

# **SPATIAL DECODING OF OSCILLATORY NEURAL ACTIVITY FOR BRAIN COMPUTER INTERFACING**

A DISSERTATION SUBMITTED TO  
THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS  
ENGINEERING  
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE  
OF BILKENT UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

By  
İbrahim Onaran  
June, 2013

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

---

Prof. Dr. A. Enis Çetin (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

---

Asst. Prof. Dr. N. Fırat İnce (Co-Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

---

Assoc. Prof. Dr. Uğur Güdükbay

I certify that I have read this thesis and that in my opinion it is fully adequate,  
in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

---

Assoc. Prof. Dr. Sinan Gezici

I certify that I have read this thesis and that in my opinion it is fully adequate,  
in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

---

Prof. Dr. Ziya İder

I certify that I have read this thesis and that in my opinion it is fully adequate,  
in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

---

Prof. Dr. Kemal Leblebicioğlu

Approved for the Graduate School of Engineering and Science:

---

Prof. Dr. Levent Onural  
Director of the Graduate School

# ABSTRACT

## SPATIAL DECODING OF OSCILLATORY NEURAL ACTIVITY FOR BRAIN COMPUTER INTERFACING

İbrahim Onaran

PhD in Electrical and Electronics Engineering

Supervisor: Prof. Dr. A. Enis Çetin

June, 2013

Neuroprosthetics (NP) aim to restore communication between people with debilitating motor impairments and their environments. To provide such a communication channel, signal processing techniques converting neurophysiological signals into neuroprosthetic commands are required. In this thesis, we develop robust systems that use the electrocorticogram (ECoG) signals of individuated finger movements and electroencephalogram (EEG) signals of hand and foot movement imageries.

We first develop a hybrid state detection algorithm for the estimation of baseline (resting) and movement states of the finger movements which can be used to trigger a free paced neuroprosthetic using the ECoG signals. The hybrid model is constructed by fusing a multiclass support vector machine (SVM) with a hidden Markov model (HMM), in which the internal hidden state observation probabilities are represented by the discriminative output of the SVM. We observe that the SVM based movement decoder improves accuracy for both large and small numbers of training dataset.

Next, we tackle the problem of classifying multichannel ECoG related to individual finger movements for a brain machine interface (BMI). For this particular problem we use common spatial pattern (CSP) method which is a popular method in BMI applications, to extract features from the multichannel neural activity through a set of spatial projections. Since we try to classify more than two classes, our algorithm extends the binary CSP algorithm to multiclass problem by constructing a redundant set of spatial projections that are tuned for paired and group-wise discrimination of finger movements. The groupings are constructed by merging the data of adjacent fingers and contrasting them to the rest, such as

the first two fingers (thumb and index) vs. the others (middle, ring and little).

In the remaining parts of the thesis, we investigate the problems of CSP method and propose techniques to overcome these problems. The CSP method generally overfits the data when the number of training trials is not sufficiently large and it is sensitive to daily variation of multichannel electrode placement, which limits its applicability for everyday use in BMI systems. The amount of channels used in projections should be limited to some adequate number to overcome these problems. We introduce a spatially sparse projection (SSP) method, taking advantage of the unconstrained minimization of a new objective function with approximated  $\ell_1$  penalty. Furthermore, we investigate the greedy  $\ell_0$  norm based channel selection algorithms and propose oscillating search (OS) method to reduce the number of channels. OS is a greedy search technique that uses backward elimination (BE), forward selection (FS) and recursive weight elimination (RWE) techniques to improve the classification accuracy and computational complexity of the algorithm in case of small amount of training data. Finally, we fuse the discriminative and the representative characteristic of the data using a baseline regularization to improve the classification accuracy of the spatial projection methods.

*Keywords:* Brain computer interfaces (BCI), brain machine interfaces (BMI), common spatial pattern, support vector machines (SVM), linear discriminant analysis (LDA), hidden Markov models (HMMs), electroencephalogram (EEG), electrocorticogram (ECoG), error correcting output codes (ECOC).

## ÖZET

# BEYİN MAKİNE ARAYÜZLERİ İÇİN SALINIMLI BEYİN İŞARETLERİNİN UZAMSAL ÇÖZÜMLEMESİ

İbrahim Onaran

Elektrik ve Elektronik Mühendisliği, Doktora

Tez Yöneticisi: Prof. Dr. A. Enis Çetin

Haziran, 2013

Nöral protezler, hareket kısıtlayıcı rahatsızlığı olan hastaların çevreleriyle olan iletişimini sağlamayı amaçlamaktadır. Bu tür bir iletişim kanalı sağlamak için nörofizyolojik işaretleri nöral protezlerin anlayacağı komutlara çeviren sinyal işleme teknikleri gerekmektedir. Bu tezde, el parmaklarının elektrokortikogram (ECoG) sinyallerini ve hayali el ve ayak hareketlerinin elektroensefalogram (EEG) sinyallerini dayanaklı sistemler geliştirmek için kullandık.

İlk önce parmakların hareketsizlik ve hareket durumlarını tahmin etmek için destek vektör makineleri (SVM) ile saklı Markov modeline (HMM) dayalı melez durum algılama yöntemi geliştirdik. Bu yöntem ECoG sinyali kullanılarak serbest tempolu bir nöral protezi tetiklemek için kullanılabilir. Bu melez model, SVM ile HMM'in birleştirilmesiyle oluşturulmuştur. HMM'nin saklı iç durum gözlemlerinin olasılıkları SVM'in ayırıcı çıktıları tarafından temsil edilmektedir. SVM tabanlı hareket çözümleyicinin hem fazla, hem de az sayıda öğretici veri için sınıflama sonucunu arttırdığı gözlemlenmiştir.

Bir sonraki adımda, bir beyin makine arayüzü (BMI) geliştirmek için ECoG sinyali kullanılarak tek tek parmak hareketlerinin sınıflandırma sorunu üzerinde çalışıldı. Bu özel sorun için BMI uygulamalarında sıkça kullanılan ortak uzamsal örüntü (CSP) metodu kullanılmıştır. CSP metodu çok kanallı nöral etkinliğinden bir dizi uzamsal izdüşüm vasıtasıyla öznitelik çıkarmakta kullanılmaktadır. İki den fazla sınıfı ayırmaya çalıştığımız için, ikili CSP metodu çoklu sınıflarda kullanılmak üzere genişletilmiştir. Bu genişletme parmakların tek olarak ve grup olarak birbirleri ile karşılaştırılmaları ile sağlanmıştır. Parmak grupları, komşu iki parmak (örneğin baş ve işaret parmakları) ve kalan parmaklar ayrı iki sınıf olacak şekilde oluşturulmuştur.

Geri kalan bölümlerde ise CSP metodunun problemleri araştırılmış ve bu problemleri çözmek için yeni teknikler ortaya konulmuştur. CSP metodu, eğitim deneme sayısı yeterli olmadığı durumlarda genellikle veriye fazla uyum göstermektedir. Ayrıca CSP metodu, BMI sistemlerinin günlük hayatta kullanılmalarını sınırlayan çoklu elektrotların yerlerindeki günlük değişimlere karşı duyarlıdır. Bu problemlerin üstesinden gelebilmek için kullanılan kanal sayısı uygun şekilde sınırlandırılmalıdır. Bu problemleri çözmek ve kanal sayısını sınırlandırmak için uzamsal olarak seyrek izdüşüm (SSP) metodu geliştirilmiştir. Bu metod, yeni bir amaç fonksiyonu ile yaklaşık olarak  $\ell_1$  norm ceza fonksiyonunun kısıtsız eniyilemesini kullanmaktadır. Ayrıca, fırsatçı  $\ell_0$  norm tabanlı kanal seçim algoritmaları incelenmiştir ve salınan arama (OS) yöntemi önerilmiştir. OS yöntemi, geri elimine etme (BE), ileri seçim (FS) ve özyinelemeli ağırlığın ortadan kaldırılması (RWE) tekniklerinin birleşiminden oluşmaktadır. Bu yöntem hesaplama karmaşıklığını azaltmak ve az miktarda eğitim verisi olduğu durumda sınıflandırma doğruluğunu artırmak için kullanılmıştır. Son olarak sınıflama doğruluğunu arttırmak için verinin ayırma ve temsil etme vasıfları hareketsizlik düzenleme metodu ile birleştirilmiştir.

*Anahtar sözcükler:* Beyin bilgisayar arayüzleri (BCI), beyin makine arayüzleri (BMI), ortak uzamsal örüntü(CSP),destek vektör makineleri (SVM), elektroensefalogram (EEG), doğrusal ayırtaç çözümleyici (LDA), saklı Markof modeli (HMM), elektrokortikogram (ECoG), hata düzeltici çıktı kodları (ECOC).

# Acknowledgement

First of all I am grateful to my advisor Prof. Dr. A. Enis Çetin for his support in every aspect of my academic life. Furthermore, his support is not limited to academic life but also extends to other parts of the life.

I am especially grateful to my co-advisor Asst. Prof. Dr. N. Firat İnce for trusting in me and accept me to his research team in Minneapolis, USA. In my opinion, this thesis would be impossible to be completed without his support and guidance.

I am grateful to Assoc. Prof. Dr. Uğur GÜDÜKBAY and Assoc. Prof. Dr. Sinan Gezici for accepting to be in my Ph.D. progress and defense committee and their support throughout my Ph.D. studies.

I am grateful to Prof. Dr. Ziya İDER and Prof. Dr. Kemal Leblebicioğlu for reading my thesis and agreeing to be in my Ph.D. defense committee.

I would like to thank Mürüvet Parlakay for answering all of my questions about the procedures of the department.

I would like to thank the Scientific and Technological Research Council of Turkey (TÜBİTAK) Science Fellowships and Grant Programmes Department (BİDEB) for my scholarship.

This research was supported in part by the National Science Foundation, award CBET-1067488, and by a grant from the University of Minnesota Interdisciplinary Informatics (UMII).

I would like thank all of my friends for their help and letting me to be a part of their life.

I would also thank my brother and sister for their support. This thesis would never be exist without the support and love of my parents.

# Contents

|          |                                                   |           |
|----------|---------------------------------------------------|-----------|
| <b>1</b> | <b>INTRODUCTION</b>                               | <b>1</b>  |
| 1.1      | Non-Invasive Signal based Studies . . . . .       | 3         |
| 1.1.1    | P300 based BCI . . . . .                          | 5         |
| 1.1.2    | Motor Imagery based BCI . . . . .                 | 6         |
| 1.2      | Invasive Signal based Studies . . . . .           | 6         |
| 1.2.1    | Single Unit Activity based Studies . . . . .      | 6         |
| 1.2.2    | Local Field Potentials based BCI . . . . .        | 7         |
| 1.2.3    | Electrocorticogram based BCI . . . . .            | 8         |
| 1.3      | Outline of the Thesis . . . . .                   | 10        |
| <b>2</b> | <b>Free paced Baseline and Movement Detection</b> | <b>14</b> |
| 2.1      | Introduction . . . . .                            | 14        |
| 2.2      | ECoG Data and Preprocessing . . . . .             | 16        |
| 2.2.1    | ECoG Dataset . . . . .                            | 16        |
| 2.2.2    | Common Spatial Patterns . . . . .                 | 18        |

|          |                                                                                                                     |           |
|----------|---------------------------------------------------------------------------------------------------------------------|-----------|
| 2.2.3    | Hybrid HMM-SVM Structure . . . . .                                                                                  | 19        |
| 2.3      | Results . . . . .                                                                                                   | 21        |
| 2.4      | Summary . . . . .                                                                                                   | 24        |
| <b>3</b> | <b>Decoding of Individual Finger Movements using Redundant Spatial Projections</b>                                  | <b>26</b> |
| 3.1      | Introduction . . . . .                                                                                              | 26        |
| 3.2      | Methods and Materials . . . . .                                                                                     | 28        |
| 3.2.1    | Multiclass CSP with Hierarchical Grouping . . . . .                                                                 | 28        |
| 3.2.2    | Support Vector Machine based Classifier . . . . .                                                                   | 29        |
| 3.2.3    | ECoG Data . . . . .                                                                                                 | 30        |
| 3.3      | Results . . . . .                                                                                                   | 32        |
| 3.4      | Summary . . . . .                                                                                                   | 35        |
| <b>4</b> | <b>Sparse Spatial Projections Using New Objective Function with <math>\ell_1</math> Norm based Penalty Function</b> | <b>36</b> |
| 4.1      | Introduction . . . . .                                                                                              | 36        |
| 4.2      | Methods and Materials . . . . .                                                                                     | 37        |
| 4.2.1    | Standard CSP and a New Objective Function . . . . .                                                                 | 37        |
| 4.2.2    | Codifference Matrix Based on Addition . . . . .                                                                     | 41        |
| 4.2.3    | ECoG & EEG Datasets . . . . .                                                                                       | 43        |
| 4.3      | Results . . . . .                                                                                                   | 44        |

|          |                                                                               |           |
|----------|-------------------------------------------------------------------------------|-----------|
| 4.4      | Summary . . . . .                                                             | 53        |
| <b>5</b> | <b>Sparse Spatial Projections Using <math>\ell_0</math> Norm based Search</b> | <b>55</b> |
| 5.1      | Introduction . . . . .                                                        | 55        |
| 5.2      | Material and Methods . . . . .                                                | 57        |
| 5.2.1    | Standard CSP Method . . . . .                                                 | 57        |
| 5.2.2    | Sparse CSP Methods . . . . .                                                  | 58        |
| 5.2.3    | The Dataset . . . . .                                                         | 63        |
| 5.3      | Results . . . . .                                                             | 64        |
| 5.4      | Summary . . . . .                                                             | 67        |
| <b>6</b> | <b>Baseline Regularized Sparse Spatial Filters</b>                            | <b>69</b> |
| 6.1      | Introduction . . . . .                                                        | 70        |
| 6.2      | Material and Methods . . . . .                                                | 72        |
| 6.2.1    | Sparse Spatial Filter . . . . .                                               | 72        |
| 6.2.2    | Recursive Weight Elimination . . . . .                                        | 74        |
| 6.2.3    | Baseline Regularized Sparse Spatial Filters . . . . .                         | 74        |
| 6.2.4    | ECoG Dataset . . . . .                                                        | 75        |
| 6.3      | Results . . . . .                                                             | 76        |
| 6.4      | Summary . . . . .                                                             | 78        |
| <b>7</b> | <b>Conclusions</b>                                                            | <b>80</b> |

|                                                                                   |            |
|-----------------------------------------------------------------------------------|------------|
| <b>APPENDIX</b>                                                                   | <b>85</b>  |
| <b>A Principle Component Analysis and Generalized Eigenvalue Decomposition</b>    | <b>85</b>  |
| <b>B Projection on to the <math>\ell_1</math> ball</b>                            | <b>89</b>  |
| B.1 Projection formulations . . . . .                                             | 89         |
| B.2 Sparse Solution . . . . .                                                     | 91         |
| <b>C Deflation of a Matrix</b>                                                    | <b>94</b>  |
| C.1 Hotelling's deflation . . . . .                                               | 94         |
| C.2 Schur Complement Deflation . . . . .                                          | 95         |
| <b>D New Objective Function and Its Properties</b>                                | <b>96</b>  |
| D.1 New objective function and its minimum points . . . . .                       | 96         |
| D.2 Gradient and Hessian of minimization function . . . . .                       | 101        |
| <b>E Projections Onto Convex Sets based Optimization of the New Cost Function</b> | <b>104</b> |

# List of Figures

|     |                                                                                                                              |    |
|-----|------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | A general diagram of the BCI system . . . . .                                                                                | 2  |
| 1.2 | Non-invasive and invasive electrode placements . . . . .                                                                     | 3  |
| 1.3 | Types of neural frequency bands . . . . .                                                                                    | 4  |
| 1.4 | A sample ECoG signal from an epilepsy patient. . . . .                                                                       | 9  |
| 2.1 | Diagram of the baseline movement detection using ECoG signals .                                                              | 16 |
| 2.2 | A sample finger position plot that describes the states of the HMM<br>model. . . . .                                         | 17 |
| 2.3 | Average accuracy vs. the number of train trial with decoding<br>length 10 . . . . .                                          | 22 |
| 2.4 | Average latency vs. decoding sequence length . . . . .                                                                       | 23 |
| 2.5 | Average accuracy vs. time for subject 3 aligned to the movement<br>onset. . . . .                                            | 24 |
| 2.6 | Average accuracy vs. time for subject 3 aligned to the movement<br>termination. . . . .                                      | 24 |
| 3.1 | Diagram of the error correcting output code (ECOC) algorithm<br>that is applied to the finger index classification . . . . . | 29 |

|      |                                                                                                                                  |    |
|------|----------------------------------------------------------------------------------------------------------------------------------|----|
| 3.2  | The average time frequency map computed from all subjects using the most reactive channel set selected for each subject. . . . . | 31 |
| 3.3  | The classification accuracies in each frequency band for three subjects. . . . .                                                 | 32 |
| 3.4  | Confusion matrices of the classification accuracies across 5 fingers for three subjects. . . . .                                 | 33 |
| 3.5  | The corresponding finger signal correlation of the classification accuracies across 5 fingers for three subjects. . . . .        | 34 |
| 4.1  | The RQ surface for a toy example . . . . .                                                                                       | 39 |
| 4.2  | Normalized IRQ values vs $\alpha$ value of the minimization function $L(\omega) = G(\omega) + \alpha\ \omega\ $ . . . . .        | 45 |
| 4.3  | $\alpha$ value of the minimization function $L(\omega) = G(\omega) + \alpha\ \omega\ $ vs the cardinality. . . . .               | 46 |
| 4.4  | The normalized IRQ values vs cardinality for each subject. . . . .                                                               | 47 |
| 4.5  | Average IRQ values for ECoG and EEG signals . . . . .                                                                            | 48 |
| 4.6  | The classification error curves of SSP and BE methods versus the cardinality. . . . .                                            | 49 |
| 4.7  | The minimum error vs. the number of trials. . . . .                                                                              | 50 |
| 4.8  | The noise to signal plus noise ratio (NSNR) vs. classification accuracy for ECoG and EEG sets . . . . .                          | 51 |
| 4.9  | The histogram of the number of electrodes with respect to the displacement induced on the test data. . . . .                     | 52 |
| 4.10 | Sample pulse noise for ECoG signal. . . . .                                                                                      | 53 |
| 4.11 | Pulse noise classification accuracy. . . . .                                                                                     | 54 |

|     |                                                                                             |     |
|-----|---------------------------------------------------------------------------------------------|-----|
| 5.1 | The average RQ of all subjects versus cardinality. . . . .                                  | 64  |
| 5.2 | The classification error curves of all methods versus the cardinality.                      | 66  |
| 5.3 | The OS and CSP filters for hand and foot movement imagery. . .                              | 67  |
| 5.4 | The average elapsed time to estimate a spatial filter vs. the cardinality. . . . .          | 68  |
| 6.1 | The average IRQ of all subjects versus cardinality for SSP and RWE methods. . . . .         | 77  |
| 6.2 | The classification error curve versus the cardinality for the SSP and RWE methods . . . . . | 78  |
| B.1 | The $\ell_1$ ball for 2-dimensional space. . . . .                                          | 90  |
| B.2 | Obtaining sparsity from $\ell_1$ ball for 2-dimensional space. . . . .                      | 92  |
| D.1 | The RQ mesh and contour graphics. . . . .                                                   | 97  |
| D.2 | The new objective function ( $G(\mathbf{w})$ ) surface and contour graphics.                | 98  |
| D.3 | The new objective function contour in terms of $R(\mathbf{w})$ and $b(\mathbf{w})$ . .      | 100 |
| D.4 | The new objective function with epsL1 penalty contour graphics. .                           | 101 |
| E.1 | Two projecting convex sets. . . . .                                                         | 106 |
| E.2 | Converging to the cost function. . . . .                                                    | 107 |

# List of Tables

|     |                                                                                                                                    |    |
|-----|------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | The state decoding accuracies of the hybrid and traditional HMM based methods. . . . .                                             | 21 |
| 3.1 | The complete list of competing classes for pairwise and redundant classifiers. . . . .                                             | 30 |
| 3.2 | The classification results for paired wise (non-redundant) and redundant decoding strategies in 65-200 Hz frequency range. . . . . | 34 |
| 4.1 | EEG dataset classification error rates (%) for each subject using SVM classifier . . . . .                                         | 46 |
| 4.2 | ECoG dataset classification error rates (%) for each subject using SVM classifier . . . . .                                        | 47 |
| 4.3 | Average test error rate and corresponding cardinality . . . . .                                                                    | 49 |
| 5.1 | EEG dataset classification error rates (%) for each subject using LDA classifier . . . . .                                         | 65 |
| 6.1 | Classification error rates (%) for each subject using SVM classifier                                                               | 77 |

# Chapter 1

## INTRODUCTION

The measurement of the electrical activity of the brain with electrodes attached on the scalp helps medical doctors to diagnose the brain diseases, exploring its functions, the effects of medication etc. In the past decade, there is a growing interest towards using brain activity to control external devices. The motivation behind this interest is to allow people that have severe motor disabilities to communicate with the outer world using solely their brain signals. The brain computer interface (BCI) constructs the communication channel between the human brain and the computer to allow paralyzed and disabled people to control a neuro-prosthetic [1,2] or directly their muscles through functional electrical stimulation (FES) [3]. Consequently, a BCI assists paralyzed or disabled people to perform essential activities in their daily life and communicate with their environment.

Components of BCI can be categorized into four different units, which are data acquisition, signal processing, classification and feedback as shown in Fig. 1. The data acquisition part of the BCI includes the electrodes that are attached to the brain or the scalp, small signal amplifiers with high impedance and the data storage unit. The signal processing component of BCI application transforms the acquired raw brain signals into useful features such as frequency band power values, some statistical parameters of the signal (mean and variance), spike count, etc [4,5]. As the last step of the BCI, these features are fed into a particular classifier to produce the control signal which drives an external device or

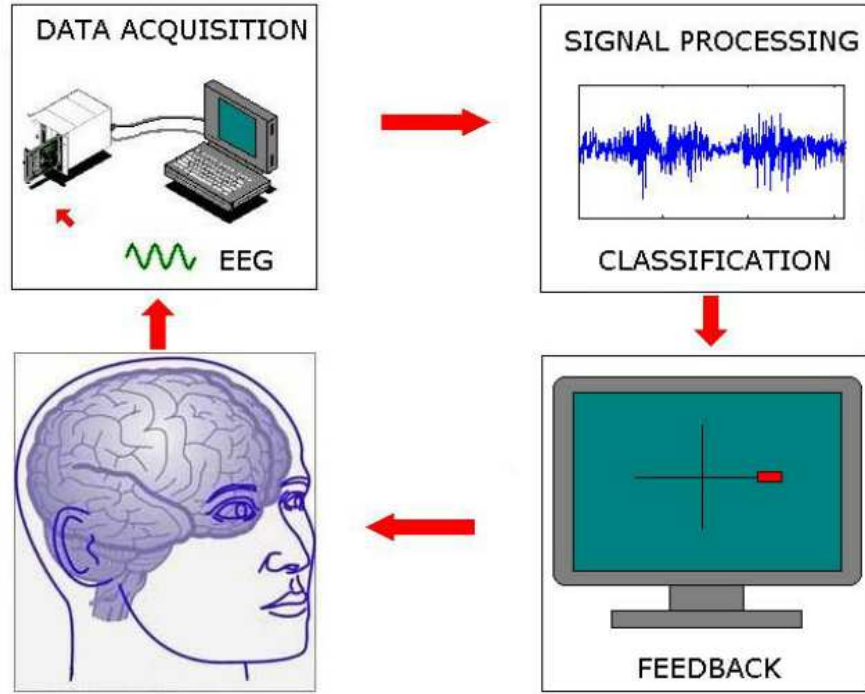


Figure 1.1: A diagram of the components of the BCI system (modified from [6])

neuroprosthetics.

There are many types of brain signals that can be utilized to establish a communication channel for BCI applications. They can be categorized into two major groups, namely invasive and non-invasive brain signals. The EEG [7–11] and magnetoencephalogram (MEG) [12] are examples of non-invasive modalities. A 10-20 EEG recording system is shown in Fig. 1.2a. In this system, the electrode locations are determined proportional to the size of the head. Electrocorticogram (ECoG) [13, 14], single unit activity (SUA) [5, 15–18] and local field potentials (LFPs) [19, 20] are the examples of the invasive neural data. Fig. 1.2b shows the general schematic of the ECoG recording grid, which is used to collect data from the surface of the brain through a set of electrodes on a grid. The feedback unit can be a robotic hand or arm [21], a cursor control on screen that can help

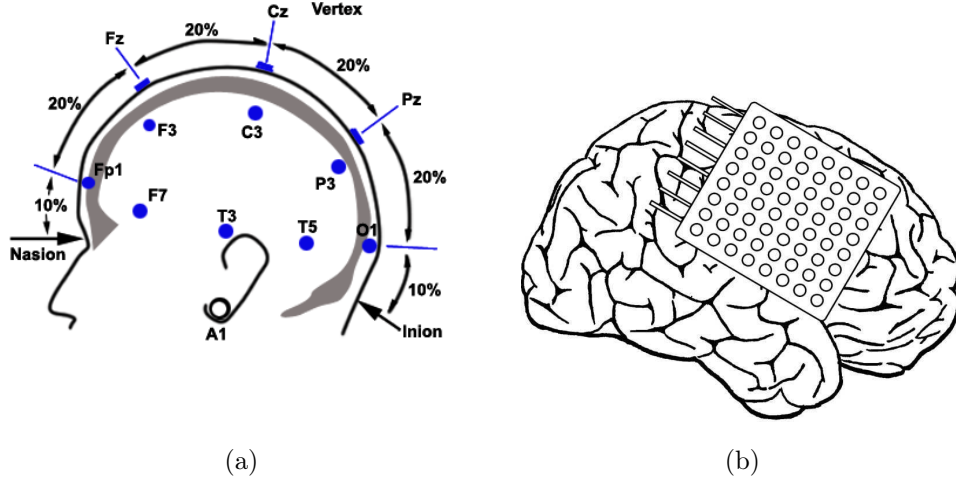


Figure 1.2: (a) The locations of the EEG electrodes in 10-20 system [23]. (b) The ECoG grid is placed on the cortex of the brain, which requires a brain surgery(modified from [24])

people to answer questions [22] or a FES generator that stimulates the paralyzed limb [3].

## 1.1 Non-Invasive Signal based Studies

The distinct characteristics of the EEG signal in human was first discovered by German psychiatrist Hans Berger in 1929 [25,26]. In an eye closing experiment, he realized that the  $\alpha$  band rhythms are decreased after opening eyes and increased in resting state. Such short lasting amplitude decrease in the oscillatory activity is called event related desynchronization (ERD). Short lasting increase in amplitude of the brain signal is called event related synchronization (ERS) [27]. ERD and ERS are the main features in EEG based BCI applications.

The EEG can be recorded from healthy subjects without any clinical risks. However it has lower spatial resolution, low signal to noise ratio (SNR), and it requires extensive user training and can easily be corrupted by various sources of noise such as electromyographic (EMG) signals, eye movements or blinks, which

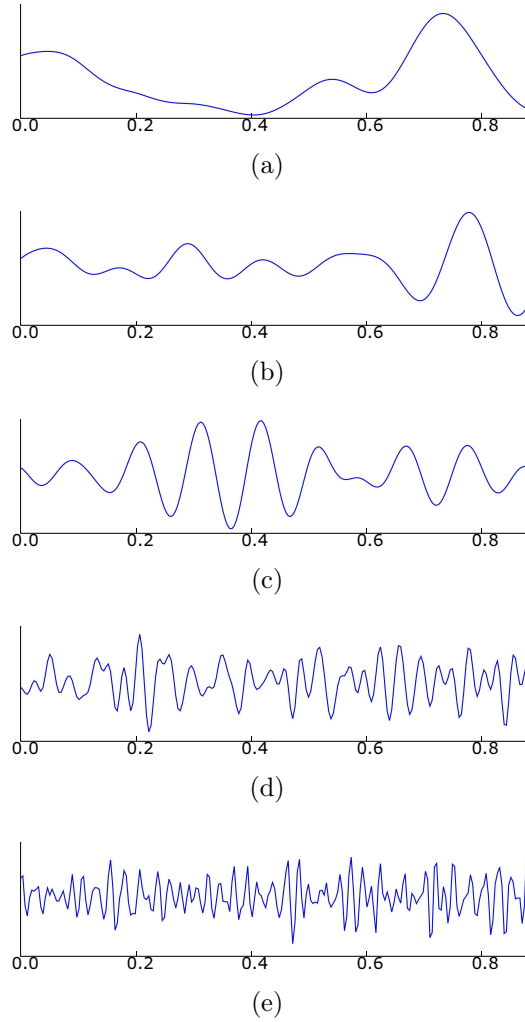


Figure 1.3: The five types of frequency bands ((a) delta, (b) theta, (c) alpha, (d) beta, (e) gamma) that is identified physiologically in human EEG signals.

causes severe artifacts [28–30].

Five physiologically different power bands are identified in human EEG/ECOG brain signals. These power bands and the corresponding frequency ranges are the delta band ( $\delta \in 0-4$  Hz), the theta band ( $\theta \in 4-8$  Hz), the alpha band ( $\alpha \in 8-13$  Hz), the beta band ( $\beta \in 14-30$  Hz) and the gamma band ( $\gamma > 30$  Hz). The typical waveforms of these frequency bands are depicted in Fig. 1.3.

After the discovery of the EEG signal, researchers used it in many areas including diagnosis of neural illnesses, understanding the functionality of the different

regions of the brain, and recently developing BCI to help people with disabilities. Two types of patterns

- i. P300 event related potential (ERP) and
- ii. motor imagery induced ERD/ERS

are widely detected and used in constructing noninvasive BCI [6, 12, 22, 31, 32].

### 1.1.1 P300 based BCI

The authors in [22] describe a system that establishes a communication between brain and the computer using event-related brain potential (ERP) with an enhanced positive-going component with a latency of about 300 ms (P300). The system is designed to help people without motor system communication ('locked-in' patients). The system displays a  $6 \times 6$  letter grid of 26 letters of alphabet and several other commands and symbols, the subject focuses on the letter or the command that he wants to express. The computer flashes each row or column of the grid at a time, which makes 12 possible position (6 rows 6 columns) for the entire grid. When the row or column of the focused grid flashes, P300 is elicited. At this point the row or the column of the letter can be determined by the computer. In this study they used four criteria, area under the P300 window, the covariance of the EEG signal, stepwise discriminant analysis (SWDA), and peak picking. The SWDA produces a score that measures the 'distance' between each epoch and the mean of a group of trials known to include a P300. For a peak picking criterion, the amplitude difference between maximum point in P300 and the minimum point prior to P300 are computed. Finally, for the covariance based feature extraction method, the average of the P300 trials is computed in 600 ms epoch, and covariances of the sub-trials are calculated. The values obtained by these four methods were used to determine the attended letter or command.

### **1.1.2 Motor Imagery based BCI**

The motor imagery (MI) is defined to be the imagination of a motor task without actually executing the task. This type of imagination can modify the neural activity in the primary sensorimotor areas like the subject is performing the actual motor task. These changes in oscillatory brain activity can be detected from the sensory motor areas with EEG electrodes attached to the scalp. ERD/ERS induced by MI are common neuro-markers to distinguish the movement and baseline (resting) states of the brain [6, 12, 31].

In [32], authors investigated the effects of MI on the EEG signals. In their work, they instruct the subjects to imagine different types of motor imagery such as imagination of left-hand, right-hand or foot movement. It is observed that the obtained neuronal activity during the real movement execution is very similar to the motor imagery EEG activity on primary sensorimotor areas. Using the MI related EEG activity, the prosthetics can be controlled with imagination without overt motor activity. The band power or adaptive autoregressive parameters are used as features which are fed into a linear discriminant analysis (LDA) based classifier to sort the motor imagery [33].

## **1.2 Invasive Signal based Studies**

### **1.2.1 Single Unit Activity based Studies**

Single unit activity (SUA) is the firing pattern of a particular neuron that is assessed through micro electrodes at very high frequencies. The firing rate of the neurons in the motor cortex is an important indicator of motor activities.

In [5], the estimation of baseline, movement planning, and movement execution states from a SUA is studied while non-human primates were executing directional hand movements in response to an externally cued paradigm. The neuronal firing rate, computed in fixed-size windows, was used as an input to

a Bayesian state estimator, with the firing rates associated with each direction and with each state modeled with a Poisson distribution. A maximum likelihood (ML) classifier then stamps each time window and the classification outputs were streamed to a finite state machine (FSM) for estimating the state of the subject. The FSM operated on ad-hoc derived transition rules. This work was extended by Achtman et al., [34], who constructed a two-stage decoder that was also based on an FSM. In contrast to [5], a growing window size was used in [34] to estimate from the neural data both the state and the direction of target.

Kemere et al., [18] used a hidden Markov model (HMM) coupled with a state-dependent Poisson firing model instead of Finite State Machine FSM [5]. These investigators demonstrated that using the a priori likelihood of the HMM states to first detect the onset of movement planning and then to calculate the ML target, results in substantial increases in performance relative to the FSM. Recent studies indicate that HMM-based solutions provide better results than FSM-based solutions that are based on ad-hoc decision rules. A common setup shared by these studies is the externally cued paradigms that were used to alter the state of the subject in a controlled manner.

The SUA from monkey subjects were utilized to decode individual finger movements in [35]. The SUA was acquired with penetrating electrodes in the M1 hand area. The firing rates from multiple electrodes were used in conjunction with an artificial neural network (ANN) to decode the finger movements. They reported 95.5% average asynchronous decoding accuracy for individuated finger and wrist movements across three monkeys.

### **1.2.2 Local Field Potentials based BCI**

In [20] the authors showed the feasibility of a high accuracy BCI based on LFPs. They recorded the LFP data from the primary motor (M1) and dorsal premotor cortex (PMd) areas of two monkey subjects. The monkey subjects are trained to move the manipulandum to control computer cursor. The task consists of a center circle and a target circle which is placed to eight different directions

around the circle. The monkeys move the cursor to the randomly selected target circle in order to obtain a liquid reward. Meanwhile the LFPs are recorded using  $10 \times 10$  Utah microelectrode arrays. In the study, common spatial pattern (CSP) is used to reduce the number of channels, after the LFP signal is sub-band filtered into five different frequency bands in which they observed systematic changes during the task. The power on the virtual channels obtained by CSP algorithm is extracted. These features are processed through a set of pairwise and groupwise classifiers. The outcomes of all classifiers are combined using the error correcting output codes (ECOC) algorithm, to yield the final direction of the motion. The results are compared with the results that are obtained from SUA recordings. The LFP and CSP+ECOC algorithm generally outperforms the SUA based classification.

In [36], LFP that is recorded from rhesus monkeys with four 4x4 penetrating electrode grids in primary motor cortex, was used to classify dexterous grasp movements. The subjects are instructed to open and close three different types of switches. They used frequency domain features of 10 visually selected LFP channels with an ANN for classification. The average classification accuracy was reported as 81% for decoding three different dexterous grasping tasks.

In [37], Huang and Andersen used local field potentials (LFPs) recorded from the parietal cortex of primates during a directional reaching task, for a state decoding application. This study demonstrated the feasibility of detecting state transitions from the oscillatory neural activity (LFPs) recorded with penetrating microelectrodes.

### 1.2.3 Electroencephalogram based BCI

Recently, human ECoG based movement detection and classification algorithms are proposed in [38]. The authors build a system that controls a prosthetic hand to perform simple movements, which can greatly improve life quality of the disabled people by allowing them to perform everyday tasks. They recorded ECoG from a subject and perform time-frequency (TF) analysis of the recording signals for the

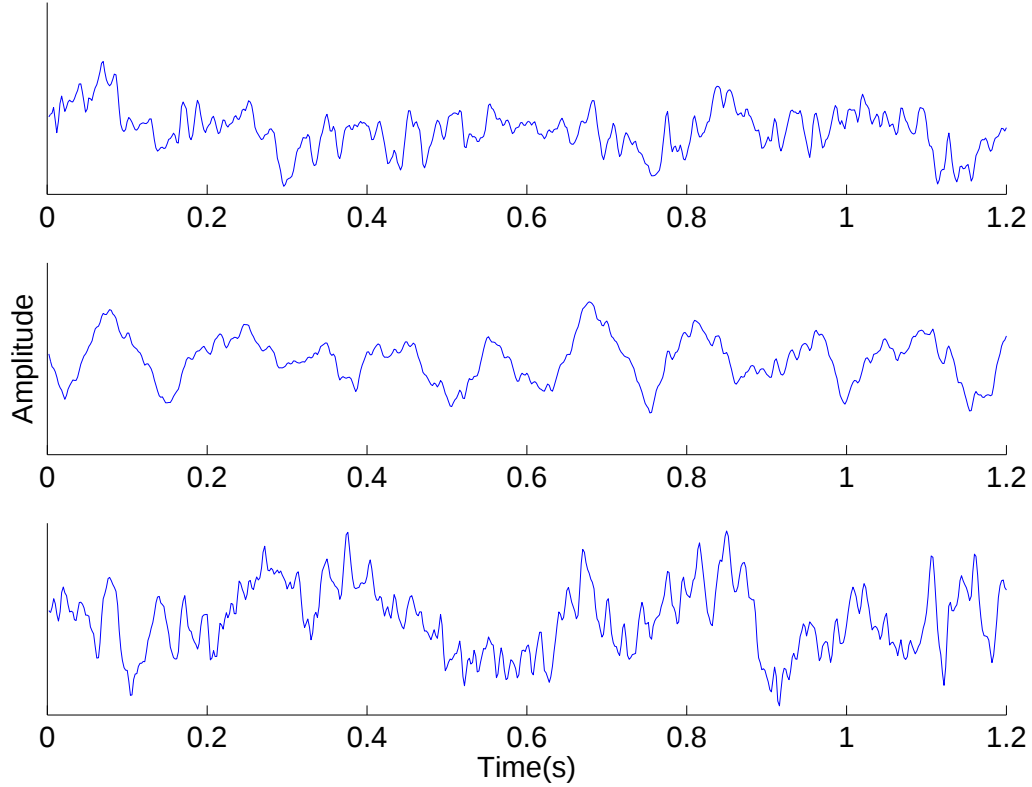


Figure 1.4: A sample ECoG signal acquired from an epilepsy patient for a few channels, while she/he is moving her/his hand fingers.

predefined channels. They observed that three frequency bands can decode rest-move states and type of simple movements such as grasping, opening and making scissor shape etc. They use a two stage classifier, the first one decodes the rest state of the patient. If patient is moving then the second stage of the classifier determines the type of the movement. The authors use linear SVM classifier to identify the state of the hand movement and resting.

The recent literature indicates that SUA is used widely in the constructions of neuroprosthetics due to its superior spatial and temporal resolution [20, 39]. On the other hand, SUA is prone to instability over time [14]. To attain more robust invasive recordings, the LFPs are recently used for BCI applications. Since the LFP represents the activity of a population of neurons within a volume of cortex, this larger listening volume makes LFPs to be acquired more reliably over time after local scaring forming around the electrode tips [19, 20, 40, 41]. The ECoG,

which is recorded from the cortex (surface) of the brain, is less invasive compared to LFP and SUA which are obtained with penetrating electrodes. Moreover, ECoG provides oscillatory activities in the brain with a higher bandwidth and spatial resolution compared to EEG signals [42]. Ability to record oscillatory activity allows us to apply existing EEG algorithms to the ECoG based BCI systems [28]. The next section describes techniques to decode multichannel oscillatory neural data recorded noninvasively with EEG and invasively with ECoG.

### 1.3 Outline of the Thesis

The functions of human hand such as grasping, lateral hip, pinch, etc. has a vital role in every aspect of the activities of the daily living. Due to interrupted neural pathways or amputation of upper limb, several people lose their hand function and have limitations in the activity of daily living. The brain controlled prosthetic hand, a neuroprosthetic, may bring many opportunities to the life of such subjects and can help them to regain their hand function. The main motivation of this thesis is to develop machine learning techniques that improve the accuracy and reliability of such a neuroprosthetic.

Construction of free-paced or self-paced BCI is one the main goals of the BCI community as it enables the user to initiate any command at will. In free-paced BMI, estimation of movement and idle states, or detection of the onset of a movement is crucial. A movement direction decoder should be initiated only when movement is detected to eliminate false and incorrect decisions in the baseline or idle stages and which lead to the erratic cursor movement seen in all BCI demonstrations. In our scheme, a new hybrid movement vs. idle state decoding system based on the fusion of SVM and HMM structures will be developed. The discriminative/generative approach accepts input features computed with common spatial patterns in different frequency bands of neural activity and returns the likelihood of one of the states of interest. To the best of our knowledge, this is the first study that explores the detection of movement execution and resting states of individual finger movements from ECoG recordings. Another novelty of

our study is that it explored the success of decoding sequential movements in a continuous fashion rather than movements in a trial-based paradigm.

In order to build a hand prosthetics, we tackle the problem of classifying multichannel ECoG related to individual finger movements. We develop and apply novel spatial projection techniques for feature extraction from multichannel ECoG recordings. For this particular aim, as the first step, we applied a recently developed hierarchical spatial projection framework of neural activity for feature extraction from ECoG [20]. The algorithm extends the binary common spatial patterns algorithm to multiclass problem by constructing a redundant set of spatial projections that are tuned for paired and group-wise discrimination of finger movements.

The recent advances in electrode design and recording technology makes it possible to record large number of BCI signals from a larger area of the brain or to get more information from smaller regions using dense electrode grids. Therefore, a dimensionality reduction algorithm needs to be employed to decrease the correlation between channels and improve the signal to noise ratio (SNR). In this scheme, the Common Spatial Pattern (CSP) algorithm is widely used due to its simplicity and lower computational complexity to extract features from high-density recordings both using noninvasive and invasive modalities [20, 43].

Despite the benefits of the CSP method, it also has a number of drawbacks. One major problem of the CSP is that it generally overfits the data when it is recorded from a large number of electrodes and when there is limited number of train trials. Moreover, the chance that CSP uses a noisy or corrupted channel linearly increases with increasing number of recording channels. Robustness over time is also a major drawback in CSP applications [44, 45]. Since all channels are used in spatial projections of the CSP, the classification accuracy may be reduced in cases when the electrode locations slightly change in different sessions. This requires almost identical electrode positions over time, which is difficult to realize [46]. The sparseness of the spatial filter might have an important role to increase the robustness and generalization capacity of the BCI system. We investigate the various types of sparse CSP methods.

The CSP method minimizes the Rayleigh quotient (RQ) of the spatial covariance matrices to achieve the variance imbalance between the classes of interest. The RQ is defined as follows:

$$R(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}}, \quad (1.1)$$

where  $\mathbf{A}$  and  $\mathbf{B}$  are the spatial covariance matrices of two different classes such as baseline and movement or two different types of hand movement etc. and the vector  $\mathbf{w}$  represents the spatial filter that we want to determine. One way to reduce the number of channels used in the projection  $\mathbf{w}$ , is to transform the CSP algorithm into a regularized optimization problem in the form of

$$L(\mathbf{w}) = R(\mathbf{w}) + \lambda \|\mathbf{w}\|_1, \quad (1.2)$$

where  $R(\mathbf{w})$  is the objective function,  $\|\mathbf{w}\|_1$  is the  $\ell_1$  norm based penalty and  $\lambda$  is a constant that controls the sparsity of the solution. In the past few years, there is growing interest in using  $\ell_1$  penalty to construct sparse solutions. However, RQ does not depend on the magnitude of the sparse filter. Therefore, RQ cannot be directly used in a norm based minimization problem, since the optimizer always minimizes the norm along the direction which RQ has been minimized. Therefore, we introduce new objective function and show that it is accurate and feasible to be employed in BCI applications.

A number of studies investigated putting the CSP into alternative optimization forms to obtain a sparse solution for it. In [47] the authors converted CSP into a quadratically constrained quadratic optimization problem with  $\ell_1$  penalty; others used an  $\ell_1/\ell_2$  [11, 44] norm based solution. These studies have reported a slight decrease or no change in the classification accuracy while decreasing the number of channels significantly. Recently, in [48] quasi  $\ell_0$  norm based criterion was used for obtaining the sparse solution which resulted an improved classification accuracy. Since  $\ell_0$  norm is non-convex, combinatorial and NP-hard, they implemented greedy solutions such as Forward Selection (FS) and Backward Elimination (BE) to decrease the computational complexity. It has been shown that BE was better than FS (less myopic) in terms of classification error and sparseness level but associated with very high complexity making it difficult to

use in rapid prototyping scenarios. We introduced oscillating search(OS) method that combines BE and FS techniques to reduce the complexity of algorithm while obtaining comparable classification accuracy.

We observed that the sparse methods are sensitive to the number of channels used (cardinality) in the spatial filter, therefore we regularized the sparse spatial filters using the baseline data. This baseline regularization technique is investigated in terms of accuracy and stability with respect to the cardinality.

The methods that we proposed can be used in case of insufficient amount data for the training. We show their effectiveness by comparing the methods that are used in general signal processing paradigms. We also aim to decode the movement and resting states of individual fingers from multichannel ECoG recordings. This is different from the previous studies that have focused on movement of the entire hand.

## Chapter 2

# Free paced Baseline and Movement Detection

### 2.1 Introduction

In BCI framework, the recorded brain activity is converted to computer commands that are used to control a neuroprosthetic or produce a feedback on the computer screen. In most of the BCI systems, the decoding algorithms that produce the computer commands are applied over a predefined segments of the neural data, which are aligned with respect to the onset of the investigated action (i.e. movement, eye blink, etc.) or external cues. Obviously, this procedure requires the locations of onset of the cue or movement to be known and limits applicability of the algorithm in real-time.

In order to build a BCI, several neuroprosthetic systems have been implemented to process invasively recorded neural signals such as single unit activity (SUA) for the control of a cursor on a computer screen, or for the control of a robotic arm [18]. In most of these systems, the decoding process was restricted to predefined time intervals in which the state of the subject was altered by external cues limiting the flexibility of the constructed system. In order to build a system

that serves a subject’s free will, the state of the brain activity needs to be determined to avoid undesired movement and to obtain accurate results for controlling an external device. For this particular purpose, in a free-paced neuroprosthetics (NP), the states that need to be estimated dynamically are generally,

- baseline (idle),
- planning, and
- movement execution.

Several attempts have been made to decode the dynamic state of the subject from neural activity [5, 34, 37].

A hybrid state detection algorithm is developed for the estimation of baseline and movement states which can be used to trigger a free paced neuroprosthetic and overcome the problems explained above. The hybrid model was constructed by fusing a multiclass support vector machine (SVM) with a hidden Markov model (HMM), where the internal hidden state observation probabilities were represented by the discriminative output of the SVM. A schematic diagram describing our signal-processing framework is depicted in Fig. 2.1. The proposed method was applied to the multichannel electrocorticogram (ECoG) recordings of BCI competition IV [49] to identify the baseline and movement states while subjects were executing individual finger movements. The results are compared to regular Gaussian mixture model (GMM)-based HMM with the same number of states as SVM-based HMM structure. Our results indicate that the proposed hybrid state estimation method out-performs the standard HMM-based solution in all subjects studied with higher latency. The average latency of the hybrid decoder was approximately 290 ms.

In this chapter, we first describe the dataset and the experimental paradigm. Next, we explain our signal-processing framework in detail. Finally, we provide experimental and discuss the results.

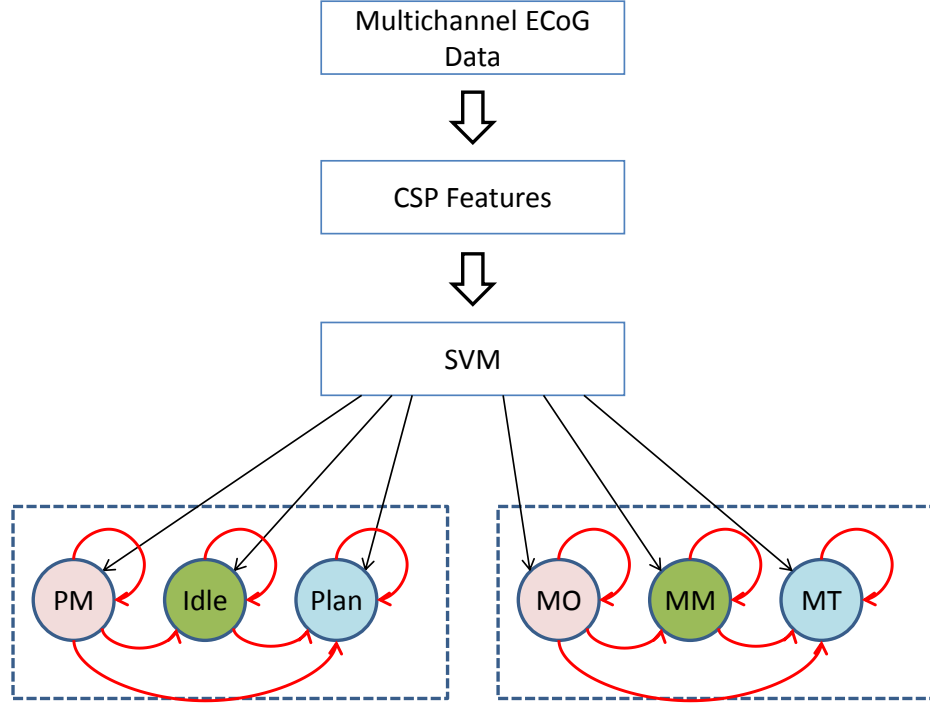


Figure 2.1: Multichannel filtered (1-4, 7-13, 16-30, and 65-200 Hz) ECoG data is fed into a CSP algorithm to reduce channel size. Each band is reduced into four virtual channels. Using the 16 dimensional CSP features, a multiclass SVM classifier is trained to distinguish between resting, planning, movement onset, mid-movement, movement termination and post-movement segments given in Fig. 2.2. These segmentations were derived by aligning the data to movement onset and movement termination. The SVM output probabilities were fed into two HMMs as observation probabilities of the hidden states. Prior and transition probabilities were computed from the training sequence using forward-backward method, where the model is restricted to left-to-right transitions only.

## 2.2 ECoG Data and Preprocessing

### 2.2.1 ECoG Dataset

We used multichannel ECoG data from BCI Competition IV, recorded during finger flexions. This data set was acquired from three epileptic patients at Harborview Hospital in Seattle, WA. The electrode grid was placed on the cortical surface. Each electrode array contained either 48 ( $8 \times 6$ ) or 64 ( $8 \times 8$ ) platinum

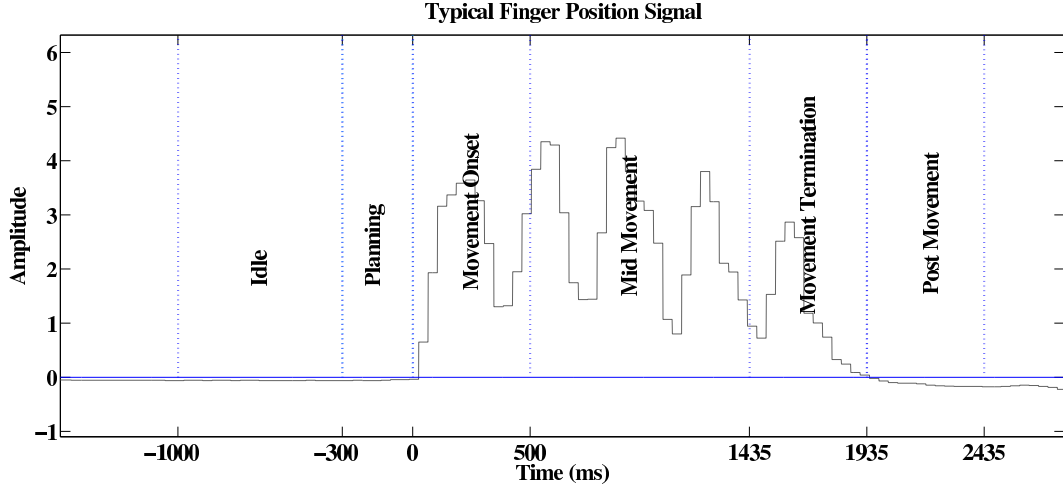


Figure 2.2: A sample finger position plot that describes the states of the HMM model.

electrodes. The diameter of each electrode on the grid is 4 mm. Electrode contacts were embedded in a silicon mat, and were spaced 1 cm apart. Synamps2 amplifiers (Neuroscan, El Paso, TX) were used to digitize and amplify the ECoG signal. The finger index to be moved was indicated with a cue on a computer monitor placed at the bedside. Each cue lasted two seconds and was followed by a two second rest period, during which the screen was blank. Subjects moved one of five fingers three to five times during a cue period, for a total of 10 minutes for each subject [49]. The movements were continuous not trial based. Only the position of the fingers was available to us and was used to distinguish between baseline (resting) and movement states. Consequently, this posed a great challenge in detection of these arbitrary movement executions as no information about the cue and go signal was available to us for our analysis. An exploratory analysis established that the duration and interval between consecutive finger movements varied dramatically. We used for analysis those segments in which each movement lasted a minimum of 1000 ms and consecutive movements were separated by at least 800 ms.

### 2.2.2 Common Spatial Patterns

As in any learning process, the generalization capacity of a model decreases with the increasing dimensionality of the input data. Moreover, the complexity and execution time of decoding algorithms increases with the number of channels of input data. Therefore, a dimensionality reduction algorithm must be employed to limit the amount of data. We applied the common spatial patterns (CSP) [50] algorithm on band-pass filtered multichannel ECoG signals in order to reduce these into a few virtual channels. Specifically, ECoG data from each subject was filtered in 1-4, 7-13, 16-30, and 65-200 Hz frequency bands. Next, each band was transformed into four virtual channels by the CSP algorithm, by taking the first and last two eigenvectors. We computed the spatial projection using

$$\mathbf{X}_{CSP}[n] = \mathbf{W}^T \mathbf{X}[n] \quad (2.1)$$

where  $\Sigma_R$  and  $\Sigma_M \in \mathbb{R}^{C \times C}$  are the covariance matrices of competing classes which are resting and movement classes respectively,  $C$  is the number of channels. The columns of  $\mathbf{W} \in \mathbb{R}^{C \times E}$  are the eigenvectors representing each CSP spatial projection and  $E$  is the number of spatial filters.  $\mathbf{X}[n] \in \mathbb{R}^C$  is the multichannel ECoG data at sample index  $n$ .

$$\Sigma_R \mathbf{W} = \Sigma_M \mathbf{W} \mathbf{\Lambda} \quad (2.2)$$

The eigenvectors of the CSP algorithm were estimated via generalized eigenvalue decomposition (GED) as shown in Eq. 2.2 by contrasting the covariance matrices of the resting and movement segments of the training data. The diagonal matrix  $\mathbf{\Lambda} \in \mathbb{R}^{E \times E}$  has corresponding eigenvalues as its diagonal entries. The covariance matrices are calculated for each trial  $i$  and normalized to its trace to reduce inter trial variability. The trial covariance matrices are then averaged to obtain the baseline ( $\Sigma_R$ ) or movement ( $\Sigma_M$ ) covariance matrices. This procedure is described in [50] in detail. Consequently, the CSP output maximized or minimized the variances of the resting and movement regions in the estimated virtual channels. The variance of each channel was computed in 250 ms windows moving with a 50 ms time step. Finally, the logarithm of the variances were concatenated

across all four frequency bands forming a 16-dimensional feature vector for each time shift.

### 2.2.3 Hybrid HMM-SVM Structure

In order to estimate resting and movement states from the recorded neural data, we built a hybrid discriminative/generative decoder based on the fusion of HMM with SVM. HMMs are widely used in speech processing and have been successfully applied to dynamic state decoding of neural data. Detailed descriptions of this method and its applications were published in the literature [51, 52]. Because it is a generative method, the HMM structure lacks discrimination capability, each model is trained independently from the other competing models. Moreover, observation probabilities are generally modeled by Gaussian Mixture models (GMM), which fail to represent the distribution of the features in high dimensional space in the presence of a low amount of training data and/or outliers. We therefore aimed to replace the observation probabilities of internal states of the HMM with the posterior probability output estimates of a multiclass SVM. Specifically, rather than using a GMM, the extracted features were fed to a multiclass SVM that was tuned to separate the distribution of the internal states. However, such an approach requires the labels of the features belonging to each state so that the SVM classifier can be trained. In this scheme, we constructed six different states by aligning the neural data with respect to movement onset and termination. These states consisted of the following six periods:

- i. Resting (baseline),
- ii. Movement planning,
- iii. Movement onset,
- iv. Mid-movement,
- v. Movement termination stage and
- vi. Post-movement stage

A schematic diagram representing these alignments and their duration is given in Fig. 2.2. Because there was no exact timing information for the planning period, we used the 400 ms window preceding each movement onset as the planning state (P). The 400 ms segment immediately following each movement termination was defined as the post-movement state (PM). The interval between PM and P was defined as the resting segment. Movement was segmented into three different states, with the first 400 ms of each movement defined as movement onset (MO). The 400 ms segment immediately preceding cessation of movement was defined as the movement termination state (MT). The interval between MO and MT was defined as the mid-movement state. We labeled the features originating from each state in the continuous training data and then fed them into the multiclass SVM for discrimination. Since the duration of the resting and mid-movement states was variable, the number of feature vectors that we extracted from these segments was much higher than for the other states, causing a bias in the decision boundary of the SVM classifier. Consequently, we reduced the number of samples for resting and mid-movement states in order to compensate for the variability in numbers of samples for each state. Specifically, the majority class was down-sampled by randomly eliminating its samples. The SVM module provides an estimated posterior probability for each state by using a one against the other classification strategy. A radial basis function was used as the kernel of the SVM. The output of the SVM module was then used in conjunction with the Forward-Backward algorithm to estimate the transition probabilities of the HMM. We used the LibSVM toolbox to implement the multiclass SVM [53] and the HMM toolbox of [52] to build the hybrid decoder. It should be noted that this procedure differs from the traditional HMM training, in which the observation and transition probabilities are altered in each iteration of the standard expectation maximization (EM) algorithm. In our case, the observation probabilities were the SVM outputs, and these were fixed during the iterative estimation of transition probabilities. The HMM had three hidden states. In each state, the observation probabilities were represented with three mixtures. Only left-to-right transitions were allowed in both hybrid HMM and traditional GMM based HMM, as shown in Fig. 2.1.

Table 2.1: The state decoding accuracies of the hybrid and traditional HMM based methods with 60 training trials using a decoding sequence length 10.

|           | Hybrid Decoder | HMM  |
|-----------|----------------|------|
| Subject 1 | 91.5           | 89.2 |
| Subject 2 | 89.2           | 88   |
| Subject 3 | 92.7           | 91.6 |
| Avg.      | 91.2           | 89.6 |

We tested our hybrid decoding system and the traditional HMM algorithm on the ECoG data derived from the three subjects of BCI Competition IV, described in Section 2.2.1. In contrast to those studies that have decoded transition from baseline to planning/movement, our challenge involved decoding transitions from movement to a resting/baseline state, as well. In order to decode the dynamic state of a subject, a sequence of observations is needed. Unlike trial based experiments, the data we used contained no predefined start and end points. In such a situation, a fixed segment of the data, which is shifted along the signal, is generally used to execute the state decoders. The use of long data segments can cause large latencies and numerical overflow of the output. Consequently, we studied the effect of different sequence lengths, for example, 5, 10, 15 and 20, on the estimation of the resting and movement states. After decoding each sequence with the constructed models, the model with the maximum posterior probability was used to determine the class of the feature sequence. Moreover, we executed several experiments with various training-set sizes, in order to examine the robustness of each algorithm against the limited amount of training data. We trained the algorithms using 10 to 70 train trials by increasing the set size by 10.

## 2.3 Results

The average classification accuracies of the hybrid and HMM methods are listed in Table 2.1. We observed that for all subjects studied, the hybrid SVM-HMM decoder provided better decoding accuracies than the traditional HMM method. On average, the detection accuracy of the hybrid method was 91.2%, whereas the

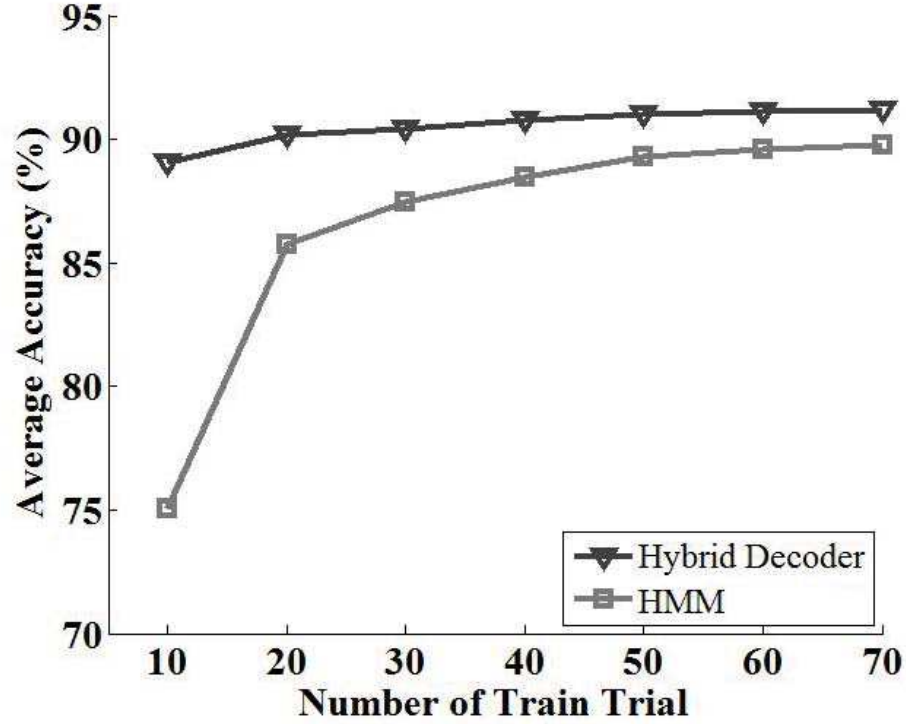


Figure 2.3: Average accuracy vs. the number of train trial with decoding length 10

HMM solution provided 89.6% decoding accuracy.

The average decoding accuracies of each method with a varying number of training trials is given in Fig. 2.3. We observed that the hybrid decoder provided superior decoding accuracies with a low number of training trials, and its performance slowly increased with increasing the training set size. In contrast, the accuracy of HMM was quite poor when using a low number of training trials. In contrast to the hybrid decoder, the accuracy of HMM rapidly improved with increasing training-set size, ultimately stabilizing after 50 training trials.

We studied decoding accuracy as a function of decoding sequence length. We observed that the decoding results were quite poor with a sequence length of five and improved rapidly by increasing the sequence length to 10. The maximum decoding results were obtained with sequence lengths of 10 and 15 in both methods, which corresponded to time windows of approximately 700 and 950 ms,

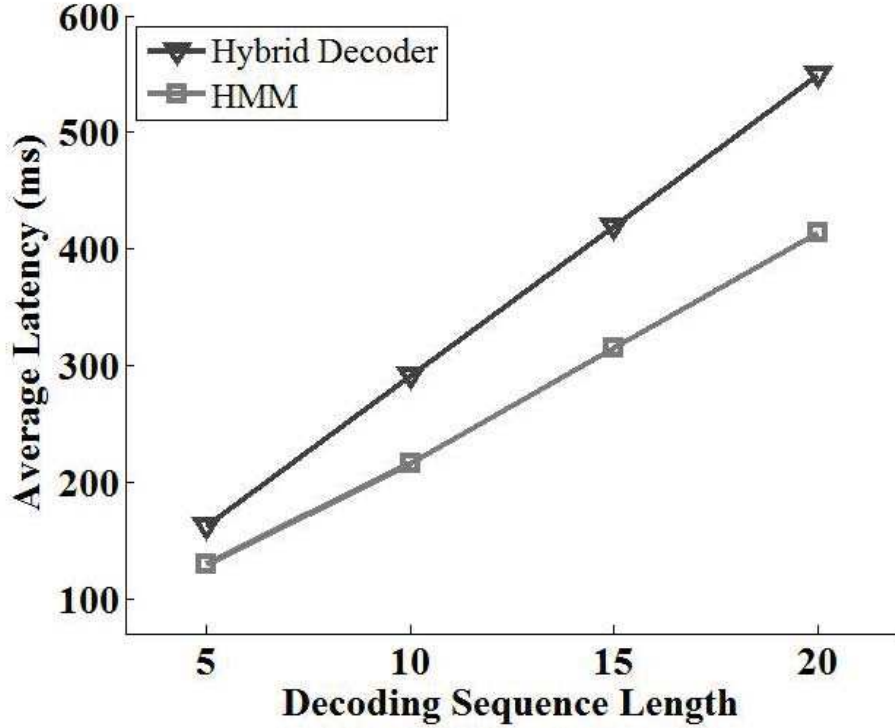


Figure 2.4: Average latency vs. decoding sequence length

respectively. The average latency of each method versus the decoding sequence length is given in Fig. 2.4. We observed that the latency of HMM was superior to the hybrid decoder. For a sequence length of 10, the latency for the hybrid and HMM were 290 and 215 ms, respectively. Although slightly better results were obtained using a sequence length of 15 with the hybrid decoder, we observed that the latency increased dramatically from 290 to 410 ms.

The temporal decoding accuracies for a representative subject at movement onset and termination are shown in Figs. 2.5 and 2.6. We observed that the decoding results at movement onset had a sharp transition compared to movement termination. We also noted that the decoding errors and latencies were higher at movement termination, as compared to movement initiation. These observations indicate that decoding state transitions from movement to resting state poses new challenges. In the subjects we studied, movement onset was associated with a burst of gamma spectrum activity, which slowly decreased towards the end of

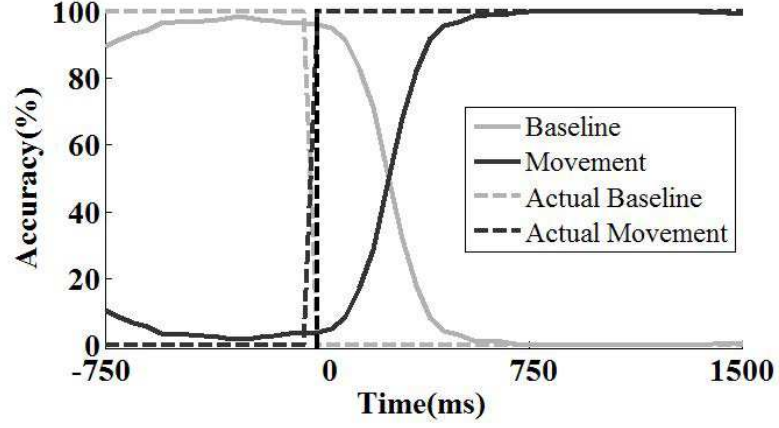


Figure 2.5: Average accuracy vs. time for subject 3 aligned to the movement onset.

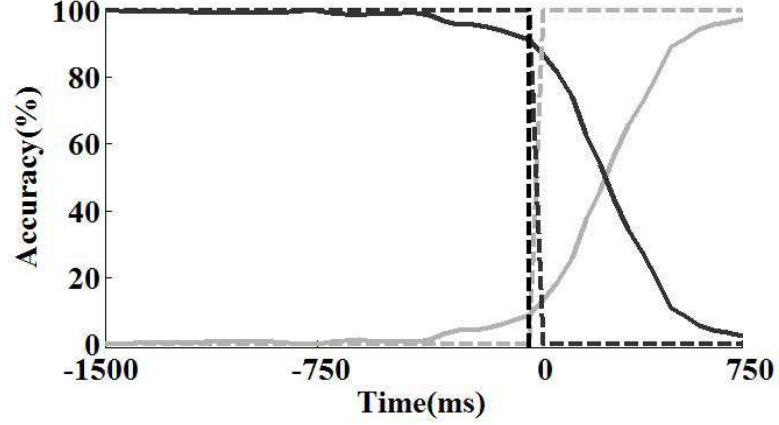


Figure 2.6: Average accuracy vs. time for subject 3 aligned to the movement termination. A hybrid decoder with a decoding sequence length of 15 was used.

the movement. There was no similar pattern observed at movement termination. This could in part explain the lower accuracy and the larger latency that characterized movement termination.

## 2.4 Summary

In this section, we report a hybrid decoder based on the fusion of SVM and HMM for dynamic state detection based on data derived from multichannel ECoG

recordings during consecutive movements of individual fingers. We have demonstrated experimentally that the latency of state decoding using ECoG data during finger movements is comparable to that obtained using SUA data during directional hand movements. We compared our method to the traditional HMM technique. The hybrid decoder out-performed the HMM technique in all three subjects studied. The main advantage of using SVM within the hybrid decoder is that the posterior probability of each state is estimated simultaneously and tuned for discrimination. This advantage might overcome the lack of discriminative capability of HMMs, as each model is trained independently from the other competing models. Moreover, the higher generalization capacity of SVM due to the large margin makes the algorithm a good candidate for applications in which a limited number of training trials exists on which to base estimates of the model parameters. However, such an approach requires supervised training in order to estimate the state discriminators, which is automatically accomplished by the traditional HMM.

## Chapter 3

# Decoding of Individual Finger Movements using Redundant Spatial Projections

### 3.1 Introduction

In the past few years a number of research groups focused on decoding individual finger movements from invasively recorded neural activity [35, 54, 55]. The motivation for such an effort was to build a hand prosthetics that can be controlled solely by brain activity in the scope of a brain machine interface (BMI). Achieving such a detailed decoding performance was possible by invasive assessment of brain activity as it provides higher spatial and temporal resolution and signal to noise ratio (SNR) compared to noninvasive techniques such as EEG as described earlier.

Recently, finger movement decoding problem was studied using human subjects where the neural activity was assessed with electrocorticography from 64 channels [56]. Shenoy and his colleagues used 3 different band features (11-40 Hz, 71-100 Hz and 101-150 Hz) for each channel. They used all 3 band powers and they represent their data with 192 features. To select a subset of features and

give decisions, a linear programming machine (LPM) which is a sparse variant of support vector machine (SVM) classifier, was utilized in their study. They also employed the original SVM classifier to give decisions. To transform the binary decisions to multiclass decisions, they used one versus all (OVA) and all versus all (AVA) strategies which are widely implemented strategies in multiclass pattern recognition algorithms. In the OVA strategy involving  $N$  classes,  $N$  different binary classifiers are trained, each of the classifiers compares the samples of a particular class against the samples of the remaining classes. To sort the unknown test data points, each test data point is fed into the classifiers and the classifier which outputs the largest value is chosen. On the other hand, in AVA method  $\binom{N}{2}$  pairs of the classes are trained to build a multiclass classifier from pairwise binary classifiers. For the test point, each of these pairwise classifiers collectively determines the final decision [57].

In this chapter we cope with the same problem of classifying of the movements of five fingers using ECoG. We employ a recently introduced redundant spatial projections based on common spatial patterns (CSP) algorithm for feature extraction from ECoG data for the identification of individual finger movements [20]. The algorithm combines the pairwise comparisons of the fingers with the comparisons of the neighbor finger groups to achieve a classification accuracy increment on the test data. We utilized a support vector machine (SVM) classifier to map the extracted spatial features into class labels. We use the ECoG data recorded from three subjects to demonstrate the efficiency of our decoding strategy.

The rest of the chapter is organized as follows. First, we explain the redundant spatial feature extraction and classification framework. Then, we represent our results, and finally give a brief discussion about the results.

## 3.2 Methods and Materials

### 3.2.1 Multiclass CSP with Hierarchical Grouping

The ECoG data is generally recorded with subdural electrode grids from epileptic patients. A majority of electrodes is likely overlap with cortical regions out of the hand area of the motor cortex. Consequently, a small number of recording channels carry finger movement related information. In any learning process the generalization capacity of the model decreases with increasing dimensionality of the input features [58, 59]. Therefore, a dimension reduction algorithm needs to be employed to decrease the dimensionality.

In this chapter, we applied the common spatial patterns (CSP) [50] algorithm on band pass filtered multichannel ECoG signals to reduce them into a few virtual channels. Reducing the number of channels also reduces the number of features, therefore the dimension of feature space is decreased to improve the generalization capability of the classifier. The spatial filtering also improves the SNR of spatially correlated ECoG data. The CSP is a subspace technique which is widely used among BMI community in binary decision problems for feature extraction. The spatial filters are a weighted linear combination of recording channels which are tuned to produce spatial projections maximizing the variance of one class and minimizing the other. We computed the spatial projection using

$$\mathbf{X}_{CSP}[n] = \mathbf{W}^T \mathbf{X}[n] \quad (3.1)$$

where the columns of  $\mathbf{W} \in \mathbb{R}^{C \times E}$  are the weight vectors representing each spatial projection,  $C$  is the number of channels and  $E$  is the number of spatial filters.  $\mathbf{X}_{CSP}[n] \in \mathbb{R}^C$  is the multichannel neurophysiological data at sample index  $n$ . The weight vectors of the CSP algorithm are estimated via generalized eigenvalue decomposition by contrasting the covariance matrices of the first class (i.e. thumb finger) and the second class (e.g., one of the finger data that is not the first class, here thumb finger) of a two class training data set.

Since we are tackling a multiclass problem, here we used the strategy of [20], to apply the CSP to the five-class finger movement data. In more detail, we

constructed several spatial filters tuned to contrast pairs of finger movements such as 1 vs. 2; 1 vs. 3; 2 vs. 4 etc. Moreover, the spatial projections were extended to the group-wise contrasts of fingers such as 1, 2 vs. 3, 4 and 5 within the same spirit of [20]. Here, we expect that the adjacent fingers will have similar neural representations which can be used in improving the SNR of the spatial covariance matrices while computing the projections. A schematic diagram of decoding algorithm is presented in Fig. 3.1 and the complete list of the classes are shown in Table 3.1.

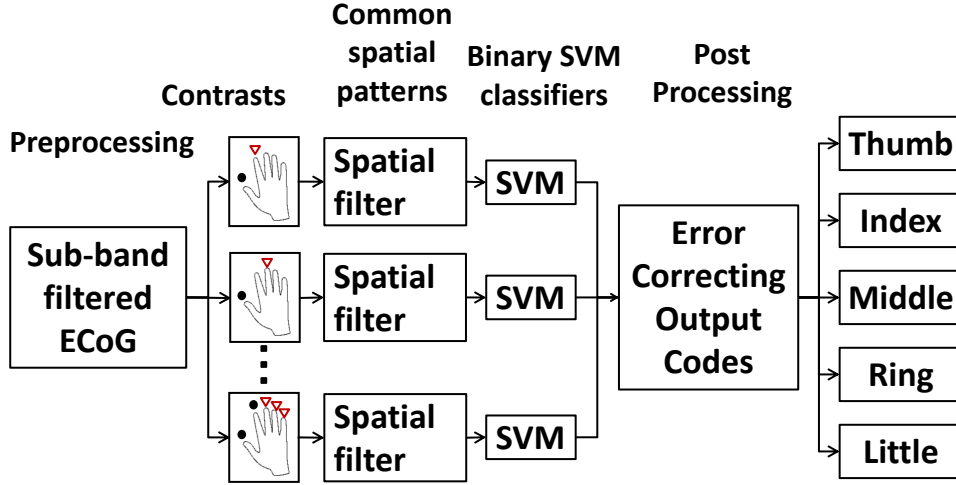


Figure 3.1: The ECoG signal is bandpass filtered and a redundant set of contrasts were constructed to compute CSP for pair wise and group wise discrimination. The resulting spatial projection features are fed into corresponding SVM classifiers. The pair-wise and group-wise SVM results are fused using ECOC strategy to get the final classification decisions.

### 3.2.2 Support Vector Machine based Classifier

For each of the spatial projection, we constructed an SVM classifier with a radial basis function (RBF) kernel and probabilistic output. To construct the classifier, we used libsvm software [53], which is a publicly available toolbox. The SVM parameters  $g$  (kernel parameter) and  $C$  (cost or regularization parameter) were set to 0.25 and 100.

Table 3.1: The complete list of competing classes for pairwise and redundant classifiers. We have 10 classifier for the pairwise classification and 5 classifiers for the redundant classification.

|           |         | Finger Indexes |         |         |         |   |   |   |   |   |   |
|-----------|---------|----------------|---------|---------|---------|---|---|---|---|---|---|
| Pairwise  | Class 1 | 1              | 1       | 1       | 1       | 2 | 2 | 2 | 3 | 3 | 4 |
|           | Class 2 | 2              | 3       | 4       | 5       | 3 | 4 | 5 | 4 | 5 | 5 |
| Redundant | Class 1 | 1, 2           | 2, 3    | 3, 4    | 4, 5    |   |   |   |   |   |   |
|           | Class 2 | 3, 4, 5        | 1, 4, 5 | 1, 2, 5 | 1, 2, 3 |   |   |   |   |   |   |

We constructed 10 pair-wise classifiers that contrast one finger movement to another. In addition, we used adjacent fingers as a hierarchy rule and contrasted two fingers versus the others with the expectation that consecutive fingers are correlated in their neural representation. For this particular setup, five spatial projections and five corresponding classifiers were constructed. In total there are 15 spatial projections (10 paired, 5 group-wise) and related SVM classifiers as shown in Table 3.1. Each classifier provides a probability output  $p$  for a feature set being one class and  $(1 - p)$  of being in the other. We employed an error correcting output code (ECOC) step to post process the outputs of redundant classifiers and provide a final decision [20, 60]. This last step was accomplished by multiplying the vector representing the log scaled classifiers output with the ECOC decoding matrix  $M$  of  $K \times L$  with entries  $m_{i,j} \in 0, 1$  where  $L(= 30)$  is two times the number of binary classifiers and  $K$  is the number of classes (i.e., 5 finger movements). The index corresponding to the maximum value of the ECOC output was selected as the predicted finger of the test data.

### 3.2.3 ECoG Data

We used multichannel ECoG data from BCI competition IV, which is described in Section 2.2.1. To reduce the data rate we low pass filtered the ECoG data with a 220 Hz cutoff frequency and down sampled it to 500 Hz. In order to identify the

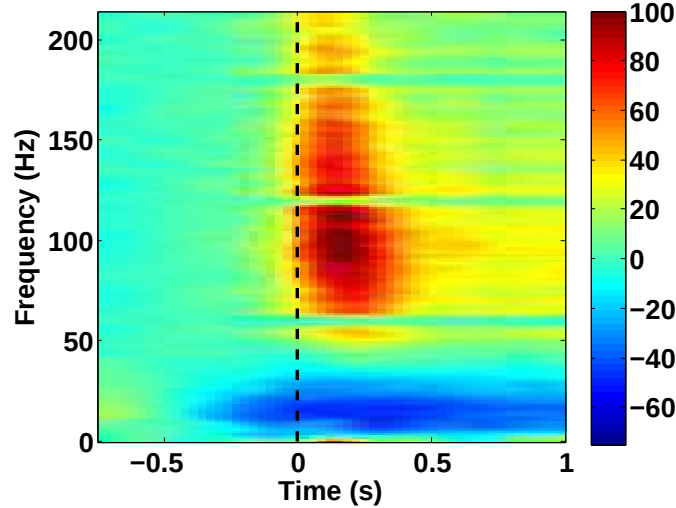


Figure 3.2: The average time frequency map computed from all subjects using the most reactive channel set selected for each subject. The t-f surface was normalized to the energy in the first 500 ms interval to identify modulated frequencies. Positive values represent energy increase and negative the decrease with respect to the baseline.

reactive frequency bands we implemented time-frequency analysis of ECoG data using short time Fourier transform. We aligned the ECoG data according to the movement onset covering a period of 750 ms before the onset and 1000 ms after it. We normalized the time-frequency plane to the energy in the first 500 ms period of the idle state. We provide a time-frequency map representing the group average of most reactive channels in each subject in Fig. 3.2. We observed a broadband energy increase in 65-200 Hz frequency band with the onset of movement. The energy in 7-32 Hz decreased before the onset of the movement. We also observed energy increase in 0-6 Hz band with the onset of the movement.

Based on these observations, the ECoG data of each subject was subband filtered in 0-6, 7-13, 14-32 and 65-200 Hz frequency bands. We used one second data following movement onset for spatial feature extraction. Next, each band was transformed into four virtual channels with CSP algorithm by taking the first and last two eigenvectors. The variance of each channel was computed in all aligned data to get 4-dimensional feature vectors for each trial. Finally, the variances are log transformed and used as input features to SVM classifiers.

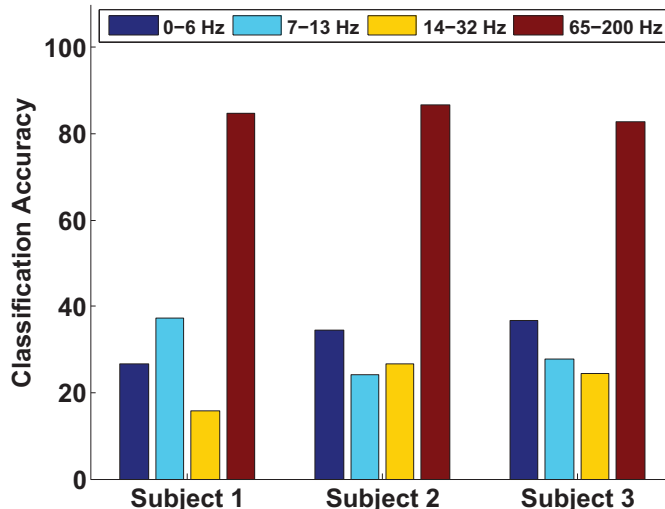


Figure 3.3: The classification accuracies in each frequency band for three subjects. The highest accuracies are obtained from 65-200 Hz band features in all subjects.

### 3.3 Results

We used a  $10 \times 10$  fold cross validation procedure to estimate the classification accuracy of our system. In Fig. 3.3, for each frequency subband we present the classification accuracies. In all subjects, the gamma (65-200 Hz) band provides the highest decoding accuracy. The average classification accuracy over all three subjects was 86.3%. In two subjects the second highest classification rate was obtained from 0-6, Hz band whereas for the first subject the  $\alpha$  band (7-13 Hz) resulted to the second highest rate. Interestingly, the 14-32 Hz band provided consistently the minimum classification accuracy on all subjects. Although this band was modulated with the movement, it did not provide any information about the index of the executed finger movement but the cognitive state.

In Figs. 3.4 and 3.5, we present the confusion matrix of our redundant classification system in gamma band and the correlation matrix of five-finger sensor data respectively. The confusion matrices show that the misclassifications generally occurred between fingers 4 and 5. We note that for subjects 2 and 3

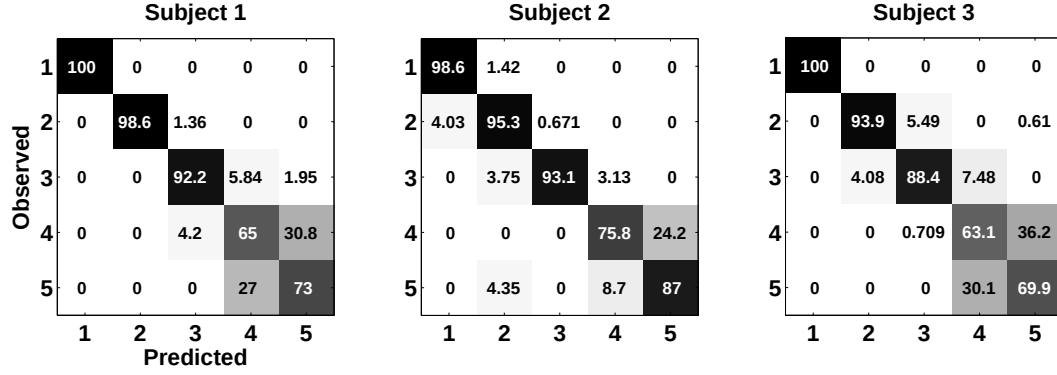


Figure 3.4: Confusion matrices for subject 1, 2 and 3. Note that the majority of misclassification occurred between the ring finger (4th) and the little finger (5th). Almost perfect separation was obtained for the thumb (1st).

the finger sensor data was also correlated between fingers 4 and 5 but not for subject 1. The misclassification for subjects 2 and 3 can be explained by the correlated movements of last two fingers. Interestingly, for the first subject despite the uncorrelated sensor data, the misclassification occurred once again between the last two fingers. This can be justified with the assumption of correlated neural representation of the adjacent fingers. The confusion matrices of other subjects also support this assumption. In contrary, for subject three, although the sensor correlation of adjacent fingers was high the misclassifications between the first four fingers were very low. This indicates that the neural representations of the first four fingers were distinguishable. However, the sensor measurements were somehow correlated which may originate from a mechanical cross talk of adjacent finger movements due to the hand anatomy of this particular subject. We note that, although very small, the misclassification occurred generally between the adjacent fingers. It should be noted that the correlated neural activity between adjacent fingers also improved the classification rates in the redundant case as the groupings improved the SNR of the common pattern shared by the adjacent fingers.

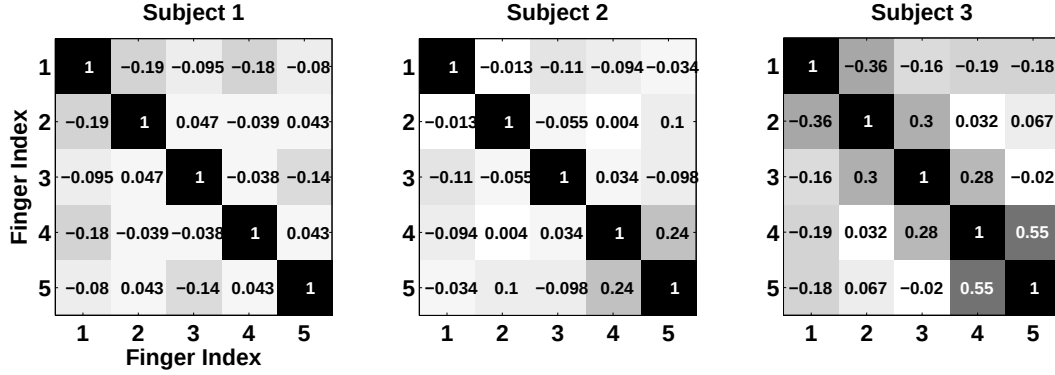


Figure 3.5: The correlation matrices of finger position sensor signals for all subjects. The color map is constructed according to absolute value of the correlation.

Table 3.2: The classification results for paired wise (non-redundant) and redundant decoding strategies in 65-200 Hz frequency range.

|           | Paired | Redundant |
|-----------|--------|-----------|
| Subject 1 | 80.7   | 86        |
| Subject 2 | 82.8   | 89.4      |
| Subject 3 | 78.9   | 83.4      |
| Mean      | 80.8   | 86.3      |

In order to quantify the gain we obtained with the redundant decoding strategy we compared it to the case where only paired (non-redundant) spatial projections and classification were executed. In Table 3.2, we show the classification accuracy for 65-200 Hz frequency bands for the redundant and paired (non-redundant) decoding strategies. We observed that in all subjects the redundant decoding strategy provided better results with respect to the paired one. These findings are in accordance with [20] where the same technique was used to decode movement direction from LFPs recorded in motor cortex. In average the paired solution provided 80.8% classification accuracy. We note that hierarchical structure noticeably increased the decoding performance of our system.

### 3.4 Summary

We report here a hybrid decoder based on the fusion of SVM and HMM for dynamic state detection based on data derived from multichannel ECoG recordings during consecutive movements of individual fingers. We have demonstrated experimentally that the latency of state decoding using ECoG data during finger movements is comparable to that obtained using SUA data during directional hand movements. We compared our method to the traditional HMM technique. The hybrid decoder outperformed the HMM technique in all three subjects studied. The main advantage of using SVM within the hybrid decoder is that the posterior probability of each state is estimated simultaneously and tuned for discrimination. This advantage might overcome the lack of discriminative capability of HMMs, as each model is trained independently from the other competing models. Moreover, the higher generalization capacity of SVM due to the large margin makes the algorithm a good candidate for applications in which a limited number of training trials exists on which to base estimates of the model parameters. However, such an approach requires supervised training in order to estimate the state discriminators, which is automatically accomplished by the traditional HMM.

# Chapter 4

## Sparse Spatial Projections Using New Objective Function with $\ell_1$ Norm based Penalty Function

### 4.1 Introduction

Common spatial pattern (CSP) method is widely used in brain machine interface (BMI) applications to extract features from the multichannel neural activity through a set of spatial projections. These spatial projections minimize the Rayleigh quotient (RQ) (Equation 4.2) as the objective function, which is the variance ratio of the classes. The CSP method easily overfits the data when the number of training trials is not sufficiently large and it is sensitive to daily variation of multichannel electrode placement, which limits its applicability for everyday use in BMI systems.

In this chapter, we construct a computationally efficient spatially sparse projection (SSP) based on a novel objective function with similar characteristics to RQ. This new objective function can be minimized in the form of (4.4) to address the drawbacks of regular CSP method. We show that our new objective function has the same minimization solution as RQ and it depends on the magnitude of the

spatial filter. The magnitude dependency of our new objective function allows us to use a continuous and differentiable function approximating  $\ell_1$  norm [61] as regularization term in an unconstrained optimization framework and can be solved using standard algorithms with low complexity. The rest of the chapter is organized as follows. In the following section, we describe our novel objective function and its relation to RQ. Then we explain its use in an unconstrained optimization problem. Next, we apply our method on the BCI competition IV ECoG dataset involving individuated movements of five fingers [49] and the BCI competition III EEG dataset IVa [62] involving imaginary foot and hand movements. We also compare our method to standard CSP and the  $\ell_0$  norm based BE solution given in [48]. Finally, we discuss our results and provide future directions.

## 4.2 Methods and Materials

### 4.2.1 Standard CSP and a New Objective Function

In the CSP framework, the spatial filters are a weighted linear combination of recording channels, which are tuned to produce spatial projections maximizing the variance of one class and minimizing the other. The spatial projection is computed using

$$\mathbf{X}[n]_{CSP} = \mathbf{W}^T \mathbf{X}[n] \quad (4.1)$$

where the columns of  $\mathbf{W}$  are the vectors representing each spatial projection,  $\mathbf{X}[n]$  is the multichannel data and  $n$  is the time sample index of the data.

$$R(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \quad (4.2)$$

Maximizing the RQ in Eq. (4.2) is identical to the following optimization problem.

$$\begin{aligned} & \underset{\mathbf{w}}{\text{maximize}} && \mathbf{w}^T \mathbf{A} \mathbf{w} \\ & \text{subject to} && \mathbf{w}^T \mathbf{B} \mathbf{w} = 1. \end{aligned} \quad (4.3)$$

After writing this optimization problem in the Lagrange form and taking the derivative with respect to  $\mathbf{w}$ , we obtain the identical problem in the form of  $\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}$  which is the generalized eigenvalue decomposition (GED). The solutions of this equation are the joint eigenvectors of  $\mathbf{A}$  and  $\mathbf{B}$  and  $\lambda$  is the associated eigenvalue of a particular eigenvector.

$$L(\mathbf{w}) = R(\mathbf{w}) + \alpha\|\mathbf{w}\| \quad (4.4)$$

The drawbacks of the CSP method that are described earlier lead us to find a way to *sparsify* the spatial filter to increase the classification accuracy and the generalization capability of the method. We assume that the discriminatory information is embedded in a few channels where the number of these channels is much smaller than the actual number of all recording channels. So the discrimination can be obtained with a sparse spatial projection, which uses only informative channels. Let us assume that the data was recorded from  $C$  channels. We are interested in obtaining a sparse spatial projection using an unconstrained minimization problem in the form of Eq. (4.4), where  $\mathbf{w}$  has only  $c$  nonzero entries,  $\text{card}(\mathbf{w}) = c$  and  $c \ll C$ . We note that the  $R(\mathbf{w})$  does not depend on the magnitude of  $\mathbf{w}$ , as shown in the following equation and Fig. 4.1. Let  $\mathbf{w}^* = \beta\mathbf{w}$ , then

$$R(\mathbf{w}^*) = R(\beta\mathbf{w}) = \frac{\beta^2\mathbf{w}^T\mathbf{A}\mathbf{w}}{\beta^2\mathbf{w}^T\mathbf{B}\mathbf{w}} = R(\mathbf{w}) \quad (4.5)$$

where  $\beta$  is any scalar which is not equal to zero. Since  $R(\mathbf{w})$  does not depend on the gain of  $\mathbf{w}$ , the optimizer arbitrarily reduces the gain of  $\mathbf{w}$  to minimize regularization term  $\alpha\|\mathbf{w}\|$  after finding the direction that minimizes  $R(\mathbf{w})$ . Thus, the solution of the optimization problem that uses  $R(\mathbf{w})$  as an objective function is essentially the same as the GED solution.

To find a sparse solution we need to have an objective function that depends on the gain of  $\mathbf{w}$ . In this scheme, we replaced  $R(\mathbf{w})$  with the following objective function.

$$G(\mathbf{w}) = \mathbf{w}^T\mathbf{A}\mathbf{w} + \frac{1}{\mathbf{w}^T\mathbf{B}\mathbf{w}} \quad (4.6)$$

This function is bounded from below and has interesting properties. Let us define  $a(\mathbf{w}) = \mathbf{w}^T\mathbf{A}\mathbf{w}$  and  $b(\mathbf{w}) = \mathbf{w}^T\mathbf{B}\mathbf{w}$ . If we define RQ in terms of  $a(\mathbf{w})$  and  $b(\mathbf{w})$

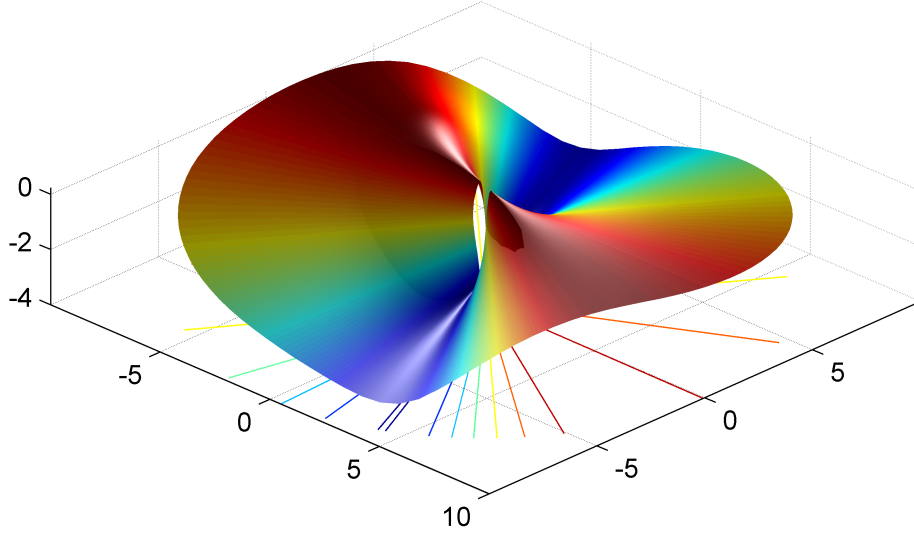


Figure 4.1: The RQ surface for two dimensional vector space. Notice that the value of the RQ does not change for a particular direction.

such that

$$R = \frac{a(\mathbf{w})}{b(\mathbf{w})}$$

then our new objective function can be expressed as

$$G(\mathbf{w}) = a(\mathbf{w}) + \frac{1}{b(\mathbf{w})} = R(\mathbf{w})b(\mathbf{w}) + \frac{1}{b(\mathbf{w})} \quad (4.7)$$

The derivative of  $G(\mathbf{w})$  with respect to  $R(\mathbf{w})$  is equal to  $b(\mathbf{w})$  which is always positive for all  $\mathbf{w}$ . This indicates that our objective function  $G(\mathbf{w})$  decreases with a decrease in  $R(\mathbf{w})$  value. After taking the derivative of  $G(\mathbf{w})$  with respect to  $b(\mathbf{w})$  and solving Equation 4.8,

$$\frac{\partial G(\mathbf{w})}{\partial b} = R(\mathbf{w}) - \frac{1}{b(\mathbf{w})^2} = 0 \quad (4.8)$$

we find that  $b(\mathbf{w})$  is equal to  $\sqrt{R(\mathbf{w})^{-1}}$ . By inserting  $b(\mathbf{w})$  value into the Eq. 4.7 we obtain the minimum value of  $G(\mathbf{w}_{min})$  as  $2\sqrt{R(\mathbf{w}_{min})}$ . This result shows that the vector  $\mathbf{w}$  that minimizes  $R(\mathbf{w})$  also minimizes  $G(\mathbf{w})$ .

We plug  $G(\mathbf{w})$  into unconstrained optimization formulation in Eq. (4.4) as the objective function. Rather than working to solve Eq. (4.4) with a non-differentiable  $\ell_1$  penalty, we replaced it with a twice differentiable smooth version

of  $\ell_1$  (epsL1) which is sufficiently close to minimizing  $\ell_1$  [61]. The main advantage of this approach is that, since epsL1 and  $G(\mathbf{w})$  are both twice differentiable we can directly apply an unconstrained optimization method to minimize  $L(\mathbf{w})$  [63]. The epsL1 is defined as

$$\|\mathbf{w}\|_{1\epsilon} = \sum_{i=1}^C \sqrt{w_i^2 + \epsilon} \quad (4.9)$$

where  $\epsilon$  is a sufficiently small parameter and  $C$  is the dimension of  $\mathbf{w}$ . The epsL1 approximates the  $\ell_1$  norm and they are identical when  $\epsilon$  is equal to zero. Twice differentiability of the epsL1 norm allows us to use it when  $w_i$  is equal to zero unlike the regular  $\ell_1$  norm which is not differentiable at zero.

The solution  $\mathbf{w}$  that minimizes the function  $L(\mathbf{w}) = G(\mathbf{w}) + \alpha\|\mathbf{w}\|$  tends to become sparse as  $\alpha$  gets bigger. The entries of  $\mathbf{w}$  generally were not exactly equal to zero, so we normalized  $\mathbf{w}$  to its maximum absolute value and eliminated the weights consequently corresponding channels that do not exceed a predefined threshold ( $=10^{-2}$ ). We used “fminunc” function of Matlab to find the solution of our unconstrained minimization problem. We computed the gradient (Eq. D.16) and Hessian (Eq. D.19) of the minimization equation analytically as shown in equations and input it to Matlab’s function to increase the speed of the solver. We computed the desired cardinality by implementing a bisection search [64] on the  $\alpha$ . The upper border of  $\alpha$  was determined initially using the

$$\frac{G(\mathbf{w}_c)}{\|\mathbf{w}_c\|}$$

ratio where  $\mathbf{w}_c$  is the full CSP solution. In case the initial upper border results a cardinality larger than the desired value, we kept doubling the  $\alpha$  parameter until we obtained a  $\alpha$  that results a cardinality which is less than or equal to the target value.

$$L(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \frac{1}{\mathbf{w}^T \mathbf{B} \mathbf{w}} + \alpha \sum_{i=1}^M \sqrt{w_i^2 + \epsilon} \quad (4.10)$$

Following the above procedure, we computed the first spatial filter  $\mathbf{w}$  that minimizes the  $G(\mathbf{w})$  which also minimizes the  $R(\mathbf{w})$ . The solution that maximizes

$R(\mathbf{w})$  is also a useful spatial filter. Therefore, we interchanged the matrix  $\mathbf{A}$  and  $\mathbf{B}$  to find a solution that maximizes  $R(\mathbf{w})$ . In order to find multiple sparse filters we deflated the covariance matrices with sparse vectors using the Schur complement deflation method described in [65].

#### 4.2.2 Codifference Matrix Based on Addition

To remove the effect of the potential pulse artifacts from the covariance matrices and increase the computational power of the algorithm by using addition instead of multiplication we define a multiplierless “vector product”. The core of this vector product is basically the multiplierless product operator. The operator  $\odot$  is basically an addition operation but the sign of the result behaves like the multiplication operation. Let  $a$  and  $b$  are two real numbers:

$$a \odot b = \begin{cases} a + b, & \text{if } a > 0 \text{ and } b > 0 \\ a - b, & \text{if } a < 0 \text{ and } b > 0 \\ -a + b, & \text{if } a > 0 \text{ and } b < 0 \\ -a - b, & \text{if } a < 0 \text{ and } b < 0 \\ 0, & \text{if } a = 0 \text{ or } b = 0 \end{cases} \quad (4.11)$$

We can also express Eq. 4.11 as follows

$$a \odot b = \text{sign}(ab) (|a| + |b|) \quad (4.12)$$

Notice that the sign of 0 is defined to be 0.

The operator  $\odot$  satisfies totality, associativity and identity properties, therefore it is a monoid function. In other words, it is a semigroup with identity property. We successfully used similar statistical methods in [66, 67]. Another similar statistical function is the average magnitude difference function (AMDF) which is widely used in speech processing to determine the periodicity of voiced sounds.

The new “vector product” is defined in a similar mannner. Let  $\mathbf{f} =$

$[f_1, \dots, f_L]^T$  and  $\mathbf{g} = [g_1, \dots, g_L]^T$  be two vectors

$$\mathbf{f}^T \odot \mathbf{g} = \sum_{l=1}^L f_l \odot g_l, \quad (4.13)$$

it is important to note that,

$$\mathbf{f}^T \odot \mathbf{f} = 2\|\mathbf{f}\|_1. \quad (4.14)$$

In order to decrease the computational cost and eliminate the pulse artifacts we introduce the codifference matrix as follows

$$\mathbf{S} = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}[n] - \boldsymbol{\mu}) \odot (\mathbf{x}[n] - \boldsymbol{\mu})^T \quad (4.15)$$

where the operator  $\odot$  acts like a vector multiplication operator, however, the scalar multiplication is replaced by an additive operator  $\odot$  imitating multiplication.

Because  $a \odot b = b \odot a$ , the codifference matrix is also symmetric as the covariance matrix. Codifference behaves similar to the covariance function. If two variables tend to vary together, codifference function produces positive results as the covariance. When two variables tend to vary inversely, codifference equation gives negative results.

One advantage of the multiplierless operator is that it reduces the effect of the outliers, such as pulse artifacts. The multiplication operation produces very large values when it acts on two large numbers. On the other hand, multiplierless product operation results relatively small values compared to the multiplication operator. This property provides a *smoothing* effect on the noise.

It is well known that  $\ell_1$  norm based algorithms are more robust to impulsive noise [68]. The new vector product defined in Eq. 4.13 is related with the  $\ell_1$  norm (see Eq. 4.14).

### 4.2.3 ECoG & EEG Datasets

We applied the SSP method on two different datasets, multiclass ECoG and two class EEG of BCI competitions IV and III, respectively.

The ECoG data was recorded from three subjects during finger flexions and extensions [49] with a sampling rate of 1 kHz. The electrode grid was placed on the surface of the brain. Each electrode array contained 48 (8x6) or 64 (8x8) platinum electrodes. The finger index to be moved was shown with a cue on a computer monitor. The subjects moved one of their five fingers 3-5 times during the cue period. The ECoG data of each subject was subband filtered in the gamma frequency band (65-200Hz) as in [69]. We used one second data following the movement onset in the analysis. The dataset contains around 146 trials for each subject.

We also used the BCI competition III dataset IVa [62]. The dataset is recorded from five subjects (aa, al, av, aw, ay) who were asked to imagine either right foot or right index finger movements. The sampling rate of the data was 1 kHz and data was recorded from 118 channels. The EEG signal was filtered in the range of 8-30 Hz. There were 140 trials available for each class. Once again, one second data following the cue was used in the analysis.

For both ECoG and EEG datasets, the signal was transformed into four spatial filters by taking first and last two eigenvectors for each CSP methods. After computing the spatial filter outputs, we calculated the energy of the signal and converted it to log scale for each sparse filter and we used them as input features to lib-SVM classifier with an RBF kernel [53]. We also investigated the efficacy of using Linear Discriminant Analysis (LDA) classifier [70] which is parameter free decision function.

Since we are tackling a multiclass problem for the ECoG dataset, we used the pairwise discrimination strategy of [20] to apply the CSP to the five-class finger movement data. In other words, we constructed sparse spatial filters tuned to contrast pairs of finger movements such as 1 vs. 2; 1 vs. 3; 2 vs. 4 etc.

In each dataset, we compared the SSP to the standard CSP and to the  $\ell_0$  norm based BE method of [48] as it provided superior results in terms of classification accuracy and reduced cardinality. We studied the classification accuracy as a function of cardinality. With the purpose of finding optimum *sparsity* level for the classification, we computed several sparse solutions, with decreasing number of cardinality on the training data. For the ECoG dataset, the sparse CSP methods were employed with  $c \in \{40, 30, 20, 15, 10, 5, 2, 1\}$ . For the EEG dataset, we computed the sparse filters with  $c \in \{80, 60, 40, 30, 20, 15, 10, 5, 2, 1\}$ . For each level we computed the corresponding RQ value. We studied the inverse of the RQ (IRQ) curve and determined the optimal cardinality where its value suddenly dropped indicating we started to lose informative channels.

For the ECoG dataset, half of the trials were used in training and the remaining half for testing. In average, we used  $15 \pm 2$  train trials per finger (the thumb, index, middle, ring and little fingers respectively). The EEG dataset contains 140 trials per class and subject. We used 70 trials in training to estimate the sparse filters, and 70 trials for testing. In both datasets, the value of the  $\epsilon$  in epsL1 regularization term was chosen to be  $10^{-6}$ .

### 4.3 Results

We observed that for the SSP method, any particular  $\alpha$  value can lead to different cardinality and normalized IRQ values for different subjects as shown in Figs. 4.2 and 4.3. In particular, this inter subject variability of IRQ did not allow us to use the same  $\alpha$  value for all subjects (See Fig. 4.2a and 4.2b). However, the variability of IRQ values of different subjects was lower when we fixed the cardinality, as shown in Figs. 4.4a and 4.4b. Consequently, due to this reduced variability and to compare our method to the BE technique, we studied the classification error as a function of cardinality. In order to decide on the optimal cardinality level to be used on the test data, the IRQ values were computed on the training data, scaled to their maximum value and averaged over subjects. In the following step, we computed the slope of the IRQ curve and normalized it to its maximum value

to get an idea about the relative change in the IRQ.

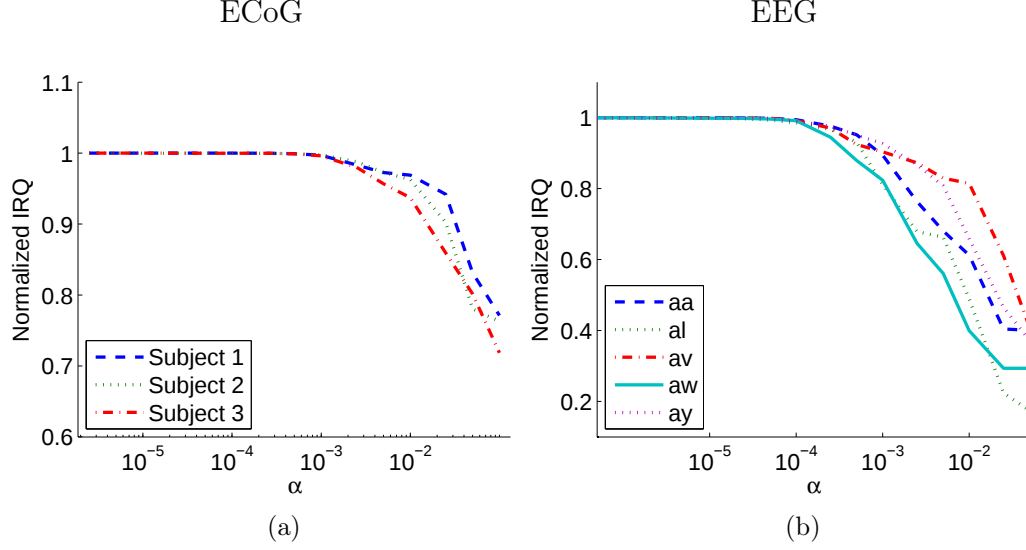


Figure 4.2: Normalized IRQ values are shown in (a) for ECoG and in (b) for EEG data.

We depicted the change in IRQ values for each cardinality as shown in Fig. 4.5a and 4.5b. As expected, decreasing the cardinality of the spatial projection resulted to a decrease in the IRQ value. To determine the optimum cardinality to be used in classification on the test data, we selected the cardinality that is below 10 % of the maximum relative change (See the dashed lines in Fig. 4.5a and 4.5b). For the ECoG dataset, the cardinality value was found to be 5 and for the EEG data set, it was found to be 15 for the SSP method. For the BE method these values are 5 and 10 respectively. These indices perfectly corresponded to the elbow of the IRQ curve, which indicates loss of informative channels. In Table 4.1 and 4.2 , we provide the classification results and selected cardinalities for the ECoG and EEG data set using different methods including SSP, CSP and  $\ell_0$  based greedy solution, BE. In order to give a flavor about the change in error rate versus the cardinality, we provided the related classification error curves in Fig. 4.6a and 4.6b. For the ECoG data selected cardinality on the training data provided minimum test errors. However, for the EEG data, although the minimum classification error was obtained at cardinality 5 for the BE method, we noticed that we identified the optimum cardinality as 10 on the training data.

On all subjects we studied, we observed that the SSP method consistently

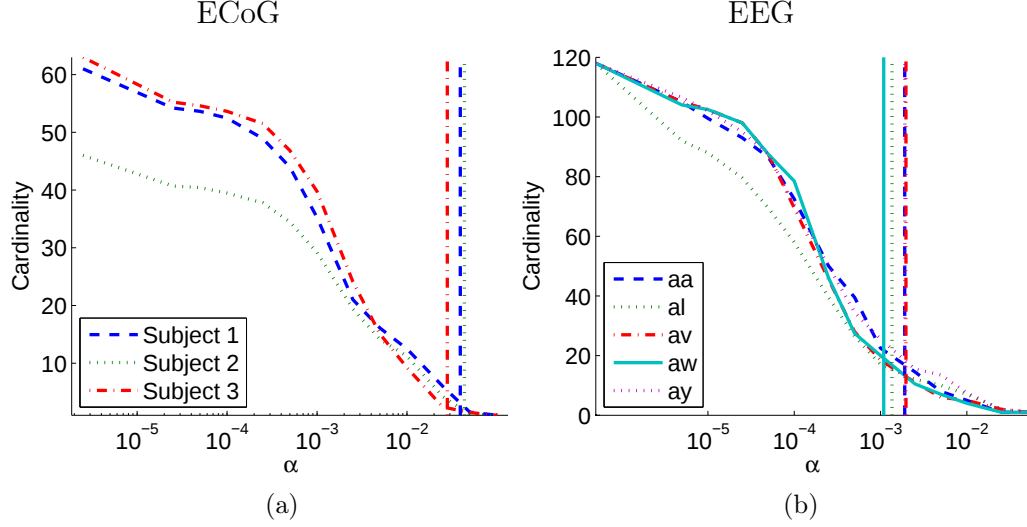


Figure 4.3: The cardinality vs  $\alpha$  value of the minimization function  $L(\omega) = G(\omega) + \alpha\|\omega\|$  (a) ECoG and (b) EEG data for each subject. The vertical lines indicate the  $\alpha$  values that are initially chosen for bisection search. The initial data point corresponds to  $\alpha = 0$  which produces the regular CSP solution.

Table 4.1: EEG dataset classification error rates (%) for each subject using SVM classifier

|     | Cardinality | aa   | al  | av   | aw  | ay   | Avg  |
|-----|-------------|------|-----|------|-----|------|------|
| BE  | 10          | 13.6 | 2.9 | 30.7 | 2.1 | 5.0  | 10.9 |
| SSP | 15          | 19.3 | 1.4 | 23.6 | 4.3 | 5.7  | 10.9 |
| CSP | 118         | 23.6 | 3.6 | 32.1 | 2.9 | 11.4 | 14.7 |

outperformed the CSP method. We noted that the minimum error rate was obtained with SSP method for ECoG data. Both SSP and BE methods used only 5 channels to achieve the minimum error rate. As expected the full CSP solution did not perform as good as the other sparse methods and likely overfitted the training data. The SSP method improved the classification error rate with an error difference of 13.2%. We obtained comparable results on EEG data using the SSP and BE methods. The error difference between regular CSP and SSP is less apparent in the EEG dataset where the difference between classification accuracies was 3.8%. This could be due to the high number of training trials used in EEG data. We studied the effect of the amount of training data on the classification accuracy and presented the results in Fig. 4.7a. When a small

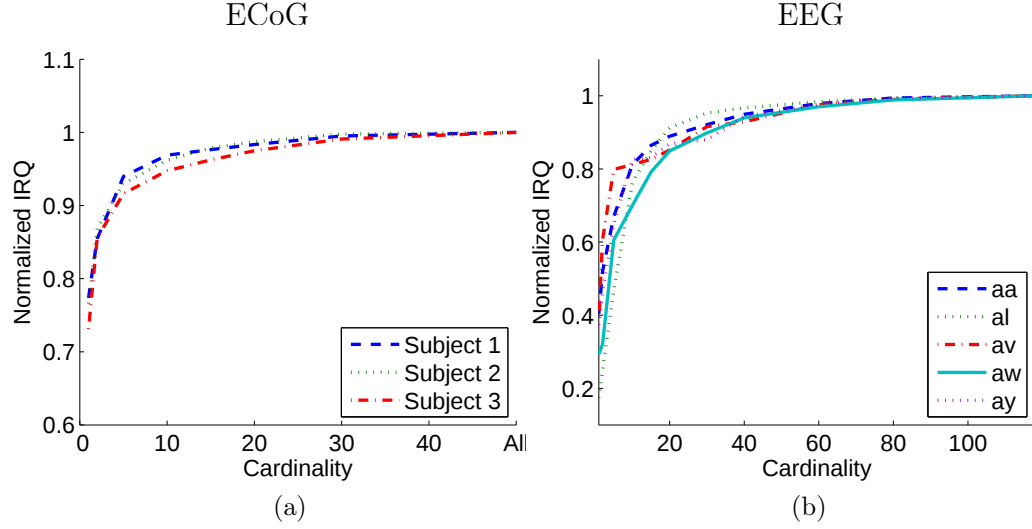


Figure 4.4: The Normalized IRQ values vs cardinality for each subject is shown in (a) and (b).

Table 4.2: ECoG dataset classification error rates (%) for each subject using SVM classifier

|     | Cardinality | Subject 1 | Subject 2 | Subject 3 | Avg |
|-----|-------------|-----------|-----------|-----------|-----|
| BE  | 5           | 19.8      | 17.1      | 16.8      | 18  |
| SSP | 5           | 18.4      | 13.4      | 18        | 17  |
| CSP | All         | 30.7      | 26        | 32.8      | 30  |

number of training trials, as low as 15 are used in the EEG dataset, the difference between the sparse and standard CSP technique was more than 6%. Interestingly, with increasing number of training trials the SSP method consistently provided better results and the difference remained between 3-4%. There was no noticeable difference between SSP and BE.

The classification results obtained with LDA classifier are given in Table 4.3. We observed that the LDA classifier which does not involve parameter selection like SVM, provided slightly higher error rates for the sparse solutions. This could be due to nonlinear decision surface and maximum margin identified by the SVM classifier. Interestingly, in both datasets, the LDA classifier resulted in lower error rates with the non-sparse CSP solution.

In order to compare the computational complexity of SSP method to the BE,

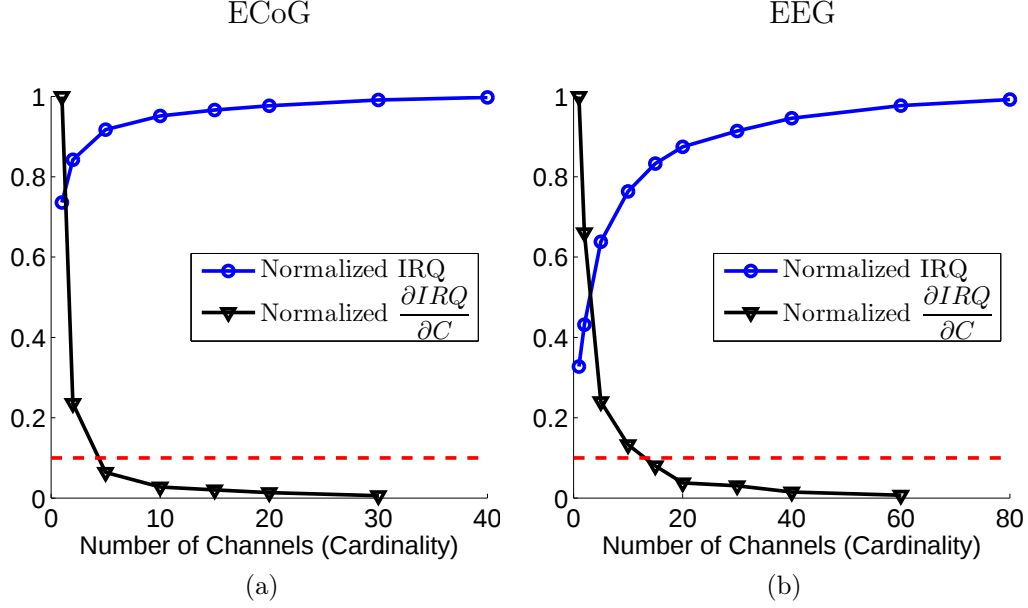


Figure 4.5: The average IRQ of all subjects versus cardinality (a) ECoG and (b) EEG data. The red line is the 10 percent threshold that determines the optimum cardinality to be used in the test data. Five channels for ECoG and 15 for EEG datasets.

we computed sparse filters with a cardinality of two from an increasing number of recording channels on simulated data. The training was performed on a regular desktop computer with 4GB of RAM and equipped with a CPU running at 2.66 Ghz. The elapsed time per filter computation increased exponentially for the BE method and linearly for the SSP method as shown in Fig. 4.7b. With 128 channels, the BE algorithm computed a single spatial filter with two nonzero entries in 90 seconds. For the SSP method with the same setup above, the elapsed time was less than a second. Although, we used the relative change in the IRQ to identify the optimum sparsity level, one can also run a typical k-fold cross validation procedure to identify the optimum level. However, in such a case training the system with BE method will take several hours which may not be feasible for BMI applications. On the other hand with the SSP method training through cross validation can be executed in a few minutes.

In order to evaluate the effect of noise and intersession variability on the performance of our approach we studied three different controlled experiments:

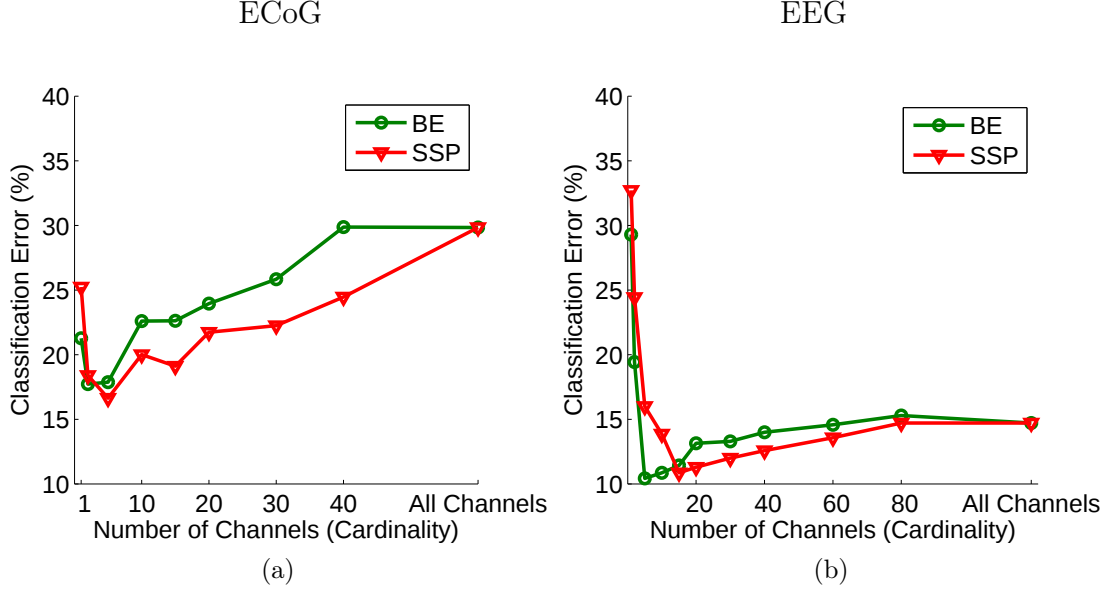


Figure 4.6: The classification error curves of SSP and BE methods versus the cardinality are given in (a) for ECoG and in (b) for EEG. The last data point corresponds to the results obtained from standard CSP which uses all channels.

Table 4.3: Average test error rate and corresponding cardinality

|                    | ECoG |      |      | EEG  |      |      |
|--------------------|------|------|------|------|------|------|
|                    | BE   | SSP  | CSP  | BE   | SSP  | CSP  |
| LDA Test Error (%) | 19.9 | 18.9 | 24.1 | 11.4 | 11.1 | 14.3 |
| SVM Test Error (%) | 17.9 | 16.6 | 29.8 | 10.9 | 10.9 | 14.7 |
| Cardinality        | 5    | 5    | All  | 10   | 15   | All  |

- i) adding Gaussian noise to a randomly selected channel in ECoG and EEG data,
- ii) simulation of electrode displacement in EEG data, and
- iii) adding pulse noise to a randomly selected channel in ECoG.

During the first experiment, we added zero mean Gaussian noise to one of the channels and calculated the resulting classification error to obtain final classification accuracy for the noisy data. This experiment was repeated for all channels and then the resulting average classification error was computed. The variance of the additional noise was increased in a controlled manner and proportional to

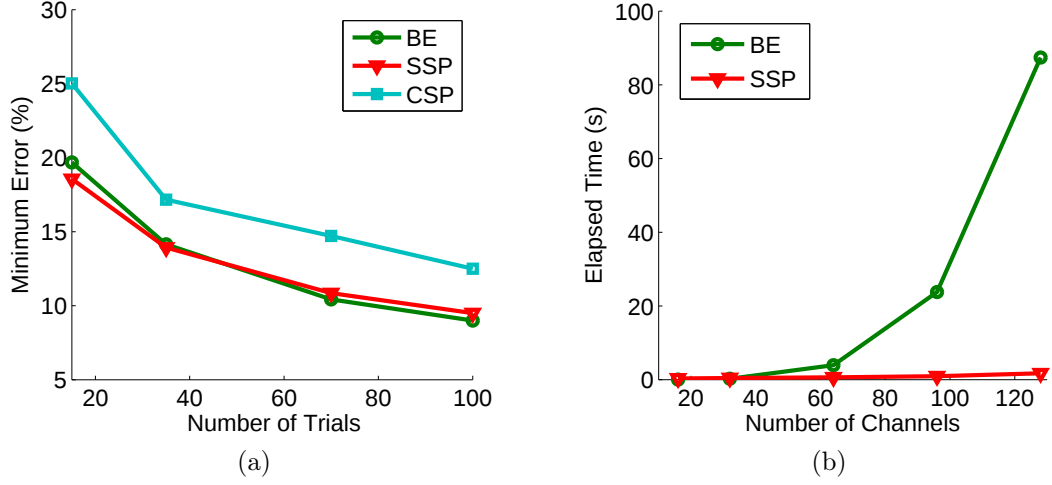


Figure 4.7: (a) The minimum error vs. the number of trials. (b) The average elapsed time to estimate a spatial filter with a cardinality of two vs. the number of total recording channels.

the average variance of all channels. The ratio of the noise to the variance of the original data was named as noise to signal plus noise ratio (NSNR) since the original signal already contaminated from different noise sources. Fig. 4.8 depicts the classification accuracy vs. NSNR which was expressed in decibel (dB) scale. It was noted that an increase in NSNR caused error to increase for all methods, however the increase in CSP method was more than the sparse filters. While using the sparse methods in the ECoG data, the classification error reached a plateau after 5 dB whereas the standard CSP error increased monotonically. A similar behavior was observed in the EEG data.

In the second experiment, in order to study intersession variability we simulated electrode displacements by interpolating the EEG test data at different positions. Since electrode locations for the ECoG data were not available, we conducted this experiment with the EEG data only. We randomly determined the direction and the amount of displacement of the new electrode locations. To be more realistic, we introduced displacement which was uniformly distributed over 118 electrodes varying between 0-50% of the distance to the nearest electrode as shown in Fig. 4.9a. The testing of algorithms was conducted on the interpolated EEG data on the displaced electrode locations. The classification results vs. cardinality is shown in Fig. 4.9b. It was noted that the error change

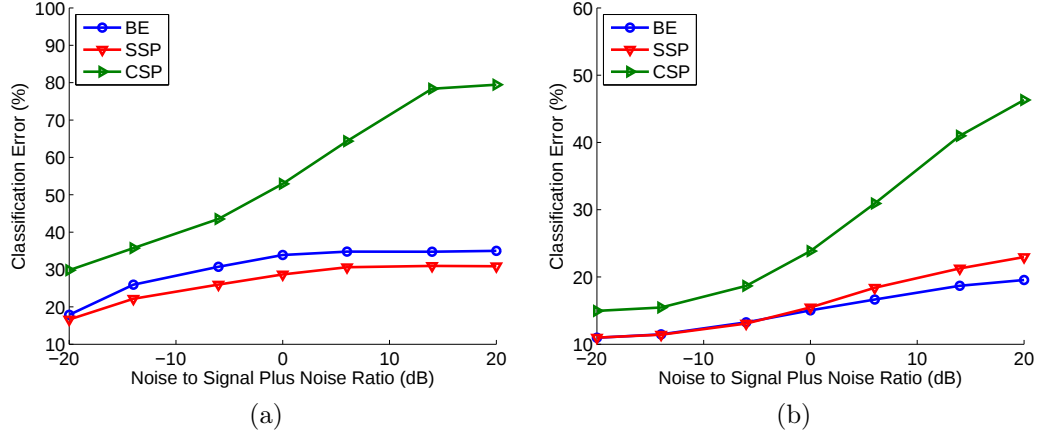


Figure 4.8: The noise to signal plus noise ratio (NSNR) vs. classification accuracy for (a) ECoG and (b) EEG sets.

for CSP, BE and SSP methods are 13%, 12% and 10%, respectively. The increase in error rate was not as severe as in the previous noise contaminated channel experiment. The sparse methods had slightly lower error increase than the standard CSP. Initially, it was expected that the sparse methods would not be this vulnerable to displacement in the electrode locations. At this point, we speculate that the smoothness of the projections have a certain advantage on the displaced electrodes. For instance, due to sparseness, it is likely that the shifted electrode locations are weighted with zero. This makes the sparse projections discard the shifted channels that fall outside the projection zone. In contrary, the standard CSP is associated with some sort of smoothness and many crucial channels, although displaced, could still be projected with similar weights due to the correlated distribution. Consequently, it should be noted that not only sparseness but also smoothness of projections could be an important parameter contributing to the generalization capability of the methods. Never the less, once again we need to highlight that, the increase in error rate was lower for the sparse methods.

In the third experiment, we tackle with the pulse noise using a smoothing CSP method. In this method, we computed the covariance matrices with the multiplierless vector product defined in section 4.2.2. We perform this experiment on the ECoG data using regular CSP filters to investigate the usefulness

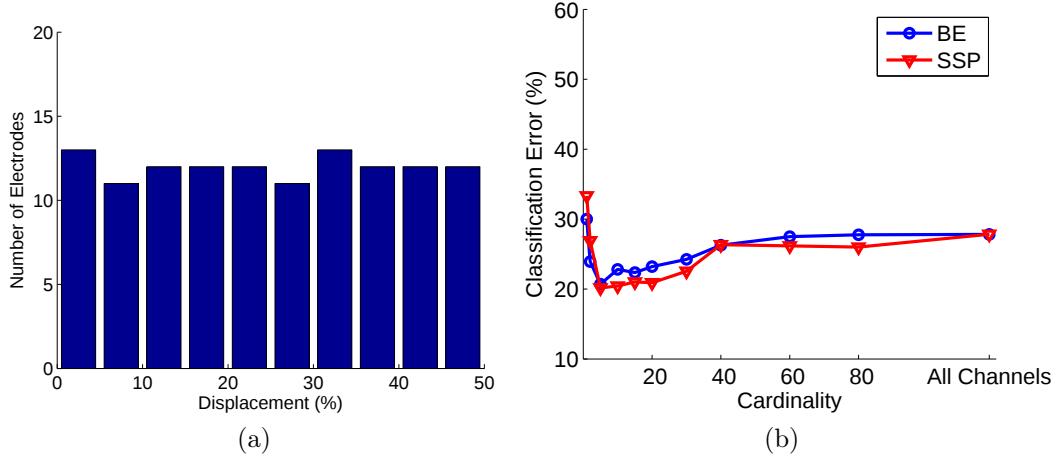


Figure 4.9: (a) The histogram of the number of electrodes with respect to the displacement induced on the test data. (b) The cardinality vs classification error on the test data with displaced electrode locations.

of the multiplierless vector product against the pulse noise. In this experiment, we contaminate one of the randomly selected channel with pulse noise that is proportional to the variance of the data. After the contamination we filter the data with a band-pass filter with 65-200 Hz. The noisy channel has impulses after bandpass filtering as shown in Fig. 4.10. We added pulses onto 10 different locations for the selected channel. We computed the covariance matrices using both multiplierless method and traditional method. The level of noise versus the classification error is plotted for both methods in Fig. 4.11. For low noise levels, traditional CSP has better results, however the classification error of the multiplierless method tends to become less than the traditional CSP method with an increase in the noise level. The reason for this difference is that the pulse noise dominates the signal energy with increase in the noise level and this problem is less severe for the multiplierless CSP, since it uses summation instead of multiplication. Furthermore, this new operator only uses summation and comparison, the computational complexity can be reduced with application specific integrated circuits (ASIC), which implements the multiplierless operation.

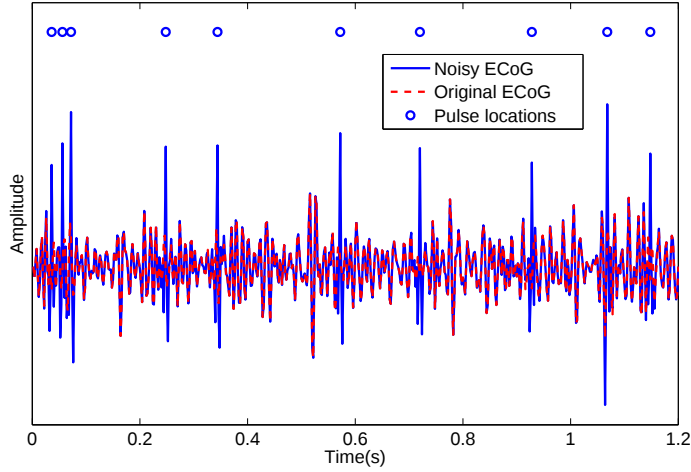


Figure 4.10: Sample of a pulse noise contaminated ECoG signal. The pulse noise is added to the randomly selected channel on 10 different locations

## 4.4 Summary

The need for the sparse filters is apparent when there is large number of recording electrodes and insufficient amount of training data. To minimize overfitting on the training data and eliminate noisy channels, we introduced a spatially sparse projection technique (SSP) based on a novel objective function. Unlike the RQ, this new objective function has a dependency on the filter magnitude. By using an approximated  $\ell_1$  norm, we computed the sparse spatial filters through an unconstrained minimization formulation with standard optimization algorithm. We applied our method to ECoG and EEG datasets and compared its efficiency to standard CSP, and to a  $\ell_0$  norm based greedy technique. The SSP method outperformed the standard CSP on both datasets and provided comparable results to  $\ell_0$  norm based method, which is associated with higher computational complexity. On the ECoG data, the SSP method provided 44% decrease in the error rate compared to standard CSP method and used only five channels in each spatial projection. The error difference between regular CSP and SSP is less apparent in the EEG dataset as SSP method provided 26% decrease in the error rate. In contrary to the ECoG data, we also observed that more channels were used to achieve minimum classification accuracy in the EEG dataset. This could

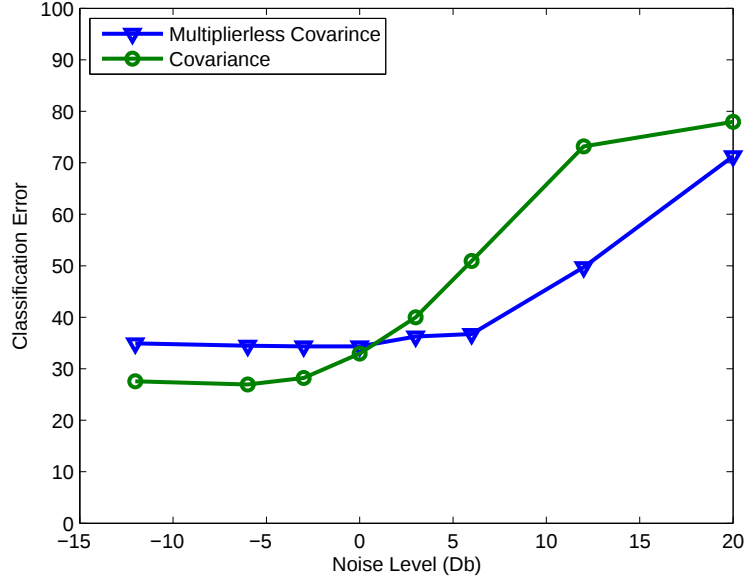


Figure 4.11: The noise to signal plus noise ratio (NSNR) vs. classification accuracy for ECoG dataset using CSP method.

be due the low spatial resolution originating from the volume conduction and low SNR of the EEG. Nevertheless, the SSP algorithm was able to reach a minimum error rate with only 15 channels. Our results indicate that SSP method can be effectively used to extract features from both EEG and ECoG datasets with large number of recording channels. We note that the sparse methods provided superior results compared to the standard CSP when there is a noisy/corrupted channel in the test data. In another setup where displaced channels were used to simulate intersession variability, we note that the sparse methods had slightly better robustness than the standard CSP. These observations indicate that the sparse spatial projection framework can be effectively used as a robust feature extraction engine of future BCI systems. Furthermore, we demonstrated the effectiveness of the new multiplierless covariance on the pulse noise contaminated ECoG data with regular CSP method. The smoothing property of the multiplierless operation improves the accuracy in case of high amplitude pulse noise contamination.

# Chapter 5

## Sparse Spatial Projections Using $\ell_0$ Norm based Search

### 5.1 Introduction

As pointed out earlier brain machine interfaces (BMI) research seeks to develop technologies that enable patients to communicate with their environment solely through the use of brain signals. Recent advances in electrode design and recording technology allow for recording of neural data from large numbers of electrodes. These large electrode arrays are used to sample a greater brain region, or to obtain more detailed information from a smaller portion of the brain using a dense setup. The increased number of recording channels demands greater computational power and has the potential to introduce irrelevant or highly-correlated channels. The common spatial pattern (CSP) algorithm is a widely used method to decrease the computational complexity of the classification algorithms as well as to decrease the correlation between channels and improve the signal-to-noise ratio (SNR) of multichannel recordings from both noninvasive and invasive modalities [20, 43].

The CSP method is a useful tool to solve the problems related to the number of channels by linearly combining channels into a few virtual channels. The CSP

method forms new virtual channels by maximizing the Rayleigh quotient (RQ) of the spatial covariance matrices. This procedure creates a variance imbalance between the classes of interest. The RQ is defined as

$$R(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \quad (5.1)$$

where  $\mathbf{A}$  and  $\mathbf{B}$  are the spatial covariance matrices of two different classes and  $\mathbf{w}$  is the spatial filter or the virtual channel.

Although useful, the CSP method has also some disadvantages. The most common problem of the CSP method is that it generally overfits the data when the number of trials is limited and when the signal is recorded from a large number of channels. Moreover, the chance of recording corrupted or noisy signal is increased with the number of recording channels. Since all channels are used in spatial projections of CSP, the classification accuracy may be reduced in situations in which electrode locations varies slightly between different recording sessions. This requires nearly identical electrode spatial location over time, which is difficult to realize [46].

To address the drawbacks of the traditional CSP method, various sparse spatial filter methods are used by researchers [11, 44, 47, 48, 71]. These methods attempted to compute sparse CSP (sCSP) filters by converting CSP into a quadratically constrained quadratic optimization problem with  $\ell_1$  penalty [47] or used an  $\ell_1/\ell_2$  norm based regularization parameter with the traditional CSP method [11, 44]. The authors of [44, 47] have reported a slight decrease or no change in the classification accuracy while decreasing the number of channels significantly. Recently, in [48] a quasi  $\ell_0$  norm based criterion was used for obtaining the sparse solution, which resulted in an improved classification accuracy. Since  $\ell_0$  norm is non-convex, and solving  $\ell_0$  norm problem is combinatorial and NP-hard [72], they implemented greedy solutions such as forward selection (FS) and backward elimination (BE) to decrease the computational complexity. It was shown that the less *myopic* BE method out-performs the FS method with a dramatically higher computational cost. In [71] recursive weight elimination (RWE) is proposed which has lower complexity and comparable classification accuracy with BE with higher cardinality, which is the number of non-zero entries in the

sparse spatial filter.

In this chapter, we fuse all these greedy techniques to obtain sparse filters yielding low classification error rates with reduced computational complexity. In this scheme, we used the oscillating search, a subset selection technique from a large set of features [73, 74]. Unlike the BE, FS or RWE, the OS does not operate in a fixed direction. We show that using the OS method one can extract sparse filters at low cardinalities with lower complexity and error rates. The rest of the chapter is organized as follows. In the following section, we describe the greedy search algorithms and their relations with the new OS algorithm. Next, we apply our method on the BCI competition III EEG dataset IVa [62] involving imaginary foot and hand movements. We also compare our method to standard CSP and other greedy search algorithms such as BE, FS and RWE. Finally, we discuss our results and provide future directions.

## 5.2 Material and Methods

### 5.2.1 Standard CSP Method

In the CSP framework, the spatial filters are a weighted linear combination of recording channels, which are tuned to produce spatial projections maximizing the variance of one class and minimizing the other. The spatial projection is computed using Equation 3.1.

Since RQ (5.1) does not depend on the magnitude of  $\mathbf{w}$ , maximizing the RQ is identical to the following optimization problem.

$$\begin{aligned} & \underset{\mathbf{w}}{\text{maximize}} && \mathbf{w}^T \mathbf{A} \mathbf{w} \\ & \text{subject to} && \mathbf{w}^T \mathbf{B} \mathbf{w} = 1. \end{aligned} \tag{5.2}$$

After writing this optimization problem in the Lagrange form:

$$L(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \alpha(\mathbf{w}^T \mathbf{B} \mathbf{w} - 1) \tag{5.3}$$

and taking the derivative of the Lagrangian with respect to  $\mathbf{w}$ , we obtain an equivalent problem in the form of

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w} \quad (5.4)$$

which is the generalized eigenvalue decomposition (GED) of the matrices  $\mathbf{A}$  and  $\mathbf{B}$ . The solutions of this equation are the joint eigenvectors of  $\mathbf{A}$  and  $\mathbf{B}$  and  $\lambda$  is the associated eigenvalue of a particular eigenvector.

### 5.2.2 Sparse CSP Methods

The drawbacks of the CSP method that are described earlier prompted us to find a way to *sparsify* the spatial filter to increase the classification accuracy and the generalization capability of the method. We assumed that the discriminatory information is embedded in a few channels where the number of these channels is much smaller than the actual number of all recording channels. So the discrimination can be obtained with a sparse spatial projection, which uses only informative channels. In this scheme assume that the data was recorded from  $C$  channels. The spatial projections  $\mathbf{w}$  has only  $c$  nonzero entries,  $\text{card}(\mathbf{w}) = c$  and  $c \ll C$ . We are interested in obtaining a sparse spatial projection using oscillating search (OS), backward elimination (BE), recursive weight elimination (RWE) and forward selection (FS).

The FS and BE are described in detail in [75]. It is known that finding a sparse solution using  $\ell_0$  norm is combinatorial and NP-hard. Therefore they suggested greedy search methods to *sparsify* the CSP filter. The methods are described below.

The covariance matrices  $\mathbf{A}$  and  $\mathbf{B}$  can be computed using the training data and assuming the sparse solution  $\mathbf{w}$  where most of the entries in  $\mathbf{w}$  are zero, is given. That means  $\mathbf{w}$  satisfy the following optimization equation.

$$\arg \max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \text{ s.t. } \|\mathbf{w}\|_0 = c. \quad (5.5)$$

It is observed that the sparse vector  $\mathbf{w}$  selects the column and rows of the

matrices  $\mathbf{A}$  and  $\mathbf{B}$ , so the Equation 5.5 can be rewritten as,

$$\arg \max_{\mathbf{w}_c} \frac{\mathbf{w}_c^T \mathbf{A}_c \mathbf{w}_c}{\mathbf{w}_c^T \mathbf{B}_c \mathbf{w}_c} \text{ s.t. } \|\mathbf{w}_c\|_0 = c, \quad (5.6)$$

where  $\mathbf{A}_c$  and  $\mathbf{B}_c$  are the matrices that have only the rows and columns those correspond to the nonzero entries of the sparse solution  $\mathbf{w}$  and  $\mathbf{w}_c$  has only nonzero entries of the sparse solution  $\mathbf{w}$ . That means the reduced matrices  $\mathbf{A}_c$  and  $\mathbf{B}_c$  are the  $c \times c$  submatrices of the original covariance matrices. The problem here is which sub-matrix should be selected. Searching all possible submatrices is infeasible and involves  $\ell_0$  norm optimization problem. FS or BE can be used to obtain suboptimal solutions of this original problem.

#### 5.2.2.1 Forward Search

The FS starts with an empty set of channels. It solves the CSP problem for all individual channels those are not in the current channel set and adds the channel that has resulted maximum variance increase to the current set. After selecting the first channel, all the remaining channels are included in the channel set to form two channel problem. The channel that has resulted maximum variance increase is selected as the second channel of the sparse solution. This procedure continues with other cardinality levels sequentially until the required cardinality level is reached. FS solves  $C - K$  CSP problems for  $(K + 1) \times (K + 1)$  matrices where  $C$  is the total number of channels and  $K$  is the number of elements in the current set of channels. In each step we increase the number of channels by adding a channel to the current set until we reach the desired cardinality  $c$ . So the computational complexity would be increased with the desired cardinality. This algorithm linearly depends on  $C$  for a particular desired cardinality, so increasing the number of channels does not affect the complexity of the algorithm as much as BE. However the results indicate that the accuracy of this method is less than the BE.

### 5.2.2.2 Backward Elimination

The BE starts with all channels and it removes the channels in the set one by one and solves the CSP for remaining channels in the set. The channel that has resulted maximum variance decrease is eliminated. The procedure continues for the remaining channels in the set until we reach the desired cardinality  $c$ . The BE method in the first step searches  $C - 1$  separate submatrices and solves GED problem for each of them to find a sparse solution whose cardinality is  $C - 1$ . Hence, a GED is solved  $C - 1$  times on  $C - 1 \times C - 1$  matrices. In each step the size of the submatrices become one less and that is also equal to the number of separate GED solutions that is performed at each step. As a result, until the desired cardinality is reached the total number of separate GED solutions dominates the computational complexity. The computational complexity is even higher when  $C$  is large and the desired cardinality is small. Since the number of CSP computations and the size of the matrices those involve in CSP solution both depend on  $C$ , the effect of  $C$  is much apparent when  $C$  is increased for BE method.

### 5.2.2.3 Recursive Weight Elimination

The RWE approach is recently introduced by [71], motivated by the work of [76], which employed a recursive feature elimination in an SVM framework.

The RWE starts with all channels and it removes the channels in the set one by one and solves the GED problem for each channel in the set. The difference between the RWE and the BE is the elimination strategy. The BE solves the GED for each submatrix by removing a channel at a time from the full set and eliminates the channel which provided the minimum drop in RQ. On the other hand RWE solves GED using all channels in the current set and finds the entry in the spatial filter that has the minimum absolute amplitude. The corresponding channel is removed from the set. The procedure continues for the remaining channels in the set until we reach the desired cardinality. Therefore, a GED is solved only once on  $K \times K$  matrix in each step where  $K$  is the current cardinality

level of sparse solution  $\mathbf{w}$  in the current step. As a result, the computational complexity of RWE is dramatically low and decreases in each step.

The RWE elimination is inspired from the projections on  $\ell_1$  ball concept [77] which is explained in appendix B in detail. In the  $\ell_1$  ball projection, a constant value is added to or subtracted from the coefficients of the spatial filter  $\mathbf{w}$  to set  $\ell_1$  norm of  $\mathbf{w}$  to a constant ( $\|\mathbf{w}\|_1 = \alpha$ ). The spatial filter coefficients that have sign changes, are eliminated by setting them to zero. This step changes the value of  $\|\mathbf{w}\|_1$ , therefore a scaling is applied so that  $\|\mathbf{w}\|_1 = \alpha$ . This last step moves our projected vector onto the  $\ell_1$  ball. When we choose  $\alpha$  such that only one component of  $\mathbf{w}$  to be zero, then projecting  $\ell_1$  ball is essentially identical to the RWE method.

#### 5.2.2.4 Oscillating Search

The oscillating search (OS) approach is motivated by the work of [73, 74]. They used OS method to select a subset of features from a large set in a computationally efficient manner. In this scheme the OS uses an upswing and a down swing procedure, by running forward addition and backward elimination, steps to modify an initial (given) set of features based on a cost criterion. The initial set is either selected randomly or using a method that requires low computational power.

Here, with the same spirit, we used OS to extract a sparse spatial filter solutions by fusing FS, BE and RWE methods. Assume that we are searching for a sparse filter with cardinality  $c$ . In order to select the initial set, we used RWE method, which is dramatically faster than BE and more accurate than FS with comparable computational complexity. After obtaining the initial set of  $c$  channels, we executed the up and down swing steps to modify it. We used the RQ as a criterion to assess the effectiveness of each identified subset. During the upswing procedure, simply, we added channels using FS to increase the number of channels. Then used the BE method to remove channels to return back to the desired cardinality  $c$ . In the downswing phase, we first eliminated channels with

BE method and then increased them back with FS to reach cardinality  $c$ . Here, the swing size,  $s$ , which is the number of channels to add or eliminate is a free parameter that needs to be set during the search procedure. If  $s$  is too small the algorithm might get easily stuck to the initially selected set. On the other hand, a large  $s$  can increase the complexity of the search dramatically. Here, for the cardinality levels  $c \leq 5$  we set  $s = c - 1$  during the downswing procedure. For higher cardinalities we set  $s = 5$ . For the upswing the  $s = 8$ . We limited the number of downswing/upswing operations to 50 in order to avoid the infinite loop.

The algorithm is summarized below, here  $c$  is desired cardinality level,  $C$  is total number of channels,  $s$  swing size and  $L$  is the number of loops that we should continue downswing or upswing phases.

- Step 1.** (*Initialization*) Select  $c$  channels using using RWE to initialize the channel set. Also set  $s$  to 1 and  $L$  to 0.
- Step 2.** (*Downswing*) Eliminate and add  $s$  channels and increase  $L$  by one, if  $L$  is 50 than set  $s$  to 1 and go to Step 4. Repeat this Step if the channel set is changed, otherwise proceed to Step 3.
- Step 3.** (*Channel set was not changed in Step 2*) Increase the swing size ( $s$ ) by one, if  $s < c$  and  $s < 4$  go to Step 2 otherwise set  $s$  to 1 and go to Step 4.
- Step 4.** (*Up Swing*) Add and eliminate  $s$  channels and increase  $L$  by one, if  $L$  is 50 than go to Step 6. Repeat this Step if the channel set is changed, otherwise go to Step 5.
- Step 5.** (*Channel set was not changed in Step 4*) Increase the swing size ( $s$ ) by one, if  $s \leq C - c$  and  $s < 10$  go to Step 4 otherwise go to Step 6.
- Step 6.** (End of OS) We completed the OS and we have a new set channels.

In order to find multiple sparse filters using the OS technique, we deflated the covariance matrices with sparse vectors using the Schur complement deflation method described in [65].

### 5.2.3 The Dataset

The performance of the oscillating search is evaluated on the BCI competition III dataset IVa [62] dataset. The dataset contains EEG signals that is recorded from five subjects aa, al, av, aw, ay while the subjects asked to imagine either foot or right index finger movements. The data recorded from 118 different electrodes at a sampling rate of 1 kHz. The recorded signal was bandpass filtered in the range of 8-30 Hz. One second data following the cue was used to compare the methods. There were 140 trials available for each subject and class.

The EEG signal was transformed into six spatial filters by taking first and last three eigenvectors for each CSP methods. After computing the spatial filter outputs, we calculated the energy of the signal and converted it to log scale and used them as input features to a linear discriminant analysis (LDA) classifier [70] which is a parameter free decision function.

We compared the OS to the standard CSP, to the  $\ell_0$  norm based BE and FS methods of [48] and RWE method of [48]. We studied the classification accuracy as a function of cardinality. With the purpose of finding optimum *sparsity* level for the classification, we computed several sparse solutions, with decreasing number of cardinality on the training data. We computed the sparse filters with  $c \in \{96, 64, 48, 32, 16, 10, 7, 5, 2\}$ . For each cardinality level we computed the corresponding RQ value. We studied the RQ curve and determined the optimal cardinality where its value suddenly dropped indicating we started to lose informative channels.

The EEG dataset contains 140 trials per class and subject. We used a 4-fold cross validation technique. We divided the data into 4 folds where each of the folds contains 35 trials per class. We used each fold for extracting the spatial filter and training the classifier. The learned system is tested on the rest of the data. The results obtained from four folds are averaged to obtain the final accuracy of the interested method.

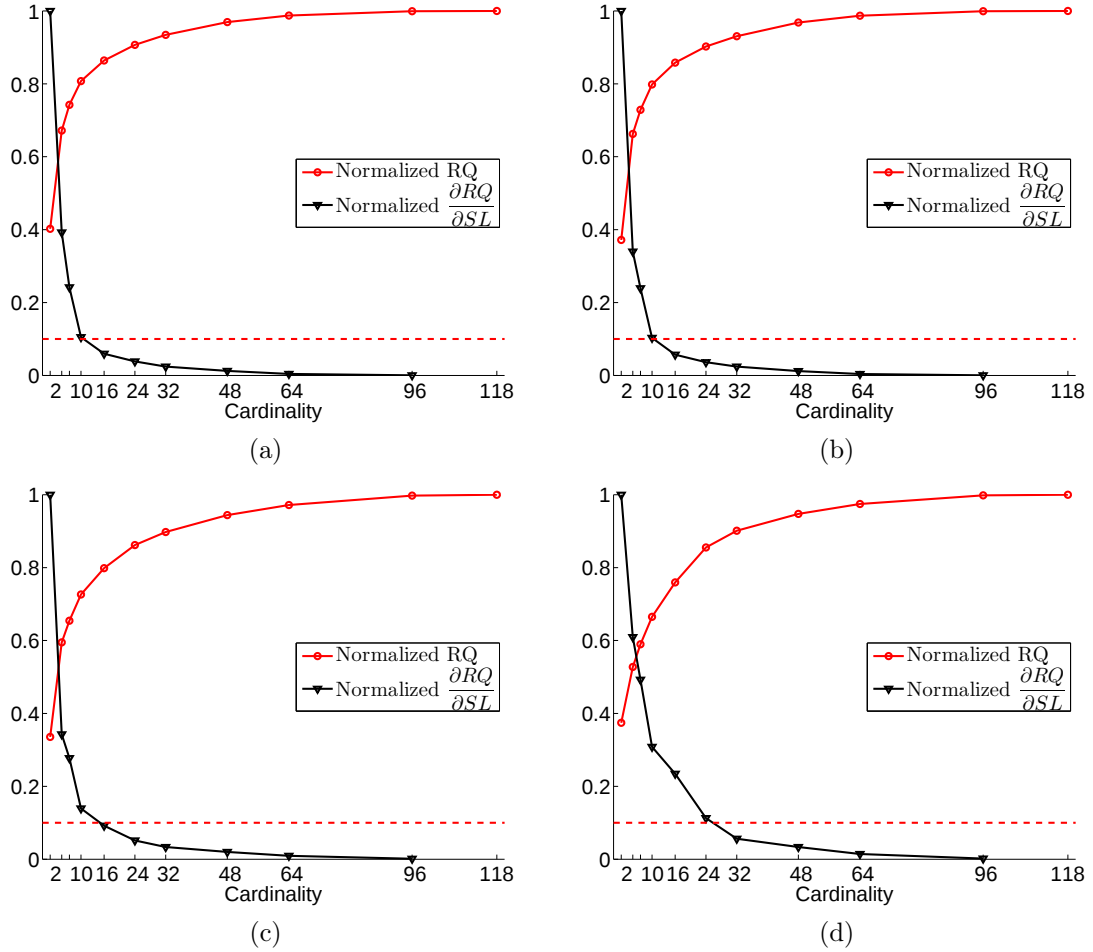


Figure 5.1: The average RQ of all subjects versus cardinality (a) OS, (b) BE, (c) RWE and (d) FS. The dashed red line is the 10 percent threshold that determines the optimum cardinality to be used in the test data. The optimum cardinality levels for OS, BE, RWE and FS methods are 10, 10, 16 and 24, respectively.

### 5.3 Results

In order to determine the optimal cardinality level to be used on the test data, the RQ values related to each cardinality level were computed on the training data, scaled to their maximum value and averaged over subjects. In the following step, we computed the slope of the RQ curve and normalized it to its maximum value to get an idea about the relative change in the RQ.

We depicted the change in RQ values for each cardinality as shown in Fig. 5.1.

As expected, decreasing the cardinality of the spatial projection resulted to a decrease in the RQ value. To determine the optimum cardinality to be used in classification on the test data, we selected the cardinality that is closest to 10% of the maximum relative change (dashed lines in Fig. 5.1). For BE and OS methods, the cardinality value was found to be 10 and for the FS and RWE methods, it was found to be 16 and 24, respectively. These indices perfectly corresponded to the elbow of the RQ curve, which indicates loss of informative channels. In Table 5.1, we provided the classification results and selected cardinalities for the EEG data set using each method. In order to give a sense of the change in error rate versus the cardinality, we provided the related classification error curves in Fig. 5.2. Although the minimum classification error was obtained at cardinality 24 for the RWE method, we noticed that we identified the optimum cardinality as 16 on the training data.

Table 5.1: EEG dataset classification error rates (%) for each subject using LDA classifier

|     | Cardinality | aa   | al   | av   | aw   | ay   | Avg  |
|-----|-------------|------|------|------|------|------|------|
| OS  | 10          | 21.1 | 3.57 | 22.3 | 7.68 | 6.96 | 12.3 |
| BE  | 10          | 21.1 | 3.21 | 24.3 | 8.57 | 6.25 | 12.7 |
| RWE | 16          | 22.1 | 3.39 | 28.8 | 8.93 | 9.64 | 14.6 |
| FS  | 24          | 20.2 | 3.21 | 29.3 | 8.75 | 7.68 | 13.8 |
| CSP | 118         | 27.9 | 5.54 | 34.3 | 10.9 | 12.7 | 18.2 |

On all subjects we studied, we observed that the sparse spatial filter methods consistently outperformed the CSP method. We noted that the minimum error rate was obtained with OS method. Both OS and BE methods used cardinality of 10 to achieve the minimum error rate. For the RWE method the optimum cardinality was 16 where for the FS method it was highest, 24. As expected the full CSP solution did not perform as good as the other sparse methods and likely overfitted the training data. The OS method improved the classification error rate with an error difference of 5.5%. We obtained comparable results using the OS and BE methods (p-value = 0.5, paired t-test) and comparable number of channels (p-value= 0.52).

Fig. 5.3 illustrates the distribution of the spatial filters obtained using the OS

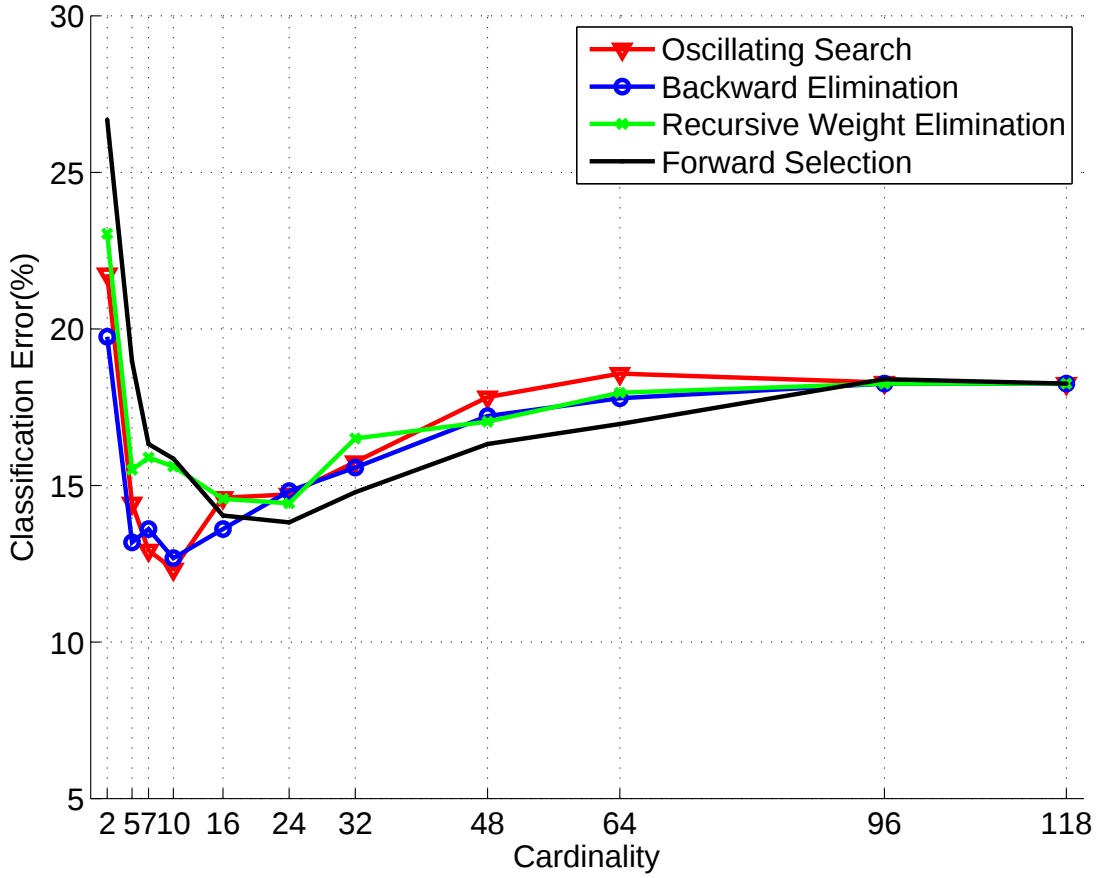


Figure 5.2: The classification error curves of all methods versus the cardinality. The last data point corresponds to the results obtained from standard CSP that uses all channels.

and CSP algorithms for each subject. We observed that the OS filter coefficients are localized on the left hemisphere and the central area, which is in accordance with the cortical regions related to right hand and the foot movement generation.

In order to compare the computational complexity of the methods, we measured the elapsed time for the extraction of a spatial filter from EEG data with cardinalities of  $c \in \{96, 64, 48, 32, 16, 10, 7, 5, 2\}$ . The training was performed on a regular desktop computer with 3 GB of RAM and equipped with a CPU running at 2.66 GHz. The elapsed time per filter computation decreased for the BE and RWE method and increased for the OS and FS method with the cardinality as shown in Fig. 5.4. The OS started with the computational time comparable to the FS and RWE at the beginning while it had the better results than them. OS

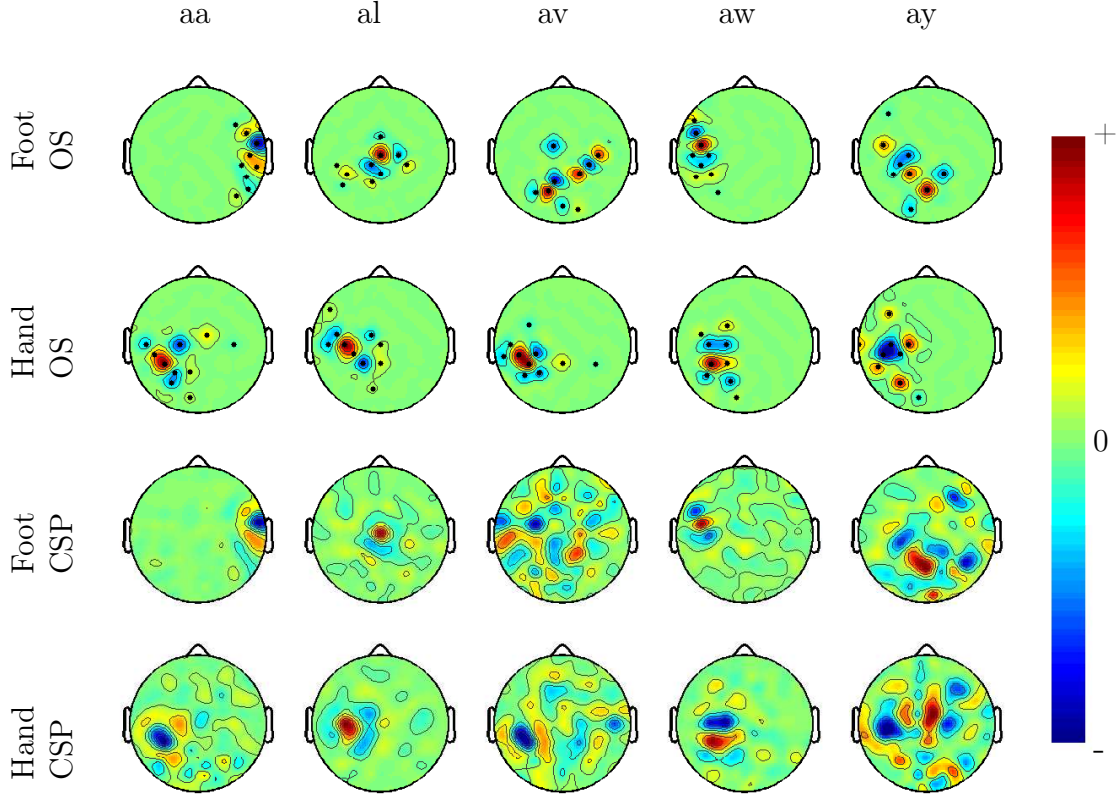


Figure 5.3: The OS and CSP filter head maps for hand and foot movement imagery.

was 20 times faster than the BE and better results than the BE at cardinality 10.

## 5.4 Summary

The CSP is a useful tool for BMI applications to reduce the work load and computational complexity of the classification algorithms. However, it generally overfits the data when the number of training trials are limited. In this chapter we adapt the method of oscillating feature search to finding spatially sparse solution in order to overcome the problems caused by the traditional CSP method. We also compared the results of this new technique with the other  $\ell_1$  norm based greedy techniques. Our results indicate that our method not only reduce the computational complexity of finding spatial patterns, it also generate comparable

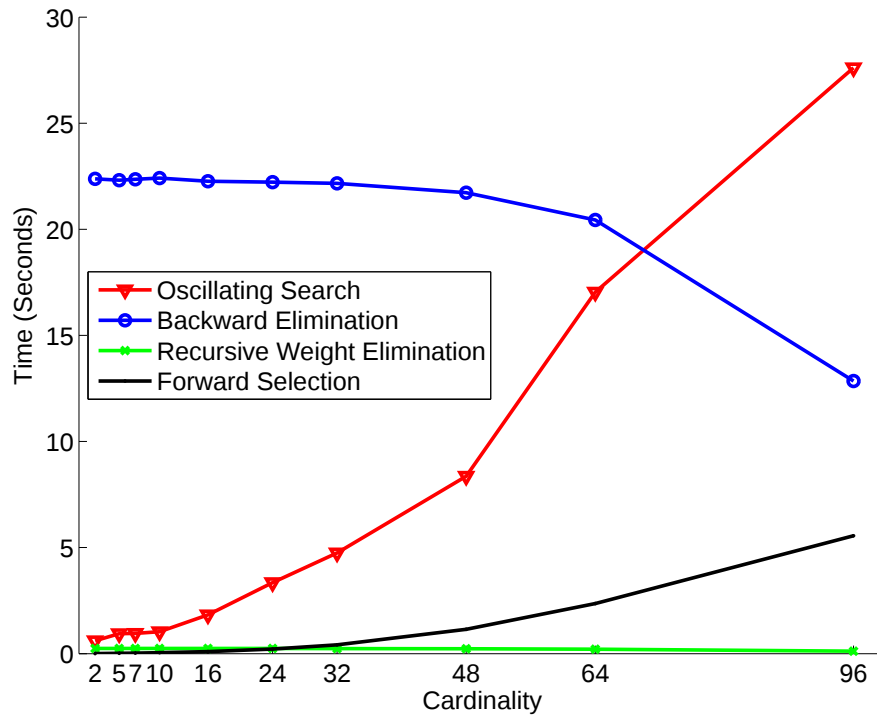


Figure 5.4: The average elapsed time to estimate a spatial filter vs. the cardinality.

results with respect to the other methods.

## Chapter 6

# Baseline Regularized Sparse Spatial Filters

The common spatial pattern (CSP) method has large number of applications in brain machine interfaces (BMI) to extract features from the multichannel neural activity through a set of linear spatial projections. These spatial projections minimize the Rayleigh quotient (RQ) as the objective function, which is the variance ratio of the classes. The CSP method easily overfits the data when the number of training trials is not sufficiently large and it is sensitive to daily variation of multichannel electrode placement, which limits its applicability for everyday use in BMI systems. To overcome these problems, the amount of channels that is used in projections, should be limited to some adequate number. We introduce a spatially sparse projection (SSP) method that renders unconstrained minimization possible via a new objective function with an approximated  $\ell_1$  penalty. We apply our new algorithm with a baseline regularization to the electrocorticogram (ECoG) data involving finger movements to gain stability with respect to the number of sparse channels.

## 6.1 Introduction

The aim of the BMI technology is to help disabled people by establishing a communication channel with their environment using only their brain signals. The recent advances in electrode design technology allow BMI applications to use increasing number of electrodes. In this scheme, the common spatial pattern (CSP) algorithm is widely used due to its simplicity and lower computational complexity to extract features from high-density recordings both using noninvasive and invasive modalities [20, 43].

The benefits of the CSP method come with some drawbacks. One major drawback of the CSP is that it generally overfits the data when it is recorded from a large number of electrodes and there is limited number of train trials. Furthermore, the chance that CSP uses a noisy or corrupted channel linearly increases with increasing number of recording channels. Another major problem is the robustness over time in CSP applications [44, 45]. Using all channels in spatial projections of CSP may reduce the classification accuracy in case the electrode locations slightly change in different sessions. In this case, CSP method requires almost identical electrode positions over time, which is difficult to realize [46]. The sparseness of the spatial filter might have an important role to increase the robustness and generalization capacity of the BMI system.

The CSP method increases or decreases the variance ratio of two classes. The variance ratio of two classes can be represented in terms of RQ of the spatial covariance matrices. The RQ is defined as

$$R(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \quad (6.1)$$

where  $\mathbf{A}$  and  $\mathbf{B}$  are the spatial covariance matrices of two different classes and  $\mathbf{w}$  is the spatial filter that we want to find. The solution of the CSP is the generalized eigenvalue decomposition of matrices  $\mathbf{A}$  and  $\mathbf{B}$ . This problem can also be solved in an unconstrained problem in the form of

$$L(\mathbf{w}) = R(\mathbf{w}) + \alpha \|\mathbf{w}\| \quad (6.2)$$

where  $R(\mathbf{w})$  is the objective function,  $\|\mathbf{w}\|$  is the  $\ell_1$  norm based penalty and  $\alpha$  is a

constant that controls the sparsity of the solution. Since RQ does not depend on the magnitude of the filter  $\mathbf{w}$ , we observed that the solution to this optimization problem is essentially scaled version of the generalized eigenvalue decomposition (GED) solution and does not depend on  $\alpha$ . Therefore, we introduced a novel objective function which has dependency on its magnitude and rise the same solution as GED when  $\alpha$  is equal to zero [78].

A number of studies investigated putting the CSP into alternative optimization forms to obtain a sparse solution for it. In [47] the authors converted CSP into a quadratically constrained quadratic optimization problem with  $\ell_1$  penalty; others used an  $\ell_1/\ell_2$  norm based solution [11, 44]. These studies have reported a slight decrease or no change in the classification accuracy while decreasing the number of channels significantly. Recently, quasi  $\ell_0$  norm based methods was used for obtaining the sparse solution which resulted an improved classification accuracy. Since  $\ell_0$  norm is non-convex, combinatorial and NP-hard, they implemented greedy solutions such as forward selection (FS), backward elimination (BE) [48] and recursive weight elimination (RWE) [71] to decrease the computational complexity. It has been shown that BE was better than RWE and FS (less myopic) in terms of classification error and sparseness level but associated with very high complexity making it difficult to use in rapid prototyping scenarios.

Selecting the sparsity level that produces high accuracy is crucial for the sparse spatial filters. We observed that the small variations in sparsity may lead to large change in the classification accuracy [78]. So a more representative sparse spatial filters needs to be constructed to eliminate large deviations on the classification accuracy.

In this chapter, we develop a baseline regularization algorithm to improve the classification accuracy and eliminate instability over the sparsity levels. The baseline regularization make the sparse spatial patterns to represent the fingers, instead of discriminating them from each other. The SSP which is computationally efficient sparse spatial projection based on a novel objective function and RWE are used to demonstrate the efficiency of the baseline regularization. The rest of the chapter is organized as follows. In the following section, we describe

our novel objective function and its relation to RQ. Then we explain its use in an unconstrained optimization problem. Next, we apply our method on the BCI competition IV ECoG dataset involving individuated movements of five fingers [49]. We also compare our method to standard CSP. Finally, we investigate the contribution of the baseline regularization to the classification accuracies by constructing a mixed generative/discriminative sparse filters.

## 6.2 Material and Methods

### 6.2.1 Sparse Spatial Filter

We sparsify the spatial filters to overcome the drawbacks of the CSP method that are described earlier and to increase the classification accuracy and the generalization capability of the method. We assume that a few channel of the data has the discriminatory information and the number of these channels is much smaller than the actual number of all recording channels. In this scheme, assume that the data was recorded from  $C$  channels. We are interested in obtaining a sparse spatial projection using an unconstrained minimization problem in the form of (6.2), where  $\mathbf{w}$  has only  $c$  nonzero entries,  $\text{card}(\mathbf{w}) = c$  and  $c \ll C$ . Since  $R(\mathbf{w})$  does not depend on the gain of  $\mathbf{w}$ , the optimizer arbitrarily reduces the gain of  $\mathbf{w}$  to minimize regularization term  $\alpha\|\mathbf{w}\|$  after finding the direction that minimizes  $R(\mathbf{w})$ . Thus, the solution of the optimization problem that uses  $R(\mathbf{w})$  as an objective function is essentially the same as the GED solution.

To find a sparse solution we need to have an objective function that depends on the gain of  $\mathbf{w}$ . In this scheme, we replaced  $R(\mathbf{w})$  with the following objective function.

$$G(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \frac{1}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \quad (6.3)$$

This function is bounded from below and has interesting properties. Let us define  $a(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w}$  and  $b(\mathbf{w}) = \mathbf{w}^T \mathbf{B} \mathbf{w}$ . If we define RQ in terms of  $a$  and  $b$

such that

$$R(\mathbf{w}) = \frac{a(\mathbf{w})}{b(\mathbf{w})}$$

then our new objective function can be expressed as

$$G(\mathbf{w}) = a(\mathbf{w}) + \frac{1}{b(\mathbf{w})} = R(\mathbf{w})b(\mathbf{w}) + \frac{1}{b(\mathbf{w})}. \quad (6.4)$$

The minimum value of  $G(\mathbf{w})$  can be calculated by taking its derivative with respect to  $b(\mathbf{w})$  along a fixed radial direction of  $\mathbf{w}$ . Since  $RQ$  depends only on the direction of  $\mathbf{w}$ , we can treat it as a constant and yielding a minimum value of  $2\sqrt{R(\mathbf{w}_{min})}$ . Hence, the direction that minimizes  $R(\mathbf{w})$  also minimizes  $G(\mathbf{w})$  and vice versa.

We put  $G(\mathbf{w})$  into unconstrained optimization formulation in (6.2) as the objective function. We placed a twice differentiable smooth version of  $\ell_1$  (epsL1) which is sufficiently close to minimizing  $\ell_1$  [61] as a regularization parameter. The main advantage of this approach is that, since epsL1 and  $G(\mathbf{w})$  are both twice differentiable we can directly apply an unconstrained optimization method to minimize  $L(\mathbf{w})$  [63]. The epsL1 is defined as

$$\|\mathbf{w}\|_{1\epsilon} = \sum_{i=1}^C \sqrt{w_i^2 + \epsilon}, \quad (6.5)$$

where  $\epsilon$  is a sufficiently small parameter and  $C$  is the dimension of  $\mathbf{w}$ . The epsL1 approximates the  $\ell_1$  norm and they are identical when  $\epsilon$  is equal to zero. Twice differentiability of the epsL1 norm allows us to use it when  $w_i$  is equal to zero unlike the regular  $\ell_1$  norm which is not differentiable at zero.

The entries of  $\mathbf{w}$  generally were not exactly equal to zero, so we normalized  $\mathbf{w}$  to its maximum absolute value and eliminated the weights consequently corresponding channels that do not exceed a predefined threshold ( $=10^{-2}$ ). We computed the desired cardinality which is the number of channels to be selected for the spatial projection by implementing a bisection search [64] on the  $\alpha$ . The upper border of  $\alpha$  was determined initially using the  $G(\mathbf{w}_{full})/\|\mathbf{w}_{full}\|$  ratio where  $\mathbf{w}_{full}$  is the full CSP solution. In case the initial upper border results a cardinality larger than the desired value, we kept doubling the  $\alpha$  parameter until we

obtained a  $\alpha$  that results a cardinality which is less than or equal to the target value.

### 6.2.2 Recursive Weight Elimination

Recursive weight elimination (RWE) is an  $\ell_0$  norm based algorithm to obtain sparse filters in very a efficient and effective way [71]. The algorithm starts with a full size covariance matrices of the traditional CSP method. Assume that the size of these covariance matrices is  $C \times C$ . In the first step, RWE solves general CSP problem and finds the weight vector  $\mathbf{w}$ . The contribution of the smallest magnitude coefficient can be ignored compared to the other coefficients, since we have a high number of channels. Assume that the index of this small coefficient is  $k$ . We can remove this coefficient by removing  $k^{th}$  row and column of the full size covariance matrices and solving the CSP on these new  $C - 1 \times C - 1$  matrices. We can decrease the number of channels to the desired cardinality level  $c$  by recursively applying this algorithm to the smaller matrices. Each cardinality reduction involves solving a traditional CSP; therefore, this method is faster than other  $\ell_0$  norm based greedy search algorithms such as BE or FS [48].

### 6.2.3 Baseline Regularized Sparse Spatial Filters

The data set consists of finger movement and baseline regions. We used the baseline data to regularize the finger to finger contrast. In other words, each multichannel finger data is contrasted with a mixture of baseline and another finger. If we assume that  $\mathbf{A}$  is the spatial covariance matrix of the first finger,  $\mathbf{C}$  is spatial covariance matrix of one of the other four fingers, and  $\mathbf{D}$  is the covariance matrix of the baseline, we find a solution to the following optimization problem,

$$L(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \frac{1}{\mathbf{w}^T (\beta \mathbf{C} + (1 - \beta) \mathbf{D}) \mathbf{w}} + \alpha \|\mathbf{w}\|, \quad (6.6)$$

where  $\beta$  is the mixing coefficient ranging from 0 to 1. We contrast a finger to another finger when  $\beta$  is 1 to obtain discriminative spatial filters. On the other hand when  $\beta$  is equal to zero, we contrast each finger with baseline which yields representative spatial filters. Therefore,  $\beta$  determines level discrimination or representation characteristic of the constructed sparse filter. Similarly, we also apply this approach to RWE method.

In this scheme, we computed the first spatial filter  $\mathbf{w}$  that minimizes the  $L(\mathbf{w})$  to obtain the sparse filter that maximize the variance of the first finger. Then we interchanged the matrices  $\mathbf{A}$  and  $\mathbf{C}$  to find the spatial filter that maximizes the variance of the other finger. In order to find multiple sparse filters we deflated the covariance matrices with these initial sparse vectors using the Schur complement deflation method described in [65]. Using this new deflated matrices, we find the second set of spatial filters and obtain a total of 4 spatial filters.

#### 6.2.4 ECoG Dataset

The ECoG data was recorded from three subjects during finger flexions and extensions [49] with a sampling rate of 1 kHz. The electrode grid was placed on the surface of the brain. Each electrode array contained 48 ( $8 \times 6$ ) or 64 ( $8 \times 8$ ) platinum electrodes. The finger index to be moved was shown with a cue on a computer monitor. The subjects moved one of their five fingers 3-5 times during the cue period. The ECoG data of each subject was subband filtered in the gamma frequency band (65-200 Hz), as in [69]. We used one second data following the movement onset and 500 ms data before the movement onset in the analysis. The dataset contains around 146 trials for each subject.

The multichannel signal was transformed into four channel signal using the spatial filters are derived using each CSP methods. After computing the spatial filter outputs, we calculated the energy of the signal and converted it to log scale for each sparse filter and we used them as input features to lib-SVM classifier with an RBF kernel [53].

Since we are tackling a multiclass problem for the ECoG dataset, we used the pairwise discrimination strategy of [20] to apply the CSP to the five-class finger movement data. In other words, we constructed sparse spatial filters tuned to contrast pairs of finger movements such as 1 vs. 2; 1 vs. 3; 2 vs. 4, and so on.

We studied the classification accuracy as a function of cardinality and the mixing parameter  $\beta$ . On the training data with the purpose of finding optimum *sparseness* level for the classification, we computed several sparse solutions, with decreasing cardinality. The sparse CSP methods were employed with  $c \in \{40, 30, 20, 15, 10, 5, 2, 1\}$ . For each cardinality, we computed the corresponding RQ value. We studied the inverse of the RQ (IRQ) curve and determined the optimal cardinality where its value suddenly dropped indicating we started to lose informative channels.

Two times two fold cross validation were run on the entire data set and the results were averaged over the folds and iterations. On the average, we used  $15 \pm 2$  train trials per finger. The value of the  $\epsilon$  in epsL1 regularization term was chosen to be  $10^{-6}$ . We used  $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$  for the baseline regularization experiments.

## 6.3 Results

We depicted the change in IRQ values for each cardinality as shown in Fig. 6.1a and 6.1b. As expected, decreasing the cardinality of the spatial projection resulted to a decrease in the IRQ value. To determine the optimum cardinality to be used in classification on the test data, we selected the cardinality that is below 10 % of the maximum relative change (see the dashed lines in Fig. 6.1). The cardinality value was found to be 5 for SSP method. For the RWE method the cardinality value was 2. These indices perfectly corresponded to the elbow of the IRQ curve, which indicates loss of informative channels. In Table 6.1, we provide the classification results and selected cardinalities using SSP, CSP and  $\ell_0$  based greedy solution, RWE with a mixing parameter  $\beta$  that provides minimum

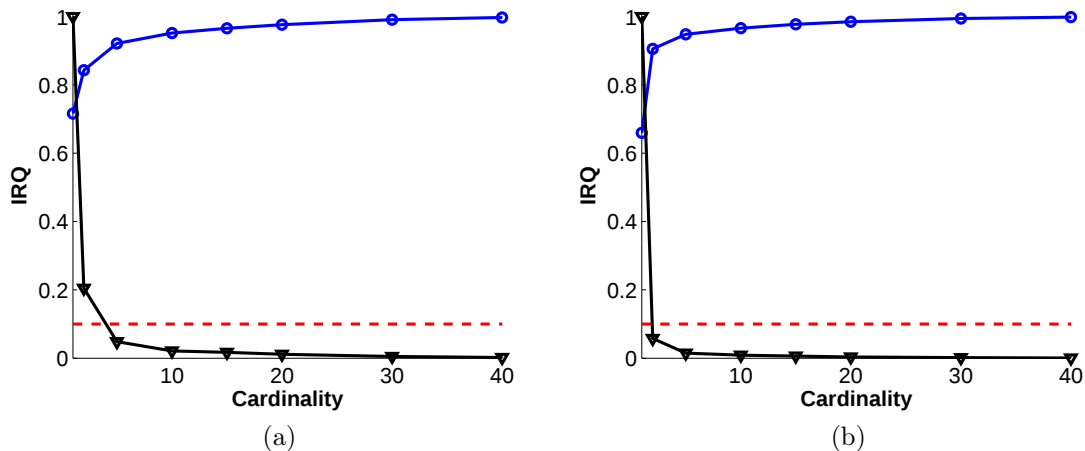


Figure 6.1: The average IRQ of all subjects versus cardinality for SSP method (a) and RWE method (b) for the  $\beta$  values 0.75 and 0.5 respectively. The dashed red line is the 10 percent threshold that determines the optimum cardinality to be used in the test data. The optimum cardinality levels are five and two respectively. The line with circle markers is IRQ curve and the line with triangle markers is derivative of the IRQ curve.

accuracy error. In order to give a flavor about the change in error rate versus the cardinality, we provided the related classification error curves in Fig. 6.2.

On all subjects we studied, we observed that the SSP method consistently outperformed the CSP method. We noted that the minimum error rate was obtained with SSP method. SSP and RWE methods used cardinality of 5 and 2 to achieve the minimum error rate respectively. As expected the full CSP solution did not perform as good as the other sparse methods and likely overfitted the training data.

We also note that the baseline regularization removes overfitting of the classifier and provides robustness to the sparsity level. In Fig. 6.2 it is shown that

Table 6.1: Classification error rates (%) for each subject using SVM classifier

|     | Cardinality | $\beta$ | Subject 1 | Subject 2 | Subject 3 | Avg   |
|-----|-------------|---------|-----------|-----------|-----------|-------|
| RWE | 2           | 0.5     | 17.7      | 14.3      | 12.9      | 14.95 |
| SSP | 5           | 0.75    | 19.6      | 12        | 12.8      | 14.79 |
| CSP | All         | 0.25    | 25.7      | 19.6      | 15.3      | 20.19 |

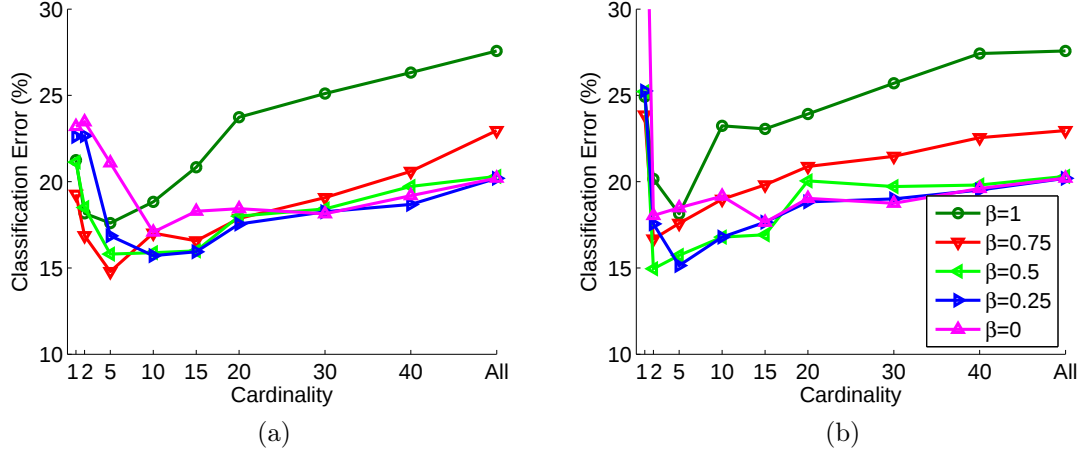


Figure 6.2: The classification error curve versus the cardinality for the SSP method (a) and the RWE method(b). The last data point corresponds to the results obtained from the standard CSP, which uses all channels.

the increase in cardinality did not affect the regularized ( $\beta \neq 1$ ) sparse filters as much as unregularized ( $\beta = 1$ ) sparse filters. On the other, hand pure generative ( $\beta = 0$ ) sparse filters accuracy error tends to increase with decreasing cardinality below the cardinality level 10.

## 6.4 Summary

In general the dimensionality of the BMI data is larger than the number of training data. This imbalance between the amount of training data and the number of channels results overfitting on the training data. To minimize overfitting and eliminate noisy channels, we introduced a spatially sparse projection technique (SSP) based on a novel objective function. By using an approximated  $\ell_1$  norm, we computed the sparse spatial filters through an unconstrained minimization formulation with standard optimization algorithm. We applied our method to ECoG dataset and compared its classification capacity to standard CSP and to an  $\ell_0$  norm based greedy technique. The sparse methods outperformed the standard CSP method. We observed that the sparse methods are sensitive to the cardinality, therefore we regularized the sparse spatial filters using the baseline

data. We study the effect of regularization on classification accuracy by implementing a baseline/movement mixing method. Our results indicate that baseline regularization improves the classification accuracies as well as it provides stability with respect to the cardinality level.

# Chapter 7

## Conclusions

In past few years, the advances in brain machine interface (BMI) technology also generates many challenging problems. One of the major problems is the robustness of BMI systems. In this thesis, we aim to solve the robustness problem by three different contributions:

- Robustness in the feature space (Chapter 2),
- Robustness obtained by redundant classifiers (Chapter 3), and
- Robustness obtained using the concept of sparsity (Chapters 4, 5 and 6).

Common problem of BMI applications is that the algorithms are trained on the trial based data. In order to construct a free paced BMI system, we propose a robust decoder for the identification of idle and movement states using the ECoG data in Chapter 2. Such a decoder operates on high number of recording channels with limited amount of data. The constructed hybrid decoder was based on the fusion of support vector machine (SVM) and hidden Markov model (HMM) for dynamic state detection. Model learning and testing was accomplished with data derived from multichannel electrocorticogram (ECoG) recordings during consecutive movements of individual fingers. The system used common spatial patterns

(CSP) features as input to the SVM classifier which provided the posterior probability of a particular internal state.

We have demonstrated experimentally that using hybrid system, the latency of state decoding using ECoG data during finger movements is comparable to that obtained using single unit activity (SUA) data during directional hand movements. The hybrid model was also compared to the traditional HMM technique. The hybrid decoder outperformed the HMM technique. The difference in state recognition accuracy was more in favor of the hybrid decoder when the amount of training data is low. The main advantage of using SVM within the hybrid decoder is that the posterior probability of each state is estimated simultaneously and tuned for discrimination. This advantage might overcome the lack of discriminative capability of HMMs, as each model is trained independently from the other competing models. Moreover, the higher generalization capacity of SVM due to the large margin makes the algorithm a good candidate for applications in which a limited number of training trials exists on which to base estimates of the model parameters. However, such an approach requires supervised training in order to estimate the state discriminators, which is automatically accomplished by the traditional HMM. Furthermore, the large margin of the SVM classifier makes the feature space more robust to the variations.

Another contribution of this thesis is to maintain classification robustness. In Chapter 3 we applied a redundant spatial projection framework based on CSP to classify ECoG data accompanying individual movements of 5-fingers. We studied the classification performance of different frequency subbands of ECoG data. We observed that the gamma (65-200 Hz) band provided the highest decoding accuracy with an average rate of 86.3% over three subjects. In all subjects we studied, the misclassifications generally occurred between fourth and fifth fingers. The overall trend of misclassifications was towards adjacent fingers. This indicates that neighboring fingers are likely represented by overlapping neural activity. Our results indicate that the redundant spatial projection framework can be successfully used in decoding finger movements for a BMI.

Recent advances in microelectronics, data acquisition and micro machined electrodes make it possible to record neural activity from dense electrode arrays. Standard signal processing and feature extraction techniques often fail in such setting due to the curse of dimensionality. The need for the sparse filters is apparent when there is large number of recording electrodes and insufficient amount of training data. With this motivation, in Chapters 4,5 and 6, we constructed robust solutions based on spatial projections to minimize overfitting on the training data and/or eliminate noisy channels for state decoding and multi class finger movement detection to build a hand neuroprosthetics.

In order to solve the problems caused by the large number of channels, a spatially sparse projection technique (SSP) based on a new objective cost function is introduced in Chapter 4 and it is applied to multichannel ECoG data which involves individual finger movements. Unlike the Rayleigh quotient (RQ), this new objective function has a dependency on the filter magnitude. By using an approximated  $\ell_1$  norm, we computed the sparse spatial filters through an unconstrained minimization formulation with standard optimization algorithm.

We applied our method to ECoG and electroencephalogram (EEG) datasets and compared its efficiency to standard CSP, and to an  $\ell_0$  norm based greedy technique. The SSP method outperformed the standard CSP on both datasets and provided comparable results to  $\ell_0$  norm based method, which suffers from higher computational complexity. On the ECoG data, the SSP method provided 44% decrease in the error rate compared to standard CSP method and used only five channels in each spatial projection. The error difference between regular CSP and SSP is less apparent in the EEG dataset as SSP method provided 26% decrease in the error rate. In contrary to the ECoG data, we also observed that more channels were used to achieve minimum classification accuracy in the EEG dataset. This could be due the low spatial resolution originating from the volume conduction and low signal to noise ratio (SNR) of the EEG. Nevertheless, the SSP algorithm was able to reach a minimum error rate with only 15 channels.

Our results indicate that SSP method can be effectively used to extract features from both EEG and ECoG datasets with moderate number of recording

channels. For those systems, with much higher number of recording channels the SSP solution might not be feasible in practical setting. To tackle with this problem, in Chapter 5, we adapted the oscillating search (OS) method which fuses recently introduced greedy sparse filter selection methods such as backward elimination (BE), forward selection (FS) and recursive weight elimination (RWE). We applied these sparse spatial filter extraction methods, as well as traditional CSP, to the EEG data IVa of brain computer interface (BCI) competition IV that involves either right hand or foot imagined movements recorded from 5 subjects over 118 channels. We observed that the OS is more accurate than all other methods and reaches the minimum classification error by using sparse filters with cardinality as low as 10. Similar classification accuracies were obtained with the BE method with the same cardinality level. However, the average filter extraction time of the OS method is 20-times faster than the BE, making OS a more feasible technique in real-life applications which require rapid training stages.

Our results indicate that the redundant spatial projection framework can be successfully used in decoding finger movements for a BMI. In those cases where a correlation or structure between class labels missing, it is difficult to construct a redundant set of feature extractors getting the benefit of larger train trials in the merged class set. For these cases, the sparse spatial projection techniques appear as a feasible solution. Our motivation when seeking for a sparse CSP solution is simple. By making each spatial filter sparse in terms of the number of channels used, there will be a decrease in the amount of variance explained by linear combination of channels but this decrease will be small since only a subset of channels are expected to contribute to real variance difference. Therefore, constructing spatial filters from a subset of all available channels yields improved generalization performance.

In Chapters 4 and 5, we used a pairwise strategy to classify the finger movements, however contribution of the baseline could help the algorithms to improve the classification accuracy. The pairwise comparison reveals the discriminative characteristics of the data, while comparing fingers to baseline associated with the representative characteristics of the data. In chapter 6, we merge the representative characteristics with the regular pairwise spatial filters. The new filters

provided better results, which shows that the representative characteristics of the data contains valuable information that improves the classification accuracy.

Our results show that the generalization capabilities of the sparse CSP methods compared to those obtained by the traditional CSP method do improve not only when there is large amount of recording channels but also when the size of the training data is small. The SSP method introduced in this study, outperformed the standard CSP on both ECoG and EEG datasets and provided comparable results to  $\ell_0$  norm based method, which is associated with higher computational complexity. In BMI applications, rapid prototyping is a crucial parameter. Long subject training yields frustration which in turns causes changes in neural patterns. Moreover, maintaining robust recording over time becomes more and more difficult if the training durations extend over hours. Consequently, BMI systems should be trained rapidly with small amount of training data such that they can be deployed rapidly in the field. In this scheme, the sparse signal processing and spatial projection techniques described in this thesis can be good candidates for constructing the neural decoding engine of future BMI systems.

# APPENDIX A

## Principle Component Analysis and Generalized Eigenvalue Decomposition

In linear algebra, the eigenvectors of a matrix are a set of special vectors that are transformed to scaled versions of themselves by the matrix. The scaling parameter is called the eigenvalue. The relation between the matrix and its eigenvector and eigenvalue is as follows,

$$\mathbf{\Sigma} \mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad (\text{A.1})$$

where  $\mathbf{\Sigma}$  is the matrix,  $\mathbf{v}_i$ s are its eigenvectors and  $\lambda_i$ s are the associated eigenvalues. The covariance matrix for a given data can be estimated as follows:

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T, \quad (\text{A.2})$$

where  $N$  is the number of points in the dataset and  $\boldsymbol{\mu}$  is the mean vector defined as:

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_i. \quad (\text{A.3})$$

In pattern recognition applications, understanding the distribution of the data helps to reduce the complexity of the classification algorithms and improves the

classification accuracy. It is generally assumed that the data come from a multivariate Gaussian distribution with covariance matrix  $\Sigma$  and the mean vector  $\boldsymbol{\mu}$ . To reveal the direction that the data is the most variant, the eigenvectors of the covariance matrix plays a crucial role. In this scheme, the eigenvectors of the covariance matrix are called principle components (PCs) and the projecting data onto a few PCs is called principle component analysis (PCA). It is shown in Eq. A.10 that the variance of the projection onto an eigenvector is the eigenvalue that is associated with this particular eigenvector. Therefore, PCA choses the eigenvectors that are associated with a few largest eigenvalues.

Let  $y = \mathbf{v}_i^T \mathbf{x}$  be projection of multivariate random vector  $\mathbf{x}$  onto the  $i^{th}$  normalized eigenvector  $\mathbf{v}_i$ ,  $\mu_y$  and  $\boldsymbol{\mu}_x$  are the means,  $\sigma$  and  $\Sigma$  covariances of random variables  $y$  and  $\mathbf{x}$  respectively.

The variance of the projected data can be expressed as

$$\sigma = \frac{1}{N} \sum_{n=1}^N (y - \mu_y)^2. \quad (\text{A.4})$$

By substituting  $\mathbf{v}_i^T \mathbf{x}$  and  $\mathbf{v}_i^T \boldsymbol{\mu}_x$  for  $y$  and  $\mu_y$ , respectively, we get

$$\sigma = \frac{1}{N} \sum_{n=1}^N (\mathbf{v}_i^T (\mathbf{x} - \boldsymbol{\mu}_x)) (\mathbf{v}_i^T (\mathbf{x} - \boldsymbol{\mu}_x))^T \quad (\text{A.5})$$

where  $\mathbf{v}_i$  is a constant vector, so we can take it out from the summation,

$$\sigma = \mathbf{v}_i^T \left( \frac{1}{N} \sum_{n=1}^N (\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T \right) \mathbf{v}_i \quad (\text{A.6})$$

We recognized the expression in parenthesis from Eq. A.2 as the covariance matrix, and the equation turns out to be,

$$\sigma = \mathbf{v}_i^T \Sigma \mathbf{v}_i \quad (\text{A.7})$$

$\mathbf{v}_i$  is the eigenvector of the covariance matrix  $\Sigma$ , from Eq. A.1

$$\sigma = \mathbf{v}_i^T \lambda_i \mathbf{v}_i, \quad (\text{A.8})$$

$$\sigma = \lambda_i |\mathbf{v}_i|^2. \quad (\text{A.9})$$

since the eigenvectors are normalized,  $|\mathbf{v}_i|^2 = 1$

$$\sigma = \lambda_i. \quad (\text{A.10})$$

From Eq. A.10, the variance of the projection is equal to the eigenvalue associated with projecting eigenvector. This concludes that the eigenvector with largest eigenvalue is the most variant projecting direction of the data.

The PCA involves all the data samples, so it is a representative transformation of the data points. However, our aim is generally discriminate data samples of different classes. For this particular purpose we want to maximize the variance of one class, while minimizing the variance of the competing class using a linear transformation like in PCA. To create variance imbalance between the competing classes, the Rayleigh quotient (RQ) shown in Eq. A.11 is maximized with projecting vector  $\mathbf{w}$  and covariance matrices of the competing classes (A and B).

$$R(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}}. \quad (\text{A.11})$$

We can rewrite this problem in the form of a optimization problem as follows,

$$\begin{aligned} & \underset{\mathbf{w}}{\text{maximize}} && \mathbf{w}^T \mathbf{A} \mathbf{w} \\ & \text{subject to} && \mathbf{w}^T \mathbf{B} \mathbf{w} = 1. \end{aligned} \quad (\text{A.12})$$

In Eq. A.13, we show the Lagrange form of this optimization problem and its derivative with respect to  $\mathbf{w}$ . We construct the Lagrangian  $L(\mathbf{w})$  and set its derivative to zero. The solution to this problem is called the generalized eigenvalue decomposition (GED).

$$\begin{aligned} L(\mathbf{w}) &= \mathbf{w}^T \mathbf{A} \mathbf{w} - \lambda \mathbf{w}^T \mathbf{B} \mathbf{w}, \\ \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} &= 2\mathbf{A} \mathbf{w} - 2\lambda \mathbf{B} \mathbf{w} = 0, \end{aligned} \quad (\text{A.13})$$

which leads to the following equation:

$$\mathbf{A} \mathbf{w} = \lambda \mathbf{B} \mathbf{w}. \quad (\text{A.14})$$

The GED can be solved using Eq. A.1 by letting  $\mathbf{\Sigma}$  be equal to  $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$  and  $\mathbf{v}_i = \mathbf{B}^{1/2}\mathbf{w}$ . The solution to the  $\lambda$  is essentially the RQ for the associated eigenvector as shown below.

$$R(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} = \frac{\lambda \mathbf{w}^T \mathbf{B} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} = \lambda. \quad (\text{A.15})$$

This concludes that projecting data onto the GED eigenvector that has the maximum eigenvalue creates the highest imbalance between the variances of the competing classes.

# APPENDIX B

## Projection on to the $\ell_1$ ball

### B.1 Projection formulations

Projecting onto  $\ell_1$  the ball is used in numerous batch and online learning sparse feature spaces such as text learning applications [77]. To project  $\mathbf{w}_{\mathbf{n}+1}$  on to the  $\ell_1$  ball, we need to solve the following optimization equation,

$$\begin{aligned} \underset{\mathbf{w}_{\mathbf{n}+1}}{\text{minimize}} \quad & \|\mathbf{w}_{\mathbf{n}+1} - \mathbf{w}_{\mathbf{n}}\|_2 \\ \text{subject to} \quad & \|\mathbf{w}_{\mathbf{n}+1}\|_1 = \alpha, \end{aligned} \tag{B.1}$$

where  $\|\cdot\|_1$  and  $\|\cdot\|_2$  represent the  $\ell_1$  and  $\ell_2$  norm of their arguments, respectively, and  $\alpha$  is the radius of the  $\ell_1$  ball. The line equation of the side of the ball that  $\mathbf{w}_{\mathbf{n}+1}$  is located can be expressed as  $\sum_{i=1}^N p_i w(i) = \alpha$  where  $p_i$  is the sign of the current vector's  $i^{th}$  component  $w_n(i)$ .

We can write this minimization problem in a Lagrange form as in Eq. B.2. We assume that the projection does not change the sign of the vector components. Differentiating the Lagrange form with respect to  $\mathbf{w}_{\mathbf{n}+1}(\mathbf{i})$  and comparing to zero (Eq. B.3) gives the optimal solution (Eq. B.4). The optimal solution is the point that minimizes the distance between  $\ell_1$  ball and the current point.

$$L(\mathbf{w}_{n+1}) = \|\mathbf{w}_{n+1} - \mathbf{w}_n\|_2^2 + \lambda \left( \sum_{i=1}^N p_i w_{n+1}(i) - \alpha \right), \quad (\text{B.2})$$

$$\frac{\partial L}{\partial w_{n+1}(i)} = 2(w_{n+1}(i) - w_n(i)) + \lambda p_i = 0, \text{ and} \quad (\text{B.3})$$

$$w_{n+1}(i) = w_n(i) - \frac{\lambda}{2} p_i. \quad (\text{B.4})$$

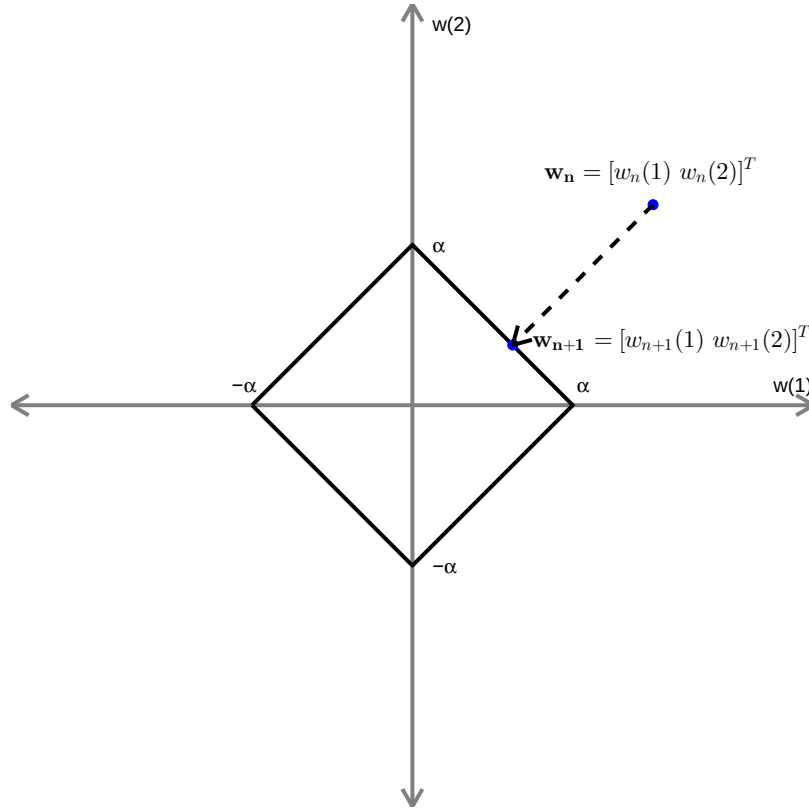


Figure B.1: The  $\ell_1$  ball for 2-dimensional space.

Using the line equation and Eq. B.4, we relate  $\lambda$  with the  $\ell_1$  ball radius  $\alpha$  as

follows.

$$\begin{aligned}
\alpha &= \sum_{i=1}^N p_i w_{n+1}(i) \\
&= \sum_{i=1}^N p_i \left( w_n(i) - \frac{\lambda}{2} p_i \right) \\
&= \sum_{i=1}^N p_i w_n(i) - \frac{\lambda}{2} \sum_{i=1}^N p_i^2 \\
&= ||\mathbf{w}_n||_1 - \frac{\lambda}{2} N.
\end{aligned} \tag{B.5}$$

The value of the  $\lambda$  can be found from Eq. B.5 and can be substituted in Eq. B.4 to find the following equation for the projected vector.

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mathbf{p} \frac{\alpha - ||\mathbf{w}_n||_1}{N}, \tag{B.6}$$

where,  $\mathbf{w}_n = [w_n(1) \ w_n(2) \ \dots \ w_n(N)]^T$ ,  $\mathbf{w}_{n+1} = [w_{n+1}(1) \ w_{n+1}(2) \ \dots \ w_{n+1}(N)]^T$ ,  $\mathbf{p} = [p_1 \ p_2 \ \dots \ p_N]^T$  and  $||\mathbf{w}_n||_1$  is the  $\ell_1$  norm of  $\mathbf{w}_n$ . According to the Eq. B.6 we subtract a constant from or add a constant to (depending on the value of  $p_i$ ) every component of the original vector. Let this constant be called  $\Theta$ , then the Eq. B.6 can be expressed as

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mathbf{p}\Theta, \tag{B.7}$$

where  $\Theta = \frac{\alpha - ||\mathbf{w}_n||_1}{N}$ .

## B.2 Sparse Solution

The original vector can not be directly projected onto the  $\ell_1$  ball, if the projection operator results sign changes of the vector components. This situation is shown in Fig. B.2. If we apply Eq. B.6 on  $\mathbf{w}_n$ , we would obtain the vector  $\mathbf{w}_a$  which is not on the  $\ell_1$  ball. To move vector  $\mathbf{w}_a$  onto the  $\ell_1$  ball, the components that have sign changes are assigned to zero (vector  $\mathbf{w}_b$ ) and then we normalize the resulting

vector such that its  $\ell_1$  norm is  $\alpha$ . Using this approach, we induce sparseness to the original vector.

$$\mathbf{w}_{n+1} = \alpha \frac{\mathbf{w}_b}{\|\mathbf{w}_b\|_1}. \quad (\text{B.8})$$

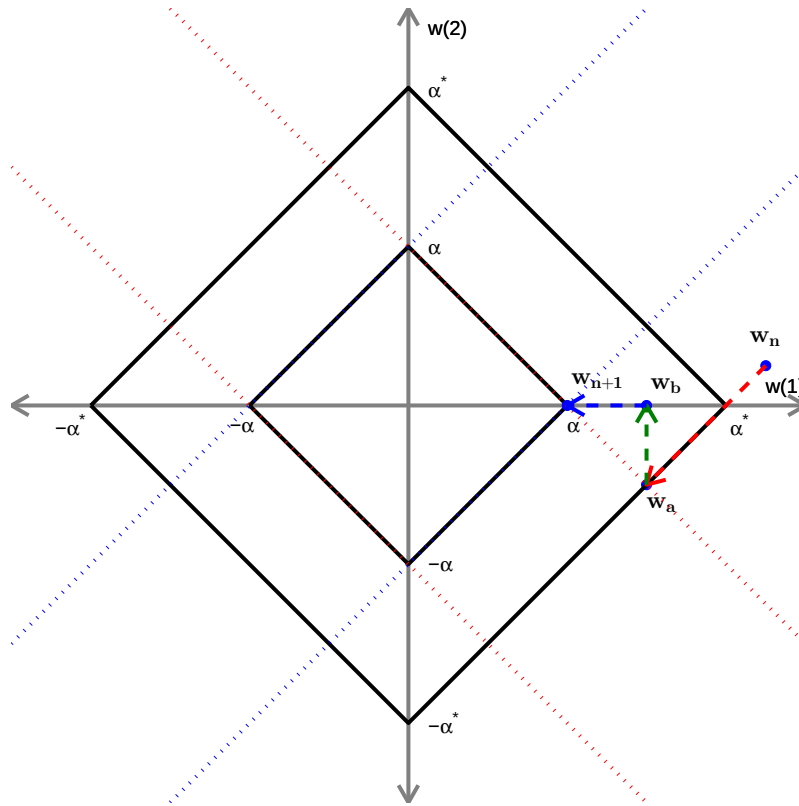


Figure B.2: Obtaining sparsity from  $\ell_1$  ball for 2-dimensional space.

Sometimes it is desired that one component of the original vector is eliminated after the projection. For this purpose, one can find a  $\alpha^*$  value such that only one component becomes zero. To find the value of  $\alpha^*$  we rearrange the Eq. B.7 as follows.

$$\begin{aligned}
w_{n+1}(i) &= w_n(i) + p_i \Theta \\
&= p_i(|w_n(i)| + \Theta).
\end{aligned} \tag{B.9}$$

In order to have no sign change the term  $(|w_n(i)| + \Theta)$  should be nonnegative for all  $i \in 1 \dots N$ . As a result of this condition,  $\alpha^*$  is found to be the maximum value for the set of expressions  $||w_n||_1 - N|w_n(i)|$  where  $i \in 1 \dots N$ . In Fig. B.2, we show that projecting onto  $\ell_1$  ball with radius  $\alpha^*$  has eliminated the second component of the original vector.

# APPENDIX C

## Deflation of a Matrix

### C.1 Hotelling's deflation

The eigenvectors ( $\mathbf{x}_i$ ) of matrix  $\mathbf{A}$  has the following property:

$$\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i, \quad (\text{C.1})$$

and

$$\mathbf{A} = \sum_{i=1}^N \lambda_i \mathbf{x}_i \mathbf{x}_i^T, \quad (\text{C.2})$$

where  $\lambda_i$  is the eigenvalue of the matrix  $\mathbf{A}$ . To remove the influence of a particular eigenvector  $\mathbf{x}_k$  from matrix  $\mathbf{A}$  we simply subtract the matrix  $\lambda_k \mathbf{x}_k^T \mathbf{x}_k$  from the original matrix  $\mathbf{A}$ . The  $k^{th}$  eigenvalue diminishes and matrix  $\mathbf{A}$  has an eigenvalue of zero instead of  $\lambda_k$ . Based on this observation, the Hotelling's deflation defined as

$$\mathbf{A}_t = \mathbf{A}_{t-1} - \mathbf{x}_t \mathbf{x}_t^T \mathbf{A}_{t-1} \mathbf{x}_t \mathbf{x}_t^T, \quad (\text{C.3})$$

where  $\mathbf{x}_t$  is the eigenvector that we want to remove the effect of it from the matrix  $\mathbf{A}_t$ . In general the Hotelling's deflation does not preserve positive semi-definiteness of the covariance matrices, therefore it can not be efficiently used to deflate the matrices with sparsified eigenvectors.

## C.2 Schur Complement Deflation

This method is based on conditional variance between the original sample space and its projected subspace. Let  $\mathbf{x} \in \mathbb{R}^p$  be unit vector and  $\mathbf{W} \in \mathbb{R}^p$  be a Gaussian random vector, then  $(\mathbf{W}, \mathbf{W}\mathbf{x})$  has covariance matrix

$$\mathbf{V} = \begin{pmatrix} \Sigma & \Sigma\mathbf{x} \\ \mathbf{x}^T\Sigma & \mathbf{x}^T\Sigma\mathbf{x} \end{pmatrix} \quad (\text{C.4})$$

where the  $\Sigma$  is the covariance matrix of  $\mathbf{W}$ . The conditional variance

$$\text{Var}(\mathbf{W}|\mathbf{W}\mathbf{x}) = \Sigma - \frac{\Sigma\mathbf{x}^T\mathbf{x}\Sigma}{\mathbf{x}^T\Sigma\mathbf{x}} \quad (\text{C.5})$$

is the Schur complement of the vector variance  $\mathbf{x}^T\Sigma\mathbf{x}$  in the full covariance matrix  $\mathbf{V}$ . Using the original covariance matrix instead of  $\Sigma$ , we end up with new deflation technique:

$$\mathbf{A}_t = \mathbf{A}_{t-1} - \frac{\mathbf{A}_{t-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{A}_{t-1}}{\mathbf{x}_t^T\mathbf{A}_{t-1}\mathbf{x}_t} \quad (\text{C.6})$$

which conditionally eliminates the effect of sparse vector  $\mathbf{x}$  from the covariance matrix  $\mathbf{A}_{t-1}$ . Schur complement deflation has the following properties:

- It preserves semi-definiteness.
- The vector  $\mathbf{x}_t$  is left and right orthogonal to the matrix  $\mathbf{A}_t$ , such that  $\mathbf{A}_t\mathbf{x}_t = \mathbf{x}_t^T\mathbf{A}_t = 0$ .
- It reduces to Eq. C.3, if  $\mathbf{x}_t$  is an eigenvector of  $\mathbf{A}_{t-1}$ .

# APPENDIX D

## New Objective Function and Its Properties

### D.1 New objective function and its minimum points

In the spatial filtering framework, the generalized Rayleigh quotient (RQ) (Eq. A.11) is used to extract the CSP solution from the covariance matrices  $\mathbf{A}$  and  $\mathbf{B}$  for two different classes A and B respectively. We want to minimize or maximize the RQ in order to increase variance imbalance between class A and class B. The minimizing spatial filter  $\mathbf{w}_{min}$  decreases the variance of class A with respect to class B, whereas the maximizing spatial filter  $\mathbf{w}_{max}$  increases the variance of class A with respect to class B. The solution of the minimization problem of the RQ is related with the generalized eigenvalue decomposition of the covariance matrices  $\mathbf{A}$  and  $\mathbf{B}$ :

$$\mathbf{A}\mathbf{w}_{min} = \lambda_{min}\mathbf{B}\mathbf{w}_{min}, \quad (\text{D.1})$$

where  $\lambda_{min}$  is the minimum valued generalized eigenvalue of the pair  $\mathbf{A}$  and  $\mathbf{B}$  and  $\mathbf{w}_{min}$  is the associated eigenvector of this eigenvalue which also minimizes the RQ. RQ is scale independent that means  $\mathbf{w}' = \alpha\mathbf{w}_{min}$  is also a solution of this problem where  $\alpha \neq 0$  and  $\alpha \in \mathbb{R}$  (Fig. D.1b). It turns out that at  $\mathbf{w}'$  the Rayleigh quotient

is equal to the corresponding eigenvalue such that  $R(\mathbf{w}') = R(\mathbf{w}_{min}) = \lambda_{min}$  as shown below.

$$R(\mathbf{w}') = \frac{\mathbf{w}'^T \mathbf{A} \mathbf{w}'}{\mathbf{w}'^T \mathbf{B} \mathbf{w}'} = \frac{\alpha^2 \mathbf{w}_{min}^T \mathbf{A} \mathbf{w}_{min}}{\alpha^2 \mathbf{w}_{min}^T \mathbf{B} \mathbf{w}_{min}} = R(\mathbf{w}_{min}). \quad (\text{D.2})$$

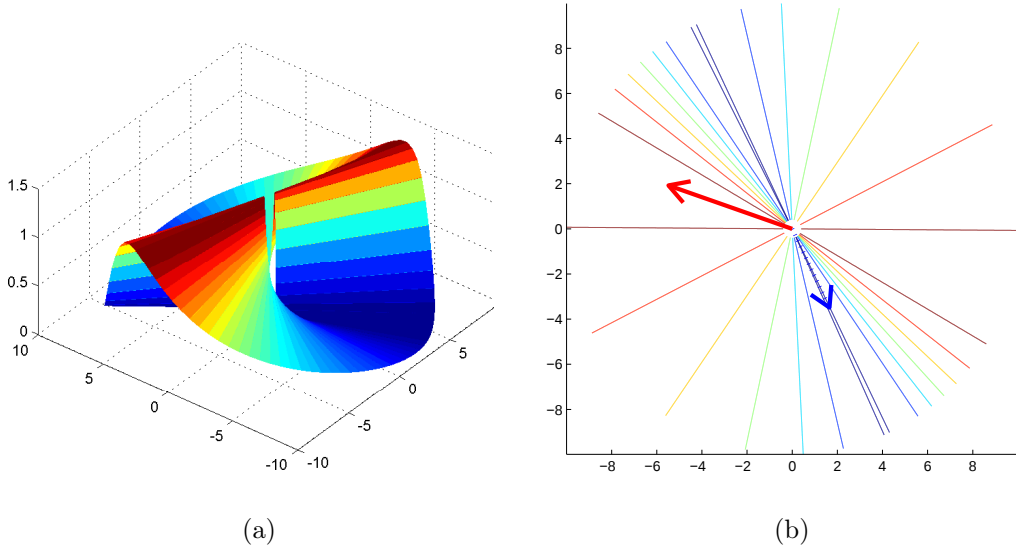


Figure D.1: (a) The RQ surface plot for matrices  $\mathbf{A}$  and  $\mathbf{B}$  defined in Eq. D.5. Notice that the RQ function has the same value on a particular direction. (b). The contour plots of the RQ curve. The blue arrow is the eigenvector that has the smallest eigenvalue and the red arrow is the eigenvector that has the largest eigenvalue.

In Chapter 4, we want to minimize  $f(\mathbf{w}) = L(\mathbf{w}) + \alpha \|\mathbf{w}\|_1$  to obtain a sparse  $\mathbf{w}$  solutions where  $\|\mathbf{w}\|_1$  is the  $\ell_1$  norm of the spatial filter  $\mathbf{w}$ . If we use Rayleigh quotient as objective function  $L(\mathbf{w})$ , the optimization algorithm arbitrarily minimizes the gain of  $\mathbf{w}$  to minimize  $\alpha \|\mathbf{w}\|_1$  after finding the direction that minimizes  $R(\mathbf{w})$ , because  $R(\mathbf{w})$  does not depend on the gain of its argument. Therefore we get the scaled down solution of Rayleigh quotient for  $f(\mathbf{w})$ . To get a sparse solution we need to have a function that depends on its argument's gain. Instead of using Rayleigh quotient we use an alternative objective function. This function is defined to be

$$G(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \frac{1}{\mathbf{w}^T \mathbf{B} \mathbf{w}}. \quad (\text{D.3})$$

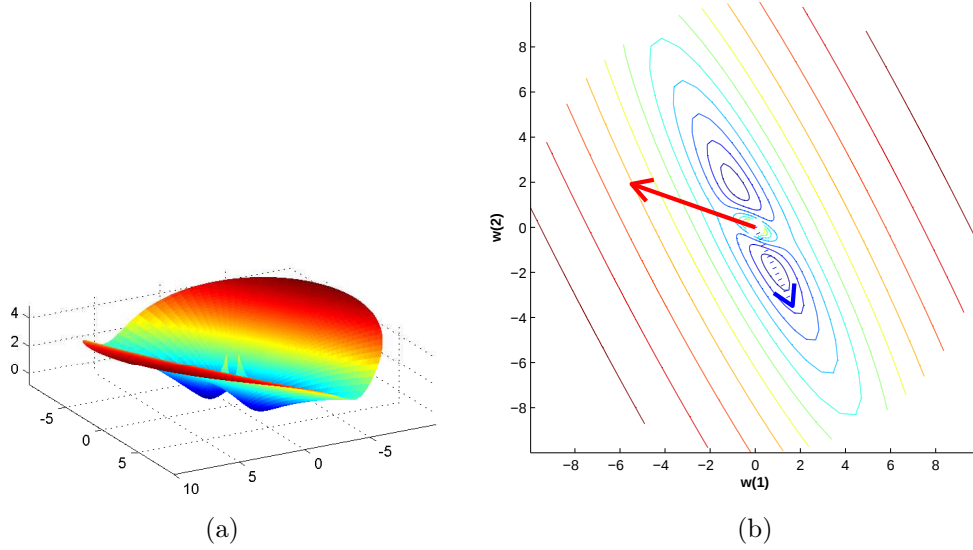


Figure D.2: (a) The surface plots of the new objective function  $G(\mathbf{w})$  for matrices  $\mathbf{A}$  and  $\mathbf{B}$  defined in Eq. D.5. Notice that this new objective function does not have the same value on a particular direction as the RQ. (b) The contour plot of the new objective function. The dotted blue arrow is the eigenvector that has the smallest eigenvalue and the solid red arrow is the eigenvector that has the largest eigenvalue. Notice that the minimum of the function is located on the direction of eigenvector that is associated with the smallest eigenvalue.

Let us define  $\mathbf{w}^T \mathbf{A} \mathbf{w}$  as “ $a(\mathbf{w})$ ” and  $\mathbf{w}^T \mathbf{B} \mathbf{w}$  as “ $b(\mathbf{w})$ ” then the Rayleigh quotient can be expressed as:

$$R(\mathbf{w}) = \frac{a(\mathbf{w})}{b(\mathbf{w})}$$

and our new objective function would be

$$G(\mathbf{w}) = R(\mathbf{w})b(\mathbf{w}) + \frac{1}{b(\mathbf{w})}, \quad (\text{D.4})$$

for the following toy problem(Fig. D.2 ).

Let

$$\mathbf{A} = \begin{bmatrix} 0.9907 & 0.4966 \\ 0.4966 & 0.3125 \end{bmatrix} \quad (\text{D.5})$$

and

$$\mathbf{B} = \begin{bmatrix} 0.9979 & 1.0263 \\ 1.0263 & 2.0656 \end{bmatrix}.$$

We would like to find the minimum values of the function  $G(\mathbf{w})$  and the location of potential minimum  $\mathbf{w}$  points. First, we take the derivative of the objective function with respect to  $\mathbf{w}$  and set this derivative to zero to obtain the minimum points of the function. The derivative of this new objective function is as follows:

$$\mathbf{h}(\mathbf{w}) = \frac{\partial G(\mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{A}\mathbf{w} - \frac{2\mathbf{B}\mathbf{w}}{(\mathbf{w}^T\mathbf{B}\mathbf{w})^2} = \mathbf{0}. \quad (\text{D.6})$$

We can transform this vector equation into a scalar equation by multiplying Eq. D.6 by  $\frac{\mathbf{w}^T}{2\mathbf{w}^T\mathbf{B}\mathbf{w}}$  from the left side. This leads to

$$0 = \frac{\mathbf{w}^T\mathbf{h}(\mathbf{w})}{\mathbf{w}^T\mathbf{B}\mathbf{w}} = \frac{\mathbf{w}^T\mathbf{A}\mathbf{w}}{\mathbf{w}^T\mathbf{B}\mathbf{w}} - \frac{1}{(\mathbf{w}^T\mathbf{B}\mathbf{w})^2} = R(\mathbf{w}) - \frac{1}{(\mathbf{w}^T\mathbf{B}\mathbf{w})^2}, \quad (\text{D.7})$$

Next, we substitute  $b(\mathbf{w})$  (which is also defined as in Eq. D.4) for  $\mathbf{w}^T\mathbf{B}\mathbf{w}$  and get the following equation:

$$b(\mathbf{w}) = \frac{1}{\sqrt{R(\mathbf{w})}}. \quad (\text{D.8})$$

Let  $\mathbf{w}_{min}$  is a solution for Eq. D.8, so the value of  $G(\mathbf{w}_{min})$  is expressed as:

$$G(\mathbf{w}_{min}) = R(\mathbf{w}_{min})b(\mathbf{w}_{min}) + \frac{1}{b(\mathbf{w}_{min})}. \quad (\text{D.9})$$

Substituting Eq. D.8 for  $b(\mathbf{w}_{min})$ , leads to

$$G(\mathbf{w}_{min}) = 2\sqrt{R(\mathbf{w}_{min})}. \quad (\text{D.10})$$

The Rayleigh quotient ( $R(\mathbf{w})$ ) can be at least the minimum GED eigenvalue of  $\mathbf{A}$  and  $\mathbf{B}$  ( $\lambda_{min}$ ), so the solution that minimizes the function  $G(\mathbf{w})$  would be on the line which is on the same direction of the eigenvector associated with the minimum eigenvalue ( $\mathbf{w}_{min}$ ).

We know that we have infinitely many solutions for the RQ minimization problem, because scaling the solution does not change the value of the RQ. This is not the case for the  $G(\mathbf{w})$  minimization problem. From Eq. D.10 we know

that the direction of the solution for  $G(\mathbf{w})$ , minimization problem is fixed, which means it is on a line; however, it should also satisfy the following equation:

$$\mathbf{w}^T \mathbf{B} \mathbf{w} = \frac{1}{\sqrt{\lambda_{min}}}, \quad (\text{D.11})$$

which forms a hyper ellipsoid in multidimensional space. Therefore, the solutions are at the intersection of a line and a hyper ellipsoid surface. This means that we have only two points that minimize the function  $G(\mathbf{w})$  in multidimensional space.

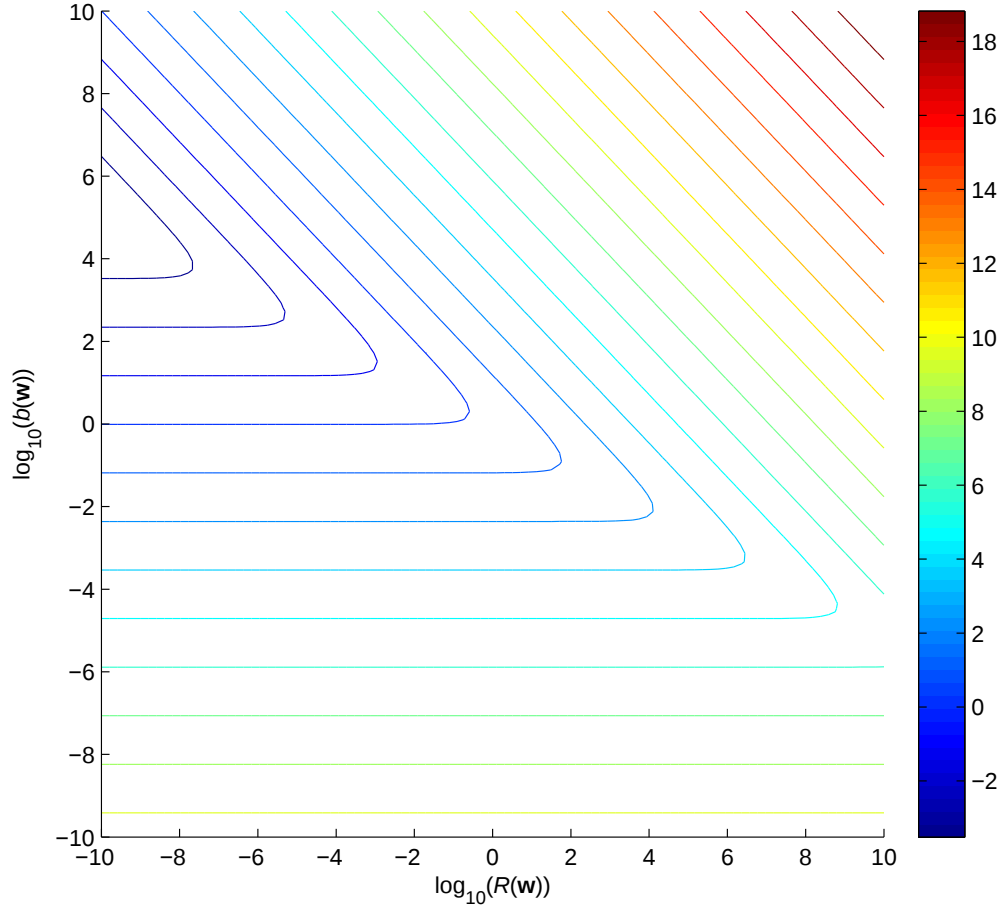


Figure D.3: The novel function ( $\log_{10}(G(\mathbf{w}))$ ) contour in terms of  $R(\mathbf{w})$  and  $b(\mathbf{w})$ . Notice that the value of the function decreases with respect to  $R(\mathbf{w})$ . So minimum  $R(\mathbf{w})$  yields a minimum value for the objective function as well.

## D.2 Gradient and Hessian of minimization function

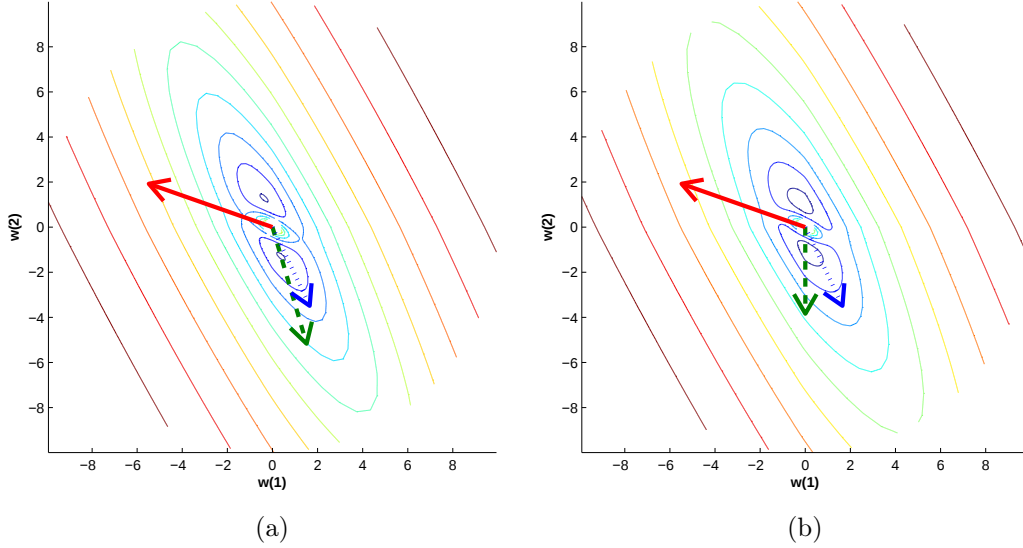


Figure D.4: The contour plot of the new objective function with epsL1 penalty and  $\alpha$  values (a) 0.25 and (b) 0.5. The dotted blue arrow represents the original solution ( $\alpha = 0$ ) and green dashed line represents the solution with epsL1 penalty. The solid red line is the original RQ solution that maximize the RQ. Notice that increasing  $\alpha$  also increase the sparsity degree of the solution by making one of components close to zero.

The new cost function with smooth  $\ell_1$  penalty can be written as

$$f(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \frac{1}{\mathbf{w}^T \mathbf{B} \mathbf{w}} + \alpha \sum_{i=1}^C \sqrt{w_i^2 + \epsilon}, \quad (\text{D.12})$$

where  $\mathbf{A}$  and  $\mathbf{B} \in \mathbb{R}^{C \times C}$  are the covariance matrices of the competing classes and  $\mathbf{w} \in \mathbb{R}^C$  is the sparse spatial filter that we want to find,  $\alpha \in \mathbb{R}$  determines the level sparsity and the function  $f \in \mathbb{R}^C \rightarrow \mathbb{R}$ .

To demonstrate how this minimization problem of the new objective function with an epsL1 norm penalty sparsify the weight vector, we used the following two dimensional toy problem,

$$\mathbf{A} = \begin{bmatrix} 0.9907 & 0.4966 \\ 0.4966 & 0.3125 \end{bmatrix} \quad (\text{D.13})$$

and

$$\mathbf{B} = \begin{bmatrix} 0.9979 & 1.0263 \\ 1.0263 & 2.0656 \end{bmatrix}$$

for various values of  $\lambda$  in Eq. D.12.

After solving Eq. D.12 using the toy problem with  $\alpha = 0$  we obtain  $\mathbf{w} = [0.4283 \ -0.9037]^T$  which is shown in Fig. D.2. This particular solution also minimizes the RQ. The solution for  $\alpha = 0.25$  is found to be  $\mathbf{w} = [0.2809 \ -0.9597]^T$  and shown in Fig. D.4a. Note that the value of the first component getting smaller with a nonzero  $\alpha$  value. Further increasing  $\alpha$  and setting it to 0.5, we find  $\mathbf{w} = [0.0015 \ -1.0000]^T$ . In this case, the first component of  $\mathbf{w}$  vanishes and becomes essentially zero.

To find minimum value for our new minimization problem expressed in Eq. D.12, Matlab's "fminunc" is used in Chapter 4. This function estimates the gradient and Hessian from the function samples if they are not provided. Computing the Hessian and the gradient using samples computationally intensive. Therefore, we computed the gradient and the Hessian of the function in Eq. D.12 as follows:

$$\frac{\partial f}{\partial \mathbf{w}} = \mathbf{H}_A \mathbf{w} - \frac{\mathbf{H}_B \mathbf{w}}{f_B^2} + \alpha \mathbf{c}, \quad (\text{D.14})$$

where the components of  $\mathbf{c}$  are

$$c_i = \frac{w_i}{\sqrt{w_i^2 + \epsilon}}. \quad (\text{D.15})$$

The matrices  $\mathbf{H}_A$  and  $\mathbf{H}_B$  are defined as:

$$\begin{aligned} \mathbf{H}_A &= \mathbf{A} + \mathbf{A}^T \text{ and} \\ \mathbf{H}_B &= \mathbf{B} + \mathbf{B}^T. \end{aligned} \quad (\text{D.16})$$

The scalar function  $f_B$  is defined as

$$f_B = \mathbf{w}^T \mathbf{B} \mathbf{w}. \quad (\text{D.17})$$

Hessian of the cost function is found to be:

$$\frac{\partial^2 f}{\partial^2 \mathbf{w}} = \mathbf{H}_A + \frac{2\mathbf{H}_B \mathbf{w} \mathbf{w}^T \mathbf{H}_B^T - f_B \mathbf{H}_B}{f_B^3} + \alpha \mathbf{H}_c, \quad (\text{D.18})$$

where the components of  $\mathbf{H}_C$  are:

$$H_c(i, j) = \begin{cases} \frac{\epsilon}{(w_i^2 + \epsilon)^{\frac{3}{2}}}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \quad (\text{D.19})$$

The new cost functions  $G(\mathbf{w})$  and  $f(\mathbf{w})$  can be also minimized using the projections onto convex sets framework. This new approach is discussed in Appendix E.

## APPENDIX E

# Projections Onto Convex Sets based Optimization of the New Cost Function

We present two new techniques based on projections onto convex space (POCS) framework for solving the optimization problems that we considered in this thesis. In general, the optimization problems are of the following form:

$$\min_{\mathbf{w} \in C} f(\mathbf{w}), \quad (\text{E.1})$$

where  $C$  is a convex set in  $\mathbb{R}^N$  and  $f(\mathbf{w})$  is the cost function. Some minimization examples that we consider are as follows:

$$f_1(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \frac{1}{\mathbf{w}^T \mathbf{B} \mathbf{w}}, \quad (\text{E.2})$$

and

$$f_1(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \frac{1}{\mathbf{w}^T \mathbf{B} \mathbf{w}} + \|\mathbf{w}\|, \quad (\text{E.3})$$

where  $\mathbf{A}$  and  $\mathbf{B}$  are the covariance matrices. A modified version of the generalized eigenvalue problem can be formulated as follows:

$$f_2(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w}, \quad (\text{E.4})$$

such that  $\mathbf{w}^T \mathbf{B} \mathbf{w} \leq 1$ . In Sparsified version of the GED is given by:

$$f_3(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w} + \|\mathbf{w}\|_1, \quad (\text{E.5})$$

such that  $\mathbf{w}^T \mathbf{B} \mathbf{w} \leq 1$ .

In problem E.2 the cost function is not convex but it can be considered convex as explained later in this section. The cost functions in E.4 and E.5 are convex.

Bregman developed projection based convex optimization methods based on the so-called Bregman distance [79, 80] to solve convex optimization problems. However it may not be easy to compute the Bregman distance in general [81]. We use Bregman's projections onto convex sets (POCS) framework to solve convex optimization problems. Bregman's POCS method is widely used for finding a common point of convex sets in many inverse problems [82–95]. The goal is simply to find a vector which is in the intersection of convex sets. In standard POCS method, there is no need to compute the Bregman distance. In each step of the algorithm, an orthogonal projection is performed onto one of the convex sets. Successive orthogonal projections converge to a vector which is in the intersection of all the convex sets. If the sets do not intersect, the iterations oscillate between members of the sets [96, 97].

Let us first consider a convex minimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^N} f(\mathbf{w}), \quad (\text{E.6})$$

where  $f : \mathbb{R}^N \rightarrow \mathbb{R}$

We increase the dimension by one to define the following sets in  $\mathbb{R}^{N+1}$  corresponding to the convex cost function  $f(\mathbf{w})$  as follows:

$$C_f = \{\underline{\mathbf{w}} = [\mathbf{w}^T \ y]^T : y \geq f(\mathbf{w})\}, \quad (\text{E.7})$$

which is the set of  $N+1$  dimensional vectors whose  $N+1^{st}$  component  $y$  is greater than  $f(\mathbf{w})$ . Another set that we use is the level set

$$C_s = \{\underline{\mathbf{w}} = [\mathbf{w}^T \ y]^T : y \leq \alpha, \ \underline{\mathbf{w}} \in \mathbb{R}^{N+1}\} \quad (\text{E.8})$$

where  $\alpha$  is a real number. Let us also assume that  $f(\mathbf{w}) \geq \alpha$  for all  $f(\mathbf{w}) \in \mathbb{R}$ . Sets  $C_f$  and  $C_s$  are graphically illustrated in Fig. E.1.

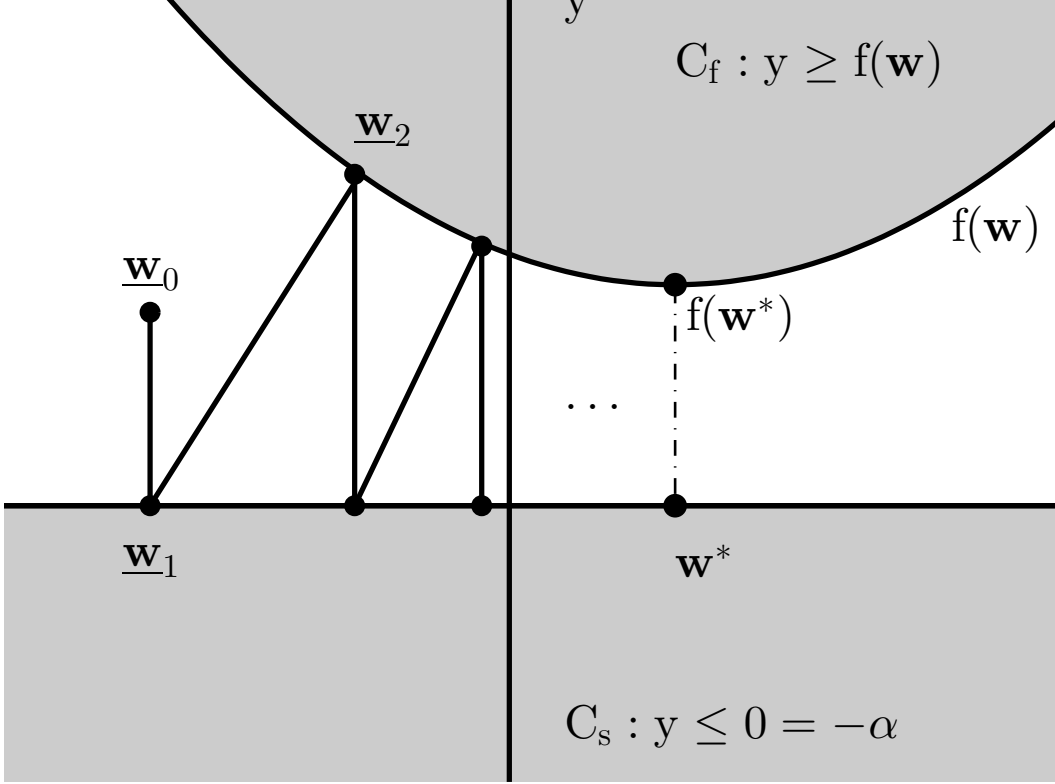


Figure E.1: We sequentially project an initial vector  $\mathbf{w}_0$  onto two convex sets  $C_f$  and  $C_s$  to find a local minimum. The minimum is located at  $\mathbf{w}^*$ .

The POCS based minimization algorithm starts with an arbitrary  $\mathbf{w}_0 = [\mathbf{w}_0^T \ y_0]^T \in \mathbb{R}^{N+1}$ . We project  $\mathbf{w}_0$  onto the set  $C_s$  to obtain the first iterate  $\mathbf{w}_1$  which will be

$$\mathbf{w}_1 = [\mathbf{w}_0^T \ 0]^T, \quad (\text{E.9})$$

where  $\alpha = 0$  is assumed as in Fig. E.1. Then we project  $\mathbf{w}_1$  onto the set  $C_f$ . The new iterate  $\mathbf{w}_2$  is determined by minimizing the distance between  $\mathbf{w}_1$  and  $C_f$ , i.e.,

$$\mathbf{w}_2 = \arg \min_{\mathbf{w} \in C_s} \|\mathbf{w}_1 - \mathbf{w}\|. \quad (\text{E.10})$$

To solve the problem in Eq. E.10 we do not need to compute the Bregman's so-called D-projection. Eq. E.10 is the ordinary orthogonal projection operation

onto the set  $C_f \in \mathbb{R}^{N+1}$ . After finding  $\underline{\mathbf{w}}_2$ , we perform the next projection onto the set  $C_s$  and obtain  $\underline{\mathbf{w}}_3$  etc. Eventually iterates oscillates between two nearest vectors of the two sets  $C_s$  and  $C_f$ . As a result:

$$\lim_{n \rightarrow \infty} \underline{\mathbf{w}}_n = [ \mathbf{w}^* \quad f(\mathbf{w}^*) ]^T, \quad (\text{E.11})$$

where  $\mathbf{w}^*$  is the N-dimensional vector minimizing  $f(\mathbf{w})$ . The proof of the equation follows from Bregman's POCS theorem [79,80] and studied by [85,97].

If the cost function  $f(\mathbf{w})$  is not a convex set, then it iterates to converge to a local minimum of the function  $f$ , as shown in Fig. E.2.

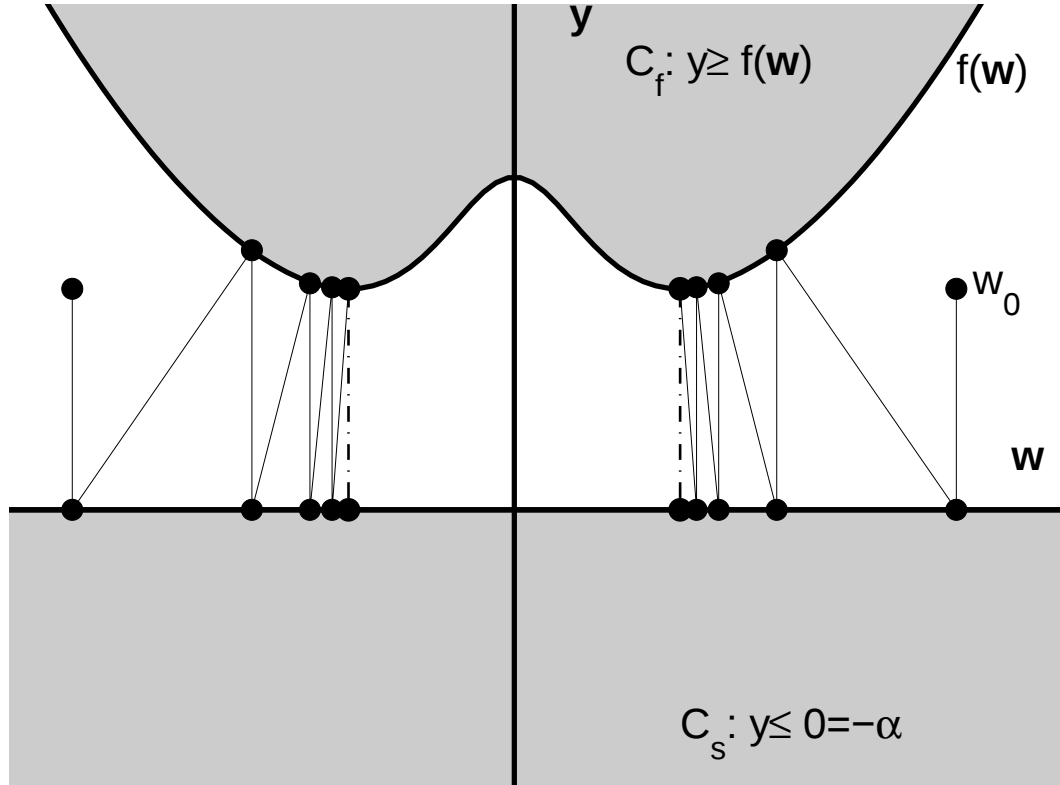


Figure E.2: The point  $\mathbf{w}_0$  converges to a local minimum by the POCS algorithm.

Since the cost functions E.2 and E.3 have only two minima with equal value, it is sufficient to find one of the solutions. Therefore the method proposed here converges to one of the solution.

# Bibliography

- [1] T. Pistohl, A. Schulze-Bonhage, A. Aertsen, C. Mehring, and T. Ball, “Decoding Natural Grasp Types from Human ECoG,” *NeuroImage*, vol. 59, no. 1, pp. 248–260, 2012. [Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/S105381191100749X>.
- [2] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz, “Cortical Control of a Prosthetic Arm for Self-Feeding,” *Nature*, vol. 453, no. 7198, pp. 1098–1101, 2008.
- [3] C. T. Moritz, S. I. Perlmutter, and E. E. Fetz, “Direct Control of Paralysed Muscles by Cortical Neurons,” *Nature*, vol. 456, pp. 639–642, 2008.
- [4] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, “Combined Optimization of Spatial and Temporal Filters for Improving Brain-Computer Interfacing,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 11, pp. 2274–2281, November 2006.
- [5] K. V. Shenoy, D. Meeker, S. Cao, S. A. Kureshi, B. Pesaran, C. A. Buneo, A. P. Batista, P. P. Mitra, J. W. Burdick, and R. A. Andersen, “Neural Prosthetic Control Signals from Plan Activity,” *NeuroReport*, vol. 14, no. 4, pp. 591–596, 2003.
- [6] N. F. İnce, “Analysis and Classification of EEG with Adapted Wavelets and Local Discriminant Bases,” Ph.D. dissertation, Cukurova University, 2005. [Online; accessed 19-June-2013]: <http://library.cu.edu.tr/tezler/5501.pdf>.

- [7] J. R. Wolpaw and D. J. McFarland, "Multichannel EEG-Based Brain-Computer Communication," *Electroencephalography and Clinical Neurophysiology*, vol. 90, no. 6, pp. 444–449, 1994.
- [8] C. W. Anderson, E. A. Stolz, and S. Shamsunder, "Multivariate Autoregressive Models for Classification of Spontaneous Electroencephalographic Signals During Mental Tasks," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 3, pp. 277–286, March 1998.
- [9] G. Pfurtscheller, C. Neuper, and J. Berger, "Source Localization Using Eventrelated Desynchronization (ERD) within the Alpha Band," *Brain Topography*, vol. 6, pp. 269–275, 1994, 10.1007/BF01211172. [Online; accessed 19-June-2013]: <http://dx.doi.org/10.1007/BF01211172>.
- [10] W. D. Penny, S. J. Roberts, and M. Stokes, "Imagined Hand Movements Identified from the EEG Mu-Rhythm," *Journal of Neuroscience Methods*, 1998.
- [11] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the Channel Selection and Classification Accuracy in EEG-Based BCI," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 6, pp. 1865–1873, June 2011.
- [12] G. Pfurtscheller and F. H. L. da Silva, "Event-Related EEG/MEG Synchronization and Desynchronization: Basic Principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999. [Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/S1388245799001418>.
- [13] S. P. Levine, "Identification of Electrocorticogram Patterns as the Basis for a Direct Brain Interface," *Journal of Clinical Neurophysiology*, vol. 16, p. 439, 1999.
- [14] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, "A Brain-Computer Interface Using Electrocorticographic Signals in Humans," *Journal of Neural Engineering*, vol. 1, no. 2, pp. 63–71, 2004.

- [15] J. K. Chapin, K. A. Moxon, R. S. Markowitz, and M. A. L. Nicolelis, "Real-Time Control of a Robot Arm Using Simultaneously Recorded Neurons in the Motor Cortex," *Nature Neuroscience*, vol. 2, no. 7, pp. 664–670, 1999.
- [16] J. Wessberg, C. R. Stambaugh, J. D. Kralik, P. D. Beck, M. Laubach, J. K. Chapin, J. Kim, S. J. Biggs, M. A. Srinivasan, and M. A. L. Nicolelis, "Letters to Nature: Real-Time Prediction of Hand Trajectory by Ensembles of Cortical Neurons in Primates," *Nature*, vol. 408, no. 16, pp. 361–365, 2002.
- [17] M. A. L. Nicolelis, "Actions from Thoughts," *Nature*, vol. 409, no. 6818, pp. 403–408, 2001.
- [18] C. Kemere, G. Santhanam, B. M. Yu, A. Afshar, S. I. Ryu, T. H. Meng, , and K. V. Shenoy, "Detecting Neural-State Transitions Using Hidden Markov Models for Motor Cortical Prostheses," *Journal of NeuroPhysiology*, vol. 100, no. 4, pp. 2442–2452, June 2008.
- [19] H. Scherberger, M. R. Jarvis, and R. A. Andersen, "Cortical Local Field Potential Encodes Movement Intentions in the Posterior Parietal Cortex," *Neuron*, vol. 46, no. 2, pp. 347–354, 2005.
- [20] N. F. Ince, R. Gupta, S. Arica, A. H. Tewfik, J. Ashe, and G. Pellizzer, "High Accuracy Decoding of Movement Target Direction in Non-Human Primates Based on Common Spatial Patterns of Local Field Potentials," *PLoS ONE*, vol. 5, no. 12, p. e14384, December 2010.
- [21] S.-y. Jung, S. kyun Kang, M.-J. Lee, and I. Moon, "Design of Robotic Hand with Tendon-Driven Three Fingers," in *International Conference on Control, Automation and Systems, 2007. ICCAS '07.*, 2007, pp. 83–86. [Online; accessed 19-June-2013]: [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=4406884&tag=1](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4406884&tag=1).
- [22] L. A. Farwell and E. Donchin, "Talking Off the Top of Your Head: Toward a Mental Prosthesis Utilizing Event-Related Brain Potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988. [Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/0013469488901496>.

- [23] “EEG 10-20 System Diagram.” 2012. [Online; accessed 19-June-2013]: [http://www.medicine.mcgill.ca/physio/vlab/biomed\\_signals/images/10-20\\_sys.gif](http://www.medicine.mcgill.ca/physio/vlab/biomed_signals/images/10-20_sys.gif).
- [24] T. N. Lal, T. Hinterberger, G. Widman, C. E. Elger, B. Schölkopf, and N. Birbaumer, “Methods towards Invasive Human Brain Computer Interfaces,” in *Advances in Neural Information Processing Systems 17*. MIT Press, 2005, pp. 737–744.
- [25] D. B. Kirsch, “There and Back Again : A Current History of Sleep Medicine,” *Chest*, vol. 139, no. 4, pp. 939–946, 2011.
- [26] R. M. Kaplan, “The Mind Reader: The Forgotten Life of Hans Berger, Discoverer of the EEG,” *Australasian Psychiatry*, vol. 19, no. 2, pp. 168–169, 2011. [Online; accessed 19-June-2013]: <http://apy.sagepub.com/content/19/2/168.short>.
- [27] G. Pfurtscheller and F. H. L. da Silva, *Event Related Desynchronization*, ser. Event-related Desynchronization. Elsevier Science Health Science Division, 1999. [Online; accessed 19-June-2013]: <http://books.google.com.tr/books?id=mIWjkgAACAAJ>.
- [28] N. F. İnce, F. Göksu, and A. H. Tewfik, “ECoG Based Brain Computer Interface with Subset Selection,” in *Biomedical Engineering Systems and Technologies*, ser. Communications in Computer and Information Science, A. Fred, J. Filipe, and H. Gamboa, Eds. Springer Berlin Heidelberg, 2009, vol. 25, pp. 357–374.
- [29] R. Flamary and A. Rakotomamonjy, “Decoding Finger Movements from ECoG Signals Using Switching Linear Models,” *CoRR*, vol. abs/1106.3395, 2011.
- [30] E. C. Leuthardt, K. J. Miller, G. Schalk, R. P. N. Rao, and J. G. Ojemann, “Electrocorticography Based Brain Computer Interface-the Seattle Experience,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 194–198, june 2006.

- [31] C. Neuper and G. Pfurtscheller, "Event-Related Dynamics of Cortical Rhythms: Frequency-Specific Features and Functional Correlates," *International Journal of Psychophysiology*, vol. 43, no. 1, pp. 41–58, 2001. [Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/S0167876001001787>.
- [32] G. Pfurtscheller and C. Neuper, "Motor Imagery and Direct Brain-Computer Communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, July 2001.
- [33] G. Pfurtscheller, C. Neuper, A. Schlogl, and K. Lugger, "Separability of EEG Signals Recorded During Right and Left Motor Imagery Using Adaptive Autoregressive Parameters," *IEEE Transactions on Rehabilitation Engineering*, vol. 6, no. 3, pp. 316–325, 1998.
- [34] N. Achtman, A. Afshar, G. Santhanam, B. M. Yu, S. I. Ryu, and K. V. Shenoy, "Free-Paced High-Performance Brain-Computer Interfaces," *Journal of Neural Engineering*, vol. 4, no. 3, p. 336, 2007.
- [35] V. Aggarwal, S. Acharya, F. Tenore, H.-C. Shin, R. Etienne-Cummings, M. H. Schieber, and N. V. Thakor, "Asynchronous Decoding of Dexterous Finger Movements Using M1 Neurons," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 1, pp. 3–14, February 2008.
- [36] M. Mollazadeh, V. Aggarwal, G. Singhal, A. Law, A. Davidson, M. Schieber, and N. Thakor, "Spectral Modulation of LFP Activity in M1 During Dexterous Finger Movements," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, August 2008, pp. 5314–5317.
- [37] E. J. Hwang and R. A. Andersen, "Brain Control of Movement Execution Onset Using Local Field Potentials in Posterior Parietal Cortex," *The Journal of Neuroscience*, vol. 29, no. 45, pp. 14 363–14 370, November 2009.
- [38] T. Yanagisawa, M. Hirata, Y. Saitoh, T. Goto, H. Kishima, R. Fukuma, H. Yokoi, Y. Kamitani, and T. Yoshimine, "Real-Time Control of a Prosthetic

Hand Using Human Electrocorticography Signals,” *Journal of Neurosurgery*, vol. 114, no. 6, pp. 1715–1722, 2011.

- [39] R. A. Andersen, S. Musallam, and B. Pesaran, “Selecting the Signals for a Brain–Machine Interface,” *Current Opinion in Neurobiology*, vol. 14, no. 6, pp. 720–726, 2004. [Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/S0959438804001588>.
- [40] R. A. Andersen, J. W. Burdick, S. Musallam, H. Scherberger, B. Pesaran, D. Meeker, B. D. Corneil, I. Fineman, Z. Nenadic, E. Branchaud, J. G. Cham, B. Greger, Y. C. Tai, and M. M. Mojarradi, “Recording Advances for Neural Prosthetics,” in *Engineering in Medicine and Biology Society, 2004. IEMBS '04. 26th Annual International Conference of the IEEE*, vol. 2, sept. 2004, pp. 5352–5355.
- [41] R. D. Flint, Z. A. Wright, and M. W. Slutzky, “Control of a Biomimetic Brain Machine Interface with Local Field Potentials: Performance and Stability of a Static Decoder over 200 Days,” in *IEEE Annual International Conference Engineering in Medicine and Biology Society (EMBC)*, September 2012, pp. 6719–6722.
- [42] T. Ball, M. Kern, I. Mutschler, A. Aertsen, and A. Schulze-Bonhage, “Signal Quality of Simultaneously Recorded Invasive and Non-Invasive EEG,” *Neuroimage*, vol. 46, no. 3, pp. 708–716, 2009.
- [43] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, “Optimizing Spatial Filters for Robust EEG Single-Trial Analysis,” *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008.
- [44] J. Farquhar, N. J. Hill, T. N. Lal, and B. Schalkopf, “Regularised CSP for Sensor Selection in BCI,” in *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course*, 2006.
- [45] B. Reuderink and M. Poel, “Robustness of the Common Spatial Patterns Algorithm in the BCI-Pipeline,” Univ. of Twente, Enschede, Tech. Rep. TR-CTI, July 2008.

- [46] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal Spatial Filtering of Single Trial EEG During Imagined Hand Movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, December 2000.
- [47] X. Yong, R. K. Ward, and G. E. Birch, "Sparse Spatial Filter Optimization for EEG Channel Reduction in Brain-Computer Interface," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*., April 2008, pp. 417–420.
- [48] F. Göksu, N. F. İnce, and A. H. Tewfik, "Sparse Common Spatial Patterns in Brain Computer Interface Applications," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 533–536.
- [49] K. J. Miller and G. Schalk, "Prediction of Finger Flexion 4th Brain-Computer Interface Data Competition," 2008. [Online; accessed 19-June-2013]: [http://www.bbci.de/competition/iv/desc\\_4.pdf](http://www.bbci.de/competition/iv/desc_4.pdf).
- [50] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial Patterns Underlying Population Differences in the Background EEG," *Brain Topography*, vol. 2, pp. 275–284, 1990.
- [51] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [52] K. Murphy, "Hidden Markov Model (HMM) Toolbox for," 1998. [Online; accessed 19-June-2013]: <http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html>.
- [53] C.-C. Chang and C.-J. Lin, "A Library for Support Vector Machines," 2001. [Online; accessed 19-June-2013]: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [54] N. Liang and L. Bougrain, "Decoding Finger Flexion Using Amplitude Modulation from Band-Specific ECoG," in *European Symposium on Artificial Neural Networks - ESANN 2009*, Bruges, Belgium, 2009, pp. 467–472.

- [55] S. Acharya, F. Tenore, V. Aggarwal, R. Etienne-Cummings, M. H. Schieber, and N. V. Thakor, “Decoding Individuated Finger Movements Using Volume-Constrained Neuronal Ensembles in the M1 Hand Area,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 1, pp. 15–23, February 2008.
- [56] P. Shenoy, K. J. Miller, J. G. Ojemann, and R. P. N. Rao, “Finger Movement Classification for an Electrocorticographic BCI,” in *Neural Engineering, 2007. CNE '07. 3rd International IEEE/EMBS Conference on*, May 2007, pp. 192–195.
- [57] R. Rifkin and A. Klautau, “In Defense of One-Vs-All Classification,” *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, December 2004. [Online; accessed 19-June-2013]: <http://dl.acm.org/citation.cfm?id=1005332.1005336>.
- [58] S. S. Keerthi, “Generalized LARS as an Effective Feature Selection Tool for Text Classification with SVMs,” in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 417–424. [Online; accessed 19-June-2013]: [http://www.keerthis.com/lars\\_icml\\_keerthi\\_05.pdf](http://www.keerthis.com/lars_icml_keerthi_05.pdf).
- [59] M. Dash and H. Liu, “Feature Selection for Classification,” *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997. [Online; accessed 19-June-2013]: <http://machine-learning.martinsewell.com/feature-selection/DashLiu1997.pdf>.
- [60] T. G. Dietterich and G. Bakiri, “Solving Multiclass Learning Problems via Error-Correcting Output Codes,” *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1995.
- [61] S.-i. Lee, H. Lee, P. Abbeel, and A. Y. Ng, “Efficient L1 Regularized Logistic Regression,” in *In AAAI*, 2006.
- [62] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, “BCI Competition III, Dataset IVa,” 2005. [Online; accessed 19-June-2013]: [http://www.bbci.de/competition/iii/desc\\_IVa.html](http://www.bbci.de/competition/iii/desc_IVa.html).
- [63] M. Schmidt, G. Fung, and R. Rosales, “Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches,” 2009.

- [64] R. L. Burden and J. D. Faires, *Numerical Analysis*, 8th ed. Thomson Brooks/ Cole, 2005.
- [65] L. Mackey, “Deflation Methods for Sparse PCA,” in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2009, pp. 1017–1024.
- [66] T. Akgül, S. Mingui, R. J. Schlahassi, and A. E. Çetin, “Characterization of Sleep Spindles Using Higher Order Statistics and Spectra,” *IEEE Trans. on Biomedical Engineering*, vol. 47, pp. 997–1009, 2000.
- [67] H. Tuna, I. Onaran, and A. E. Çetin, “Image Description Using a Multiplier-Less Operator,” *Signal Processing Letters, IEEE*, vol. 16, no. 9, pp. 751–753, 2009.
- [68] J. A. Cadzow, “Minimum  $\ell_1$ ,  $\ell_2$ , and  $\ell_\infty$  Norm Approximate Solutions to an Overdetermined System of Linear Equations,” *Digital Signal Processing*, vol. 12, no. 4, pp. 524–560, 2002. [Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/S1051200401904099>.
- [69] İ. Onaran, N. F. İnce, and A. E. Çetin, “Classification of Multichannel ECoG Related to Individual Finger Movements with Redundant Spatial Projections,” in *International IEEE Engineering in Medicine and Biology Society (EMBS) Conference*, August 2011.
- [70] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2000.
- [71] F. Göksu, F. İnce, and İ. Onaran, “Sparse Common Spatial Patterns with Recursive Weight Elimination,” in *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, November 2011, pp. 117–121.
- [72] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004. [Online; accessed 03-June-2013]: <http://www.worldcat.org/isbn/0521833787>.

- [73] P. Somol, J. Novovicová, J. Grim, and P. Pudil, “Dynamic Oscillating Search Algorithm for Feature Selection,” in *19th International Conference on Pattern Recognition (ICPR)*. IEEE, 2008, pp. 1–4.
- [74] P. Somol and P. Pudil, “Oscillating Search Algorithms for Feature Selection,” in *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 2. IEEE, 2000, pp. 406–409.
- [75] B. Moghaddam, Y. Weiss, and S. Avidan, “Generalized Spectral Bounds for Sparse LDA,” in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML ’06. New York, NY, USA: ACM, 2006, pp. 641–648.
- [76] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer Classification Using Support Vector Machines,” *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.
- [77] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, “Efficient Projections Onto the  $\ell_1$ -Ball for Learning in High Dimensions,” in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML ’08. New York, NY, USA: ACM, 2008, pp. 272–279. [Online; accessed 19-June-2013]: <http://doi.acm.org/10.1145/1390156.1390191>.
- [78] İ. Onaran, N. F. İnce, and A. E. Çetin, “Sparse Spatial Filter via a Novel Objective Function Minimization with Smooth  $\ell_1$  Regularization,” *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 282 – 288, 2013. [Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/S1746809412001103>.
- [79] L. Bregman, “Finding the Common Point of Convex Sets by the Method of Successive Projection.(Russian),” in *Dokl. Akad. Nauk SSSR*, vol. 162, 1965, pp. 487–490.
- [80] L. M. Bregman, “The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming,” (*USSR*) *Computational Mathematics and Mathematical Physics*, vol. 7, no. 3, pp. 200–217, 1967. [Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/0041555367900407>.

- [81] K. Köse, “Signal and Image Processing Algorithms Using Interval Convex Programming and Sparsity,” Ph.D. dissertation, Bilkent University, 2012. [Online; accessed 19-June-2013]: <http://signal.ee.bilkent.edu.tr/Theses/kkoseThesis.pdf>.
- [82] A. E. Çetin, H. Özaktas, and H. M. Ozaktas, “Resolution Enhancement of Low Resolution Wavefields with POCS Algorithm,” *Electronics Letters*, vol. 39, no. 25, pp. 1808–1810, 2003.
- [83] D. C. Youla and H. Webb, “Image Restoration by the Method of Convex Projections: Part 1 Theory,” *IEEE Transactions on Medical Imaging*, vol. 1, no. 2, pp. 81–94, 1982.
- [84] G. T. Herman, “Image Reconstruction from Projections,” *Real-Time Imaging*, vol. 1, no. 1, pp. 3–18, 1995.
- [85] Y. Censor, W. Chen, P. L. Combettes, R. Davidi, and G. Herman, “On the Effectiveness of Projection Methods for Convex Feasibility Problems with Linear Inequality Constraints,” *Computational Optimization and Applications*, vol. 51, no. 3, pp. 1065–1088, 2012. [Online; accessed 19-June-2013]: <http://dx.doi.org/10.1007/s10589-011-9401-7>.
- [86] K. Slavakis, S. Theodoridis, and I. Yamada, “Online Kernel-Based Classification Using Adaptive Projection Algorithms,” *IEEE Transactions on Signal Processing*, vol. 56, pp. 2781–2796, 2008.
- [87] Y. Censor and A. Lent, “An Iterative Row-Action Method for Interval Convex Programming,” *Journal of Optimization Theory and Applications*, vol. 34, no. 3, pp. 321–353, 1981.
- [88] K. Slavakis, S. Theodoridis, and I. Yamada, “Adaptive Constrained Learning in Reproducing Kernel Hilbert Spaces: the Robust Beamforming Case,” *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4744–4764, dec 2009. [Online; accessed 19-June-2013]: <http://dx.doi.org/10.1109/TSP.2009.2027771>.
- [89] K. S. Theodoridis and I. Yamada, “Adaptive Learning in a World of Projections,” *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 97–123, 2011.

- [90] H. J. Trussell and M. R. Civanlar, "The Landweber Iteration and Projection Onto Convex Set," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 6, pp. 1632–1634, 1985.
- [91] P. L. Combettes and J. Pesquet, "Image Restoration Subject to a Total Variation Constraint," *IEEE Transactions on Image Processing*, vol. 13, pp. 1213–1222, 2004.
- [92] P. L. Combettes, "The Foundations of Set Theoretic Estimation," *Proceedings of the IEEE*, vol. 81, no. 2, pp. 182–208, February 1993.
- [93] C. W. Kim, R. Ansari, and A. E. Çetin, "A Class of Linear-Phase Regular Biorthogonal Wavelets," in *ICASSP '92*, 1992, pp. 673–676.
- [94] I. Sezan and H. Stark, "Image Restoration by the Method of Convex Projections: Part 2-Applications and Numerical Results," *IEEE Transactions on Medical Imaging*, vol. 1, no. 2, pp. 95–101, 1982.
- [95] H. T. Lent, "An Iterative Method for the Extrapolation of Band-Limited Functions," *Journal of Mathematical Analysis and Applications*, 83 (2), pp.1981, vol. 83, pp. 554–565, 1981.
- [96] A. E. Çetin, O. N. Gerek, and Y. Yardımcı, "Equiripple FIR Filter Design by the FFT Algorithm," *Signal Processing Magazine, IEEE*, vol. 14, no. 2, pp. 60–64, 1997.
- [97] L. G. Gubin, B. T. Polyak, and E. V. Raik, "The Method of Projections for Finding the Common Point of Convex Sets," *(USSR) Computational Mathematics and Mathematical Physics*, vol. 7, no. 6, pp. 1–24, 1967.  
[Online; accessed 19-June-2013]: <http://www.sciencedirect.com/science/article/pii/0041555367901139>.