# ONE STEP FEATURE EXTRACTION AND CLASSIFICATION WITH REGULARIZATION FOR BCI

A thesis submitted in partial fulfillment of the requirements for the award of

**Master of Technology**
in
**COMMUNICATION ENGINEERING**

**of the Calicut University**

by


**ARUN K. BHARATHAN**
(Reg. No. : NSALCCM005)



N.S.S. COLLEGE OF ENGINEERING, PALAKKAD

Electronics and Communication Engineering Department


AUGUST 2013

## Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Place: Signature:

Date : Name: ARUN K. BHARATHAN

## Acknowledgement

First of all, I would like to thank the Almighty for giving me the strength to complete the thesis work.

I would like to express my heartfelt gratitude to my Project Guide Prof. Nandakumar P., Professor at Electronics and Communication Engineering Department, for his constant support and guidance throughout the thesis. He is so humble and down to earth and has inspired a lot.

I would like to appreciate the guidance given by other staff members of the department, especially Dr. T. Sudha, Professor and M.Tech coordinator, Electronics and Communication Engineering Department, and Prof. Kala L., Associate Professor, of Electronics and Communication Engineering Department, N.S.S College of Engineering, Palakkad, for their time and valuable feedback.

My sincere thanks to the Head of the Department Dr. R. Sindhu, Professor and former Head of the Department, Prof. Abdul Kareem M., Associate Professor for their valuable support throughout the duration of the project.

I would like to thank my colleague Mr. Aneesh M. Koya, for his support and his timely assistance with this thesis. I would also like to appreciate the support and encouragement given by my family and colleagues.

# Abstract

Brain computer interfaces (BCIs) have the potential to offer humans a new and innovative nonmuscular modality through which to communicate directly via their brain activity with the environment. These systems rely on the acquisition and interpretation of the commands encoded in neurophysiological signals without using the conventional muscular output pathways of the central nervous system (CNS). Brain imaging technologies such as EEG, fMRI and MEG are used to observe this neurophysiological activity. Electroencephalograph (EEG) is the only practical noninvasive, cheap and real-time capable imaging technology for use in a BCI system. BCIs propose to offer people who suffer from neuromuscular disorders, whom lack any voluntary motor movement, with the only possibility of communication and control.

This thesis firstly addresses the issues for using EEG as a BCI input modality by reviewing the methods for EEG acquisition and analysis. The components and methodologies for a BCI system framework and the state of the art in this technology are then presented. Feature extraction and classification are the main stages in the BCI system. The feature extraction stage identifies discriminative information in the brain signals that have been recorded. The classification stage classifies the signals by taking the feature vectors into account.

A one step process is used which combines the discrimination and classification stages with a linear classifier employing a regularization scheme based on the spectral $\ell 1$-norm and Tikhonov regularization of its coefficient matrix. The spectral regularization not only provides a principled way of complexity control but also enables automatic determination of the rank of the coefficient matrix. Using the Linear Matrix Inequality technique, we formulate the inference task as a single convex optimization problem. This method is applied to the motor-imagery EEG classification problem. The method not only improves upon conventional methods in the classification performance but also determines a subspace in the signal that concentrates discriminative information without any additional feature extraction step.

# Contents

**Acknowledgement** iv

**Abstract** v

**Contents** vi

**List of Figures** ix

**List of Tables** x

**1 Introduction** 1

1.1 Background . . . . . 1

1.2 Motivation and Problem Statement . . . . . 3

1.3 Aim and Objectives . . . . . 3

1.4 Thesis Outline . . . . . 4

**2 Introduction To Brain Computer Interfaces** 6

2.1 Introduction . . . . . 6

2.2 Overview of Brain Computer Interfaces . . . . . 7

2.3 Types of BCIs . . . . . 8

2.4 Signal Acquisition-Neuro imaging techniques . . . . . 9

2.4.1 Electroencephalogram . . . . . 10

2.4.2 Electrocorticogram . . . . . 12

2.4.3 Magnetoencephalography (MEG) . . . . . 12

2.4.4 Functional Magnetic Resonance Imaging (fMRI) . . . . . 13

2.4.5 Near Infrared Spectroscopy (NIRS) . . . . . 13

2.5 Control Signal Types in BCIs . . . . . 13

2.5.1 Visual Evoked Potentials (VEPs) . . . . . 14

2.5.2 Slow Cortical Potentials (SCPs) . . . . . 14

2.5.3 P300 Evoked Potentials . . . . . 15

2.5.4 Sensorimotor Rhythms (mu and beta rhythms) . . . . . 16

2.6 Features Extraction and Selection . . . . . 16
2.7 Artifacts in BCIs . . . . . 18
2.8 Classification Algorithms . . . . . 19
2.9 Applications . . . . . 19
2.9.1 Spelling Devices . . . . . 20
2.9.2 Environment Control . . . . . 20
2.9.3 Wheelchair Control . . . . . 20
2.9.4 Neuromotor Prostheses . . . . . 20
2.9.5 Gaming and Virtual Reality . . . . . 21

**3 Feature Extraction and Classification 22**
3.1 Introduction . . . . . 22
3.2 Common Spatial Pattern Algorithm . . . . . 22
3.2.1 Geometric Approach . . . . . 24
3.2.2 Generalized Eigen value problem . . . . . 26
3.2.3 Optimization Approach . . . . . 28
3.3 Linear Discriminant Analysis . . . . . 28
3.3.1 Mathematical Operation . . . . . 30
3.3.2 Limitations of LDA . . . . . 34

**4 A One Step Feature Extraction and Classification with Regularization for BCI 35**
4.1 Introduction . . . . . 35
4.2 Signal Analysis Framework . . . . . 35
4.2.1 Discriminative Learning . . . . . 36
4.2.2 Detector Function . . . . . 39
4.2.3 Regularization . . . . . 40
4.2.4 Discussion on Discriminative Approach . . . . . 40
4.3 Implementation . . . . . 42

**5 Results and Discussion 44**
5.1 Introduction . . . . . 44
5.2 Dataset Description . . . . . 44
5.2.1 BCI Competition IV Dataset 4a . . . . . 44

5.2.2 BCI Competition III Dataset 1 . . . . . 45
5.3 Signal Pre-processing and Predictor Model . . . . . 46
5.4 Results . . . . . 47
5.5 CSP with Tikhonov Regularization . . . . . 50
5.6 Novelty in our work . . . . . 50
5.7 Conclusion . . . . . 52

**6 Conclusion** **54**

**Bibliography** **55**

# List of Figures


1.1 Brain Computer Interface . . . . . 2

2.1 Visualization of Brain Rhythms . . . . . 11
2.2 Two element feature vectors for all exciting trials in red and non-exciting trials in green . . . . . 17

3.1 Geometric approach before and after CSP filtering . . . . . 24
3.2 Spatial Pattern . . . . . 27
3.3 Spatial Filter . . . . . 27
3.4 From the 3D scatter plots it is clear that LDA outperforms PCA in terms of class discrimination n this is one example where the discriminatory information is not aligned with the direction of maximum variance . . . . . 30
3.5 The two classes are not well separated when projected onto this line . . . . 31
3.6 This line succeeded in separating the two classes and in the mean time reducing the dimensionality of our problem from two features $(x1, x2)$ to only a scalar value $y$. . . . . 31
3.7 Befor Linear Discrimination . . . . . 32
3.8 After Linear Discrimination. . . . . 33

4.1 Logistic Function . . . . . 38

5.1 Timing scheme of the BCI paradigm . . . . . 45
5.2 Electrode Positions . . . . . 45
5.3 Low rank weight matrix for ECoG data . . . . . 48
5.4 Spatial filter and spatial pattern . . . . . 49
5.5 Spatial filter captured by CSP . . . . . 50
5.6 Spatial filter captured by one step process with DS . . . . . 50
5.7 Spatial filter learned by one step process for TRCSP . . . . . 51
5.8 Spatial filter learned by TRCSP by Eigen decomposition . . . . . 51

# List of Tables

5.1 Result for Dataset 2a . . . . . 48
5.2 Comparison with CSP . . . . . 48
5.3 Performance Comparison of TRCSP Using One Step Process and by eigen decomposition . . . . . 52

# Chapter 1

# Introduction

## 1.1 Background

A brain-computer interface (BCI) is a device that can read brain signals and convert them into control and communication signals. BCIs are often directed at assisting, augmenting, or repairing human cognitive or sensory-motor functions. A standard BCI system consists of a signal acquisition, signal enhancement, feature extraction, classification and the control interface stages as shown in Fig. 1.1. This chapter gives a basic introduction to BCI, types of neuroimaging modalities used in the signal acquisition step and types of electro-physiological control signals that determine the user intentions.

BCI design represents a new frontier in science and technology that requires multidisciplinary skills from fields such as neuroscience, engineering, computer science, psychology and clinical rehabilitation to achieve the goal of developing an alternative communication medium. Despite the technological developments, there remain numerous obstacles to build an efficient BCI. The biggest challenges are related to accuracy, speed and usability. If a disabled person can move his/her eyes or even a single muscle in a controlled way, the interfaces based on eye-gaze or EMG switch technology are more efficient than any of the BCIs that exist today. The maximum information transfer rate of current BCI systems is typically 25 bits per minute.

The past two decades have seen an explosion of scientific interest in a completely different and novel approach of interacting with a computer. Inspired by the social recognition of people who suffer from severe neuromuscular disabilities, an interdisciplinary field of research has been created to offer direct human computer interaction via signals generated by the brain itself. BCI technology, as it is known, is a revolutionary communication channel that enables users to control computer applications through thoughts alone. The development of the cognitive neuroscience field has been instigated by recent advances in

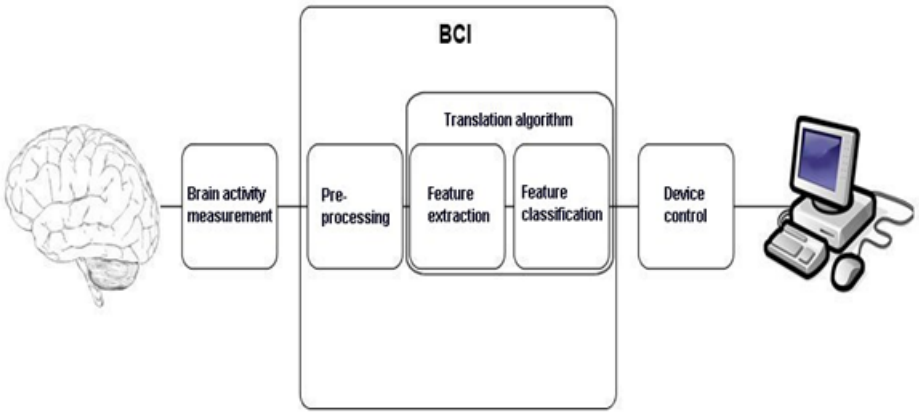


Figure 1.1: Brain Computer Interface

brain imaging technologies such as Electroencephalography (EEG), Magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI).

EEG is an imperfect and distorted indicator of brain activity, yet the fact that it can be acquired cheaply, is non-invasive and demonstrates direct functional correlations with high temporal resolution makes it the only practical direct brain computer communication channel. It is a new and challenging medium for us to exploit in a similar manner to the other communication modalities such as voice or vision. The endless potential of tapping into human brain signals may see the fantasies of science fiction writers becoming reality in the future.

Many complex processes and systems would operate on the basis of thought in the future. Currently, the field of BCI is in infancy stage and would require deeper insights on how to capture the right signals and then process them suitably. The advancements are limited to recognition of certain words, expressions, moods etc. Efforts are being made to recognize the objects as they are seen by the brain. These efforts will bring in newer dimensions in the understanding of brain functioning, damage and repair. It is possible to recognize the thoughts of the human brain by capturing the right signals from the brain in future. But present operating systems and interfaces are not suitable for working with thought based system. In future new OS may be developed for the BCI applications.

The focus of this thesis is on the signal enhancement or pre-processing, feature extraction and feature classification stage. The BCI competition data sets were used for the analysis of algorithm. A combined feature extraction and feature classification method is employed in our work.

## 1.2 Motivation and Problem Statement

Focusing on the EEG as the BCI input modality, the goal was to develop a deep understanding of the neurophysiological processes that could be exploited to implement a BCI system. After performing a state-of-the-art review of BCI systems, it was envisaged to design and implement a system. A basic knowledge of EEG waveform characteristics, signal processing methodologies for feature extraction and classification is a prerequisite before attempting to design and implement a BCI system.

The common spatial pattern (CSP) method is used for the extraction of motor imagery tasks in BCI. It is followed by a classification stage using linear discriminant analysis stage. In CSP, the usual notion is to take a pair of eigen vectors from both the sides of coefficient matrix. This is not an optimal method for noise contaminated signals. An alternate method to overcome this problem is to use a discriminative model which captures a low rank structure controlled by regularization and thereby achieving a simple classification. This regularizer reduces singular values of the coefficient matrix.

## 1.3 Aim and Objectives

Based on a short segment of EEG called a trial, the signal analysis in BCI aims to predict the brain state of a user out of prescribed options (e.g. foot Vs. left-hand motor imagery Vs. rest). In machine learning terms, this is a multi-class classification problem. The challenge in EEG-based BCI is the low spatial resolution caused by volume conduction, the high artefact and the outlier content of the signal and the mass of data that makes the application of conventional statistical analysis difficult. Therefore many studies have focused on how to extract a small number of task informative features from the data that can be fed into some relatively simple classifiers; commonly used are linear spatial filtering methods or independent component analysis coupled with heuristic frequency band selection or band weighting.

One of the shortcomings of the feature extraction approaches is the strong and hard-to-control inductive bias. It limits their application to rather specific experimental paradigms for which they are developed. Another approach is the discriminative approach that tries to optimize the classifier coefficients from the training data under a unified criterion. The theoretical advantage of the discriminative approach is that the coefficients (e.g.,

spatial filter and temporal filter) are jointly optimized under a single criterion. Moreover, inductive bias can be controlled in a principled manner through regularization. However many previous studies had to solve non-convex optimization problems, which can be challenging because of multiple local minima and difficulty in terminating the learning algorithms.

Here, we combine the probabilistic data-fit criteria with sparse regularizers. The proposed regularizers naturally induce sparse or factorized models through a convex optimization problem; moreover the number of components is automatically determined; in addition it is shown that the decoding model can be instantly converted into a loss function that is used for the training of the classifier; thus no intermediate goal such as binary classification needs to be imposed. Finally, it has been shown that the different second order informations in the signal can be combined and selected in a systematic manner through the dual spectral (DS) regularization. The issue of complexity control, feature extraction, and the interpretability of the resulting model is thus tackled in a unified and systematic manner under the roof of a convex regularized empirical risk minimization problem.

## 1.4 Thesis Outline

This thesis presents the fundamental knowledge behind developing an Electroencephalogram based BCI as well as a state-of-the-art review of BCI research. The thesis concludes by looking in to the future of BCI technology.

The **Chapter two** aims at reviewing the main BCI designs and their applications. This chapter starts by giving some definitions related to BCI. Then it reviews the methods and techniques used to design a BCI. As such, it details with the different processing steps composing a BCI, i.e., measurements of brain activity, preprocessing, feature extraction and classification. Finally, some BCI applications and the already developed prototypes are discussed, by emphasising virtual reality applications.

The **Chapter three** addresses the problem of feature extraction and classification. CSP method for feature extraction and LDA method for classification.

**Chapter four** presents the technique of one step feature extraction and classification, its algorithm and implementation.

**Chapter five** deals with results, discussion and modification of the technique.

The conclusion in **Chapter six** briefly summarizes the outcome of the thesis and provides suggestions for future works.

# Chapter 2

# Introduction To Brain Computer Interfaces

## 2.1 Introduction

Brain Computer Interface (BCIs) [1] started with Hans Berger's inventing of electrical activity of the human brain and the development of electroencephalography (EEG). In 1924 Berger recorded an EEG signals from a human brain for the first time. By analyzing EEG signals Berger was able to identify oscillatory activity in the brain, such as the alpha wave $8 - 12$ Hz, also known as Berger's wave. The first recording device used by Berger was very elementary, which was in the early stages of development and was required to insert silver wires under the scalp of the patients. In later stages, those were replaced by silver foils that were attached to the patients head by rubber bandages. More sophisticated measuring devices such as the Siemens double-coil recording galvanometer, which displayed electric voltages as small as one ten thousandth of a volt, led to success. Berger analyzed the interrelation of alternations in his EEG wave diagrams with brain diseases. EEGs permitted completely new possibilities for the research of human brain activities.

In this chapter we review the aspects of BCI research mentioned above and highlight recent developments and open problems. The review is ordered by the steps that are needed for brain computer communication. We start with methods for measuring brain activity (Section 2.3) and then give a description of the neurophysiologic signals that can be used in BCI systems (Section 2.4). The translation of signals into commands with the help of signal processing and classification methods is described in Section 2.5. Finally, applications that can be controlled with a BCI are described in Section 2.6.

## 2.2 Overview of Brain Computer Interfaces

Any natural form of communication or control requires peripheral nerves and muscles. The process begins with the user's intent. This intent triggers a complex process in which certain brain areas are activated, and hence signals are sent via the peripheral nervous system to the corresponding muscles, which in turn perform the movement necessary for the communication or control task. The activity resulting from this process is often called motor output or efferent output. Efferent means conveying signal from central to peripheral nervous system and further to an effecter such as muscle. The efferent pathway is necessary for motor control, the reverse of it afferent pathway (sensory pathway) is for learning motor skills and dexterous tasks, such as typing or playing a musical instrument. But, a BCI offers alternate pathway to natural communication and control, a man made system that bypasses body's normal path way from central nervous system. Instead of depending on peripheral nervous system and muscles, a BCI directly measures brain activity associated with an action from user and translates this brain activity into corresponding control signals. This translation involves signal processing and pattern recognition which is done by a computer. Since the measured activity originates directly from the brain and not from the peripheral systems or muscles, the system is called a Brain Computer Interface [2].

A Brain Computer Interface (BCI) provides a communication path between human brain and the computer system [3]. The major goal of BCI research is to develop a system that allows disabled people to communicate with other persons and helps to interact with the external environments. This area includes components like, comparison of invasive and non invasive technologies to measure brain activity, evaluation of control signals (i.e. patterns of brain activity that can be used for communication), development of algorithms for translation of brain signals into computer commands, and the development of new BCI applications.

A BCI is an artificial intelligence system that can recognize a certain set of patterns in brain signals following five consecutive stages: signal acquisition, pre-processing or signal enhancement, feature extraction, classification, and the control interface. The signal acquisition stage captures the brain signals and may also perform noise reduction and artifact processing. The pre-processing stage prepares the signals in a suitable form for further processing. The feature extraction stage identifies discriminative information

in the brain signals that have been recorded. Once measured, the signal is mapped onto a vector containing effective and discriminant features from the observed signals. The extraction of this interesting information is a very challenging task. Brain signals are mixed with other signals coming from a finite set of brain activities that overlap in both time and space. Moreover, the signal is not usually stationary and may also be distorted by artifacts such as electromyography (EMG) or electrooculography (EOG). The feature vector must also be of a low dimension, in order to reduce feature extraction stage complexity, but without relevant information loss. The classification stage classifies the signals taking the feature vectors into account. The choice of good discriminative features is therefore essential to achieve effective pattern recognition, in order to decipher the user's intentions. Finally the control interface stage translates the classified signals into meaningful commands for any connected device, such as a wheelchair or a computer.

## 2.3 Types of BCIs

The BCIs can be categorized into

1. exogenous or endogenous

2. synchronous (cue-paced) or asynchronous (self-paced)

According to the nature of the signals used as input, BCI systems can be classified as either exogenous or endogenous. Exogenous BCI uses the neuron activity elicited in the brain by an external stimulus such as visual or auditory evoked potentials. Exogenous systems do not require extensive training since their control signals, SSVEPs and P300, can be easily and quickly set-up. Besides, the signal controls can be realized with only one EEG channel and can achieve a high information transfer rate of up to 60 bits/min. On the other hand, endogenous BCI is based on self-regulation of brain rhythms and potentials without external stimuli. Through neurofeedback training, the users learn to generate specific brain patterns which may be decoded by the BCI such as modulations in the sensorimotor rhythms or the Slow Cortical Potentials. The advantage of an endogenous BCI is that the user can operate the BCI at free will and move a cursor to any point in a two-dimensional space, while an exogenous BCI may constrain the user to the choices presented. Also, endogenous BCI are especially useful for users with advanced stages of ALS or whose sensory organs are affected [4].

According to the input data processing modality, BCI systems can be classified as synchronous or asynchronous. Synchronous BCIs analyze brain signals during predefined time windows. Any brain signal outside the predefined window is ignored. Therefore, the user is only allowed to send commands during specific periods determined by the BCI system. For example, the standard Graz BCI represents a synchronous BCI system. The advantage of a synchronous BCI system is that the onset of mental activity is known in advance and associated with a specific cue. Moreover, the patients may also perform blinks and other eye movements, which would generate artifacts, if the BCI did not analyze the brain signals to avoid their misleading effects. This simplifies the design and evaluation of synchronous BCI. Asynchronous BCIs continuously analyze brain signals no matter when the user acts. They offer a more natural mode of human-machine interaction than synchronous BCI. However, asynchronous BCIs are more computation demanding and complex.

## 2.4 Signal Acquisition-Neuro imaging techniques

BCIs use brain signals to gather information on user intentions.. Two types of brain activities [5] may be monitored:

1. electrophysiological
2. hemodynamic

Electrophysiological activity is generated by electro-chemical transmitters exchanging information between the neurons. Electrophysiological activity is measured by electroencephalography, electrocorticography, magnetoencephalography, and electrical signal acquisition in single neurons. The hemodynamic response is a process in which the blood releases glucose to active neurons at a greater rate than in the area of inactive neurons. The glucose and oxygen delivered through the blood stream results in a surplus of oxyhemoglobin in the veins of the active area, and in a distinguishable change of the local ratio of oxyhemoglobin to deoxyhemoglobin. These changes can be quantified by neuroimaging methods such as functional magnetic resonance and near infrared spectroscopy. These kinds of methods are categorized as indirect, because they measure the hemodynamic response, which, in contrast to electrophysiological activity, is not directly related to neuronal activity [2].

### 2.4.1 Electroencephalogram

EEG measures electric brain activity caused by the flow of electric currents during synaptic excitations of the dendrites in the neurons. EEG signals are easily recorded in a non-invasive manner through electrodes placed on the scalp, for that reason it is by far the most widespread recording modality. However, it provides very poor quality signals as the signals have to cross the scalp, skull, and many other layers. This means that EEG signals in the electrodes are weak, and of poor spatial resolution. This technique is moreover severely affected by background noise generated either inside the brain or externally over the scalp.

The EEG recording system consists of electrodes, amplifiers, A/D converter, and a recording device. The electrodes acquire the signal from the scalp, the amplifiers process the analogue signal to enlarge the amplitude of the EEG signals so that the A/D converter can digitalize the signal in a more accurate way. Finally, the recording device, which may be a personal computer or similar, stores, and displays the data.

A minimal configuration for EEG measurement therefore consists of one measurement, one reference, and one ground electrode. Multi-channel configurations can comprise up to 256 measurement electrodes. These electrodes are usually made of silver chloride (AgCl). Electrode-scalp contact impedance should be between $1k\Omega$ and $10k\Omega$ to record signal accurately. The electrode-tissue interface is not only resistive but also capacitive and it therefore behaves as a low pass filter. EEG gel creates a conductive path between the skin and each electrode that reduces the impedance. Use of the gel is cumbersome, however, as continued maintenance is required to assure a relatively good quality signal. Electrodes that do not need to use gels, called 'dry' electrodes, have been made with other materials such as titanium and stainless-steel.

EEG comprises a set of signals which may be classified according to their frequency as shown in Fig. 2.1. According to the distribution over the scalp or biological significance, well-known frequency bands have been defined. These frequency bands are referred to as delta ($\delta$), theta ($\theta$), alpha ($\alpha$), beta ($\beta$), and gamma ($\gamma$) from low to high, respectively. The delta ($\delta$) band lies below 4 Hz, delta rhythms are usually only observed in adults in deep sleep state and are unusual in adults in an awake state. Due to low frequency, it is easy to confuse delta waves with artifact signals, which are caused by the large muscles of the neck or jaw.

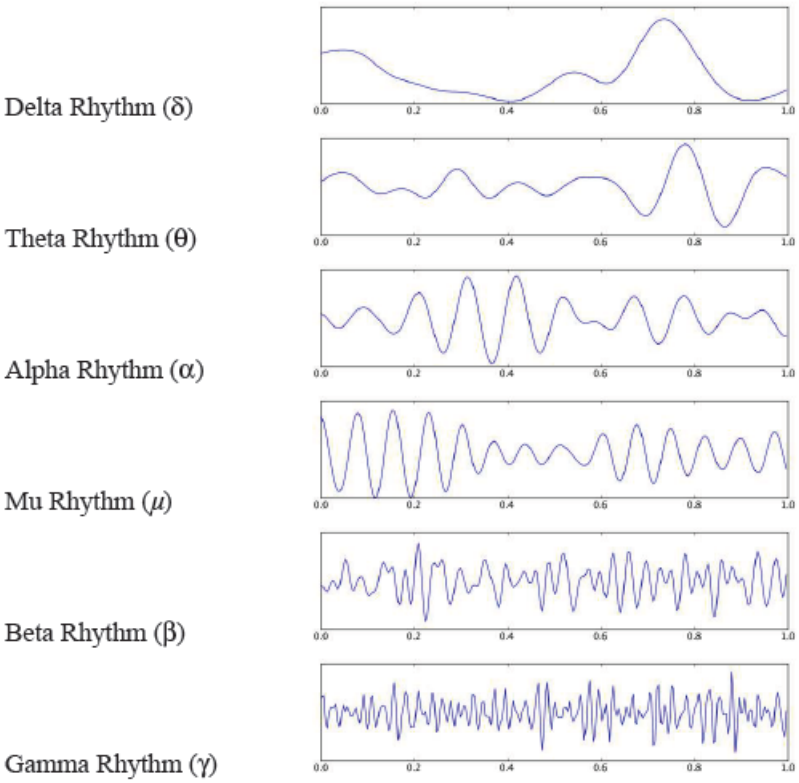


Figure 2.1: Visualization of Brain Rhythms

Theta ($\theta$) waves lie within the 4 to 7 Hz range. Theta band has been associated with meditative concentration and a wide range of cognitive processes such as mental calculation, maze task demands, or conscious awareness.

Alpha ($\alpha$) rhythms are found over the occipital region in the brain. These waves lie within the 8 to 12 Hz range. Their amplitude increases when the eyes close and the body relaxes and they attenuate when the eyes open and mental effort is made. These rhythms primarily reflect visual processing in the occipital brain region and may also be related to the memory brain function, that alpha activity may be associated with mental effort. Increasing mental effort causes a suppression of alpha activity, particularly from the frontal areas. Mu rhythms may be found in the same range as alpha rhythms, mu rhythms are strongly connected to motor activities.

Beta ($\beta$) rhythms, within the 12 to 30 Hz range, are recorded in the frontal and central regions of the brain and are associated with motor activities. Beta rhythms are desynchronized during real movement or motor imagery. Beta waves are characterized by their symmetrical distribution when there is no motor activity. However, in case of active movement, the beta waves attenuate, and their symmetrical distribution changes.

Gamma ($\gamma$) rhythms belong to the frequency range from 30 to 100 Hz. The presence of gamma waves in the brain activity of a healthy adult is related to certain motor functions or perceptions, among others Gamma rhythms are less commonly used in EEG-based BCI systems, because artifacts such as electromyography (EMG) or electrooculography (EOG) are likely to affect them. Nevertheless, this range is attracting growing attention in BCI research because, compared to traditional beta and alpha signals, gamma activity may increase the information transfer rate and offer higher spatial resolution.

### 2.4.2 Electrocorticogram

ECoG is a technique that measures electrical activity in the cerebral cortex by means of electrodes placed directly on the surface of the brain. Compared to EEG, ECoG provides higher temporal and spatial resolution as well as higher amplitudes and a lower vulnerability to artifacts such as blinks and eye movement. However, ECoG is an invasive recording modality which requires a craniotomy to implant an electrode grid, entailing significant health hazards. ECoG has been used for the analysis of alpha and beta waves or gamma waves produced during voluntary motor action.

### 2.4.3 Magnetoencephalography (MEG)

MEG is a non-invasive imaging technique that registers the brains magnetic activity by means of magnetic induction. MEG measures the intracellular currents flowing through dendrites which produce magnetic fields that are measurable outside of the head. The neurophysiological processes that produce MEG signals are identical to those that produce EEG signals. The advantage of MEG is that magnetic fields are less distorted by the skull and scalp than electric fields. Magnetic fields are detected by superconducting quantum interference devices, which are extremely sensitive to magnetic disturbances produced by neural activity. MEG requires effective shielding from electromagnetic interferences.MEG provides signals with higher spatiotemporal resolution than EEG, which reduces the training time needed to control a BCI and speeds up reliable communications. MEG has also been successfully used to localize active regions inside the brain. In spite of these advantageous features, MEG is not often used in BCI design because MEG technology is too bulky and expensive to become an acquisition modality suitable for everyday use.

### 2.4.4 Functional Magnetic Resonance Imaging (fMRI)

fMRI is a non-invasive neuroimaging technique which detects changes in local cerebral blood volume, cerebral blood flow and oxygenation levels during neural activation by means of electromagnetic fields. fMRI is generally performed using MRI scanners which apply electromagnetic fields of strength in the order of $3T$ or $7T$. The main advantage of the use of fMRI is high space resolution. For that reason, fMRI have been applied for localizing active regions inside the brain. However, fMRI has a low temporal resolution of about 1 or 2 seconds. Additionally, the hemodynamic response introduces a physiological delay from 3 to 6 seconds. fMRI appears unsuitable for rapid communication in BCI systems and is highly susceptible to head motion artifacts. fMRI requires overly bulky and expensive hardware.

### 2.4.5 Near Infrared Spectroscopy (NIRS)

NIRS is an optical spectroscopy method that employs infrared light to characterize non-invasively acquired fluctuations in cerebral metabolism during neural activity. Infrared light penetrates the skull to a depth of approximately 13 cm below its surface, where the intensity of the attenuated light allows alterations in oxyhemoglobin and deoxyhemoglobin concentrations to be measured. Due to shallow light penetration in the brain, this optical neuroimaging technique is limited to the outer cortical layer. In a similar way to fMRI, one of the major limitations of NIRS is the nature of the hemodynamic response, because vascular changes occur a certain number of seconds after its associated neural activity. The spatial resolution of NIRS is quite low, in the order of 1 cm. Nevertheless, NIRS offers low cost, high portability, and an acceptable temporal resolution in the order of 100 milliseconds.

## 2.5 Control Signal Types in BCIs

The purpose of a BCI is to interpret user intentions by means of monitoring cerebral activity. Brain signals involve numerous simultaneous phenomena related to cognitive tasks. Most of them are still incomprehensible and their origins are unknown. However, the physiological phenomena of some brain signals have been decoded in such way that people may learn to modulate them at will, to enable the BCI systems to interpret their

intentions. These signals are regarded as possible control signals in BCIs.

Numerous studies have described a vast group of brain signals that might serve as control signals in BCI systems. The control signals that are employed in current BCI systems such as visual evoked potentials, slow cortical potentials P300 evoked potentials, and sensorimotor rhythms are discussed [6].

### 2.5.1 Visual Evoked Potentials (VEPs)

VEPs are brain activity modulations that occur in the visual cortex after receiving a visual stimulus [7]. These modulations are relatively easy to detect since the amplitude of VEPs increases enormously as the stimulus is moved closer to the central visual field. VEPs may be classified according to three different criteria: (i) by the morphology of the optical stimuli, (ii) by the frequency of visual stimulation; and (iii) by field stimulation. According to the first criterion, VEPs may be caused by using flash stimulation or using graphic patterns such as checkerboard lattice, gate, and random-dot map. According to the frequency, VEPs can also be classified as transient VEPs (TVEPs) and as steady-state VEPs (SSVEPs). TVEPs occur when the frequency of visual stimulation is below 6 Hz, while SSVEPs occur in reaction to stimuli of a higher frequency. Lastly, according to the third criterion, VEPs can be divided into whole field VEPs, half field VEPs, and part field VEPs depending on the area of on-screen stimulus. For instance, if only half of the screen displays graphics, the other half will not display any visual stimulation and the person will look at the centre of the screen, which will induce a half field VEP. SSVEP-based BCIs allow users to select a target by means of an eye-gaze. The user visually fixes attention on a target and the BCI identifies the target through SSVEP features analysis.

### 2.5.2 Slow Cortical Potentials (SCPs)

SCPs are slow voltage shifts in the EEG that last a second to several seconds. SCPs belong to the part of the EEG signals below 1 Hz. SCPs are associated with changes in the level of cortical activity. Negative SCPs correlate with increased neuronal activity, whereas positive SCPs coincide with decreased activity in individual cells. These brain signals can be self-regulated by both healthy users and paralyzed patients to control external devices by means of a BCI. SCP shifts can be used to move a cursor and select the targets presented on a computer screen.

People can be trained to generate voluntary SCP changes using a thought-translation device. The thought-translation device is a tool used for self-regulation SCP training, which shows visual-auditory marks so that the user can learn to shift the SCP. The thought-translation device typically comprises a cursor on a screen in such a way that the vertical position of the cursor constantly reflects the amplitude of SCP shifts. Success in SCP self-regulation training depends on numerous factors, such as the patient's psychological and physical state, motivation, social context, or the trainer-patient relationship. Therefore, the value of SCPs as a suitable control signal for each subject can only be determined on the basis of initial trials. Other factors, such as sleep quality, pain, and mood also have an influence on self-regulation performance. Their effects are not identical for all patients and further investigation is certainly needed to establish general rules on this matter.

### 2.5.3 P300 Evoked Potentials

P300 evoked potentials are positive peaks in the EEG due to infrequent auditory, visual, or somatosensory stimuli. The P300 responses are elicited about $300ms$ after attending to an oddball stimulus among several frequent stimuli. Some studies have proven that the less probable the stimulus, the larger the amplitude of the response peak. The use of P300-based BCIs does not require training. However, the performance may be reduced because the user gets used to the infrequent stimulus and consequently P300 amplitude is decreased.

A typical application of a BCI based on visual P300 evoked potentials comprises a matrix of letters, numbers, or other symbols or commands. The rows or columns of this matrix are flashed at random while the EEG is monitored. The user gazes at the desired symbol and counts how many times the row or column containing the desired choice flashes. P300 is elicited only when the desired row or column flashes. Thus, the BCI uses this effect to determine the target symbol. Due to the low signal-to-noise ratio in EEG signals, the detection of target symbols from a single trial is very difficult. The rows or columns must be flashed several times for each choice. The epochs corresponding to each row or column are averaged over the trials, in order to improve their accuracy. However, these repetitions decrease the number of choices per minute, e.g., with 15 repetitions, only two characters are spelled per minute. Although most of the applications based on P300

evoked potentials employ visual stimuli, auditory stimuli have been used for people with visual impairment.

### 2.5.4 Sensorimotor Rhythms (mu and beta rhythms)

Sensorimotor rhythms comprise mu and beta rhythms, which are oscillations in the brain activity localized in the mu band $7-13$ Hz, also known as the Rolandic band, and beta band $13-30$ Hz, respectively. Both rhythms are associated in such a way that some beta rhythms are harmonic mu rhythms, although some beta rhythms may also be independent. The amplitude of the sensorimotor rhythms varies when cerebral activity is related to any motor task although actual movement is not required to modulate the amplitude of sensorimotor rhythms. Similar modulation patterns in the motor rhythms are produced as a result of mental rehearsal of a motor act without any motor output. Sensorimotor rhythms have been used to control BCIs, because people can learn to generate these modulations voluntarily in the sensorimotor rhythms. Sensorimotor rhythms can endure two kinds of amplitude modulations known as event-related desynchronization (ERD) and event-related synchronization (ERS) that are generated sensory stimulation, motor behaviour, and mental imagery. ERD involves an amplitude suppression of the rhythm and ERS implies amplitude enhancement. The mu band ERD starts $2.5s$ before movement on-set, reaches the maximal ERD shortly after movement-onset, and recovers its original level within a few seconds. In contrast, the beta rhythm shows a short ERD during the movement initiation of movement, followed by ERS that reaches the maximum after movement execution. This ERS occurs while the mu rhythm is still attenuated. Gamma rhythms reveal an ERS shortly before movement-onset.

## 2.6 Features Extraction and Selection

Different thinking activities result in different patterns of brain signals. BCI is seen as a pattern recognition system that classifies each pattern into a class according to its features. BCI extracts some features from brain signals that reflect similarities to a certain class as well as differences from the rest of the classes. The features are measured or derived from the properties of the signals which contain the discriminative information needed to distinguish their different types [8].

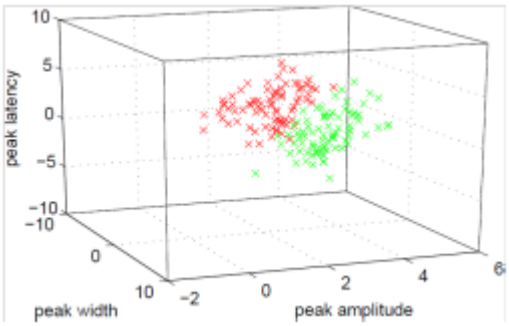


Figure 2.2: Two element feature vectors for all exciting trials in red and non-exciting trials in green

The design of a suitable set of features is a challenging issue. The information of interest in brain signals is hidden in a highly noisy environment, and brain signals comprise a large number of simultaneous sources. A signal that may be of interest could be overlapped in time and space by multiple signals from different brain tasks. For that reason, in many cases, it is not enough to use simple methods such as a band pass filter to extract the desired band power.

Brain signals can be measured through multiples channels. Not all information provided by the measured channels is generally relevant for understanding the underlying phenomena of interest. Dimensionality reduction techniques such as principal component analysis (PCA) or independent component analysis (ICA) can be applied to reduce the dimension of the original data, removing the irrelevant and redundant information. Computational costs are thereby reduced. Brain signals are inherently non-stationary. Time information about when a certain feature occurs should be obtained. Some approaches divide the signals into short segments and the parameters can be estimated from each segment. However, the segment length affects the accuracy of estimated features. FFT performs very poorly with short data segments. Wavelet transform or adaptive autoregressive components are preferred to reveal the non-stationary time variations of brain signals. A two class feature space is shown in Fig. 2.2.

Multiples features can be extracted from several channels and from several time segments before being concatenated into a single feature vector. One of the major difficulties in BCI design is choosing relevant features from the vast number of possible features. High dimensional feature vectors are not desirable due to the "curse of dimensionality" in training classification algorithms. The feature selection may be attempted examining all possible subsets of the features. However, the number of possibilities grows exponentially,

making an exhaustive search impractical for even a moderate number of features. Some more efficient optimization algorithms can be applied with the aim of minimizing the number of features while maximizing the classification performance.

For dimensionality reduction principle Component analysis or Independent Component analysis are used. AutoRegressive Components (AR), Matched Filtering (MF), Wavelet Transform (WT), Common Spatial Pattern (CSP), Genetic Algorithm (GA), Sequential Selection were used as feature extraction and selection methods. Among them CSP is commonly used and its detailed explanation will be given in next chapter.

## 2.7 Artifacts in BCIs

Artifacts are undesirable signals that contaminate brain activity and are mostly of non-cerebral origin. Since the shape of neurological phenomenon is affected, artifacts may reduce the performance of BCI-based systems. artifacts may be classified into two major categories:

- physiological artifacts
- non-physiological or technical artifacts

Physiological artifacts are usually due to muscular, ocular and heart activity, known as electromyography (EMG), electrooculography (EOG), and electrocardiography (ECG) artifacts respectively. EMG artifacts, which imply typically large disturbances in brain signals, come from electrical activity caused by muscle contractions, which occur when patients are talking, chewing or swallowing. EOG artifacts are produced by blinking and other eye movements. Blinking makes generally high-amplitude patterns over brain signals in contrast to eye movements which produce low-frequency patterns. These electrical patterns are due to the potential difference between the cornea and the retina, as their respective charges are positive and negative. For that reason, the electric field around the eye changes when this dipole moves. EOG artifacts mostly affect the frontal area, because they are approximately attenuated according to the square of the distance. Finally, ECG artifacts, which reflect heart activity, introduce a rhythmic signal into brain activity.

Technical artifacts are mainly attributed to power-line noises or changes in electrode impedances, which can usually be avoided by proper filtering or shielding. Therefore, the

BCI community focuses principally on physiological artifacts, given that their reduction during brain activity acquisition is a much more challenging issue than non-physiological artifact handling. Common methods for removing artifacts in EEG are linear filtering, linear combination and regression, ICA and PCA.

## 2.8 Classification Algorithms

The aim of the classification step in a BCI system is recognition of a user's intentions on the basis of a feature vector that characterizes the brain activity provided by the feature step [9]. Either regression or classification algorithms can be used to achieve this goal, but using classification algorithms is currently the most popular approach. Classification algorithms have traditionally been calibrated by users through supervised learning using a labelled data set. It is assumed that the classifier is able to detect the patterns of the brain signal recorded in online sessions with feedback. However, this assumption results in a reduction in the performance of BCI systems, because the brain signals are inherently non-stationary. The patterns observed in the experimental samples during calibration sessions may be different from those recorded during the online session. On the other hand, progressive mental training of the users or even changes in concentration, attentiveness, or motivation may affect the brain signals. Therefore, adaptive algorithms are essential for improving BCI accuracy. Adaptation to non-stationary signals is particularly necessary in asynchronous and non-invasive BCIs. K-Nearest Neighbour Classifier (k-NNC), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Bayesian Statistical Classifier, Artificial Neural Network (ANN) were the main classification algorithms used.

## 2.9 Applications

Any device that can be connected to a computer or to a microcontroller could be controlled with a BCI. In practice however, the set of devices and applications that can be controlled with a BCI is limited. Some of the applications possible with current BCIs are described [1].

### 2.9.1 Spelling Devices

Spelling devices allow severely disabled users to communicate with their environment by sequentially selecting symbols from the alphabet. One of the first spelling devices mentioned in the BCI literature is the P300 speller. A SCP based system in which the alphabet is split into two halves and subjects can select one halve by producing positive or negative SCPs. The selected halve is then again split into two halves and this process is repeated recursively until only one symbol remains.

### 2.9.2 Environment Control

Environment control systems allow controlling electrical appliances with BCI. P300 and SSVEP based synchronous BCI system are already developed. Development of asynchronous BCI systems is probably the most important research topic to advance the area of environment control systems.

### 2.9.3 Wheelchair Control

Disabled subjects are almost always bound to wheelchairs. A BCI can potentially be used to steer a wheelchair. Steering a wheelchair is a complex task and wheelchair control has to be extremely reliable, the possible movements of the wheelchair are strongly constrained in current prototype systems. Both P300 and oscillatory process based wheelchair control systems have developed.

### 2.9.4 Neuromotor Prostheses

The idea underlying research on neuromotor prostheses is to use a BCI for controlling movement of limbs and to restore motor function in amputees. Different types of neuromotor prostheses can be envisioned depending on the information transfer rate a BCI provides. If neuronal ensemble activity is used as control signal, high information transfer rates are achieved and 3D robotic arms can be controlled. If an EEG based BCI is used, only simple control tasks can be accomplished.

### 2.9.5 Gaming and Virtual Reality

Besides the applications targeted towards disabled subjects, prototypes of gaming and virtual reality applications have been described. Examples for such applications are the control of a spaceship with oscillatory brain activity, the control of an animated character in an immersive 3D gaming environment with SSVEPs.

# Chapter 3

# Feature Extraction and Classification

## 3.1 Introduction

The goal a BCI system is to generate a control signal based on the recorded EEG signal from scalp. There is usually a set of control signal that need to be generated based on the captured EEG signal. So we need to classify the EEG signals in order to generate a proper control signal according to thought. But the EEG signal will be highly contaminated with various artifacts like EOG, scalp movement, eye blinks etc., such artifacts are to be removed before classification. This is fulfilled in pre-processing stage, and kind of pre-processing used may vary depending on goal of the application.

After pre-processing, the data is classified. Classification techniques like LDA, Gaussian Mixture Modelling (GMM) etc., may be directly applied to raw signal, But the raw signal is very high dimensional, there are too many parameters to fine tune so the direct classification usually will not work well. So additional mapping called feature extraction, from raw signal segments on to feature vectors. These feature vectors will be low dimensional and will easily be handled by machine learning stage.

## 3.2 Common Spatial Pattern Algorithm

Using spatial filters, mental activities such as imagined movement can be extracted from the EEG signals. A technique to compute these spatial filters is the common spatial patterns (CSP) algorithm [10] [11]. The features exploited by this technique in its original form are Event- related Synchronization and Desynchronization (ERD\ERS) localized in the sensori-motor cortex, but the method is not restricted to these applications. The CSP algorithm uses labelled trials to produce a transformation that maximizes the variance for one class while minimizing the variance for the other class. The difference in variance can

be used to classify a fragment of EEG signals into one of two classes. CSP was originally introduced in [12] and first applied to EEG in [13]. Due to its simplicity, speed and relative robustness, CSP is the foremost technique for oscillatory processes, and it can be used to get a quick estimate of whether the data contains information of interest or not. It was reported that CSP algorithm was used more popular than others such as standard ear-reference, common average reference (CAR), small Laplacian (3 cm to set of surrounding electrodes), large Laplacian (6 cm to set of surrounding electrodes), Principal Component Analysis (PCA) and Independent Components Analysis (ICA). CSP uses log-variance features over a single non-adapted frequency range (which may have multiple peaks), and neither temporal structure (variations) in the signal is captured, nor are interactions between frequency bands. The major strength of the CSP algorithm is its adaptive spatial filter that it can capture.

Prior to calculation of the spatial filters, the reference EEG signal is filtered in an $(8 - 30)$ Hz band. The frequency band was chosen because it encompasses the alpha and beta frequency bands, which have been shown to be most important for movement classification. Furthermore, in a recent movement study [14], it was shown that a broad frequency band (e.g., $8 - 30$ Hz) gives better classification results compared to narrow bands. The spatial filtering projects the channels of the original signal down to a small set of (usually $4 - 6$ Hz) replacement channel, where the (linear) mapping is optimized such that the variance in these channels is maximally informative with respect to the prediction task. CSP can also be applied to independent components to rate their importance or for better artifact robustness. A wide range of classifiers can be used with CSP features, the most commonly used one being LDA. There exists a large collection of CSP variants and extensions, mostly to give better control over spectral filtering, including multiband CSP [15], Spectrally Weighted CSP [16] [17], Invariant CSP, Common Spatio-Spectral Patterns (CSSP), Common Sparse Spectral Spatial Pattern (CSSSP), Regularized CSP (RCSP), and several others.

The complexity of EEG signals has restricted a further exploring for the overfitting of CSP. The complexity mainly comes from three aspects, the non-stationarity of the signals, the unclear internal mechanisms, and the large variance of the classification results. First, the non-stationarity of the EEG signals refers to the non-stationary in the single trial and between trials. Second, in BCI application, the EEG signals are often seen as the output of

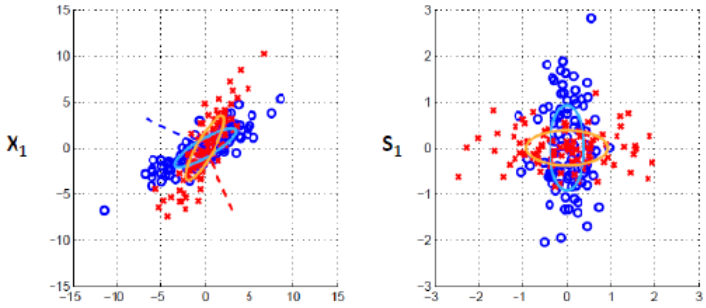


Figure 3.1: Geometric approach before and after CSP filtering

a black-box. Due to the complexity of the brain, it is still not very clear how the signals are produced from a well-defined paradigm (e.g. motor imagery). Third, the classification results are of high variation. Since the discovery of ERD, it has been found that the classification accuracy over subjects is quite different. But what influences the results is still unknown. To get a thorough understanding for some methods, a large number of data sets with hundreds of trials from tens of subjects are needed, which increases the cost of investigation.

The aim of CSP in a 2 class problem is to design a pair of spatial filters (i.e., spatial transforms) such that the filtered signal's variance is maximal for one class while minimal for the other (Fig. 3.1). The method used to design such spatial filters is based on the simultaneous diagonalization of two covariance matrices. This method, called the method of common spatial patterns, has been introduced to EEG analysis for detection of abnormal EEG and was recently applied successfully to the classification of movement-related EEG.

The CSP filters can be obtained from the per-class signal covariance matrices in 3 ways.

1. Geometric Approach
2. Generalized Eigen value problem
3. Optimization approach

### 3.2.1 Geometric Approach

A more intuitive approach is a three-step procedure:

1. Determine a whitening transform U for the average of both covariance matrices using PCA.

2. Apply it to one of the matrices and calculate its principal components P.

3. The spatial filter operation $W$ is to first whiten by $U$ and then transform by $P^{-1}$, i.e., W = $P^{-1}U$ so then $Z = WX$.

The raw EEG data of a single trial $d \times N$ is represented as a matrix $X$ , where $d$ is the number of channels (i.e., recording electrodes) and $N$ is the number of samples per channel. The normalized spatial covariance of the EEG can be obtained from

$$\sum = \frac{XX^T}{trace\,(XX^T)} \tag{3.1}$$

where $T$ denotes the transpose operator and trace $\left(XX^T\right)$ is the sum of the diagonal elements of $XX^T$. For each of the two distributions to be separated the spatial covariance is calculated by averaging over the trials of each group. i.e. per-class average covariance matrices

$$\sum^{(c)} = \langle \sum_t \rangle^c \quad c \in (+1, -1) \tag{3.2}$$

The composite spatial covariance is given as

$$\sum = \sum^{(+1)} + \sum^{(-1)} \tag{3.3}$$

where both $\sum^{(+1)}$ and $\sum^{(-1)}$ are $d \times d$ channel covariances matrices. $\sum$ can be factored as $\sum = U\lambda U^T$, where $U$ is the matrix of eigenvectors and $\lambda$ is the diagonal matrix of eigenvalues. The eigenvalues are assumed to be sorted in descending order.

The whitening transformation

$$P = \sqrt{\lambda^{-1}}U^T \tag{3.4}$$

Equalizes the variances in the space spanned by $U$, i.e., all eigen values of $P \sum P^T$ are equal to one. If $\sum^{(+1)}$ and $\sum^{(-1)}$ are transformed as

$$S^{(+1)} = P\Sigma^{(+1)}P^T \quad and \quad S^{(-1)} = P\Sigma^{(-1)}P^T \tag{3.5}$$

where $S^{(+1)}$ and $S^{(-1)}$ share common eigen vectors .i.e.

$$S^{(+1)} = B\lambda_{(+1)}B^T \tag{3.6}$$

$$S^{(-1)} = B\lambda_{(-1)}B^T \tag{3.7}$$

$$\lambda_{(+1)} + \lambda_{(-1)} = I \tag{3.8}$$

where I is the identity matrix. Since the sum of two corresponding eigenvalues is always one, the eigenvector with largest eigenvalue for $S^{(+1)}$ has the smallest eigenvalue for $S^{(-1)}$ and vice versa. This property makes the eigenvectors B useful for classification of the two distributions. The projection of whitened EEG onto the first and last eigenvectors in (i.e., the eigenvectors corresponding to the largest$\lambda_{(+1)}$ and $\lambda_{(-1)}$ will give feature vectors that are optimal for discriminating two populations of EEG in the least squares sense.

With the projection matrix $W = (B^{-1}P)^T$, the decomposition (mapping) of a trial X is given as

$$Z = WX \tag{3.9}$$

The columns of $W^{-1}$ are the common spatial patterns and can be seen as time-invariant EEG source distribution vectors.

The features used for classification are obtained by decomposing (filtering) the EEG according to Eqn. 3.10. For each direction of imagined movement, the variances of only a small number $'m'$ of signals most suitable for discrimination are used for the construction of the classifier. The signals $Z_P\,(1, 2.....2m)$ that maximize the difference of variance of left versus right motor imagery EEG are the ones that are associated with the largest eigenvalues $\lambda_{(+1)}$ and $\lambda_{(-1)}$. These signals are the $'m'$ first and last rows of $Z$ due to the calculation of $W$.

$$f_p = \log\left(\frac{var\,(Z_p)}{\sum_{i=1}^{2m} var\,(Z_p)}\right) \tag{3.10}$$

The feature vectors $f_p$ of left and right trials are used to calculate a linear classifier. The log-transformation serves to approximate normal distribution of the data.

### 3.2.2 Generalized Eigen value problem

Given a set of $t$ trial segments $X_t \in R^{d\times N}$ , where $d$ is the number of channels and $N$ is the number of samples, per-trial channel covariance matrices $\Sigma_t = X_t X_t^T \in R^{d\times d}$,

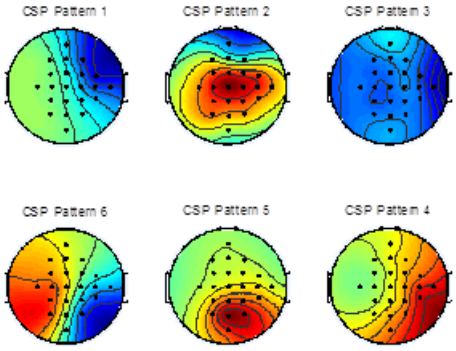


Figure 3.2: Spatial Pattern

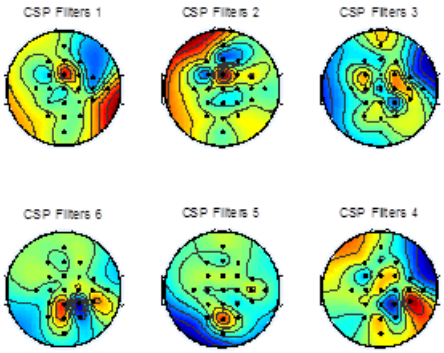


Figure 3.3: Spatial Filter

and per-class average covariance matrices $\Sigma^{(c)} = \langle \Sigma_t \rangle^c$. The generalized eigen value problem [18] [19] is to determine the nontrivial solutions of the equation

$$\Sigma^{(+1)} X = \lambda \left( \Sigma^{(+1)} + \Sigma^{(-1)} \right) X \tag{3.11}$$

where both $\Sigma^{(+1)}$ and $\Sigma^{(-1)}$ are $d \times d$ channel covariance matrices and $\lambda$ is a diagonal matrix of Eigen values. The values of $\lambda$ that satisfy the equation are the generalized eigenvalues and the corresponding values of $X$ are the generalized right eigenvectors. Since, $\left( \Sigma^{(+1)} + \Sigma^{(-1)} \right)$ is non-singular, the problem could be solved by reducing it to a standard eigenvalue problem.

$$\left( \Sigma^{(+1)} + \Sigma^{(-1)} \right)^{-1} \Sigma^{(+1)} X = \lambda X \tag{3.12}$$

The columns of $X$ give spatial filter coefficients and rows of $X^{-1}$ gives spatial pattern. Usually we take $1-3$ pair of filter or patterns from either side as in Fig. 3.3 and Fig. 3.2.

### 3.2.3 Optimization Approach

In the optimization approach [20] [21], given a set of $t$ trial segments $X_t \in R^{d \times N}$, per-trial channel covariance matrices $\Sigma_t = X_t X_t^T \in R^{d \times d}$, and per-class average covariance matrices $\Sigma^{(c)} = \langle \Sigma_t \rangle^c$, optimize the spatial filter $w_c$ for class $c$ as:

$$w_c = \max_w w^T \Sigma^{(c)} w \tag{3.13}$$

$$s.t\ w^T \left( \Sigma^{(+1)} + \Sigma^{(-1)} \right) w = 1$$

where $w^T \sum w$ gives the variance in direction w. This is a quadratically constrained quadratic program, which can be solved using Sequential Quadratic Program (SQP) method from fmincon of Matlab.

## 3.3 Linear Discriminant Analysis

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant are methods used in statistics, pattern recognition and machine learning to find a linear com-

bination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification. LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. In the other two methods however, the dependent variable is a numerical quantity, while for LDA it is a categorical variable (i.e. the class label). Logistic regression and probit regression are more similar to LDA, as they also explain a categorical variable. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method. Rather than the ANOVA categorical independent variables and a continuous dependent variable, discriminant analysis has continuous independent variables and a categorical dependent variable.

LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made. LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis [22] [23].

The terms Fisher's linear discriminant and LDA are often used interchangeably, although Fisher's original article actually describes a slightly different discriminant, which does not make some of the assumptions of LDA such as normally distributed classes or equal class covariances. There are many possible techniques for classification of data. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two commonly used techniques for data classification and dimensionality reduction. Linear Discriminant Analysis easily handles the case where the within-class frequencies are unequal and their performances have been examined on randomly generated test data. This

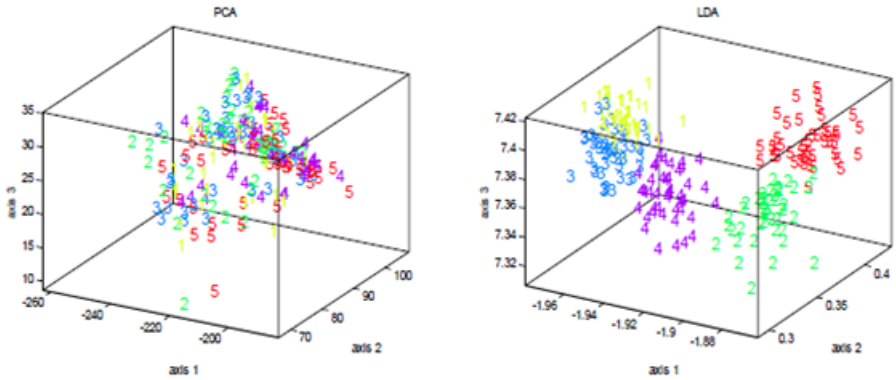


Figure 3.4: From the 3D scatter plots it is clear that LDA outperforms PCA in terms of class discrimination n this is one example where the discriminatory information is not aligned with the direction of maximum variance

method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. The use of Linear Discriminant Analysis for data classification is applied to classification problem in speech recognition. The prime difference between LDA and PCA is that PCA does more of feature classification and LDA does data classification. In PCA, the shape and location of the original data sets changes when transformed to a different space whereas LDA doesnt change the location but only tries to provide more class separability and draw a decision region between the given classes. This method also helps to better understand the distribution of the feature data. A comparison of PCA and LDA in three dimensional space is shown in Fig. 3.4.

### 3.3.1 Mathematical Operation

Assume we have $m$-dimensional samples $\{x_1, x_2, , x_N\}$, $N1$ of which belong to $w_1$ and $N2$ belong to $w_2$. We seek to obtain a scalar $y$ by projecting the samples $x$ onto a line(C-1 space, C = 2).

$$y = w^T x, \quad x = \begin{pmatrix} x_1 \\ . \\ . \\ x_m \end{pmatrix} \quad w = \begin{pmatrix} w_1 \\ . \\ . \\ w_m \end{pmatrix} \tag{3.14}$$

Of all the possible lines we would like to select the one that maximizes the separability of the scalars. Let us consider $\mu_1$ and $\mu_2$ are the means of the two classes.

In order to find a good projection vector, we need to define a measure of separation

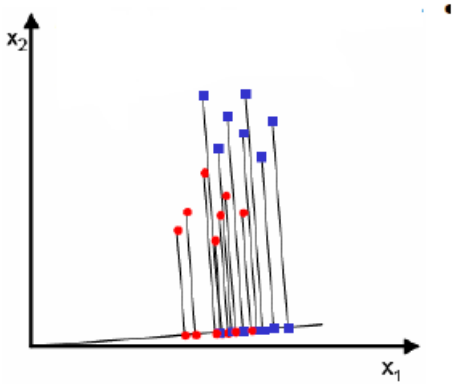


Figure 3.5: The two classes are not well separated when projected onto this line

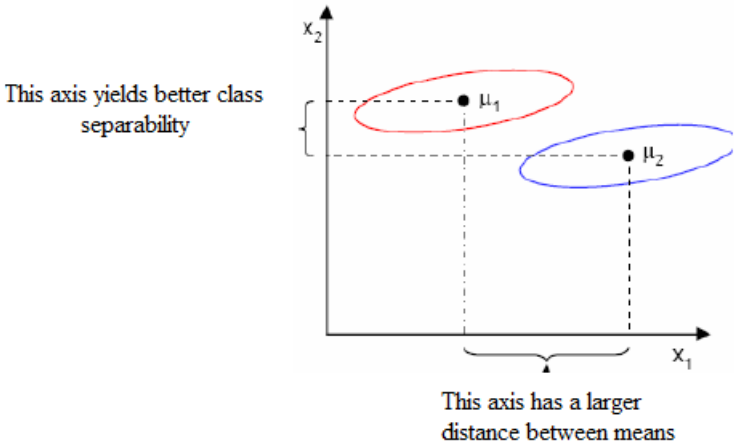


Figure 3.6: This line succeeded in separating the two classes and in the mean time reducing the dimensionality of our problem from two features $(x1, x2)$ to only a scalar value $y$.

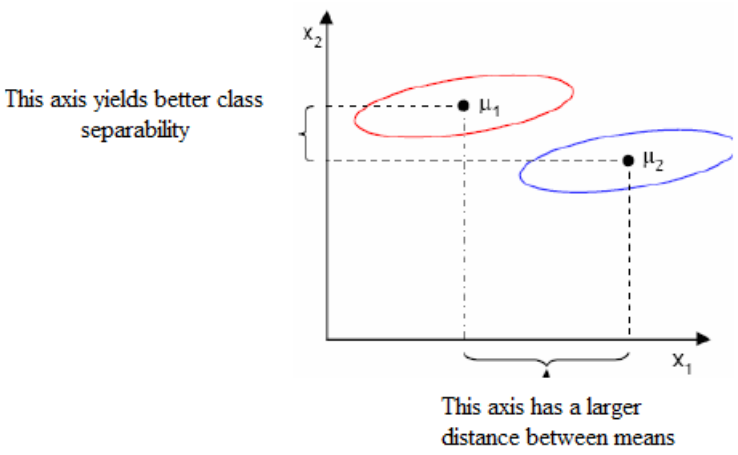


Figure 3.7: Befor Linear Discrimination

between the projections. An example of projection using a bad projection vector is shown in Fig. 3.5, and using a good projection vector is shown in Fig. 3.6.

$$\tilde{\mu_1} = \frac{1}{N} \sum_{y \in w_i} y = w^T \mu_i \tag{3.15}$$

$\tilde{\mu_1}$ represents the mean vector of each class in $y$ feature space. We could then choose the distance between the projected means as our objective function.

$$J(w) = \mid \tilde{\mu_1} - \tilde{\mu_2} \mid = \mid w^T \mu_1 - w^T \mu_2 \mid = \mid w^T (\mu_1 - \mu_2) \mid \tag{3.16}$$

However, the distance between the projected means is not a very good measure since it does not take into account the standard deviation within the classes.

The solution proposed by Fisher is to maximize a function that represents the difference between the means, normalized by a measure of the within-class variability, or the so-called 'scatter'. An equivalent of the variance such as scatter is calculated for each of the classes. Fig. 3.7 shows before linear discrimination of two classes and Fig. 3.8 shows after linear discrimination of two classes.

For each class we define the scatter, an equivalent of the variance and is given by

$$\tilde{S_l^2} = \sum_{y \in w_i} (y - \mu_i)^2 \tag{3.17}$$

where $\tilde{S_l^2}$ measures the variability within class $w_i$ after projecting it on the $y$-space. Thus $\tilde{S_1^2} + \tilde{S_2^2}$ measures the variability within the two classes at hand after projection, hence

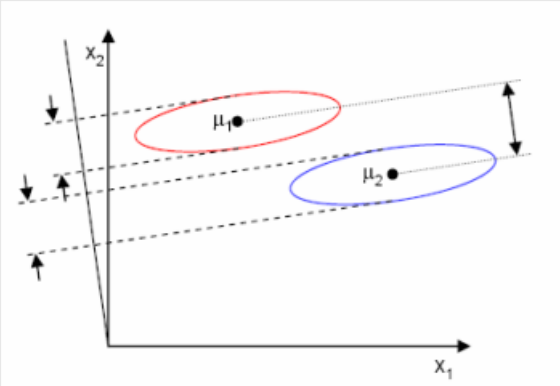


Figure 3.8: After Linear Discrimination.

it is called within-class scatter of the projected samples. The Fisher linear discriminant is defined as the linear function $w^T x$ that maximizes the criterion function:

$$J(w) = \frac{|\tilde{\mu_1} - \tilde{\mu_2}|^2}{\tilde{S_1^2} + \tilde{S_2^2}} \tag{3.18}$$

Therefore, looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible.

In order to find the optimum projection $w*$, express $J(w)$ as an explicit function of $w$. The scatter matrices are proportional to the covariance matrices and $S_B$ is the "between classes scatter matrix" and $S_W$ is the "within classes scatter matrix". A measure of the scatter in multivariate feature space x which are denoted as scatter matrices is defined as;

$$S_i = \sum_{x \in w_i} (x - \mu_i)(x - \mu_i)^T \tag{3.19}$$

$$S_w = S_1 + S_2 \tag{3.20}$$

$$S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{3.21}$$

where $S_i$ is the covariance matrix of class $w_i$, and finally express the Fisher criterion in terms of $S_W$ and $S_B$ as:

$$J(w) = \frac{|\tilde{\mu_1} - \tilde{\mu_2}|^2}{\tilde{S_1^2} + \tilde{S_2^2}} = \frac{w^T S_B w}{w^T S_W w} \tag{3.22}$$

Hence $J(w)$ is a measure of the difference between class means (encoded in the between-class scatter matrix) normalized by a measure of the within-class scatter matrix.

After differentiation and equating to zero, to find the maximum of $J(w)$,

$$S_W^{-1} S_B w - J(w) w = 0 \tag{3.23}$$

On solving the generalized eigen value problem, we get

$$S_W^{-1} S_B w = \lambda w \quad where \quad \lambda = J(w) = scalar$$

yields

$$w^* = arg \max_w \left( \frac{w^T S_B w}{w^T S_W w} \right) = S_W^{-1} (\mu_1 - \mu_2) \tag{3.24}$$

This is known as Fishers Linear Discriminant, although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension.

The problem of maximizing $J$ can be transformed into the following constrained optimization problem as

$$\min_w -\frac{1}{2} w^T S_B w \quad s.t \quad w^T S_W w = 1 \tag{3.25}$$

### 3.3.2 Limitations of LDA

- LDA produces at most $C-1$ feature projections If the classification error estimates establish that more features are needed, some other method must be employed to provide those additional features
- LDA is a parametric method (it assumes unimodal Gaussian likelihoods) If the distributions are significantly non-Gaussian, the LDA projections may not preserve complex structure in the data needed for classification.
- LDA will also fail if discriminatory information is not in the mean but in the variance of the data

# Chapter 4

# A One Step Feature Extraction and Classification with Regularization for BCI

## 4.1 Introduction

A regularized discriminative framework for EEG analysis is a framework for analysis of electroencephalography (EEG) signal that unifies various tasks such as feature extraction, feature selection, feature combination, and feature classification [14]. These tasks are often independently tackled under a regularized empirical risk minimization problem. The features are automatically learned, selected and combined through a convex optimization problem.

The framework is applied to two typical BCI problems, namely the ECoG dataset and the prediction of limb movement based on motor imagery data. In both datasets, the proposed approach shows competitive performance against conventional methods, while at the same time the results are easily accessible to neurophysiological interpretation.

## 4.2 Signal Analysis Framework

In this section a discriminative learning framework for BCI is presented. The framework consists of three components. The first one is a probabilistic predictor model that is used for both decoding the intention of a user and learning the predictor model from a collection of trials. The second component is the design of a detector function. The last component is how to appropriately control the complexity of the detector function. These three issues are presented under Discriminative learning, Detector function, and

Regularization sections, respectively.

### 4.2.1 Discriminative Learning

In any BCI system, the goal of signal analysis is to construct a function that predicts the intention of a user from his/her brain signal. In our discriminative approach we are interested in the whole function from the brain signal to the probability distribution over possible user intention, which we call a predictor. When we deal with this type of probabilistic predictor we are facing two tasks. First, how to decode the intention of a user given the brain signal and the predictor. Second, how to learn the predictor from a collection of labelled examples. The answers to these questions are derived naturally from probability theory and statistics.

Let $X \in \chi$ be the input brain signal and let $q(Y|X)$ be the predictor, which assigns probabilities to the user's command $Y \in y$ given the brain signal $X$. The task of decoding is to find the most likely command $\hat{y}$ given the input $X$ and the predictor $q$ as follows:

$$\tilde{y} = arg \max_{y \in Y} q\left(Y = y \mid X\right) \tag{4.1}$$

The task of learning is to find a predictor from a suitably chosen collection of candidates, which we call a model, and we assume that a model is parametrized by a parameter vector $\theta \in \Theta$. The predictor specified by $\theta$ is denote as $q_\theta$; thus the model is a set $\{q_\theta : C \in \Theta\}$. In order to say how a predictor $q_\theta$ compares to another predictor $q_\theta^{'}$, it is necessary to define a loss function. We can consider the probability that the predictor assigns to each possible user intention $y$ as the pay off the predictor can obtain if the actual intention coincides with the prediction; the predictor can choose its strategy between equally distributing the probability mass over all the possible outcomes and concentrating it on a single output that is based on the brain signal $X$. This pay off is commonly measured in the logarithmic scale. The loss function is thus defined as the negative logarithmic pay off (or the Shannon information content in information theory) as follows:

$$\ell_L\left(\left(X, y\right), \theta\right) = -\log q_\theta\left(Y = y \mid X\right) \tag{4.2}$$

where, $X$ is the brain signal and $y$ is the true intention of the user. Thus the loss

is smaller if the predictor predicts the actual intention of the user with high confidence. Suppose we are given a collection of input signal $X_i$ and true intention $y_i$, which we denote $\{X_i, y_i\}_{i=1}^n$. It is reasonable to choose the parameter $\theta$ that minimizes the empirical average of losses

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((X_i, y_i), \theta\right) \tag{4.3}$$

However, often the complexity of the class of predictors $q_\theta$ is very large and the minimization of $L_n(\theta)$ leads to over fitting due to small sample size. Therefore, we learn the parameter $\theta$ by solving the following constrained minimization problem:

$$\min_{\theta \in \emptyset} L_n(\theta) \;\; s.t \quad \Omega(\theta) \leq C \tag{4.4}$$

The second term $\Omega(\theta)$ is called the regularizer and it measures the complexity of the parameter configuration $\theta$. $C$ is a hyper-parameter that controls the complexity of the model and is selected by cross validation. A complexity function induces a nested sequence of subsets $\Theta_C := \{\theta \in \Theta : \Omega(\theta) \leq C\}$, which is parameterized by the bound C on the complexity; i.e., $C1 < C2 < C3 < ...$ implies $\Theta_{C1} < \Theta_{C2} < \Theta_{C3}...$ and vice versa. Therefore we can consider a sequence of predictors that we obtain through the learning framework at monotonically increasing level of complexity.

If we suppose that the training examples $\{X_i, y_i\}_{i=1}^n$ are sampled independently and identically from some probability distribution $p(X, Y)$, the above function $L_n(\theta)$ can be considered as the empirical version of the following function.

$$L(\theta) = D\left(p(Y \mid X) \parallel q_\theta(Y \mid X)\right) + H\left(p(Y \mid X)\right) \tag{4.5}$$

where $D(p \parallel q)$ is the Kullback-Leibler divergence between two probability distributions $p$ and $q$; the second term is the conditional entropy of $Y$ given $X$ and is a constant that does not depend on the model parameter $\theta$.

### Logistic Model

The logistic regression [24] model is a popular model in a binary decision setting. The logistic model assumes the user command $Y$ to be either one of the two possibilities; e.g., $Y = -1$ and $Y = +1$ for left and right-hand movement, respectively. The logistic

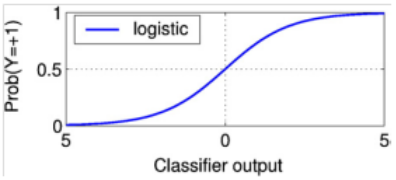


Figure 4.1: Logistic Function

predictor $q_\theta$ is defined through a latent function $f_\theta$; we define a real valued function $f_\theta$ which outputs a positive number if $Y = +1$ is more likely than $Y = -1$ and vice versa. Then a logistic function u(z)=1/(1+exp(-z)) is applied to the output $f_\theta(X)$ to convert it into the probability of $Y = +1$ given $X$; similarly applying the logistic function to $-f(X)$ gives the probability of $Y = -1$ given $X$. Thus we have the following expression for the predictor:

$$q_\theta\left(Y = y \mid X\right) = \frac{1}{1 + \exp\left(-y f_\theta\left(X\right)\right)} \quad \left(y \in \{1, -1\}\right) \tag{4.6}$$

In fact, under the predictor $q_\theta$ defined above, the log likelihood ratio of $Y = +1$ to $Y = -1$ given $X$ is precisely the latent function value $f_\theta(X)$ as follows.

$$\log \frac{q_\theta\left(Y = +1 \mid X\right)}{q_\theta\left(Y = -1 \mid X\right)} = f_\theta\left(X\right) \tag{4.7}$$

The loss function for the logistic model is called the logistic loss (shown in Fig. 4.1)and can be written as follows:

$$\ell_L\left(\left(X, y\right), \theta\right) = \log\left(1 + \exp\left(-y f_\theta\left(X\right)\right)\right) \tag{4.8}$$

which is obtained by taking the negative logarithm of Eqn. 4.7. As shown above, it is often a useful strategy to construct a model as a combination of a class of functions that converts the input signal into a scalar value and a link function that converts this value into the probability of the command $Y$. The function $f_\theta$ is called a detector because in the BCI context it captures some characteristic spatio-temporal activity in the brain; a class of functions parametrized by $\theta \in \Theta$ is called a detector model.

### 4.2.2 Detector Function

We use the following linear detector function:

$$f_\theta(X) = \langle W, X \rangle + b \tag{4.9}$$

$\theta(W, b)$ is a matrix of some appropriate size and $b \in R$ is a bias term. $\langle W, X \rangle = \sum_{ij} W(i,j)X(i,j)$ is the inner product between two matrices $X$ and $W$. $W(i,j)$ denotes the $(i,j)_{th}$ element of a matrix $W$.

$X \in S_+^C$ is covariance of appropriately filtered EEG signal, i.e., $X$ and $W$ are both $C \times C$ matrices. The detector is called the second-order detector in this case. This model can be used to detect slow change in the cortical potential also, in that case we use first order feature.

Since we are interested in the second-order information such as variance and covariance, we set $X$ as a block diagonal concatenation of these terms as in Eqn. 4.10:

$$\begin{pmatrix} \frac{1}{\eta_{(1)}}\Xi^{(1)} & & \\ & \frac{1}{\eta_{(2)}}\Xi^{(2)} & \\ & & \frac{1}{\eta_{(k)}}\Xi^{(k)} \end{pmatrix} \tag{4.10}$$

where $\Xi^{(1)}$ is the first-order term ($X$ in the above first-order model) and $\Xi^{(k)} = cov(X^k)$ is the covariance matrix of a short segment of band-pass filtered EEG $X(k)$ for $k = 1, ., .K$. Here we consider $K$ second-order terms that are filtered by different (possibly overlapped) band-pass filters. We call $X$ the augmented input matrix and the corresponding $W$ the augmented weight matrix. The normalization factor $\eta*$ is introduced in order to prevent biasing the selection of terms with large power or large size; it is defined as the square root of the total variance of each block element, i.e.,

$$\eta_* = \sqrt{\sum_k var\left(\Xi\left(k\right)\right)} \tag{4.11}$$

Where $k \in \{1, 2, ., K\}$. This choice is motivated by the common practice in the $\ell_1$-regularization to standardize each feature to unit variance. In fact, when all the block diagonal matrices are $1 \times 1$, the dual spectral (DS) regularization reduces to $\ell_1$ regularization (lasso) and the above $\eta*$ reduces to the standard deviation of each feature. It can

be shown that when we learn the augmented weight matrix $W$ under suitable regularization, the weight matrix turns out to have the same block diagonal structure as the input $X$. This model is called the second-order detector. This model can be used to detect oscillatory features such as event-related desynchronization which is useful in detecting real or imagined movement. In these tasks it is known that both the slow change in the cortical potential and the event related desynchronization are useful features to predict the movement. We combine these features in the block diagonal form.

### 4.2.3 Regularization

In this section we present Dual Spectral regularizer. The DS regularizer is defined as the linear sum of singular-values of the weight matrix $W$, which is called the DS norm [25] [26].

$$\Omega_{DS}(\theta) = \sum_{j=1}^{r} \sigma_j(W) \tag{4.12}$$

where $\sigma_j(W)$ is the $j^{th}$ singular value of the weight matrix $W$ and $r$ is the rank of $W$. The DS regularization can be considered as another generalization of the $\ell_1$-regularization; it induces sparsity in the singular-value spectrum of the weight matrix $W$. That is, it induces low-rank matrix $W$. Similarly to group-lasso, when a singular component is switched off, all the degrees of freedom associated to that component (i.e., left and right singular vectors) are simultaneously switched off. However in contrast to group-lasso regularizer, there is no notion of any group a priori. The DS regularization automatically tunes the feature detectors as well as the rank of $W$.

In machine learning literature, the low-rank enforcing property of the dual spectral norm is well known and has been used in applications such as collaborative filtering , multi-class classification multi-output prediction. The DS regularizer induces the positive semidefinite cone constraint. In fact, mathematically these cones are understood as generalizations of the positive-orthant cone induced by the $\ell_1$ (lasso) regularizer.

### 4.2.4 Discussion on Discriminative Approach

Major difference in various discriminative approaches arises in the parameterization of the detector function $f_\theta(X)$. Detector function used here is discussed below.

### Second Order Feature based BCI

One of the most successful approach in motor-imagery based BCI is common spatial pattern (CSP). A commonly used form of CSP based detector model can be written as in Eqn. 4.2.4.

$$f_\theta(X) = \sum_{j=1}^{J} \beta_j \log\left(w_j^T X^T B_j B_j^T X w_j\right) + \beta_0 \tag{4.13}$$

where $X \in R^{T\times C}$ is a short segment of multi-channel EEG measurement with $C$ channels and $T$ sampled time-points; $B_j \in R^{T\times T}$ are temporal filters, $w_j \in R^C$ are spatial filters, $\{\beta_j\}_{j=1}^J$ are weighting coefficients of the $J$ features, and $\beta_0$ is a bias term. CSP is a dimensionality reduction method based on a generalized eigen value problem. In the conventional CSP based approach, thus the classifier is trained in three steps. First, the temporal filter coefficients $B_j$ is chosen a priori or based on some heuristics. Second, the spatial filter is obtained from solving the generalized eigen value. DS regularization with following detector model is assumed:

$$f_\theta(X) = \langle W, X^T X\rangle + b \tag{4.14}$$

The above is obtained from CSP equation by omitting the logarithm, and the temporal filter coefficient $B_j$ (assumed to be constant), and denoting $W = \sum_{j=1}^{j} \beta_j w_j w_j^T$. It is demonstrated that using the DS regularization, good classification performance is obtained with only a few spatial filters $w_j$. In DS regularization the rank may be chosen as 4 or 5 which is typically the same as with CSP approach. The common practice in CSP is taking 3 filters from right end and left end of the spatial filter matrix and it is not optimal.

Models that learn good spatial filter, temporal filter and LDA weights all in one optimization problem were proposed but they suffer from local minima problem due to non-smooth function. Another method which enforces rank constraint was also proposed, with Frobenius matrix regularizer and hinge loss function.

## 4.3 Implementation

The logit of the posterior class probability can be modelled with a linear function as:

$$\log \frac{P\left(y=+1 \mid X\right)}{P\left(y=-1 \mid X\right)} = Tr[W^T X] + b \tag{4.15}$$

Negative log-likelihood of Eq 4.8 is minimized with a dual spectral norm penalization term, which is written as follows:

$$\min_{W \in R^{R \times C}, b \in R, z \in R^n} \sum_{i=1}^{n} \ell_{LR}\left(z_i\right) + \lambda \|W\|_1, \tag{4.16}$$

$$y_i \left(Tr[W^T X_i] + b\right) = z_i, \quad i = 1, ...., n$$

where $z_i (i = 1, ..., n)$ are called the latent variables, $\lambda$ is the regularization constant, and $Tr[\bullet]$ denotes the trace. Here the logistic loss $\ell_{LR}$ and the spectral $\ell_1$- norm of a matrix $\|W\|_1$ are defined as follows:

$$\ell_{LR}\left(z\right) = \log\left(1 + \exp\left(-z\right)\right) \tag{4.17}$$

$$\|W\|_1 = \sum_{c=1}^{r} \sigma[W] \tag{4.18}$$

Using two auxiliary positive semidefinite matrix $Q_1 \in S_+^R$ and $Q_2 \in S_+^C$, we can rewrite the spectral $\ell_1$-norm, which is convex but a non-differentiable function, with linear matrix inequality (LMI), as follows:

$$\min_{W \in R^{R \times C}, b \in R, z \in R^n, Q_1 \in S_+^C, Q_2 \in S_+^R} \sum_{i=1}^{n} \ell_{LR}\left(z_i\right) + \lambda\left(Tr[Q_1] + Tr[Q_2]\right), \quad i = 1, ...., n \tag{4.19}$$

$$\begin{pmatrix} \mathrm{Q}_1 & -\frac{1}{2}W \\ -\frac{1}{2}W^T & \mathrm{Q}_2 \end{pmatrix} \succeq 0$$

The transformation has been done because to change non-smooth convex dual spectral norm to a smooth convex optimization problem [27]. By minimizing some alternate variables, the rank of $W$ is minimized indirectly,thus minimizing our objective function

, where $S_+^C$, denotes symmetric positive semidefinite matrices. In our case since $W$ is symmetric $Q_1 = Q_2$, and the minimum is attained at $\frac{1}{2}USU^T$, ($\frac{1}{2}$ is for notational convenience), which specifies its convexity and on finding the determinant of above LMI, we get $(U + W) \succeq 0$ and $(U - W) \succeq 0$,i.e, the matrix $U + W$ and $U - W$ are symmetric positive semidefinite [21].

We can implement this primal problem in package like CVX [28] which is much slower. The alternate is specialized package which optimize the dual of the problem like DAL [29] package and packages which implement ADMM (alternating direction method of multipliers).

# Chapter 5

# Results and Discussion

## 5.1 Introduction

In this chapter the results achieved with this method is discussed. The algorithm is applied to various binary classification problem and resulting plots and tables are shown. For plotting topoplot function from eeglab toolbox [30] is used. A comparison of this algorithm with CSP is also given. A modification of this method with another regularization term is also given.

The data set from BCI completion IV dataset $2a$ [31] and dataset 1 of BCI competition III [32] is used. The data set consists of motor imagery EEG data from 9 subjects and ECoG recording from 1 subject respectively. The dataset description of respective data sets is given below.

## 5.2 Dataset Description

### 5.2.1 BCI Competition IV Dataset 4a

The cue-based BCI paradigm consisted of four different motor imagery tasks, namely the imagination of movement of the left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4). Two sessions on different days were recorded for each subject. Each session is comprised of 6 runs separated by short breaks. One run consists of 48 trials (12 for each of the four possible classes), yielding a total of 288 trials per session.

The subjects were sitting in a comfortable armchair in front of a computer screen. At the beginning of a trial ($t = 0s$), a fixation cross appeared on the black screen. In addition, a short acoustic warning tone was presented. After two seconds ($t = 2s$), a cue in the form of an arrow pointing either to the left, right, down or up (corresponding to one of the four classes left hand, right hand, foot or tongue) appeared and stayed on the

screen for 1.25$s$. This prompted the subjects to perform the desired motor imagery task. No feedback was provided. The subjects were asked to carry out the motor imagery task until the fixation cross disappeared from the screen at $t = 6s$. A short break followed where the screen was black again. The paradigm is illustrated in Fig. 5.1.

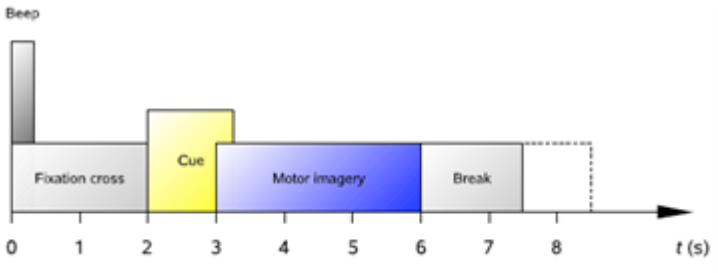


Figure 5.1: Timing scheme of the BCI paradigm

Twenty-two Ag/AgCl electrodes (with inter-electrode distances of 3.5$cm$) were used to record the EEG; the montage is shown in Fig. 5.2. All signals were recorded monopolarly with the left mastoid serving as reference and the right mastoid as ground. The signals were sampled with 250 Hz and band pass-filtered between 0.5 Hz and 100 Hz. The sensitivity of the amplifier was set to 100 $\mu V$.

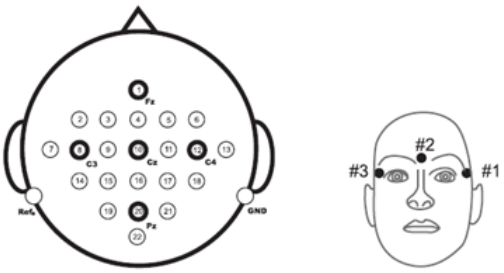


Figure 5.2: Electrode Positions

An additional 50 Hz notch filter was enabled to suppress line noise. In addition to the 22 EEG channels, 3 monopolar EOG channels were recorded and also sampled with 250 Hz (see Fig. 5.2). They were band pass filtered between 0.5 Hz and 100 Hz (with the 50 Hz notch filter enabled), and the sensitivity of the amplifier was set to 1 mV. The EOG channels are provided for the subsequent application of artifact processing methods.

### 5.2.2 BCI Competition III Dataset 1

During the BCI experiment, a subject had to perform imagined movements of either the left small finger or the tongue. The time series of the electrical brain activity was picked

up during these trials using a $8 \times 8$ ECoG platinum electrode grid which was placed on the contra lateral (right) motor cortex. The grid was assumed to cover the right motor cortex completely, but due to its size (approx. $8 \times 8$cm) it partly covered also surrounding cortex areas. All recordings were performed with a sampling rate of 1000 Hz. After amplification the recorded potentials were stored as microvolt values. Every trial consisted of either an imagined tongue or an imagined finger movement and was recorded for 3 seconds duration. To avoid visually evoked potentials being reflected by the data, the recording intervals started 0.5 seconds after the visual cue had ended. The test and training data were taken with one week gap.

## 5.3 Signal Pre-processing and Predictor Model

The dataset $2a$ is 4 class classification problem with trials corrupted by EOG artifacts. The data set consists of motor imagery EEG data from 9 subjects. Data set IIa, from BCI competition IV comprises EEG signals from 9 subjects who performed left hand, right hand, foot and tongue motor imagery. EEG signals were recorded using 22 electrodes. But here we focus on binary classification. All the pair wise combination of the performed classes produced 6 ($4C_2$ combinations) datasets per subject with 144 trials per class, so a total of $6 \times 9 \times 144$ trials for training. The numbers of trials for testing were also same. The ECoG data set consists of 278 training trials and 100 testing trials, each trials were 3 seconds long. The ECoG data set filtering was done using filtfilt command in matlab, because it avoids start-up transients. For the dataset $2a$ the signal is band-pass filtered at $7 - 30$ Hz using a butterworth filter and the interval $500 - 3500$ ms after the appearance of visual cue is cut out from the continuous EEG signal. Here a trial is $S \in R^{C \times T}$ where $C$ is the number of electrodes and $T$ is the number of sampled time points. Both datasets were down sampled to 100 Hz for subsequent processing. The input matrix $X$ is defined as $X = SS^T$ where $X$ is the channel covariance matrix. Input data $S$ was not normalized, because it reduced performance. For dataset $2a$ all the 22 channels were filtered with RLS filter [33] using the 3 additional EOG channels provided. Only 22 channels were used for the testing and training purpose. Since the number of channel is only 22 the algorithm works faster. For ECoG data all the 64 channels were used. The logistic predictor model is used here, since the problem is binary as given by Eqn. 4.6, thus the decoding is carried

out by simply taking the sign of the detector function as in Eqn. 5.1 .

$$\hat{y} = \begin{cases} +1 & \text{if } f_\theta(S) \geq 0 \\ -1 & \text{if } f_\theta(S) < 0 \end{cases} \tag{5.1}$$

For the learning of the detector function the logistic loss function (Eqn. 4.8) was used in Eqn. 4.4. For the detector function we use a single second order detector working on complete alpha and beta band $7-30$ Hz and two other second order detectors each working on alpha $7-15$ Hz and beta band $15-30$ Hz separately. Since splitting into more bands gave poor results it was not used in subsequent analysis. The second-order term is band-pass filtered at $7-30$ Hz and pre-processed with a spatial whitening matrix $\sum^{s-\frac{1}{2}}$ i.e., X= $\sum^{s-\frac{1}{2}} cov(S^{BP}) \sum^{s-\frac{1}{2}}$. A dual spectral regularizer is used with it. $\sum^s$ indicates average of covariance matrices. Our aim is to simultaneously learn and select few informative spatio-temporal filters in a systematic manner. We used $2 \times 10$ fold cross-validation for the selection of the regularization constant.

## 5.4 Results

The result of this signal analysis framework applied to the motor imagery datasets are given below. The table shows classification accuracy (in percentage) of test data of BCI competition dataset $2a$ for DS regularizer. The first column in table 5.1 shows the subject label and first row shows the combination of 2 different motor imagery tasks. $1, 2, 3, 4$ indicates left hand, right hand, foot and tongue motor imagery respectively, choosing a binary class from 4 class resulted in six combinations. For some subjects low rank ($< 3$) weight matrix performs best and for some high rank structure is required for good prediction. The model is chosen using cross validation on training data.

The Table 5.2 shows comparison of dual spectral regularizer to CSP for left hand and right hand motor imagery (results in percentage), for CSP analysis, only 3 filter pairs are retained for feature extraction and LDA is used for classification. The DS regularizer with an auto tuned coefficient matrix easily outperforms CSP. For the ECoG data set DS regularizer was able to achieve 91% accuracy which is equivalent to winner's classification accuracy. The Fig. 5.3 shows the learned low rank coefficient matrix W using DS regularizer. The raw CSP performance is only 67% for this dataset. However

Table 5.1: Result for Dataset 2a

| | [1 2] | [1 3] | [1 4] | [2 3] | [2 4] | [3 4] |
|---|---|---|---|---|---|---|
| A1 | 86.81 | 85.42 | 99.31 | 91.67 | 100 | 67.36 |
| A2 | 73.61 | 77.78 | 68.75 | 78.47 | 82.64 | 72.92 |
| A3 | 98.61 | 93.75 | 84.03 | 98.61 | 97.22 | 86.11 |
| A4 | 78.47 | 90.97 | 86.11 | 95.14 | 87.50 | 75.00 |
| A5 | 70.18 | 77.78 | 81.25 | 75.00 | 79.86 | 80.56 |
| A6 | 69.44 | 79.86 | 72.22 | 74.30 | 75.00 | 75.69 |
| A7 | 84.72 | 97.92 | 97.92 | 97.92 | 97.92 | 84.03 |
| A8 | 99.31 | 94.44 | 97.22 | 95.14 | 97.22 | 95.83 |
| A9 | 92.36 | 95.14 | 100 | 90.97 | 96.53 | 97.92 |

Table 5.2: Comparison with CSP

| | DS | CSP |
|---|---|---|
| A1 | 86.81 | 88.89 |
| A2 | 73.61 | 51.39 |
| A3 | 98.61 | 96.53 |
| A4 | 78.47 | 70.14 |
| A5 | 70.18 | 54.86 |
| A6 | 69.44 | 71.53 |
| A7 | 84.72 | 81.25 |
| A8 | 99.31 | 93.75 |
| A9 | 92.36 | 93.75 |

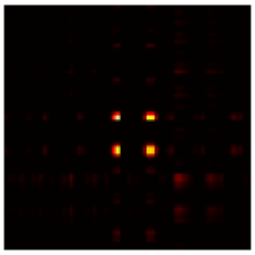

Figure 5.3: Low rank weight matrix for ECoG data

the performance of DS regularizer is reduced if we allow further flexibility by dealing with the alpha-band $7-15$ Hz and beta-band $15-30$ Hz separately Fig. 5.4. The first model captures alpha activity; the second model captures beta activity. One possible explanation is over fitting. In all the topoplots shown red indicates strong positivity and violet indicates strong negativity and the black dots indicate electrode position. The Fig. 5.4 shows the spatial filter captured by $7-30$ Hz, $7-15$ Hz, and $15-30$ Hz respectively, by the detector models for left hand and right hand motor imagery in the top row. The bottom row shows spatial pattern, since the block weight matrix associated to the second-order component is symmetric; we show the eigenvalues of this matrix as spatial pattern. The DS regularizer tries to reduce the rank of matrix $W$ by discarding the irrelevant rows and columns, which finally results in features being concentrated in certain areas.

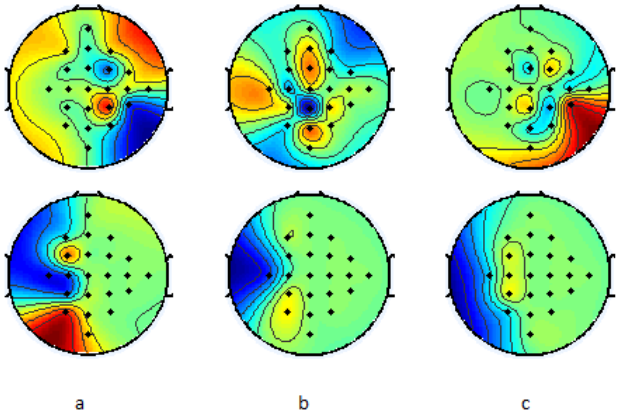


Figure 5.4: Spatial filter and spatial pattern

The Fig. 5.4 shows spatial filter captured for left hand, right hand trial. The spatial filters are obtained by whitening right or left singular vector after singular value decomposition $w_j = \sum^{s-1/2} V(:, j)$. We can see from the Fig. 5.4 the features captured by alpha and beta band combined, alpha and beta models for the same task. The second-order models can be applied efficiently online with DS regularization because the coefficient matrix is typically low rank.

Fig. 5.5 shows the two spatial filters captured by CSP corresponding to left hand and right hand movements. The two spatial filters captured by one step process with DS regularization method is shown in Fig. 5.6. The one step process with DS regularization picked up the discriminative information whereas CSP captured noise along with it.

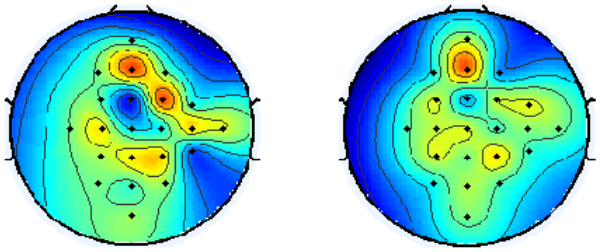

Figure 5.5: Spatial filter captured by CSP

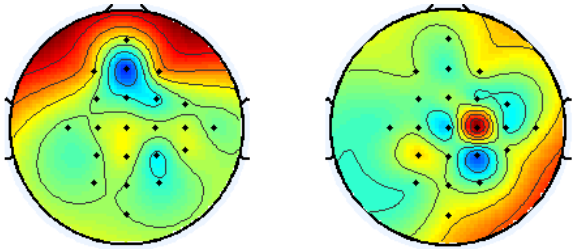

Figure 5.6: Spatial filter captured by one step process with DS

## 5.5 CSP with Tikhonov Regularization

This algorithm is based on the regularization of the CSP objective function using quadratic penalties. Tikhonov Regularization is a classical form of regularization, initially introduced for regression problems, and which consists in penalizing solutions with large weights. The penalty term is then $P(w) = \|w\|_2 = w^T w = w^T I w$. Such regularization is expected to constrain the solution to filters with a small norm, hence mitigating the influence of artifacts and outliers. There is a weighted version of Tikhonov Regularization called Weighted Tikhonov regularization [34]; here instead $I$, a channel prior weight matrix $K$ is used. The diagonal matrix $D$ can be assumed in several ways. With this quadratic penalty the CSP objective function be can written as in Eqn 5.2 .

$$J(w) = \frac{w^T C_1 w}{w^T C_2 w + \alpha w^T I w} \tag{5.2}$$

where $\alpha$ is regularization parameter that we need to fine tune using cross-validation.

## 5.6 Novelty in our work

Using the signal analysis framework built, Tikhonov regularization can be incorporated in to our model. No previous work has been reported on incorporating Tikhonov regularization in to one step signal analysis framework.In the one step classification procedure, the

weight matrix $W$ is assumed to be a diagonal matrix formed by multiplication of spatial filter matrices $w^T w$. By minimizing the trace of this matrix $W$, the sum of squared euclidean norms of each spatial filters are minimized. This regularization can be employed by changing the regularizing term to $trace(W)$ and putting the constraint W is a diagonal matrix in Eqn. 4.16. The constraint $W$ is a PSD matrix is removed, because of inferior performance. Also channel priors can be introduced by changing the regularizing term to $trace(W * D)$, where $D$ is the diagonal matrix with channel priors. The Fig. 5.7 and Fig. 5.8 shows the filter pair learned with this regularization for left and right hand motor imagery and by TRCSP using Eigen decomposition from [34], both looks similar.

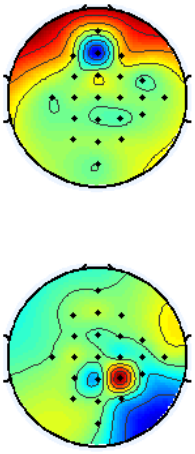

Figure 5.7: Spatial filter learned by one step process for TRCSP

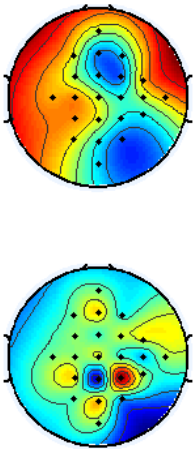

Figure 5.8: Spatial filter learned by TRCSP by Eigen decomposition

From the Fig. 5.7 the spatial filters learned by this method is not as concentrated as spatial filters learned by DS regularization. From table 5.3 we can see both the algorithm has comparable performance. The table shows classification accuracies in percentage

Table 5.3: Performance Comparison of TRCSP Using One Step Process and by eigen decomposition

| | One Step TRCSP | TRCSP |
|---|---|---|
| A1 | 85.68 | 88.89 |
| A2 | 60.42 | 54.17 |
| A3 | 90.72 | 96.53 |
| A4 | 73.61 | 70.83 |
| A5 | 58.33 | 62.50 |
| A6 | 70.56 | 67.36 |
| A7 | 81.94 | 81.25 |
| A8 | 89.56 | 95.87 |
| A9 | 90.00 | 91.67 |

for test data. If a comparison with DS regularization is made, the DS regularization outperforms TRCSP slightly, for this dataset in most of the cases, for some case this method works better. So, this points to requirement of subject specific learning algorithms instead of a fixed one. Also, one thing to note that here we cannot say completely that weight matrix $W$ is SPSD due to $w^T w$ alone, this matrix also consists of weighting parameters for classification.

This regularized objective function is easy to implement, since it is smooth and convex. But again raw implementation is CPU intensive and slower because of large number of variables to optimize for dataset 4a 22 variables in $W \in S_+^{22}$ (since $W$ is diagonal) plus one bias term. For a second order solver to make a step, gradient of this 23 variables and inverse of Hessian matrix (second derivative) has to be calculated. There are second order solvers for optimization which works very fast when gradient is supplied like Limited memory quasi-Newton method, here $H^{-1}$ is approximated incrementally from supplied gradient. L-BFGS (Limited-Memory Broyden-Fletcher-Goldfarb-Shanno) [35] is a practical variant of it. In our case objective function is smooth and gradient can be supplied so this method can be used.

## 5.7 Conclusion

The spectral $\ell_1$-regularization in the weight matrix space provides not only a principled way of complexity control but also an elegant low rank solution that concentrates all the discriminative information in a few components. Using the LMI technique the inference task is formulated as a single convex optimization problem. Here the method is applied

to the single trial EEG classification problem in the context of BCI. This method significantly outperforms conventional methods. Moreover, the method automatically produces a decomposition of the signal into small number of components; the number of components is automatically selected. This method sidesteps the explicit rank constraints, which usually result in non-convex optimization, by using the $\ell_1$-regularization on the singular values of the weight matrix. The one step feature extraction and discrimination with TRCSP learns an efficient model which combats overfitting and results in better classification accuracy for test data.

# Chapter 6

# Conclusion

The framework used in this thesis can be applied to different types of regularizers as well as the predictor model can be changed with other predictor models like hinge loss, which inturn induces sparsity. The test trials can be classified by taking sign of the sum of inner product between $W$ and covariance matrix $S$ and bias term. This makes it easier to use in online systems, with low rank structure this can be done with less computation. Since the predictor output gives amount of surprise or entropy for each input, this can be used to implement multiclass variant of this in one step. We may also apply the method to other multiple-sensor recordings e.g., fMRI signals or computer vision problems, regression problems etc.

The key idea in our approach is to focus on directly predicting the intention of a user. This enabled us to approach decoding and learning in a unified and systematic manner and to avoid intermediate steps. Note that this idea applies not only to other BCI paradigms including invasive BCIs but also to general discriminative neurophysiological paradigms even beyond EEG.

# Bibliography


[1] Graimann, Bernhard, Brendan Allison, and Gert Pfurtscheller. "Braincomputer interfaces: A gentle introduction." In Brain-Computer Interfaces, pp. 1-27. Springer Berlin Heidelberg, 2010.

[2] Nicolas-Alonso, Luis Fernando, and Jaime Gomez-Gil. "Brain computer interfaces, a review." Sensors 12, no. 2 (2012): 1211-1275

[3] Pfurtscheller, Gert, and F. H. Lopes da Silva. "Event-related EEG/MEG synchronization and desynchronization: basic principles." Clinical neurophysiology 110, no. 11 (1999): 1842-1857.

[4] Yang, B. H., G. Z. Yan, and R. G. Yan. "[A review of brain-computer interfaces (BCIs)]." Zhongguo yi liao qi xie za zhi= Chinese journal of medical instrumentation 29, no. 5 (2005): 353.

[5] Makeig, Scott, Stefan Debener, Julie Onton, and Arnaud Delorme. "Mining event-related brain dynamics." Trends in cognitive sciences 8, no. 5 (2004): 204-210.

[6] Musallam, Sam, B. D. Corneil, Bradley Greger, Hans Scherberger, and R. A. Andersen. "Cognitive control signals for neural prosthetics." Science 305, no. 5681 (2004): 258-262.

[7] Blankertz, Benjamin, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Mller. "Single-trial analysis and classification of ERP componentsa tutorial." NeuroImage 56, no. 2 (2011): 814-825.

[8] Lotte, Fabien, Marco Congedo, Anatole Lcuyer, Fabrice Lamarche, and Bruno Arnaldi. "A review of classification algorithms for EEG-based raincomputer interfaces." Journal of neural engineering 4 (2007).

[9] Bashashati, Ali, Mehrdad Fatourechi, Rabab K. Ward, and Gary E. Birch. "A survey of signal processing algorithms in braincomputer interfaces based on electrical brain signals." Journal of Neural engineering 4, no. 2 (2007): R32

[10] Blankertz, Benjamin, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and K-R. Muller. "Optimizing spatial filters for robust EEG single-trial analysis." Signal Processing Magazine, IEEE 25, no. 1 (2008): 41-56.

[11] Blankertz, Benjamin, Guido Dornhege, Matthias Krauledat, Klaus-Robert Mller, and Gabriel Curio. "The non-invasive Berlin braincomputer interface: fast acquisition of effective performance in untrained subjects." NeuroImage 37, no. 2 (2007): 539-550.

[12] Fukunaga, Keinosuke. Introduction to statistical pattern recognition. Access Online via Elsevier, 1990.

[13] Ramoser, Herbert, Johannes Muller-Gerking, and Gert Pfurtscheller. "Optimal spatial filtering of single trial EEG during imagined hand movement." Rehabilitation Engineering, IEEE Transactions on 8, no. 4 (2000): 441-446.

[14] Tomioka, Ryota, and Klaus-Robert Mller. "A regularized discriminative framework for EEG analysis with application to braincomputer interface." Neuroimage 49, no. 1 (2010): 415-432.

[15] Ang, Kai Keng, Zhang Yang Chin, Haihong Zhang, and Cuntai Guan. "Filter bank common spatial pattern (FBCSP) in brain-computer interface." In Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pp. 2390-2397. IEEE, 2008.

[16] Tomioka, Ryota, Guido Dornhege, Kazuyuki Aihara, and K-R. Mller. "An iterative algorithm for spatio-temporal filter optimization." In Verlag der Technischen Universitt Graz. 2006.

[17] Tomioka, Ryota, Guido Dornhege, Guido Nolte, Kazuyuki Aihara, and Klaus-Robert Mller. "Optimizing spectral filters for single trial EEG classification." In Pattern Recognition, pp. 414-423. Springer Berlin Heidelberg, 2006.

[18] Melzer, Thomas. "SVD and its application to generalized eigenvalue problems." Vienna University of Technology (2004).

[19] Lewis, Adrian S., and Michael L. Overton. "Eigenvalue optimization." Acta numerica 5, no. 1 (1996): 149-190.

[20] Boyd, Stephen Poythress, and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

[21] Vandenberghe, Lieven, and Stephen Boyd. "Semidefinite programming." SIAM review 38, no. 1 (1996): 49-95.

[22] Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." Annals of eugenics 7, no. 2 (1936): 179-188.

[23] Welling, Max. "Fisher linear discriminant analysis." Department of Computer Science, University of Toronto (2005).

[24] Tomioka, Ryota, Kazuyuki Aihara, and Klaus-Robert Mller. "Logistic regression for single trial EEG classification." Advances in neural information processing systems 19 (2007): 1377-1384.

[25] Srebro, Nathan, and Adi Shraibman. "Rank, trace-norm and max-norm." In Learning Theory, pp. 545-560. Springer Berlin Heidelberg, 2005.

[26] Rennie, Jason DM. "The Relation Between the Spectral and Trace Norms." (2006).

[27] Fazel, Maryam, Haitham Hindi, and Stephen P. Boyd. "A rank minimization heuristic with application to minimum order system approximation." In American Control Conference, 2001. Proceedings of the 2001, vol. 6, pp. 4734-4739. IEEE, 2001.

[28] Grant, Michael, Stephen Boyd, and Yinyu Ye. "CVX: Matlab software for disciplined convex programming." (2008).

[29] Ryota Tomioka Masashi Sugiyama, "Dual Augmented Lagrangian Method for Efficient Sparse Reconstruction", IEEE Signal Proccesing Letters, 16 (12) pp. 1067-1070, 2009.

[30] Delorme, Arnaud, Tim Mullen, Christian Kothe, Zeynep Akalin Acar, Nima Bigdely-Shamlo, Andrey Vankov, and Scott Makeig. "EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing." Computational intelligence and neuroscience 2011 (2011): 10.

[31] Data sets 2a: 4-class motor imagery (description) provided by the Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University

of Technology, (Clemens Brunner, Robert Leeb, Gernot Mller-Putz, Alois Schlgl, Gert Pfurtscheller)

[32] Data set I: motor imagery in ECoG recordings, session-to-session transfer (description I) provided by Eberhard-Karls-Universitt Tbingen, Germany, Dept. of Computer Engineering and Dept. of Medical Psychology and Behavioral Neurobiology (Niels Birbaumer), and Max-Planck-Institute for Biological Cybernetics, Tbingen, Germany (Bernhard Schkopf), and Universitt Bonn, Germany, Dept. of Epileptology

[33] He, P., G. Wilson, and C. Russell. "Removal of ocular artifacts from electro-encephalogram by adaptive filtering." Medical and biological engineering and computing 42, no. 3 (2004): 407-412.

[34] Lotte, Fabien, and Cuntai Guan. "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms." Biomedical Engineering, IEEE Transactions on 58, no. 2 (2011): 355-362.

[35] Nash, Stephen G., and Jorge Nocedal. "A numerical study of the limited memory BFGS method and the truncated-Newton method for large scale optimization." SIAM Journal on Optimization 1, no. 3 (1991): 358-372.