification of inlier points as it can be done with the γ-index (see [53]), which is defined as follows: Let $\mathbf{w}$ be a point in CSP-space, and let $\mathrm{nn}_1(\mathbf{w}),\ldots,\mathrm{nn}_k(\mathbf{w})$ be the $k = 5$ nearest neighbors of $\mathbf{w}$, according to the distance m. The average distance of $\mathbf{w}$ to its neighbors is then called the γ-index of $\mathbf{w}$, i.e.

$$\gamma(\mathbf{w}) = \frac{1}{k}\sum_{i=1}^{k} m(\mathbf{w}, \mathrm{nn}_i(\mathbf{w})).$$

If $\mathbf{w}$ lies in a densely populated region of the hypersphere, then the average distance to its neighbors is small, whereas if it lies in a sparse region, the average distance is high. The data points with the smallest γ are good candidates for prototypical CSP filters since they are similar to other filters in the comparison set. This suggests that these filters are good solutions in a number of experiments and are therefore robust against changes in the data such as outliers, variations in background noise etc. (see also section 4.1). Only the CSP filter with the lowest γ-index can clearly be regarded as “inlier”-point of a cluster. In order to find other regions of the filter space which are also densely populated, we applied a heuristic which is presented in the next paragraph.

Finding Cluster Prototypes

We first calculated the γ-index of each filter to obtain a ranking according to the distance function explained above. The lowest γ-index indicates that the corresponding filter is inside a region with many other filter examples and should therefore be chosen as cluster prototype. The same applies to the second-to-lowest γ-index, but in this case it would not be recommendable to select this filter, since it is highly probable that the filter is from the same region as the first one. To ensure that we also sample prototypes from other clusters, an

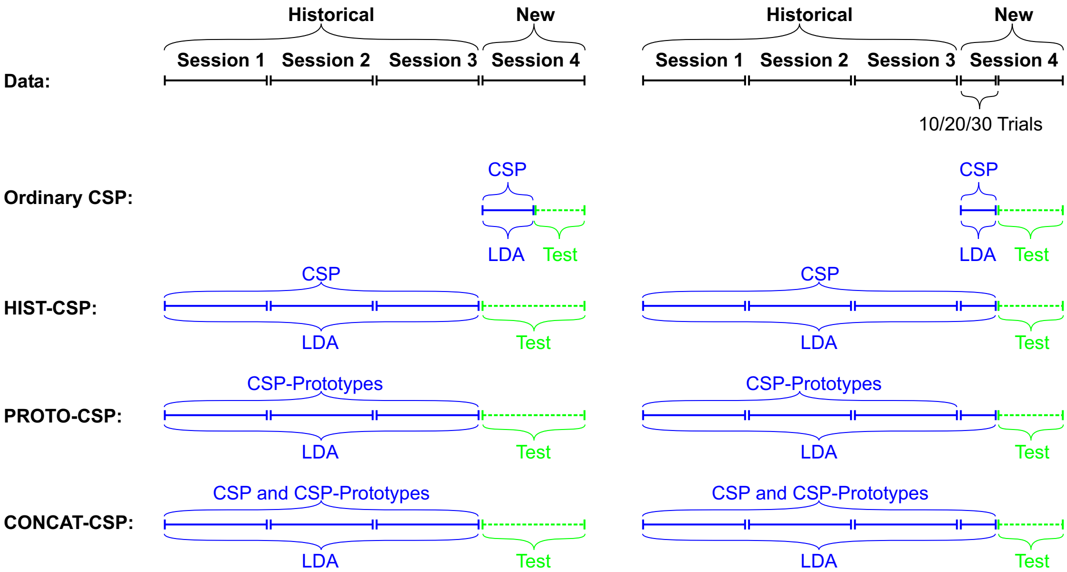


Figure 6.9.: Overview of the presented training and testing modes for the example of four available sessions. The left part shows a comparison of ordinary CSP with three methods that do not require calibration. The validation scheme in the right part compares CSP with three adaptive methods. See text for details.

incremental procedure of choosing and re-weighting was applied to determine a predefined number of cluster prototype filters.

The search starts with one prototype only, that is chosen as the filter with the minimal overall γ-index. The chosen filter point is removed from the set of all filter points. Then the average distance of each remaining filter to its neighbors is re-weighted by the inverse of the distance to the removed point, as explained in [83]. Due to this re-weighting, all points in the vicinity of the chosen cluster prototype receive a larger γ-index. The re-weighting is driven by the assumption that these neighboring points belong to the same cluster with high probability. Due to their increased γ-index, they are less likely chosen as prototypes in the next iteration. The iterative procedure ends, when a predefined number of cluster prototypes has been determined.

6.2.3. Competing Analysis Methods: How Much Calibration Is Needed?

Fig. 6.9 shows an overview of the validation methods used for the algorithms under study. The left part shows validation methods which mimick the following BCI scenario: a new session starts and no data has been collected yet. The top row represents data of all sessions in original order. Later rows describe different data splits for the training of the CSP filters and LDA (both depicted in blue solid lines) and for the testing of the trained algorithms on unseen data (green dashed lines). The ordinary CSP method does not take any historical data from prior sessions into account (second row). It uses training data only from the first half of the current session. This serves as a baseline to show the general quality of the data, since half of the session data is generally enough to train a classifier that is well adapted to the second half of the session. Note that this evaluation only corresponds to a real BCI scenario where many calibration trials of the same day are available.

Zero training methods

This is contrasted to the following rows, which show the exclusive use of historic data in order to calculate LDA and one single set of CSP filters from the collected data of all prior sessions (third row), or calculate one set of CSP filters for each historic session and derive prototypical filters from this collection as described in Section 6.2.2 (fourth row), or use a combination of row three and four that results in a concatenation of CSP filters and derived CSP prototypes (fifth row). Feature concatenation is an effective method that has been shown to improve CSP-based classifiers considerably (see [35]).

Adaptive training methods

The right part of Fig. 6.9 expands the training sets for rows three, four and five for the first 10, 20 or 30 trials per class of the data of the new session. In the methods of row 4 and 5, only LDA profits from the new data, whereas CSP prototypes are calculated exclusively on historic data as before. This approach is compared against the ordinary CSP approach that now only uses the same small amount of training data from the new session.

This scheme, as well as the one presented in the previous paragraph, has been cross-validated such that each available session was used as a test session instead of the last one.

6.2.4. Results

The underlying question of this work is how strongly the distributions of EEG data are affected by changes that occur between experimental sessions. As a practical consequence, the question arises whether information gathered from previous experimental sessions can prove its value in a new session. In an ideal case existing CSP filters and LDA classifiers could be used to start the feedback phase of the new session immediately, without the need to collect new calibration data.

I checked for the validity of this scenario based on the data described in Section 6.2.1. Table 6.4 shows the classification results for the different classification methods under the Zero-training validation scheme. For subjects *al*, *ay* and *zq*, the classification error of CONCAT is of the same magnitude as the ordinary (training-based) CSP-approach. For the other three subjects, CONCAT outperforms the methods HISTand PROTO. Although the ideal case is not reached for every subject, the table shows that our proposed methods provide a decent step towards the goal of Zero-training for BCI.

Another way to at least reduce the necessary preparation time for a new experimental session is to record only very few new trials and combine them with data from previous sessions in order to get a quicker start. I simulate this strategy by allowing the new methods HIST, PROTO and CONCAT to take a look also on the first 10, 20 or 30 trials per class of the new session. The baseline to compare their performance would be a BCI system trained only on these initial trials. In Fig. 6.10, this comparison is depicted. Here the influence of the number of initial training trials becomes visible. If no new data is available, the ordinary classification approach of course can not produce any output, whereas the history-based methods, e.g. CONCAT already generates a stable estimation of the class labels. All

Subjects	*aw*	*al*	*cm*	*zp*	*ay*	*zq*
Classes	LF	RF	LF	LR	LR	LR
Ordinary CSP	5.0	2.7	11.8	16.2	11.7	6.2
HIST	10.1	2.9	23.0	26.0	13.3	**6.9**
PROTO	9.9	3.1	21.5	26.2	**10.0**	11.4
CONCAT	**8.9**	**2.7**	**19.5**	**23.7**	12.4	7.4
Sessions	13	7	4	4	5	4

Table 6.4.: Results of Zero-Training modes. All classification errors are given in %. While the ordinary CSP method uses half of the new session for training, the three methods HIST, PROTO and CONCAT exclusively use historic data for the calculation of CSP filters and LDA. (as described on the left side of Fig. 6.9). Amongst them, CONCAT performs best in four of the six subjects. For subjects *al*, *ay* and *zq* its result is even comparable to that of ordinary CSP.

methods gain performance in terms of smaller test errors as more and more trials are added. Only after training on at least 30 trials per class, ordinary CSP reaches the classification level that CONCAT had already shown without any training data of the current session.

Fig. 6.11 shows some prototypical CSP filters as detected by IBICA clustering for subject *al* and left hand vs. foot motor imagery. All filters have small support (i.e., many entries are close to 0), and the few large entries are located on neurophysiologically important areas: Filters 1–2 and 4–6 cover the motor cortices corresponding to imagined hand movements, while filter 3 focuses on the central foot area. This shows that the cluster centers are spatial filters that meet the neurophysiological expectations, since they are able to capture the frequency power modulations over relevant electrodes, while masking out unimportant or noisy channels.

6.2.5. Discussion

This work shows that experienced BCI subjects do not necessarily need to perform a new calibration period in a new experiment. By analyzing the CSP parameter space, I could reveal an appropriate characterization of CSP filters. Finding clusters of CSP parameters for old sessions, novel prototypical CSP filters can be derived, for which the neurophysiological validity could be shown exemplarily. The concatenation of these prototype filters with some CSP filters trained on the same amount of data results in a classifier that not only performs comparable to the presented ordinary CSP approach (trained on a large amount of data from the same session) in half of the subjects, but also outperforms ordinary CSP considerably when only few data points are at hand. This means that experienced subjects are predictable to an extent that they do not require calibration anymore. The presented data clearly show that the distributions of the CSP filters are changing from session to session, which corresponds to nonstationary time series on a long timescale. However, the newly introduced perspective of data mining *on the parameters* has led to a method for the extraction of very robust features which can also be expected to work on a new, unseen data set.

Advanced BCI systems (e.g. BBCI) have the ability to dispense with extensive subject training and now allow to infer a blueprint of the subject's volition from a short calibration session of approximately 30 min. This became possible through the use of modern machine

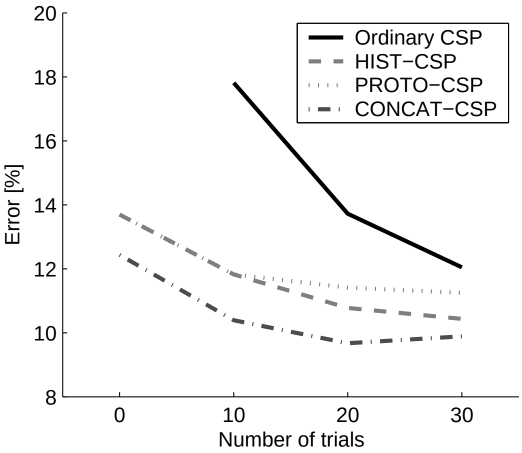


Figure 6.10.: Incorporating more and more data from the current session (10, 20 or 30 trials per class), the classification error decreases for all of the four methods described on the right side of Fig. 6.9. The three methods HIST, PROTO and CONCAT clearly outperform ordinary CSP. Interestingly the best zero-training method CONCAT is only outperformed by ordinary CSP if the latter has a head start of 30 trials per class.

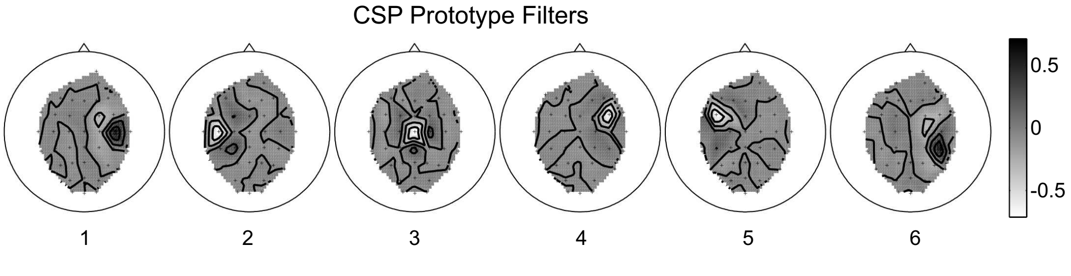


Figure 6.11.: First six CSP prototype filters determined by IBICA for *al*.

learning technology. The next step along this line to make BCI more practical is to strive for zero calibration time. Certainly it will not be realistic to achieve this goal for arbitrary BCI novices, rather in this study I have concentrated on experienced BCI users (with 4 and more sessions) and discussed algorithms to re-use their classifiers from prior sessions.

As all the data presented in this section was analyzed offline, it is still an open question how well the results will transfer to the online scenario. Therefore, I will now put the methods covered here into action and will present the results obtained by an online experiment with the CONCAT-classifier in the next section.

6.3. Towards Zero Training for Brain-Computer Interfacing

In the previous Section 6.2, I have presented a method for the comparison of different spatial filters. This led to the identification of particularly stable CSP filters which can be expected to perform well on future sessions. This development opens up a new field for further investigations: In the case of long-term BCI users, who repeatedly perform BCI sessions with the same mental tasks, rich datasets of previous sessions are accessible. While the standard machine learning approach only focuses on the current day, the previous section has demonstrated in an offline analysis, that also data from other sessions than the current one can be used to set up a classifier with a high performance right from the start. As a proof of concept, the offline analysis has shown that the CONCAT method is even superior to the standard CSP approach with up to 30 trials of calibration data.

The transfer of these results to an online application can be jeopardized by many different factors. Although the classification setup will be exactly as in the offline simulation, the subjects can now be influenced by the feedback, which might put them into a different psychological state. Motivation and task involvement as well as frustration in periods of low performance can hardly be simulated in offline measurements.

A further problem for the transfer of the CONCAT classifier to an online environment might be the fact that CONCAT is only trained on calibration data, while it has been shown that there can be a substantial shift of the features when going from online to offline data. Therefore, one can expect that a bias adaptation will be necessary for some of the subjects, as it was suggested in Section 5.1.

The superior method from the last section, CONCAT, is now tested against the standard approach where spatial filters and classifiers are trained anew on the calibration data of a new session.

The study is presented in the following order: In Section 6.3.2, I introduce an experimental setting that allows for the comparison of our CONCAT approach and the ordinary approach including calibration. In Section 6.3.4, I show the results of this comparison, discuss our findings (Section 6.3.5) and put them into perspective.

6.3.1. Features and Classification

The online experiments will be performed analogously to the methods presented in Section 6.2. Therefore, the classification will rely on the discrimination of imagined hand and foot movements, and spatial filters will be required to extract the most discriminative signals from the EEG signal. Here I will describe generally, how spatial filters are used to calculate

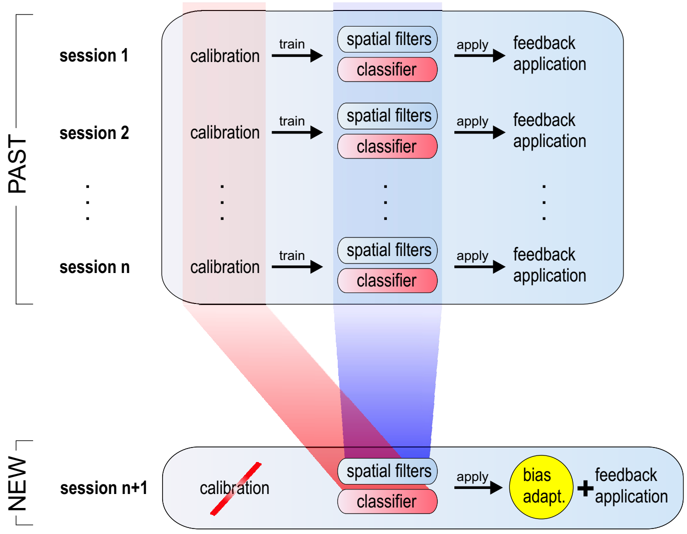


Figure 6.12.: Scheme of the CONCAT training procedure. Session 1 to session n shows a standard BCI procedure: spatial filter and classifiers are learned anew from a calibration recording (e.g. with CSP and LDA) before they are applied during a feedback application at the same day. The new CONCAT method eliminates the calibration recording: spatial filters and a classifer are predetermined before session $n+1$ starts. The spatial filters for session $n+1$ are extracted from old spatial filters (blue), the classifier for session $n+1$ is calculated from old calibration recordings (red). The feedback application of session $n+1$ is preceded by a bias adaptation (yellow).

features for classification, and how the ongoing EEG is translated into a control signal. This method applies to both classical CSP and the proposed method.

The EEG signals of the calibration measurement are band-pass filtered (subject-specific frequency band, see Section 6.3.2 and Table 6.5) and spatially filtered with the selected CSP filters. From these signals the log-variance is calculated in each trial of the calibration data (interval is selected subject-specfically, typically 750 to 3500 ms relative to the presentation of the visual cue). This procedure results in a feature vector with dimensionality equal to the number of selected CSP filters (which was in this study 6 for classical CSP and 12 for the proposed method, see Section 6.3.3). For classification least squares regression (LSR) was used.

For online operation, features are calculated in the same way every 40 ms from the most recent segment of EEG (sliding windows of 1000 ms width). CSP filters calculated from the initial calibration measurement are not adapted during online operation. Nevertheless, the system allows stable performance even for several hours ([94, 17]). But for optimal feedback the bias of the classifier might need to be adjusted for feedback. Since the mental state of the user is very much different during the feedback phase compared to the calibration phase, also the non-task-related brain activity differs. For a thorough investigation of this issue cf. [71, 126, 69], or see Section 5.1 of this work. With regard to this study, the issue is discussed in Section 6.3.2.

	#chan-	#past	#train		Frequency band		Interval	
Subject	nels	sessions	trials	Classes	(CSP)	(CONCAT)	(CSP)	(CONCAT)
zq	46	7	845	LR	[9 14]	[9 25]	[810 4460]	[500 3000]
ay	46	4	324	LR	[8 22]	[9 25]	[710 2650]	[500 3000]
zp	46	5	704	LR	[10 25]	[9 25]	[2750 5000]	[500 3000]
al	44	9	684	FR	[11 25]	[9 25]	[1600 4690]	[500 3000]
aw	44	13	1075	LF	[11 17]	[10 25]	[1500 4500]	[500 3000]
zk	46	7	240	LR	[8 31]	[9 25]	[920 4390]	[500 3000]

Table 6.5.: Subject-specific parameters. The first until third column report the number of sensors and sessions, as well as the number of trials per class which were available in total from these previous sessions. The fourth column indicates the two motor imagery classes that have been used (L: left hand, R: right hand; F: right foot). The frequency band for CSP analysis was chosen for each subject individually. For original CSP (column 5) it was chosen on data of the actual session. For CONCAT (column 6) it was chosen on previously available sessions. The same holds for the time window used for the training of the classifier, denoted in milliseconds after stimulus presentation: for CSP (column 7), the window was optimized on the training data, while for CONCAT, a fixed window was used for all subjects.

6.3.2. Experimental Setup

To demonstrate the feasibility of the CONCAT approach, a BCI feedback study was designed to compare the proposed approach with the classical CSP approach in terms of feedback performance. The specific construction of the two classification setups is described in Section 6.3.3.

The BCI experiments were performed with 6 healthy subjects, 5 male and one female, aged 26–41. These were all the subjects who had performed at least 4 BCI sessions before with the Berlin Brain-Computer Interface (BBCI). The large amount of past experimental data is a prerequisite for the extraction of prototypical CSP filters as described in Section 6.2.2, since the cluster density in the CSP filter space can only be estimated with a sufficient number of sample points.

The feedback consisted of the visual presentation of a computer cursor which was controlled by the output of one of two different classifiers. The first three feedback runs were done with the pre-computed CONCAT-classifier, see Section 6.3.3. After the completion of the third run, an ordinary CSP classifier was trained as described in Section 6.3.3, and in the next 8 runs, either the CONCAT or the ordinary CSP classifier was used for feedback; the order was randomly chosen and unknown to the subject. Due to the high impact that a modulation of the oscillatory activity in the visual cortex can have on the classification of bandpower-based classifiers (see Chapter 5), I enforced a difference in the visual workload by switching from ordinary “Fixed-Duration” cursor control (blocks I–II and IV) to “Fixed-Duration” Feedback of Results (blocks III and V), where the cursor was invisible (see Section 2.3.5 for details).

The EEG data were bandpass-filtered to a subject-specific frequency band (see Table 6.5), and spatial filters, as described in Section 6.2.2 and Section 2.2.1, were applied. Finally, the

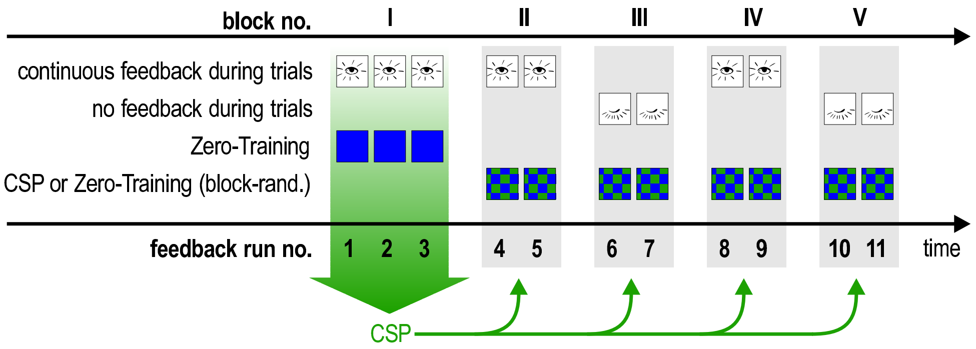

block no.
I
II
III
IV
V
continuous feedback during trials
no feedback during trials
Zero-Training
CSP or Zero-Training (block-rand.)
feedback run no.
1
2
3
4
5
6
7
8
9
10
11
time
CSP

Figure 6.13.: This figure shows a schematic overview of the applied paradigms for each run of the feedback session. Block I, II and IV were conducted with regular Cursor Control feedback (with a fixed duration), whereas the cursor was invisible during blocks III and V. In block I, the predefined classifier was applied, and the sequential order of the classifiers (either regular CSP or CONCAT) was randomized for block II–V. CSP was trained using the data of block I.

band power of the spatially and temporally filtered signals was estimated by calculating the logarithm of the squared sum of the filter outputs. These features were fed into a linear classifier. I used least squares regression (LSR), in order to force the classwise mean of the linear classifier output to be +1 and -1, respectively. Details on LSR-classifiers are given in Section 2.2.2.

At a rate of 25 Hz, graded classifier outputs were calculated for the last 1000 ms, and averaged over 8 samples. A scalar factor was multiplied to the result, and finally a real-valued bias term was added.

Guided by our experience with nonstationary bias, a bias adaptation was performed at the beginning of every run. Therefore, the subject controlled the cursor for 20 trials (10 per class), and the bias was adapted at the end of this period. The procedure corresponds to the initial calibration of the bias as presented in Section 5.1. In the following 100 trials (50 per class), the subject received feedback in a "Cursor Control" feedback application.

6.3.3. Construction of Classifiers

Here I will describe the determination of the spatial filters and classifier for the proposed approach and the calculation of filters and classifier for the classical CSP approach on data recorded at the beginning of the session. The feedback performance of these two approaches is compared using the experimental design described in Section 6.3.2 and results are reported in Section 6.3.4. Most of these settings are chosen as straightforward consequences from the offline analysis presented in Section 6.2.

The Zero-Training Filters and Classifier

The clustering approach for prototypical CSP filters relies on the same distance function and training procedure as presented in Section 6.2: spatial filters are clustered according to their non-euclidean distance in the parameter space, and cluster centers are chosen as representatives for especially stable filters.

For each subject, data from a number of past sessions (past data) has been available (see Table 6.5). Based on this data, a set of spatial filters and the CONCAT classifier was constructed individually for each subject. This preparation could take place days before the planned feedback experiment, as only historic data is involved for the construction of CONCAT. For every subject, I performed the following:

I first calculated for each class the three filters with the three largest eigenvalues for each historic session of the subject using the CSP algorithm from Section 2.2.1. Those top three filters of both classes and all past sessions of that subject, amounting to 6 prototype filters (Section 6.2.2), constituted the first 6 dimensions of the feature space. In addition to these prototypical filters, I also pooled all the data from past experiments of that subject and calculated ordinary CSP filters on this collection of historic data sets. The resulting filters (3 per class) were appended to the 6 prototype filters. Filtering the EEG data of the pooled data set (all past sessions of the subject) resulted in a 12-dimensional feature space. Finally, a linear classifier was calculated on the features using Least Squares Regression (LSR).

The Ordinary CSP Filters and Classifier

For each subject, I also built a set of ordinary CSP filters and a corresponding classifier. In contrast to the CONCAT solution, this setup can not be prepared beforehand. The construction is done on the fly during a new experimental session and does not involve data from past sessions. This corresponds to the standard classification scenario as presented in Section 2.3, and will be refered to as CSP in the following.

For the training of this regular CSP classifier, I first recorded three runs of feedback data (with feedback provided by the output of the CONCAT-classifier), totalling to more than 150 trials per class. According to the cross-validation error on this data, the optimal frequency band was selected, as well as some additional parameters like length and starting point of the training time interval for estimating the band power. The Common Spatial Patterns were computed on this data and two spatial filters were chosen for each class. These parameters were chosen as described in Section 2.3. Then a linear classifier (LSR) was trained using filtered data from the first three runs.

6.3.4. Results

Feedback Performance

The first three runs of feedback showed that all subjects under study were able to operate the BCI with the pre-computed classifier at a high accuracy, where only 10 trials per class from the current day were required to update the classification scenario. Fig. 6.15 shows, for each subject, the percentage of successful (“hit”) trials from each run. After the third run, the subjects could not know in advance, which one of the two classifiers was used for the generation of the feedback.

For subjects *zq*, *al* and *zk*, the CSP feedback performed better than the CONCAT feedback. In *ay* and *aw*, the feedback performance on the four blocks is very similar with both classifiers, whereas in subject *zp*, the CONCAT feedback even outperformed the CSP feedback. Note that if the initial three runs are further taken into account for a more exact estimation

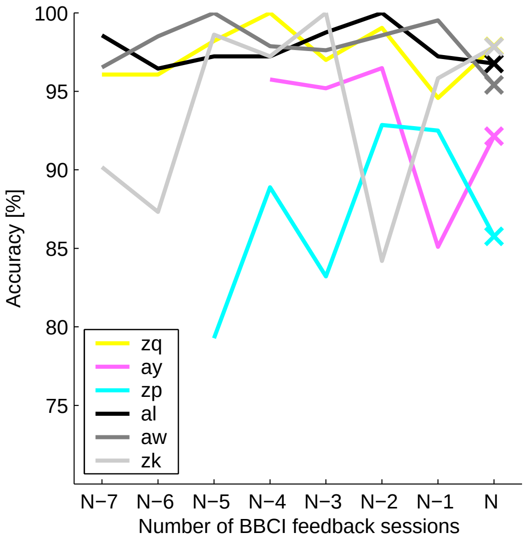


Figure 6.14.: The discriminability of the calibration data for each previous session ($N-7,\dots,N-1$) as calculated by the cross-validation error of the CSP algorithm. Frequency band and time window were specifically optimized for each session and each subject. The cross-validation error on session N is calculated on the three runs from block I, with the settings from table 6.5.

of the feedback performance of CONCAT, Subject *zq*'s performance with CONCAT can no longer be found to be inferior to the performance with CSP.

The performance over all subjects is shown in Fig. 6.16, where the feedback performance in each run of the four blocks is collected in a single boxplot for each classifier. The CSP performance is slightly higher on average, although this difference is not significant: a Wilcoxon ranking test was performed, at a significance level of $p = 0.05$.

Adaptation of Classifier Bias

The bias was updated at the beginning of every run. I can now check if this update was necessary for the accuracy of the classifiers. For run i and classifier j and movement class k, let μ_{ijk} be the mean of the classifier output of the corresponding 50 trials. Then the value $\hat{b}_{ij} := \frac{b_{ij}}{\mu_{ij2}-\mu_{ij1}}$ relates the optimal bias b_{ij} for run i and classifier j with the actual distance between the class means. A value of 1 would correspond to shifting the decision boundary by the entire inter-means distance. The results of this calculation are shown in Fig. 6.17. For most subjects, the required shift is moderate ($\hat{b}_{ij} < 0.5$), but for subjects *zp* and *zk*, the CONCAT classifier requires a strong update of the bias, since the absolute values exceed 1. The CSP classifier, trained on data from the same day, is not as susceptible to bias shift as the CONCAT classifier, since the change is comparatively small also for these two subjects. This finding supports the hypothesis from Section 5.1 that a bias-shift is required for classifiers that are trained on calibration data without visual feedback (such as the CONCAT-classifier), whereas the shift *within* the session is comparatively smaller. The latter is the case for the CSP-classifier which is trained on online BCI data with visual feedback.

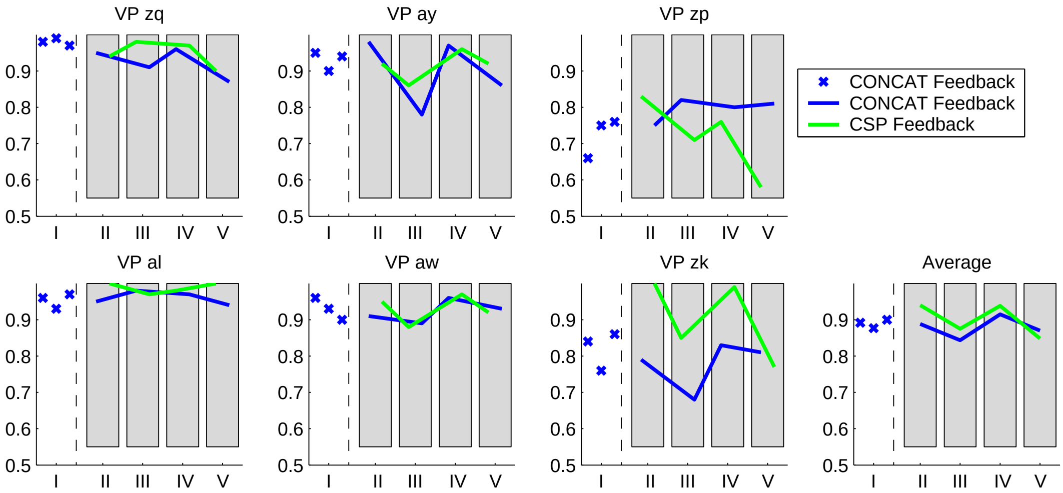


Figure 6.15.: The feedback results for each of the six subjects. The feedback accuracy is denoted for the 100 trials of each run. The initial three runs, here marked as "I", were done with the CONCAT classifier, and in the following the order of the classifiers was randomly permuted in each block of two runs, here denoted as "II–V". The shift of the blue curve relative to the green curve within the shaded areas indicates the order of the classifiers within the block.

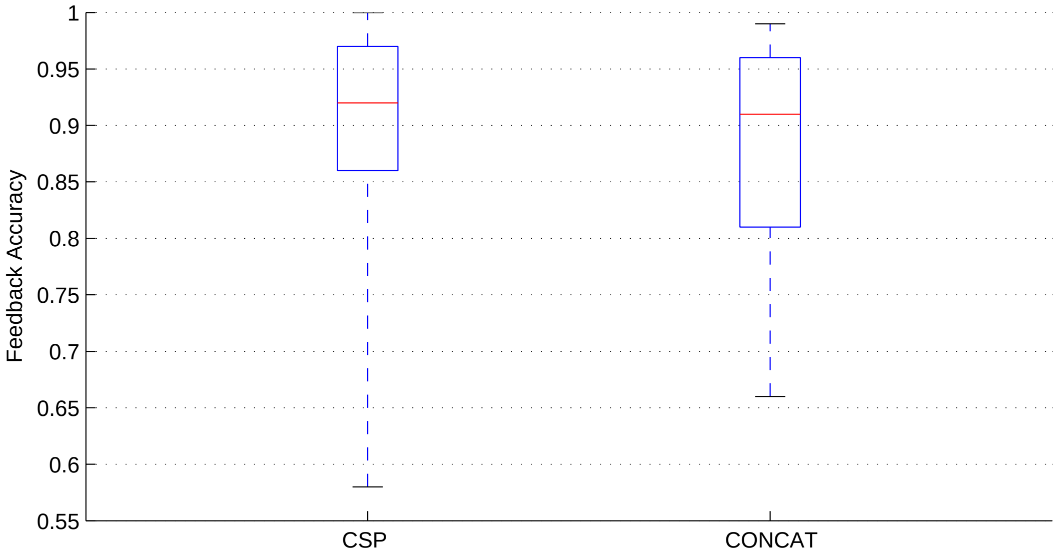


Figure 6.16.: This figure shows the feedback performance of the CSP and the CONCAT classifier over all subjects. The median of the CSP feedback accuracy is slightly higher. This difference is not significant (Wilcoxon ranking test, $p < 0.05$).

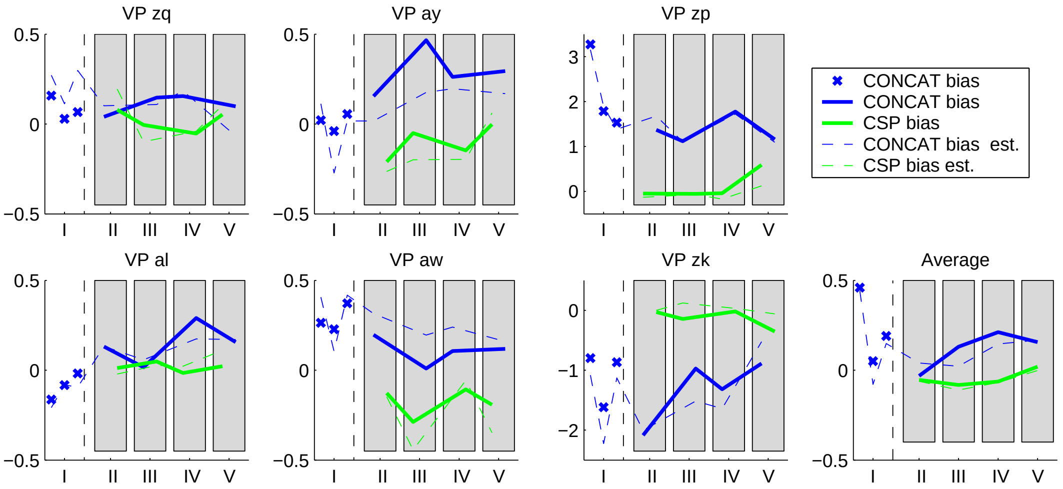


Figure 6.17.: At the beginning of each run, the bias for the classifier was adapted using 10 trials per movement imagination class. The plot shows the optimal bias update, as calculated on the following 100 trials. This value is normalized by the difference between the classifier output class means. The solid lines show the optimal bias for CSP (green) and CONCAT (blue) classifier separately. The dashed lines indicate the bias, as it was actually calculated on the initial 20 trials by the adaptation procedure during the feedback.

Besides the check for necessity of the bias update, Fig. 6.17 also provides a comparison of the "optimal" bias with the actual bias, both calculated with the same normalization. The dashed lines indicate the bias, as it was computed on the initial 20 trials during the feedback. From this figure, it is evident that the estimated and the optimal bias coincide quite well. Although the estimation error is sometimes not neglectable (as for subjects *aw* and *zk*),the dashed and the corresponding solid lines are highly correlated. If the classifier would not have been adapted (corresponding to setting the bias to 0 in Fig. 6.17), the error would be larger in nearly all runs than with the proposed adaptation strategy. This proves that the update procedure is in fact stable and useful in combination with the CONCAT-classifier.

Fig. 6.18 exemplifies the effect of the bias shift for subject *zp*. In the left part, the classifiers are calculated for each of the 1100 trials of the feedback, without adding any bias term. While CSP classification (on the x-axis) shows a good separability of the data into positive and negative values (for right hand and left hand movement, respectively), the CONCAT classifier assigns negative values to almost every point, resulting in a poor classification rate (near 50%, corresponding to chance level accuracy). This effect can be alleviated by estimating the bias on the 20 initial trials that were performed previous to every run. The right part of the figure shows the result: both CSP and CONCAT classification rate now are comparable. Note that an improvement of classification accuracy by bias adaptation was highly significant for two subjects.

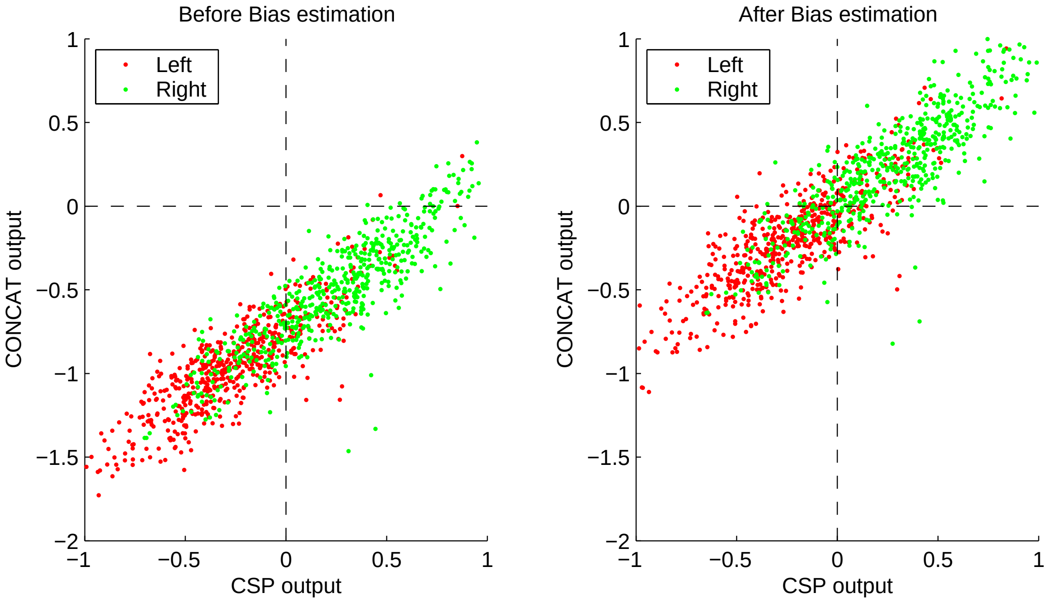


Figure 6.18.: This figure shows the effect of the bias estimation for subject *zp*. In the left part of the figure, both CONCAT and the CSP classifier are computed on the 1100 trials of the feedback session, without adding a bias term. While the CSP method performs already quite well, the output of CONCAT (on the y-axis) is negative for almost all samples, which would correspond to a classification error near 50%. The right part of the figure shows the output on the same trials, after an initial bias adaptation on the 20 initial trials per run. For the CSP classification, the bias is not changing the result significantly, but CONCAT clearly profits from the bias update.

Discriminabilty owed to Each Prototype Filter

Here I investigate each prototype CSP filter with respect to the discriminability of the corresponding log variance feature and relate it to its γ-index, see Section 6.2.2. For the evaluation of the discriminability of each features, I use as measure the area under the ROC-curve (AUC, see e.g. [42] and Section 3.1). This value is 0.5 for features that are uncorrelated with the class affiliation and 1 for features that are perfectly separable. I regarded the γ-index, calculated on the previous sessions, as a quality prediction for the performance of the feature in the online application of the classifier. Fig. 6.19 confirms this hypothesis by showing that there is in fact a strong negative correlation between the γ-index and the AUC-value of the features. The higher the density of the CSP filters, accumulated over many sessions, at a particular point, the higher the discriminability of the corresponding log variance feature in the current online session. Note that below a γ-value of 0.7, only features of the three subjects with the overall highest feedback performances (subjects *al*, *zq* and *aw*) can be found. These features, on the other hand, have the highest AUC-values.

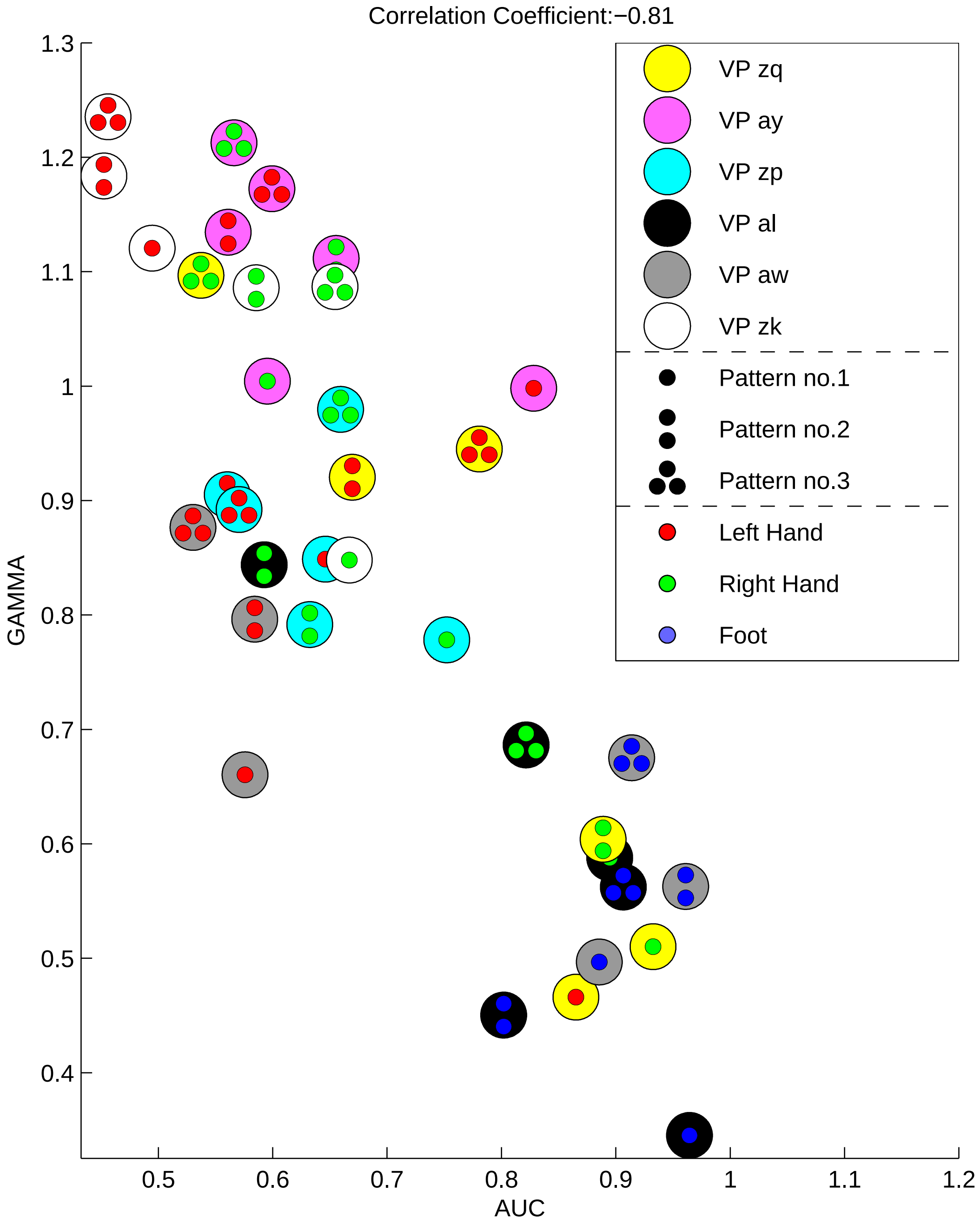


Figure 6.19.: This figure compares the γ-index of a prototypical CSP filter, as calculated on previous sessions, with the discriminability of this feature in the feedback session. The filters with the lowest γ-index have the highest performance. This correlation is highly significant ($p < 0.01$).

6.3.5. Discussion

The final validation of BCI algorithms can only be provided in online experiments. However, in contrast to offline evaluation, only one classifier can be applied to the same data set. This makes a comparison especially hard, since the differences between different data sets (high inter-subject and inter-session variability) add to the variability of the performance. It is therefore required to record all data sets under similar conditions. All data for one subject was recorded on the same day, which clearly limits the possible number of runs that could be performed. I evaluated the performance of this new classifier with the standard CSP method that is used for the classification of band power features in imaginary movements ([18]). In order to keep the subjects equally motivated under both conditions, they were not informed which classifier was used for which part of the experiment and instructed them to keep trying to hit the ordered targets on the screen, irrespective of the possibly degraded performance of the current classifier.

The aim of this study was to construct and evaluate a classification method that can be applied without a lengthy calibration measurement. While the features I chose have proven to be quite discriminative for the classification task at hand, the bias adaptation was indispensable for two of the six subjects (and did not degrade the performance for the other subjects). Possible explanations for the shift of the bias from one session to another include the differences in electrode impedances as well as physiological effects like superimposed occipital α-rhythm, see Chapter 5 and [126, 71, 69]. The number of trials per class that are initially used for the adaptation period has to be chosen according to a trade-off between the total duration of the adaptation period and the precision of the estimation. After preliminary off-line evaluations I found 10 trials per class to be a quite balanced choice. Note that this number might as well be adjusted according to the predicted feedback accuracy for the subject. Bias parameter estimation is clearly expected to degrade with a more variable feedback discriminability during the adaptation period, and the presented findings support this expectation. Therefore, if a low feedback performance for a subject can be expected, one could easily increase the number of trials used for adaptation. It is on the other hand desirable to keep the total duration of the adaptation period very short, since the goal is to operate real-world BCI applications right from the start, where knowledge of class labels is not available and even the equality of class distributions are not always reasonable assumptions.

In this study, the training data for the CSP-classifier are different from the usual calibration data: in the normal case, no feedback is given during the presentation of stimuli. Also, the visual scene now resembles more closely the feedback setup (see Section 2.3), i.e., the targets are on the left and right side of the screen and change the color to indicate the next movement task. Although one might suspect that this could degrade the classification performance of the CSP classifier due to the higher complexity of the mental task, this is not the case. Fig. 6.14 shows the development of the cross-validation error over the previous experiments for each subject. Parameters like the frequency band and the time interval were subject-specifically optimized in each session. The last point (session N) denotes the experiment from this study, where the first three runs were taken into account. This corresponds to the data on which the CSP classifier was trained. The cross-validation performance for this session is of the same magnitude as the previous performance and hence does not reveal a systematic disadvantage for the CSP method. On the contrary, the following application of the classifier might even benefit from the fact that the task difference between the training data and the test data is relatively small.

For the training of the CONCAT classifier, some of the parameters were not specifically optimized, such as the frequency band, the training window for parameter estimation on the previous sessions, and the movement type combination used for the feedback. The settings that were applied here were fixed beforehand. It has been shown in recent publications [38, 22], that the optimization of spatial and temporal parameters can result in significantly better classification accuracy. Therefore, selecting these highly subject-dependent parameters on the same day's training data for the CSP classifier may have resulted in a slight advantage for this method, but I decided for the optimization in order to have the best possible classifier as a comparison.

Only in subject *zk*, the CSP classifier clearly outperforms the CONCAT classifier. The reason might be due to the amount of training data which was present from previous sessions: while the training sessions for all other subjects contained more than 100 trials per class, only 35 trials per class and session were recorded for subject *zk*, see also table 6.5. This led to a higher variability in the collection of CSP filters; it also explains the low γ-index for all features of subject *zk*, see Fig. 6.19.

For subject *zk*, the γ-values for the CONCAT-features are slightly higher than for subject *zp*. From the feedback performance in Fig. 6.15, one can even see a slow positive trend for the CONCAT classifier throughout the day. The trend in the performance for the CSP classifier, on the other hand, is degrading over time. Subject *zp* reported that she was trying to control the feedback with different strategies over time, always switching to the mental imagery that seemed most reliable at each point in time. This variability in the mental strategies, induced by the feedback presentation, is reflected in the brain signals. Fig. 6.20 shows the evolution of the scalp topographies related to the discriminability of the band power features in each electrode. I calculated the band power features for the 100 feedback trials in each run and calculated the r^2-values between left and right hand imagery class, as a measure of linear discriminability. The figure shows that towards the end of the session, the features on the right motor cortex are more discriminative than the features initially on the left motor cortex. The feedback performance of the CSP classifier appears to be more susceptible to this shift, while the CONCAT classifier is based on a broader basis of spatial filters, which can account for the variability in the signals. A possible remedy for the degrading performance is the adaptive estimation of the linear hyperplane of the classifiers, [71, 146]. Using an adaptation period as short as 10 trials per class, however, the adaptation of the hyperplane for CONCAT fails for almost every subject, as an offline evaluation on the given shows. This is mainly due to the fact that for a linear classifier, the number of parameters to be estimated grows quadratically with the number of feature dimensions. Since the CONCAT feature space has 12 dimensions (6 “prototype” filters and 6 “CSP” filters), 20 trials are too little data. Similar results have been shown in Section 5.2 (see also [71]) for classical CSP; the suggested bias update requires only the estimation of one single parameter and is therefore more robust. If, however, the feature discrimination performance is changing over time like in subject *zp*, this bias update might not be sufficient any more. Other options, like a continuous adaptation of the bias throughout the feedback run, require at least the a posteriori knowledge of all the labels of this run, which can not be granted in all feedback applications. Moreover, in Chapter 5 (see also [126]), this adaptation scheme did not prove to be superior to the initial adaptation of the bias.

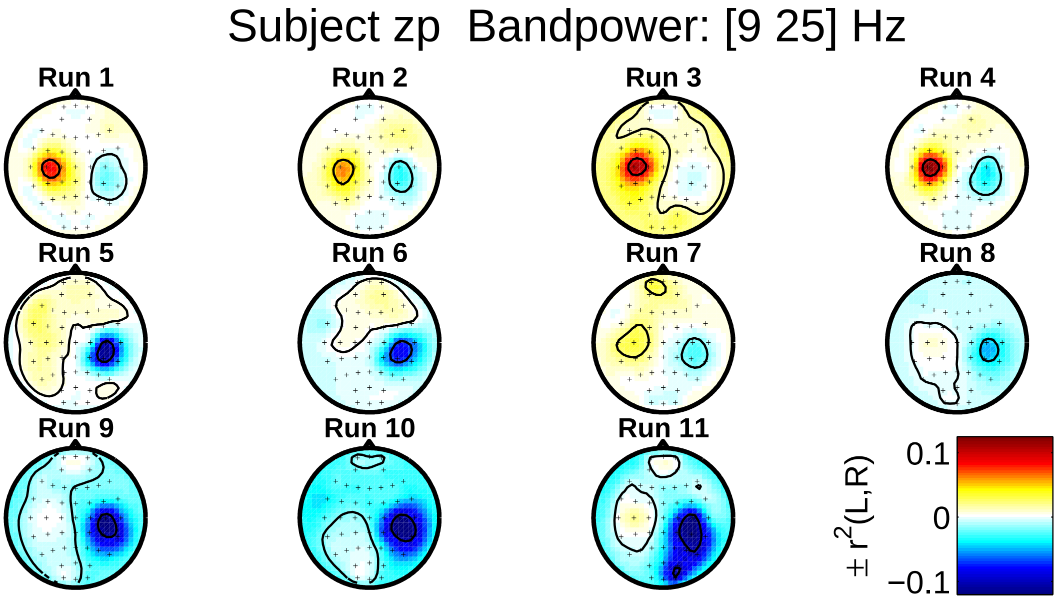


Figure 6.20.: For each feedback run of the session, this figure shows the scalp topographies of class discriminability on band power features for subject *zp*. After bandpass filtering to the frequency band of 10–25 Hz, the log-bandpower was calculated for each electrode in the window 500–3000 ms after the presentation of the stimulus. Finally, signed r^2-values were calculated as a measure of class discriminability.

Fig. 6.19 suggests a good prediction accuracy for prototypical CSP filters with a low γ-index. However, since the features of some subjects (e.g. *zk* and *zp*) appear to form distinct clusters for each class, one should consider some reasonable normalization between these values. The γ-index, as formulated above, depends mainly on the number of dimensions and on the number of samples, since if the number of dimensions (in this case: the number of electrodes) is fixed, the maximally possible γ-index is a monotonic decreasing function in the number of samples. Not only the maximal, but also the expected minimal γ-index under randomly drawn samples will differ. Therefore, I estimated this value by a simulation: the number of dimensions and samples were chosen for every subject according to Table 6.5. The minimal γ-value was calculated and averaged over 1000 repetitions. The results are displayed in Table 6.6. Since the values range from 1.12 for subject *aw* to 1.22 for subject *ay*, the correlation found in Fig. 6.19 is not influenced if each γ-value is normalized by the expected minimal γ-value. Note that for subjects *zk* and *ay*, some of the γ-values are close to 1 after normalization; this corresponds to a “cluster” density which is expected to occur even in random samples. These features, in turn, have very low AUC-values.

With respect to the cumbersome electrode preparation great advancements could be achieved in the meantime. In [117], a novel dry EEG recording technology was presented which does not need preparation with conductive gel. In the reported study with good BCI subjects, feedback performance was comparable to the approach with conventional EEG caps for most subjects. Note that this system only uses 6 electrodes and can thus be miniaturized to

Subject	Expected Minimal γ
zq	1.17 ± 0.02
ay	1.22 ± 0.02
zp	1.20 ± 0.02
al	1.15± 0.02
aw	1.12± 0.02
zk	1.17 ± 0.02

Table 6.6.: This table shows the minimal γ-index for a collection of randomly drawn points, together with the standard deviation. For this calculation, the same dimensionality (corresponding to the number of electrodes) and the same number of points (corresponding to three times the number of experiments) was used.

run with a tiny EEG amplifier and a pocket PC.

This study has successfully transfered the results obtained in Section 6.2 to the online scenario. For the majority of subjects, the new classifier performed with a similar accuracy like the standard machine learning approach which was trained on three runs of feedback data from the same day. The theoretical considerations concerning the distance measure in the space of spatial filters were justified with this promising result. By analyzing the amount of variability from session to session, I have introduced a new method which completely overcomes the tedious calibration period. Especially in the case of paralyzed or completely locked-in patients, who rely on communication devices on a daily basis, this method is particularly appealing, since it lets the subjects initiate the communication right away.

The study also revealed that for some of the subjects, the bias had to undergo substantial adaptation. This was not surprising, since the findings of Chapter 5 already suggested that the output of classifiers trained on calibration data often needs a shift during the feedback period. The method of an initial bias adaptation, which was also developed in that chapter, proved to be extremely effective, since it decreased the error for the bias substantially.

After the analysis of the degree of nonstationarity across sessions, the presented approach is the successful combination of methods which account for this nonstationary behaviour. The result is a single method, which not only shows a stable performance throughout an entire session, but also requires minimal calibration time for the next session.

7. Conclusion and Outlook

In this work, I have presented a new framework for the assessment of nonstationarity. The concept of time series with probability distributions which change over time can be found in many research fields where data are measured with sensors (such as audio and video data analysis, speech recognition, biomedical or meteorological data), but it can also occur in more abstract processes such as stock market rates or network traffic analysis. In all these cases, nonstationarity can lead to serious problems if methods for signal processing or classification are applied to the data under the hypothesis of stationary distributions. I have applied the presented framework to the field of EEG data. In this scenario, I could demonstrate the power of these methods by visualizing and interpreting the data.

A variety of visualization tools was introduced in Chapters 3 and 5 for the differences between the brain signals of two distinct time intervals. These tools have found their way into the BCI research community: they were first presented in [126] and later adapted in [144]. By application to data from online BCI feedback experiments, I could show that a source for nonstationarity on many timescales is the modulation of occipital alpha during different states of visual input (see Chapter 5 and Section 6.1). This is an unprecedented discovery with the methods of applied machine learning and points out the impact of the shift on the classification performance. In this sense, I have exemplified that the analysis of nonstationarity in a machine learning context can also lead to neurophysiological insights.

Once the reasons for the change of the distributions over time are known, it makes sense to consider remedies against their influence on the classification performance. I have suggested various methods for adapting the classifiers over the course of an experiment, and have shown that they can be readily applied in online experiments. The key ingredient, a bias adaptation, is a very robust method and also turned out to be an important prerequisite for the transfer of classifiers across sessions. However, the discovery that bandpower features can actually undergo a shift within a single experimental session has led to a series of publications which suggest other means of adaptation for this scenario ([132, 133, 16]).

With the same approach, namely with an analysis of the variability of the optimal parameters, I developed and implemented a new method which reduces the calibration period of usually 20–40 minutes substantially (see Chapter 6). After attaching the electrodes, subjects can immediately receive feedback and use BCI applications at high information transfer rates. In the same spirit as [117], where a method is presented to overcome the need for transductive gel for EEG measurements (“dry electrode cap”), this method enables longterm BCI users to start BCI sessions with almost no preparation time. For daily applications, this is a crucial requirement and will help in the realization of BCI devices for severely disabled users. The development of this novel approach has paved the way for revolutionizing modern rehabilitation for the disabled. The applicability of devices of this kind makes it also attractive for healthy users to use BCIs as additional input channel for man-machine interaction. Computer games and the direct control of machines can only be useful and applicable, if the calibration time of the devices is reduced to a minimum, while preserving maximal precision. Combined with a “dry” cap, my development is a large step towards this goal. This achievement as well as other work (see [9, 10, 11, 12, 13, 14, 17, 18, 39, 38, 41, 61, 65, 66, 67, 68, 69, 70, 90, 95, 96, 97, 126, 132, 133]), has contributed to the BBCI’s international success. Note that the method is by no means limited to its application in BCI, despite the potential it shows in this field. It can be regarded as a general tool for machine learning and signal processing.

Future research will have to transfer the tools provided in this thesis to other scenarios, such as the transfer of classifier parameters from subject to subject. Although the variability across subjects can easily be regarded within the same framework as the variability from session to session, it is out of the scope of this work. However, with this approach, BCI research can be conceivable for a wider range of applications, by reducing the calibration time for naive subjects, such as it has been introduced in this work for longterm BCI users. It is, moreover, not only a straight-forward, but also highly promising idea to apply these methods to other neurophysiological paradigms or multi-class applications.

Apart from the question of robustification for BCI, it is a task with high potential to apply these methods to other areas where machine learning methods are affected by the nonstationarity in the data. Future research should strive for the robustification of general time series, in order to make machine learning applications more usable.

A. Appendix

A.1. Delta

The δ-index of a point in a given data set is a measure for its outlierness, as it was used in Section 4.1.

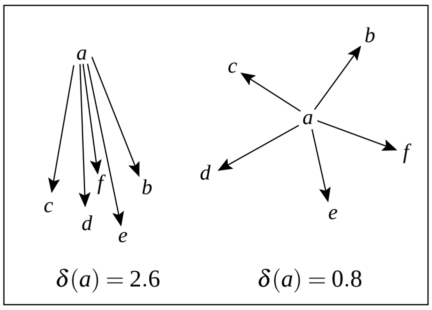


Figure A.1.: In the left example a is an outlier an thus its δ index is large. In the right example it is part of a larger group so its δ index is small. Both examples assume $k = 5$.

Consider n data points $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$ in d-dimensional space with the euclidean norm $||x|| = \sqrt{x^\top x}$. We denote the k nearest neighbors of $x \in \mathbb{R}^d$ among the given set by

$$\mathrm{nn}_1(x), \dots, \mathrm{nn}_k(x) \in \{x_1, \dots, x_n\} \subset \mathbb{R}^d.$$

The outlier index $\delta(x)$ is defined to be the length of the mean of the vectors pointing from x to its k nearest neighbors, i.e.,

$$\delta(x) = ||\frac{1}{k} \sum_{j=1}^{k} (x - \mathrm{nn}_j(x))||.$$

As shown in Figure A.1, δ is large if the neighbors are all in the same direction, which is usually the case for outliers.

A.2. Gamma

If the data under study are taken from an arbitrary metric space, it is not granted that an addition operation is defined for this space. This means that an outlier index can not be defined according to the definition of δ, since this requires subtraction, addition and scalar multiplication to be defined. In the following definition of the γ-index, this problem is solved by applying the averaging *after* the application of the metric.

Let $\mathbf{w} \in (\mathscr{S}, m)$ be a point in an arbitrary metric space $\mathscr{S}$, and let $\mathrm{nn}_1(\mathbf{w}), \dots, \mathrm{nn}_k(\mathbf{w})$ be the k nearest neighbors of $\mathbf{w}$, according to the metric m. The average distance of $\mathbf{w}$ to its neighbors is then called the γ-index of $\mathbf{w}$, i.e.

$$\gamma(\mathbf{w}) = \frac{1}{k} \sum_{i=1}^{k} m(\mathbf{w}, \mathrm{nn}_i(\mathbf{w})).$$

In this form, the γ index was applied to the space of CSP filters, which has an inherently non-euclidean metric (see Section 6.2).

A.3. Adaptation: Implementation details

A.3.1. The BBCI software package

The Berlin Brain-Computer Interface is an inter-coordinated package of hardware and software solutions, designed to meet a large variety of requirements for brain-computer interfacing. Its implementation is specifically tailored for modularity, i.e., the components can be modified and replaced without losing functionality. I will give an overview in the following, but for a more detailed description, see [34].

Most of the BBCI online toolbox is written in MATLAB [92], since this allows for a fast and intuitive modification of the signal processing and classification routines involved. This requirement is crucial for the ongoing experimental research in the BBCI project. Since, on the other hand, the graphical output of MATLAB is not optimized for realtime applications, the online toolbox was divided into several parts which communicate via the network-protocols TCP [148] and UDP [149], to distribute the workload on different processors. This modular setup even makes it possible to distribute the components to different machines, connected over local area network or internet.

The single parts of the toolbox can be grouped into mainly four components:

1. Acquisition: The EEG data are recorded with a BrainVision Recorder, obtained from the company Brain Products GmbH. The included software also provides a TCP-server, which makes the data available at a rate of 25 Hz, i.e., in blocks of 40 ms length. The data are given with the associated channel labels and with blocknumbers to avoid loss of data.

2. Signal processing and classification: This unit is the core of the BBCI toolbox, since it encompasses the routines which can be implemented using machine learning techniques. The data are first fetched from the TCP server (as described above), convoluted with spatial and temporal filters and then written into a buffer of appropriate length. The following feature extraction as well as the classification method depend strongly on the applied BCI paradigm and the pre-defined parameters. After application of the classifier, simple post-processing steps, such as the application of a scalar factor or a real-valued bias term, can be performed. The resulting output value is sent to the graphical feedback unit via UDP.

3. Graphical output: Again, the type of the presented feedback application depends on the BCI paradigm. In any case, the feedback unit will transform the classification

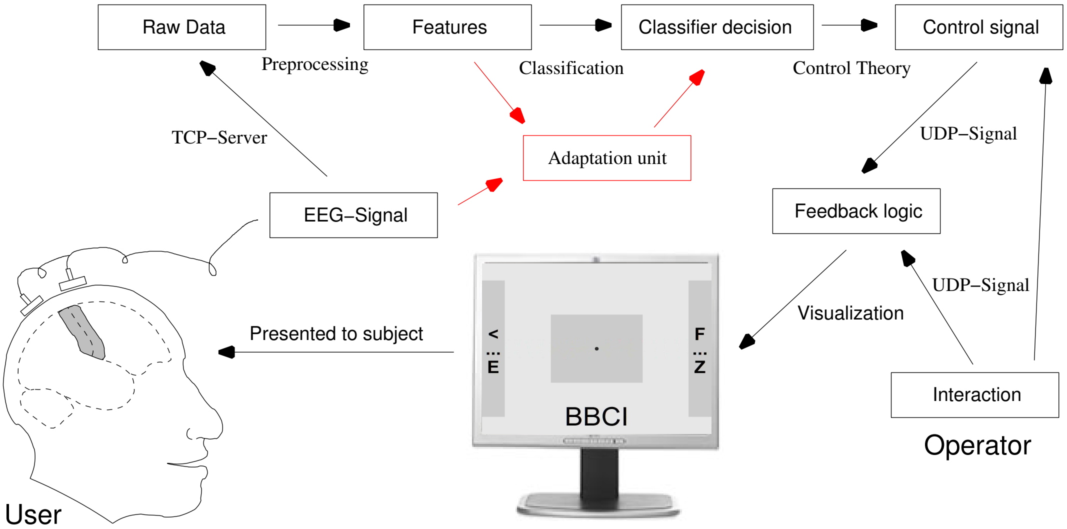


Figure A.2.: The figure shows an overview of the BBCI framework. The red part denotes modifications which were required for the implementation of adaptive classifiers.

values it receives into graphical events on a computer screen. In the case of a rate-controlled "cursor"-feedback, the incoming values are used to manipulate the horizontal position of a cursor on the screen; a positive value will move the cursor to the right, a negative value will move it to the left side of the screen.

4. Operator interaction: All parts of the feedback loop can be controlled by an operator. A graphical user interface (GUI) is provided which enables the operator to send control parameters to the classification unit and to the graphical unit.

A.3.2. The Adaptation unit

Figure A.2 demonstrates the interaction of the adaptation unit with the various other parts of the BBCI online toolbox. The demands for the adaptation unit were as follows:

1. Access to parameters, i.e., single parts of the classifier.
2. Possibility to exchange the entire classifier.
3. Receive control signals from feedback applications, e.g. beginning and end of adaptation periods.
4. Receive control signals from the GUI.
5. Display the exchanged parts of the classifier on the GUI, for control purposes on behalf of the experimenter.

Since one of the crucial requirements is the access to all classifier parameters, the adaptation was integrated into the classification unit of the BBCI online toolbox. In this fashion, the adaptation unit has full access to the entire MATLAB workspace which includes the loaded variables and the classification parameters.

In the same framework, the adaptation routine can read out the feature variables as well as the classification output values in the ongoing feedback presentation. An analysis of these values can result in a reasonable update of the parameters. By listening to marker signals which are accessible on the TCP server of the acquisition device, the adaptation routine is responsive to specific start and end triggers sent by the feedback routine.

For the communication with the GUI, a new UDP communication channel is established, enabling the adaptation routine to send control signals to the GUI, which can modify some of the values stored here. The GUI, on the other hand, is now equipped with a new thread which regularly checks for communication packets originating from the adaptation unit.

List of Figures

List of Tables

Bibliography

[1] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, New York, 3rd edition, 1994.

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

[3] M. Bensch, W. Rosenstiel, and M. Bogdan. Phase synchronisation in meg for brain-computer interfaces. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006*, pages 18–19. Verlag der Technischen Universität Graz, 2006.

[4] H. Berger. Über das Elektroenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 99(6):555–574, 1933.

[5] N. Birbaumer, T. Elbert, A. G. M. Canavan, and B. Rockstroh. Slow potentials of the cerebral cortex and behavior. *Physiological Review*, 70(1):1–41, 1990.

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

[7] G. E. Birch, P. D. Lawrence, and R. D. Hare. Single-trial processing of event-related potentials using outlier information. *IEEE Transactions on Biomedical Engineering*, 40(1):59–73, 1993.

[8] 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 (NIPS 01)*, volume 14, pages 157–164, 2002.

[9] B. Blankertz, G. Dornhege, M. Krauledat, V. Kunzmann, F. Losch, G. Curio, and K.-R. Müller. The berlin brain-computer interface: Machine-learning based detection of user specific brain states. In G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors, *Toward Brain-Computer Interfacing*, pages 85–101. MIT press, Cambridge, MA, 2007.

[10] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio. The Berlin Brain-Computer Interface: Report from the feedback sessions. Technical Report 1, Fraunhofer FIRST, 2005.

[11] 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. URL http://dx.doi.org/10.1016/j.neuroimage.2007.01.051.

[12] 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 Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):147–152, 2006. URL http://dx.doi.org/10.1109/TNSRE.2006.875557.

[13] B. Blankertz, G. Dornhege, M. Krauledat, M. Schröder, J. Williamson, R. Murray-Smith, and K.-R. Müller. The Berlin Brain-Computer Interface presents the novel mental typewriter Hex-o-Spell. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006*, pages 108–109. Verlag der Technischen Universität Graz, 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(6):581–607, 2006.

[15] B. Blankertz, G. Dornhege, C. Schäfer, R. Krepki, J. Kohlmorgen, K.-R. Müller, V. Kunzmann, F. Losch, and G. Curio. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):127–131, 2003. URL http://dx.doi.org/10.1109/TNSRE.2003.814456.

[16] 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 Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008. in press.

[17] B. Blankertz, M. Krauledat, G. Dornhege, J. Williamson, R. Murray-Smith, and K.-R. Müller. A note on brain actuated spelling with the Berlin Brain-Computer Interface. In C. Stephanidis, editor, *Universal Access in HCI, Part II, HCII 2007*, volume 4555 of *LNCS*, pages 759–768, Berlin Heidelberg, 2007. Springer.

[18] B. Blankertz, F. Losch, M. Krauledat, G. Dornhege, G. Curio, and K.-R. Müller. The Berlin Brain-Computer Interface: Accurate performance from first-session in BCI-naive subjects. *IEEE Transactions on Biomedical Engineering*, 2008. in press.

[19] B. Blankertz, K.-R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer. The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials. *IEEE Transactions on Biomedical Engineering*, 51(6):1044–1051, 2004. URL http://dx.doi.org/10.1109/TBME.2004.826692.

[20] B. Blankertz, K.-R. Müller, D. 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 approachs to actual BCI problems. *IEEE Transac-*

tions on Neural Systems and Rehabilitation Engineering, 14(2):153–159, 2006. URL `http://dx.doi.org/10.1109/TNSRE.2006.875642`.

[21] B. Blankertz, C. Schäfer, G. Dornhege, and G. Curio. Single trial detection of EEG error potentials: A tool for increasing BCI transmission rates. In *Artificial Neural Networks – ICANN 2002*, pages 1137–1143, 2002.

[22] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1):41–56, Jan. 2008. URL `http://dx.doi.org/10.1109/MSP.2008.4408441`.

[23] C. Brunner, R. Scherer, B. Graimann, G. Supp, and G. Pfurtscheller. Online control of a brain-computer interface using phase synchronization. *IEEE Transactions on Biomedical Engineering*, 53(12):2501–2506, 2006.

[24] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[25] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. *IEE Proceedings F*, 140(6):362–370, 1993.

[26] J. M. Carmena, M. A. Lebedev, R. E. Crist, J. E. O'Doherty, D. M. Santucci, D. F. Dimitrov, P. G. Patil, C. S. Henriquez, and M. A. Nicolelis. Learning to control a brain-machine interface for reaching and grasping by primates. *Public Library of Science Biology*, E42, 2003.

[27] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 2001.

[28] R. J. Croft and R. J. Barry. Removal of ocular artifact from the EEG: a review. *Neuropsychologie Clinique*, 30:5–19, 2000.

[29] R. Q. Cui, D. Huter, W. Lang, and L. Deecke. Neuroimage of voluntary movement: topography of the Bereitschaftspotential, a 64-channel DC current source density study. *Neuroimage*, 9(1):124–134, 1999.

[30] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

[31] E. Donchin, K. M. Spencer, and R. Wijesinghe. The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering*, 8(2):174–179, June 2000.

[32] J. P. Donoghue and J. N. Sanes. Motor areas of the cerebral cortex. *Journal of Clinical Neurophysiology*, 11:382–396, 1994.

[33] J. L. Doob. *Stochastic Processes*. John Wiley & Sons, 1953.

[34] G. Dornhege. *Increasing Information Transfer Rates for Brain-Computer Interfacing*. PhD thesis, University of Potsdam, 2006.

[35] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Combining features for BCI. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Inf. Proc. Systems (NIPS 02)*, volume 15, pages 1115–1122, 2003.

[36] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. *IEEE Transactions on Biomedical Engineering*, 51(6):993–1002, June 2004. URL `http://dx.doi.org/10.1109/TBME.2004.827088`.

[37] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Increase information transfer rates in BCI by CSP extension to multi-class. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 733–740. MIT Press, Cambridge, MA, 2004.

[38] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE Transactions on Biomedical Engineering*, 53(11):2274–2281, 2006. URL `http://dx.doi.org/10.1109/TBME.2006.883649`.

[39] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller. Optimizing spatio-temporal filters for improving brain-computer interfacing. In *Advances in Neural Inf. Proc. Systems (NIPS 05)*, volume 18, pages 315–322, Cambridge, MA, 2006. MIT Press.

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

[41] G. Dornhege, M. Krauledat, K.-R. Müller, and B. Blankertz. General signal processing and machine learning tools for BCI. In G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors, *Toward Brain-Computer Interfacing*, pages 207–233. MIT Press, Cambridge, MA, 2007.

[42] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley & Sons, 2nd edition edition, 2001.

[43] D. J. Edell, V. V. Toi, V. M. McNeil, and L. D. Clark. Factors influencing the biocompatibility of insertable silicon microshafts in cerebral cortex. *IEEE Transactions on Biomedical Engineering*, 39(6):635–643, 1992.

[44] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein. ERP components on reaction errors and their functional significance: a tutorial. *Biological Psychology*, 51(2-3):87–107, 2000.

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

[46] P. Ferrez and J. Millán. You are wrong! – automatic detection of interaction errors from brain waves. In *19th International Joint Conference on Artificial Intelligence*, pages 1413–1418, 2005.

[47] J. L. Fleiss. *Statistical Methods for Rates and Proportions*. Wiley & Sons, 2nd edition edition, 1981.

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

[49] A. Furdea. Toward an auditory P300 speller. Talk at NIPS 2006 workshop *Current Trends in Brain-Computer Interfacing*, 2006.

[50] S. Gonzalez Andino, R. Grave de Peralta Menendez, G. Thut, J. Millán, P. Morier, and T. Landis. Very high frequency oscillations (VHFO) as a predictor of movement intentions. *NeuroImage*, 32(1):170–179, 2006.

[51] H. Gray. *Anatomy of the Human Body*. Lea & Febiger, 1918.

[52] F. R. Hampel, E. M. Rochetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics*. Wiley, 1986.

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

[54] J. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523–534, 2006.

[55] C. S. Herrmann. Human eeg responses to 1-100 hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Experimental Brain Research*, 137(3-4):346–353, 2001. URL `http://dx.doi.org/10.1007/s002210100682`.

[56] N. Hill, T. Lal, M. Schröder, T. Hinterberger, N. Birbaumer, and B. Schölkopf. Selective attention to auditory stimuli: A brain-computer interface paradigm. page 102, Kirchentellinsfurt, Germany, 2004. Knirsch Verlag.

[57] L. Hochberg, M. Serruya, G. Friehs, J. Mukand, M. Saleh, A. Caplan, A. Branner, D. Chen, R. Penn, and J. Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, July 2006.

[58] P. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[59] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[60] H. H. Jasper. The ten-twenty electrode system of the International Federation. *Electroencephalography and clinical neurophysiology*, 10:371–375, 1958.

[61] M. Kawanabe, M. Krauledat, and B. Blankertz. A bayesian approach for adaptive BCI classification. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006*, pages 54–55. Verlag der Technischen Universität Graz, 2006.

[62] J. Kohlmorgen, K.-R. Müller, and K. Pawelzik. Segmentation and identification of drifting dynamical systems. In *Neural Networks for Signal Processing VII*, pages 326–335. IEEE, 1997.

[63] Z. J. Koles and A. C. K. Soong. EEG source localization: implementing the spatio-temporal decomposition approach. *Electroencephalography and Clinical Neurophysiology*, 107:343–352, 1998.

[64] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1993.

[65] M. Krauledat, B. Blankertz, G. Dornhege, M. Schröder, G. Curio, and K.-R. Müller. On-line differentiation of neuroelectric activities: algorithms and applications. In *Proceedings of the 28th Annual International Conference IEEE EMBS on Biomedicine*, New York City, 2006.

[66] M. Krauledat, G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. The Berlin brain-computer interface for rapid response. *Biomedizinische Technik*, 49(1):61–62, 2004.

[67] M. Krauledat, G. Dornhege, B. Blankertz, F. Losch, G. Curio, and K.-R. Müller. Improving speed and accuracy of brain-computer interfaces using readiness potential features. In *Conference of the IEEE Engineering in Medicine and Biology Society*, volume 4, pages 4511–4515, 2004. URL `http://dx.doi.org/10.1109/IEMBS.2004.1404253`.

[68] M. Krauledat, G. Dornhege, B. Blankertz, and K.-R. Müller. Robustifying EEG data analysis by removing outliers. *Chaos and Complexity Letters*, 2(3):259–274, 2007.

[69] M. Krauledat, F. Losch, and G. Curio. Brain state differences between calibration and application session influence BCI classification accuracy. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006*, pages 60–61. Verlag der Technischen Universität Graz, 2006.

[70] 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 Information Processing Systems 19*, pages 753–760, Cambridge, MA, 2007. MIT Press.

[71] M. Krauledat, P. Shenoy, B. Blankertz, R. P. N. Rao, and K.-R. Müller. Adaptation in CSP-based BCI systems. In G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors, *Toward Brain-Computer Interfacing*, pages 305–309. MIT Press, Cambridge, MA, 2007.

[72] R. Krepki. *Brain-Computer Interfaces: Design and Implementation of an Online BCI System of the Control in Gaming Applications and Virtual Limbs*. PhD thesis, Technische Universität Berlin, Fakultät IV – Elektrotechnik und Informatik, 2004.

[73] J. Kronegg and T. Pun. Measuring the performance of brain-computer interfaces using the information-transfer rate. In *3rd Int. Brain-Computer Interface Technology meeting, Rensselaerville, NY*, 2005.

[74] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw. A comparison of classification techniques for the P300 speller. *Journal of Neural Engineering*, 3(4):299–305, Dec 2006.

[75] T. N. Lal, T. Hinterberger, G. Widman, M. Schröder, N. J. Hill, W. Rosenstiel, C. E. Elger, B. Schölkopf, and N. Birbaumer. Methods towards invasive human brain computer interfaces. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 737–744. MIT Press, Cambridge, MA, 2005.

[76] W. Lang, O. Zilch, C. Koska, G. Lindinger, and L. Deecke. Negative cortical DC shifts preceding and accompanying simple and complex sequential movements. *Experimental Brain Research*, 74(1):99–104, 1989.

[77] P. Laskov, C. Schäfer, I. Kotenko, and K.-R. Müller. Intrusion detection in unlabeled data with quarter-sphere support vector machines (extended version). *Praxis der Informationsverarbeitung und Kommunikation*, 27:228–236, 2004.

[78] H. Lee, R. V. Bellamkonda, W. Sun, and M. E. Levenston. Biomechanical analysis of silicon microelectrodeinduced strain in the brain. *Journal of Neural Engineering*, 2(4):81–89, 2005.

[79] R. Leeb, G. Bauernfeind, S. Wriessnegger, H. Scharfetter, and G. Pfurtscheller. First steps towards the NIRS-based Graz-BCI. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006*, pages 102–103. Verlag der Technischen Universität Graz, 2006.

[80] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller. Spatio-spectral filters for improving classification of single trial EEG. *IEEE Transactions on Biomedical Engineering*, 52(9):1541–1548, 2005. URL `http://dx.doi.org/10.1109/TBME.2005.851521`.

[81] E. C. Leuthardt, K. J. Miller, G. Schalk, R. P. N. Rao, and J. G. Ojemann. Electrocorticography-based brain computer interface–the seattle experience. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):194–198, 2006.

[82] D. J. McFarland, C. W. Anderson, K.-R. Müller, A. Schlögl, and D. J. Krusienski. BCI meeting 2005 – workshop on BCI signal processing: Feature extraction and translation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):135–138, 2006.

[83] F. C. Meinecke, S. Harmeling, and K.-R. Müller. Robust ICA for super-gaussian sources. In C. G. Puntonet and A. Prieto, editors, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2004)*, 2004.

[84] F. C. Meinecke, S. Harmeling, and K.-R. Müller. Inlier-based ICA with an application to super-imposed images. *Int. J. of Imaging Systems and Technology*, 2005.

[85] F. C. Meinecke, A. Ziehe, J. Kurths, and K.-R. Müller. Measuring Phase Synchronization of Superimposed Signals. *Physical Review Letters*, 94(8):084102, 2005.

[86] P. Meinicke, M. Kaper, F. Hoppe, M. Heumann, and H. Ritter. Improving transfer rates in brain computer interfacing: A case study. In S. T. S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1107–1114, 2003.

[87] J. Mellinger, G. Schalk, C. Braun, H. Preissl, W. Rosenstiel, N. Birbaumer, and A. Kübler. An MEG-based brain-computer interface (BCI). *NeuroImage*, 36(3): 581–593, 2007.

[88] M. Middendorf, G. McMillan, G. Calhoun, and K. S. Jones. Brain-computer interface based on the steady-state visual-evoked response. *IEEE Transactions on Rehabilitation Engineering*, 8(2):211–214, June 2000.

[89] J. Millán. On the need for on-line learning in brain-computer interfaces. In *Proceedings of the International Joint Conference on Neural Networks*, Budapest, Hungary, July 2004. IDIAP-RR 03-30.

[90] J. Millán, A. Buttfield, C. Vidaurre, M. Krauledat, A. Schlögl, P. Shenoy, B. Blankertz, R. P. N. Rao, R. Cabeza, G. Pfurtscheller, and K.-R. Müller. Adaptation in brain-computer interfaces. In G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors, *Toward Brain-Computer Interfacing*, pages 303–326. MIT Press, Cambridge, MA, 2007.

[91] J. Millán, F. Renkens, J. Mouriño, and W. Gerstner. Brain-actuated interaction. *Artificial Intelligence*, 159:241–259, 2004.

[92] C. B. Moler. MATLAB — an interactive matrix laboratory. Technical Report 369, University of New Mexico. Dept. of Computer Science, 1980.

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

[94] K.-R. Müller and B. Blankertz. Toward noninvasive brain-computer interfaces. *IEEE Signal Processing Magazine*, 23(5):125–128, September 2006.

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

[96] K.-R. Müller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz. Machine learning and applications for brain-computer interfacing. In M. J. Smith and G. Salvendy, editors, *Human Interface, Part I, HCII 2007*, volume 4557 of *LNCS*, pages 705–714, Berlin Heidelberg, 2007. Springer. in press.

[97] K.-R. Müller, M. Krauledat, G. Dornhege, S. Jähnichen, G. Curio, and B. Blankertz. A note on the Berlin Brain-Computer Interface. In G. Hommel and S. Huanye, editors, *Human Interaction with Machines: Proceedings of the 6th International Workshop held at the Shanghai Jiao Tong University*, pages 51–60, 2006.

[98] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Neural Networks*, 12(2):181–201, May 2001.

[99] G. R. Müller-Putz, R. Scherer, C. Brauneis, and G. Pfurtscheller. Steady-State Visual Evoked Potential (SSVEP) based communication: impact of harmonic frequency components. *Journal of Neural Engineering*, 2:123–130, 2005.

[100] G. R. Müller-Putz, R. Scherer, C. Neuper, and G. Pfurtscheller. Steady-state somatosensori evoked potentials: Suitable brain signals for brain-computer interfaces? *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(1):30–37, 2006.

[101] G. R. Müller-Putz, R. Scherer, G. Pfurtscheller, and R. Rupp. EEG-based neuroprothesis control: a step towards clinical practice. *Neuroscience Letters*, 382:169–174, 2005.

[102] N. Murata, K.-R. Müller, A. Ziehe, and S. i. Amari. Adaptive on-line learning in changing environments. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 599. The MIT Press, 1997.

[103] D. G. Nair, K. L. Purcott, A. Fuchs, F. Steinberg, and J. A. Kelso. Cortical and cerebellar activity of the human brain during imagined and executed unimanual and bimanual action sequences: a functional MRI study. *Brain Research: Cognitive Brain Research*, 15(3):250–260, 2003.

[104] M. A. Nicolelis, A. A. Ghazanfar, C. R. Stambaugh, L. M. Oliveira, M. Lambach, J. Chapin, R. J. Nelson, and J. H. Kaas. Simultaneous encoding of tactile information by three primate cortical areas. *Nature Neuroscience*, 7:621–630, 1998.

[105] G. Nolte, F. C. Meinecke, A. Ziehe, and K.-R. Müller. Identifying interactions in mixed and noisy complex systems. *Physical Review E*, 73, 2006.

[106] G. Nolte, A. Ziehe, F. C. Meinecke, and K.-R. Müller. Analyzing coupled brain sources: Distinguishing true from spurious interaction. In *Advances in Neural Inf. Proc. Systems (NIPS 05)*, volume 18, 2006. accepted.

[107] B. Obermaier, G. R. Müller, and G. Pfurtscheller. "Virtual keyboard" controlled by spontaneous EEG activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(4):422–426, 2003.

[108] A. V. Oppenheim and R. W. Schafer. *Discrete-time signal processing*. Prentice Hall Signal Processing Series. Prentice Hall, 1989.

[109] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, pages 320–327, May, 2000.

[110] L. Parra, C. Spence, A. Gerson, and P. Sajda. Response error correction - a demonstration of improved human-machine performance using real-time EEG monitoring. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):173–177, 2003.

[111] K. Pawelzik, J. Kohlmorgen, and K.-R. Müller. Annealed competition of experts for a segmentation and classification of switching dynamics. *Neural Computation*, 8: 340–356, 1996.

[112] G. Pfurtscheller. Graphical display and statistical evaluation of event-related desynchronization (ERD). *Electroencephalography and Clinical Neurophysiology*, 43: 757–760, 1977.

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

[114] G. Pfurtscheller and F. H. L. da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, Nov 1999.

[115] M. E. Phelps. Emission computed tomography. *Seminars in Nuclear Medicine*, 7(4): 337–365, 1977.

[116] A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization – A Universal Concept in Nonlinear Sciences*. Cambridge University Press, 2001.

[117] 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), 2007. URL `http://dx.doi.org/10.1371/journal.pone.0000637`.

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

[119] P. Sajda, A. Gerson, K.-R. Müller, B. Blankertz, and L. Parra. A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):184–185, 2003. URL `http://dx.doi.org/10.1109/TNSRE.2003.814453`.

[120] G. Schalk, J. R. Wolpaw, D. J. McFarland, and G. Pfurtscheller. EEG-based communication: presence of an error potential. *Clinical Neurophysiology*, 111:2138–2144, 2000.

[121] A. Schlögl, J. Fortin, W. Habenbacher, and M. Akay. Adaptive mean and trend removal of heart rate variability using kalman filtering. In *Conference of the IEEE Engineering in Medicine and Biology Society*, 2001.

[122] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[123] B. Schölkopf and A. J. Smola. *Learning with kernels*. MIT Press, Cambridge, MA, 2002.

[124] M. Schröder, T. N. Lal, T. Hinterberger, M. Bogdan, N. J. Hill, N. Birbaumer, W. Rosenstiel, and B. Schölkopf. Robust EEG channel selection across subjects for brain computer interfaces. *EURASIP Journal on Applied Signal Processing, Special Issue: Trends in Brain Computer Interfaces*, 19:3103–3112, 2005.

[125] E. W. Sellers, D. Krusienski, D. McFarland, and J. Wolpaw. Noninvasive brain-computer interface research at the wadsworth center. In G. Dornhege, J. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors, *Toward brain-computer Interfacing*, pages 31–42. MIT Press, 2007.

[126] 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. URL `http://dx.doi.org/10.1088/1741-2560/3/1/R02`.

[127] F. Silva, T. H. van Lierop, C. F. Schrijer, and W. S. van Leeuwen. Organization of thalamic and cortical alpha rhythm: Spectra and coherences. *Electroencephalography and Clinical Neurophysiology*, 35:627–640, 1973.

[128] R. Sitaram, A. Caria, R. Veit, K. Uludag, T. Gaber, A. Kübler, and N. Birbaumer. Functional magnetic resonance imaging based BCI for neurorehabilitation. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006*, pages 104–106. Verlag der Technischen Universität Graz, 2006.

[129] R. Sitaram, H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu, and N. Birbaumer. Temporal classification of multi-channel near infrared spectroscopy signals of motor imagery for developing a brain-computer interface. *NeuroImage*, 34(4):1416–1427, 2007.

[130] A. Stuart and K. Ord. *Distribution Theory*, volume 1 of *Kendall's Advanced Theory of Statistics*. Wiley, 1994.

[131] J. Subbaroyan, D. C. Martin, and D. R. Kipke. A finite-element model of the mechanical effects of implantable microelectrodes in the cerebral cortex. *Journal of Neural Engineering*, 2(4):103–113, 2005.

[132] M. Sugiyama, B. Blankertz, M. Krauledat, G. Dornhege, and K.-R. Müller. Importance-weighted cross-validation for covariate shift. In *Proc. DAGM, LNCS 4174*, pages 354–363. Springer-Verlag, 2006.

[133] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:1027–1061, 2007.

[134] S. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics and Decisions*, 23(4):249–279, 2005.

[135] S. Sutton, M. Braren, J. Zubin, and E. R. John. Evoked-potential correlates of stimulus uncertainty. *Science*, 150(700):1187–1188, 1965.

[136] P. Sykacek, S. Roberts, and M. Stokes. Adaptive BCI based on variational Bayesian Kalman filtering: an empirical evaluation. *IEEE Transactions on Biomedical Engineering*, 51(5):719–729, 2004.

[137] D. Tax and R. Duin. Uniform object generation for optimizing one-class classifiers. *Journal for Machine Learning Research*, pages 155–173, 2001.

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

[139] R. Tomioka, G. Dornhege, K. Aihara, and K.-R. Müller. An iterative algorithm for spatio-temporal filter optimization. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006*, pages 22–23. Verlag der Technischen Universität Graz, 2006.

[140] R. Tomioka, G. Dornhege, G. Nolte, K. Aihara, and K.-R. Müller. Optimizing spectral filters for single trial EEG classification. In *Proc. DAGM, LNCS 4174*, pages 414–423. Springer-Verlag, 2006.

[141] R. Tomioka, J. Hill, B. Blankertz, and K. Aihara. Adapting spatial filtering methods for nonstationary BCIs. In *Proceedings of 2006 Workshop on Information-Based Induction Sciences (IBIS2006)*, pages 65–70, 2006.

[142] V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, New York, 1995.

[143] V. Vapnik. *Statistical learning theory*. John Wiley, New York, 1998.

[144] C. Vidaurre, R. Scherer, R. Cabeza, A. Schlögl, and G. Pfurtscheller. Study of discriminant analysis applied to motor imagery bipolar data. *Medical & Biological Engineering & Computing*, 45(1):61–68, 2007.

[145] C. Vidaurre, A. Schlögl, R. Cabeza, and G. Pfurtscheller. About adaptive classifiers for brain computer interfaces. *Biomedizinische Technik*, 49(1):85–86, 2004.

[146] C. Vidaurre, A. Schlögl, R. Cabeza, R. Scherer, and G. Pfurtscheller. A fully online adaptive BCI. *IEEE Transactions on Biomedical Engineering*, 53(6):1214–1219, 2006.

[147] N. Weiskopf, K. Mathiak, S.W.Bock, F. Scharnowski, R. Veit, W. Grodd, R. Goebel, and N. Birbaumer. Principles of a brain-computer interface (BCI) based on real-time functional Magnetic Resonance Imaging (fMRI). *IEEE Transactions on Biomedical Engineering*, 51:966–970, 2004.

[148] Wikipedia. Transmission control protocol — wikipedia, the free encyclopedia, 2007. URL `http://en.wikipedia.org/w/index.php?title=Transmission_Control_Protocol&oldid=164630423`. [Online; accessed 15-October-2007].

[149] Wikipedia. User datagram protocol — wikipedia, the free encyclopedia, 2007. URL `http://en.wikipedia.org/w/index.php?title=User_Datagram_Protocol&oldid=162087765`. [Online; accessed 15-October-2007].

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

[151] J. R. Wolpaw and D. J. McFarland. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17849–17854, 2004.

[152] J. R. Wolpaw, D. J. McFarland, and T. M. Vaughan. Brain-computer interface research at the Wadsworth Center. *IEEE Transactions on Rehabilitation Engineering*, 8(2):222–226, 2000.

[153] A. Ziehe and K.-R. Müller. TDSEP – an efficient algorithm for blind separation using time structure. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks, ICANN'98*, Perspectives in Neural Computing, pages 675 – 680, Berlin, 1998. Springer Verlag.