The Bilinear Brain: Bilinear Methods for EEG Analysis and Brain Computer Interfaces

by

Christoforos Christoforou

A dissertation submitted to the Graduate Faculty in Computer Science
in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York
2009

THE CITY UNIVERSITY OF NEW YORK
DEPARTMENT OF COMPUTER SCIENCE

This manuscript has been read and accepted for the
Graduate Faculty in Computer Science in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Robert M. Haralick

Date	Chair of Examining Committee

Theodore Brown

Date	Chair of Examining Committee

Lucas Parra

Ioannis Stamos

Paul Sajda

Supervisory Committee

City University Of New York

Abstract

The Bilinear Brain: Bilinear methods for EEG Analysis and Brain Computer Interfaces

by

Christoforos Christoforou

Chair of the Supervisory Committee:
Distinguished Professor Robert M. Haralick

Analysis of electro-encephalographic (EEG) signals has been proven an extremely useful research tool for studying the neural correlates of behavior. Single-trial analysis of these signals is essential for the development of non-invasive Brain Computer Interfaces. Analysis of these signals is often expressed as a single-trial classification problem. The goal is to infer the underling cognitive state of an individual using purely EEG signals. The high dimensional space of EEG observations and the low signal-to-noise ratio (SNR) -often -20db or less - as well as the inter-subject variability and limited observations available for training, make the single-trial classification of EEG an extremely challenging computational task. To address these challenges we introduce concepts from Multi-linear Algebra and incorporate EEG domain knowledge. More precisely, we formulate the problem in a matrix space and introduce a bilinear combination of a matrix to reduce the space dimensions. Thus the title of this dissertation: "The Bilinear Brain". We also address the issue of inter-subject variability by defining a model that is partially subject-invariant. We develop two

classification algorithms based on the Bilinear model. We term the first algorithm Second Order Bilinear Discriminant Analysis (SOBDA). It combines first order and second order statistics of the observation space. The second algorithm we term Bilinear Feature Based Discriminant (BFBD) and addresses the issue of inter-subject variability. We evaluate our methods on both simulated and real human EEG data-sets and show that our method outperforms state-of-the-art methods on different experimental paradigms.

ACKNOWLEDGMENTS

First and foremost I would like to thank my two advisors Lucas Parra and Robert M. Haralick, both of whom played a major role in my accomplishment. Professor Parra introduced me to the world of applied EEG research and provided me with the opportunity to explore and appreciate the significance of applied science. He gave me the freedom of choosing and exploring a topic that suited my scientific aspirations, while providing the inspiration, guidance and support to tackle every challenge along the way. As a member of his lab I uncovered opportunities I never dreamed of. I spent two weeks attending the Machine Learning Summer School in Canberra, Australia and two months visiting and working in the leading lab on Brain Computer Interfacing in Berlin, Germany. Additionally, I had the opportunity to participate in various project meetings across the United States. All this was made possible thanks to his moral support and of course the financial support of his lab.

Professor Haralick helped me sustain my identity as a computer scientist. I considered his lab the 'incubation chamber of thought' and looked forward to our weekly lab meetings where he encouraged us to develop our thought process, deepen our understanding and expand our horizons. I truly appreciate his perspective and am grateful to him for challenging me to think outside the box and seek different ways of approaching scientific problem solving. I will try to channel his insight throughout my life and will always remember his words "how you know you really understand something is when you can see it from many different perspectives and feel comfortable with it."

My sincere thanks to Professor Paul Sajda for his feedback throughout this work, which played a pivotal role in understanding and interpreting the neurological significance of the results of this research. I admire his laidback attitude towards life and his excellent communication and leadership skills. I aspire to improve on these important skills in my own life.

I would like to thank my friend and colleague Dr. Mads for his significant cooperation. Not a very environmentally friendly encounter since we used up thousands of napkins at the coffee shop around the corner to write down formulas and develop new ideas, but a very productive collaboration nonetheless.

A special thanks goes to Marios Philiastides, or I should say Dr. Marios, for providing necessary data sets used in this thesis. It was an exciting surprise to find a fellow Cypriot working on a similar research topic as myself. I am sure our roads will cross again in the future and I would feel fortunate to work with you again.

I would also like to extend my gratitude to Dr. Stamos for reading and providing feedback on this thesis, as well as to all the others whose names I have not mentioned, but was fortunate enough to have worked with during my Ph.D. years.

Following a comment from my fiancee Maria Theodorou who pointed out "When are you going to thank me for....?", and then provided a long-long list of things, I would like to take this opportunity to thank my soon-to-be wife, for her support, love and understanding during this long and sometimes stressful period of completing this dissertation.

Last, but certainly not least, I would like to thank my parents, Andreas and Angeliki Christoforou for their unconditional love and support throughout the years.

DEDICATION

To my parents, Andrea and Angeliki

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1
INTRODUCTION

1.1 Motivation of Research

The work presented in this dissertation is motivated by the analysis of functional brain imaging signals recorded via electroencephalography (EEG). EEG is measured across time and typically at multiple scalp locations, providing a spatio-temporal dataset of the underlying neural activity. In addition, these measurements are often taken over multiple repetitions or trials, where trials may differ in the type of stimulus presented, the task given to the subject, or the subject's response. Analysis of these signals is often expressed as a single-trial classification problem. The goal for the classifier is to determine from EEG data which stimulus was presented or how the subject responded. Many of these classification techniques were originally developed in the context of Brain Computer Interfaces (BCI) but are now more widely used to interpret activity associated with neural processing.

In the case of BCI algorithms [43, 2, 4, 5] the aim is to decode brain activity on a single-trial basis to provide a direct control pathway between a user's intentions and a computer. Such an interface could provide "locked in patients" a more direct and natural control over a neuroprosthesis or other computer applications [2]. Furthermore, by providing an additional communication channel for healthy individuals, BCI systems increases productivity and efficiency in high-throughput tasks [20, 31].

Single-trial Discriminant analysis has been used as a research tool to study the neural correlates of behavior. By extracting activity that differs maximally between two experimental conditions, the typically low signal-to-noise ratio of EEG can be overcome. The resulting Discriminant components can be used to identify the spatial origin and time course of stimulus/response specific activity, while the improved SNR can be leveraged to correlate variability of neural activity across trials to behavioral variability and behavioral performance [34, 20, 35]

Single-trial classification of EEG signals is a challenging task. The first challenge in the analysis EEG of signals is the fact that EEG observations lie in a high dimensional space. Learning regularities in the data in such high dimensions is hard because of the *curse of dimensionality*. Further, the signal to noise ratio in the aforementioned datasets is extremely low and the number of samples available to be used in learning a classifier is usually small. Finally, it is well known that neurological data suffer from inter-subject variability and sometimes intra-subject variability.

1.2 Problem Statement and Our Approach

The goal of this dissertation is to develop methods to analyze EEG signals at the single-trial level. Single-trial analysis of this multi-sensor modality is often formulated as a binary classification problem. The observation to be classified is a segment of EEG signal, and the classes correspond to two mental states that are assume to cause this observation. The underling assumption is that differences in mental states reflect themselves in the EEG signals.

From a mathematical point of view single-trial EEG observation can be expressed by a matrix $\mathbf{X} \in \mathbb{R}^{D \times T}$ where D denotes the number of sensors and T the number of temporal samples of the EEG. Multiple trials are then represented as a set of such matrices. Each such trial in the set is associated with some underlining mental state(e.g. right or left hand imaginary movement,

stimulus versus control conditions, etc.); that state constitutes the label for that EEG segment. Having defined the transition form EEG to the proper mathematical structure we can formalize the single trial classification of EEG as follow.

Problem Statement: Single-trial EEG Classification Given a set of training examples $\mathcal{D} = \{\mathbf{X}_n, y_n\}_{n=1}^{N}, \mathbf{X} \in \mathbb{R}^{D \times T}, y \in \{-1, 1\}$, where $\mathbf{X}_n$ corresponds to the EEG signal of D channels and T sample points and y_n indicates one of two conditions or classes, the task is to predict the class label y for an new trial $\mathbf{X} \notin \mathcal{D}$.

Our approach for addressing this problem is to exploit the spatial, temporal and spectra; structure of the EEG signals. EEG is a multi sensor modality, with each sensor having a spatial relation (proximity) with neighboring sensors. This imposes some correlation in the measurements. Similarly, as a *temporal* signal there are correlations over time, while neurophysiological constrains restrict the *spectral* profile of the signals. This structure implies that the these extremely high dimensional data do not spread the entire $D \times T$ dimensional space but rather lie in smaller area within that space. We will utilize concepts from bilinear algebra to take advantage of these regularities in the data - thus the title of this thesis: The Bilinear Brain.

The extremely low signal-to-noise ratio of the data mandates the use of domain knowledge. This corresponds to information about the characteristics of the brain signals as well as the anatomical and functional structure of the brain. This knowledge is accumulated through various neuro-imaging experiments. Some of this knowledge is generic and applies across many cognitive task experiments while others can more specific to certain tasks. In our approach we exploit this information by selecting specific models and by allowing the experimentalist to introduce prior probabilities and selecting appropriate modeling parameters .

The inter-sessions and inter-subject variability in the EEG signals restrict the analysis on EEG data obtain from a single session. This is a problem for most methods in EEG analysis since it restricts the number of samples that one can obtain for training a given model. In our approach, we alleviate this problem by distinguishing between subject-specific and subject invariant parameters. This us to utilize data from multiple session to learn the subject-invariant part of the model, while fine tuning the subject-specific part using single-session data.

1.3 Overview and Contributions of the Thesis

1.3.1 Overview of the thesis

This thesis is organized as follow. Chapter 2 provides a brief overview of the principles of EEG, methodology to analyze the data and the neurological significance and challenges of EEG analysis. In Chapter 3 we review current research advances in EEG single-trial classification and introduces the notion of spatio-temporal 'projection' of EEG and the the Bilinear Discriminant Component Analysis (BDCA) algorithm. Chapter 4 illustrates a new Second Order Bilinear Discriminant Analysis (SOBDA) framework as an unified model for EEG analysis. Chapter 5 presents the Bilinear Feature Based Discriminant (BFBD) to address inter-subject variability in EEG. Chapter 6, summarizes the achievements of this research and identifies possible improvements as well as future direction in related research.

1.3.2 Novel contributions of the thesis

Our purpose is to develop new algorithms for the single-trial analysis of EEG signals. The novel contributions of this research are:

1. **Spatio-Spectro-Temporal single-trial classification -** In this dissertation we introduce discriminant models that combine spatial, spectral and temporal information in EEG analysis. As already pointed out in the previous section, the multi-sensor nature of EEG signals define the spatial structure of the modality. The time signals, by definition, have a temporal structure, while the underlining rhythmic activation of neurons determines the spectral characteristics of EEG. Out method exploits these regularities to explore the part of the observation space that concentrates most of the information of interest while ignoring much of the space that is likely to be spanned by noise. We apply concepts from bilinear algebra in order exploit this structure in EEG. This is the first time in EEG analysis that a single model utilizes all three aspects of EEG structure in an optimization framework.

2. **Phase-locked and oscillatory feature combination -** We present a unified approach to classifying single-trial EEG data based on a combination of first-order(phase-locked) and second-order (oscillatory) features, both of which may be informative in real data-sets in areas such as Brain Computer Interfacing. Our model essentially uses the Event Related Potentials (ERP) and Event Related (De)-synchronization (ERS/ERD) characteristics of EEG, by formulating a convex combination of the two. Rather than committing a priori to one or the other feature, our method extracts these features directly from the data and combines them in an optimal way to maximize the discriminability of the two classes. To the best of our knowledge this is the first method that combines these to major features of EEG analysis in a single discriminant model.

3. **Generalized EEG analysis framework -** We formulate a generic framework to single-trial EEG analysis. We show that a number of different previously proposed approaches in single-trial EEG analysis are special cases of the current framework. Specifically, fixing some of the parameters of our model corresponds to introducing the various assumption of different methods. This enables us to gain a deeper understanding of the underlining assumptions of the various methods. Our method not only provides this unified view of EEG analysis but also suggests ways of future research and new approaches. More than a neat mathematical formulation, our work provides the conceptual framework by which EEG analysis can be done. A number of new ideas for EEG analysis result from this formulation.

4. **Inter-subject EEG analysis -** In this work we consider the problem of inter-subject and inter-session variability of EEG. Different sessions of EEG recordings vary dramatically even for the same subject. This variation is caused by the experimental setup procedure, variation in electrode position and conductivity, the anatomical structure of the brain, and the individual's alertness at the time of the experiment, to name a few. Because of this variability, one is restricted in applying single-trial EEG analysis separately to data obtained from each session. This restricts the training set sample in a classification model, since there are is a limit on the number of trials that can be performed by a subject in a single session. . In this work we address this inter-subject and inter-session variability by modeling both session-invariant and session-specific characteristics of EEG. By incorporating in a single model both of these characteristics we can use recordings of EEG from multiple experiments and multiple subjects to train the subject/session invariant part of our model. Then we can use data obtained from individual session to fine tune the subject/session specific components

of our model.

Parts of this dissertation are based on work published in [10],[9],[17] and [31]

Chapter 2

SINGLE-TRIAL ANALYSIS OF EEG

2.1 Overview of EEG data

The electroencephalogram measures the difference in voltage between two electrodes on the scalp; these voltage differences are typically measured between each recording electrode and one or more reference electrodes. Recording electrodes are often arranged on the scalp according to the 10-20 system of electrode placements figure (2.2). Population of neurons firing in synchrony create electrical currents causing potential differences in the encephalogram. The exact underlining physics of EEG is outside the scope of this dissertation; the interested reader is referenced to [28]. For our purposes we consider EEG as a non-invasive modality reflecting brain activity associated with various perceptual and cognitive tasks.

A typical EEG experiment would involves a subject performing some task that could trigger a particular mental state in the brain. Depending on the experiment the subject might be asked to attend to a visual or auditory stimulus or to perform some mental task for some fraction of time following a visual /auditory cue. By engaging to the task, the subject activates neurons in some areas of the brain. this neural activity is reflected in the EEG. The signals are recorded starting from the time of the cue and for some fraction later while the subject is performing the task. This constitutes a single EEG trial. The temporal length of a single trial depends on the experiment and could vary from 500ms to 4000ms. Each such trial is labeled according to the task the subject was performing at that time. Usually in our experiments multiple trials are recorded for two different

Figure 2.1: Segment of continues EEG signals. A total of 32 channels is shown for a time window of 6 seconds. The EEG between the purple and red vertical lines, indicate a single-trial segment for one of the underlying mental states. The EEG signals between the purple and green vertical lines is another trial that corresponds to a different mental state.

tasks. The goal of computational analysis is to be able to infer the underlining task of the subject solely by looking at the EEG signals. A segment of continuously recorded EEG signals from 32 electrodes is shown in figure (2.1). The beginning of each trial is noted in the figure by a vertical line.

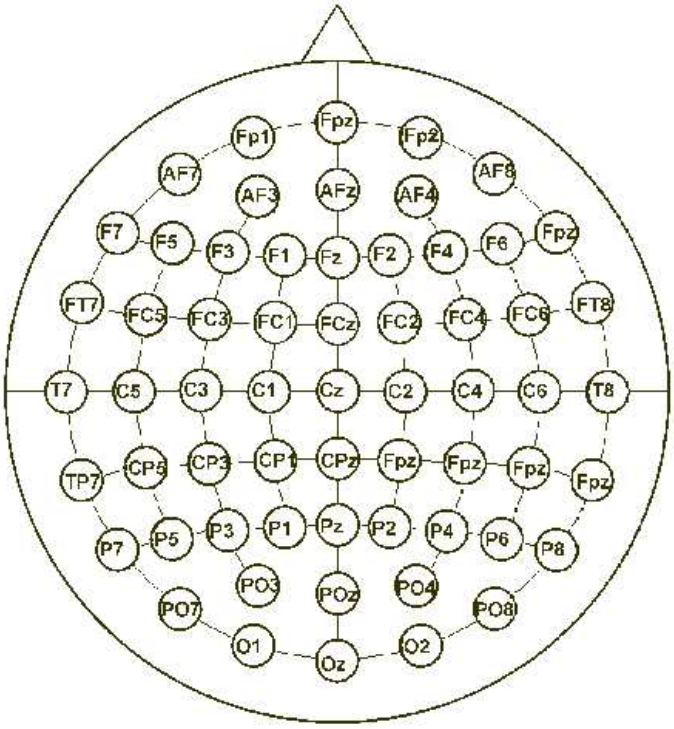

Figure 2.2: 10-20 international electrode placement system.

2.2 Features of EEG signals

There are two major groups of features that are used in traditional EEG analysis: Event Related Potentials (ERP) and Event Related Synchronization/Desynchronization (ERS/ERD). Since these two concepts will play an important role in motivating the algorithmic development of this thesis we introduce them is some detail.

2.2.1 Event Related Potentials.

Event Related Potentials are the measured EEG signals that are evoked by observable events (e.g., stimulus presentation, behavioral response) and reproduced from one trial to the next with precise timing. They are generated by large populations of neurons that fire time-locked to some event. Traditionally, ERPs are obtained by averaging of event-locked EEG over hundreds of trails. The waveform of ERPs is usually composed of one or more peaks and is labeled by the latency and polarity of those peaks. An ERP with a positive peak at 100ms is called 'P100', a negative peak at 200ms is called 'N200'. What constitutes a distinct ERP remains a controversial issue, even though several such waveforms have been named and associated with specific mental tasks. Below we briefly mention some of the ERPs that we will refer to later in the thesis.

- The P300 component of the EEG is a positive potential that occurs in the context of the "oddball paradigms," where a series of standard stimuli is randomly interleaved with non-standard stimuli, termed "deviants." After the presentation of each deviant, P300 appears as a large positive potential, which occurs with a relative latency of about 300 ms to the stimulus. It has been called "Attention reorienting response" as it elicited when ever an event captures a subjects attention.

- The N170 component, also known as the face-selective component, is a negative peak at 170 ms after a brief presentation of an image containing a face and it is not evoked when presented with non-face objects.

- Lateralized Readiness Potential (LRP) also known as "Bereitschaftpotential" is a slow negative potential that builds up during the preparation of hand movements, such as keystrokes with the fingers.

2.2.2 Event Related Synchronization/Desynchonization.

Some of the event-related responses of the EEG consist of either increase or decrease in the power of this oscillatory activity. This is due to the increase or decrease of the synchrony of population of neurons firing. These phenomena are called "Event Related Synchronization" and "Event Related De-synchronization". They are observed during the execution of a variety of mental states and mental tasks such as sensory and semantic processing, memory and movement tasks.

EEG has been decomposed into a series of fixed broad spectral bands based mainly on historical reasons. These include the delta band (0.5 Hz- 4Hz), theta band (4Hz - 8Hz), alpha band (8Hz - 13Hz), beta band (13Hz-30Hz) and gamma band (>30Hz). We briefly mention some of the more prominent ERS/ERD phenomena that are associated with the data we are using for the evaluation of the algorithms presented in this thesis.

The alpha band (8Hz - 13Hz) is a very strong signal in the parietal and occipital region of the cortex that is known to be modulated by sensory processing and changing attentional state. When a subject is not attending a visual stimulus or is disengaged from a visual task then the power of the alpha band increases. A similar signal, referred to as the μ-rhythm over motor areas is

attenuated during the execution or during imagined hand or foot movements. The μ-rhythm is in the same frequency band as the alpha band however the spatial distribution of the ERD for motor tasks is centered at the corresponding motor cortices. While one imagines a left hand movement the μ-rhythm over the right motor cortex decreases. Similarly, while thinking about a right hand movement the μ-rhythm over the left motor cortex is decreased.

2.3 Single-trial classification of EEG - Importance for Neuroscience

Single-trial classification of EEG signals finds applicability in the area of Brain Computer Interfaces as well as identifying the relationship between neuronal activity and psychophysical judgment. Below we will review some of these applications.

From the BCI perspective the goal is to decode brain activity as reliably and as fast as possible. Specifically, the interface must identify neurophysiological activity that is associated with a subject's choices or decisions. These decision signals can then be used to control applications such as text input programs, wheelchairs or neuroprosthesis. At the core of BCI systems usually lies a binary classification algorithm that operates on single trials EEG signals. The classifier is trained such that it discriminates between two mental states as recording by EEG.

The relationship between neuronal activity and behavioral choices has long been of great interest to systems neuroscientist. This association is usually achieved by comparing the so called psychometric and neurometric functions. The psychometric function captures the behavioral performance of a subject during an experiment against a 'difficulty level' variable. For example in a visual task of identifying faces vs car in an image, the 'difficulty level' variable can be the percentage of noise in the stimuli, while the behavioral performance is shown the by percentage of correctly identified images. Neurometric function on the other hand measures the performance of a classifier applied on

some neurological data, against the same 'difficulty' variable. The ground truth is considered the subjects response on a give stimulus rather than the true stimulus. How well the neurometric functions follow the psychometric function is what determines the relationship of the neural activity and the mental task. In experiments involving monkeys, the neural signal used is usually the variability in the firing rate in specific neurons. To obtain these measurements, electrodes need to be surgically placed into the monkey's brain. For experiments that involve humans however, one needs to rely on non-invasive neural modalities such as EEG. Hence, single-trial classification of EEG signals, with various stimulus noise levels can be used with neuromatric/psychometric function comparison method to directly associate neural activity with behavioral performance[35].

2.4 Computational Challenges of EEG

Single-trial classification of EEG signals is a challenging task. EEG data lies in a high dimensional space. A few seconds of recording with 64 to 128 sensors, at a sampling rate of 1000Hz. This results in observations of few hundred thousands dimensions. Learning regularities in the data in such high dimensions is hard because of the *curse of dimensionality*.

Further, the signal to noise ratio in the EEG signals is extremely low, often -20dB or less and the number of samples available to be used in learning a classifier is usually small - typically 50-100 samples per class.

It is well known than EEG data suffer from inter-subject variability and even intra subject variability in certain cases. The spatial properties of brain signals depend on the anatomical structure of the brain of each individual. Everyone has a unique brain. Hence, spatial regularities in the EEG signal are subject-specific. The inter-subject variability has been noted extensively in the literature [30]. Further, due to variation in experimental conditions and equipment an

inter-session variability has been noted. This variation is usually due to external factors such as electrode/sensor position and conductivity. Finally, in certain applications the variability of the regularity itself changes during a session as the distribution of the data changes from training to testing session [39], for instance, due to fatigue an loss of attention. All these challenges need to be addressed in order to have a successful single-trial EEG classifier.

Chapter 3

BACKGROUND

As it was motivated in the previous chapter, both in the context of Brain Computer Interfaces as well as in the context of EEG analysis, it is of great interest to be able to employ a single-trial classification of EEG signals. Applying an off-the-shelf classifier on the EEG signals may not be sufficient because of the extremely noisy observations and very large dimensions of the signal. So it is critical to employ domain knowledge when one formulates the EEG classification problem. This domain knowledge traditionally is employed by following a two step process: a signal processing/feature extraction step, and the actual classification applied on the resulting feature space. For both steps a large variety of methods have been proposed and each method is tightly coupled with the exact experimental paradigms for which they were designed. In this section we will first review some of those signal processing and feature extraction methods as they have been applied in different paradigms of EEG analysis. Then we will review classification methods we employed.

3.1 Signal Processing and Feature Extraction

It is difficult for classification algorithms to extract relevant information if the dimensionality of the data is large compared to the number of available samples. This problem is referred to as the "Curse of Dimensionality" in the Pattern Recognition field. By reducing the dimensions of the space in a 'sensible' way i.e. by eliminating dimensions with little or no Discriminant information and

retaining dimensions that capture discriminant information, the problem becomes more tractable. Hence dimensionality reduction, and feature extraction is an extremely important step for most of the current analysis methods. We will focus on the most commonly used methods for feature extraction in the content of EEG.

3.1.1 Spectral Filtering

It is often the case in EEG analysis that neurophysiological models associated with an experimental paradigm suggest that the signal of interest is mainly located at specific frequency bands. In section 2.2, we noted one such example. A common approach for capturing such information is to use digital frequency filters. Based on the desired frequency band one designs two sequences $\mathbf{a}$ and $\mathbf{b}$ with length N_a and N_b for the frequency band of interest. These sequences can be obtained using the filter design techniques of Butterworth or Chebyshev. The filtered signal $\mathbf{x}_{filter}$ is obtained from the original signal $\mathbf{x}$ as

$$\mathbf{a}(1)\mathbf{x}_{filter}(t) = \sum_{i=1}^{N_b} \mathbf{b}(i)\mathbf{x}(t-i-1) - \sum_{i=2}^{N_a} \mathbf{a}(i)\mathbf{x}_{filter}(t-i+1) \tag{3.1}$$

An alternative to time-domain filter, is frequency domain filters. Using Fast Fourier Transform (FFT) one can transform a temporal signal to its frequency domain. Filtering is achieved by selecting the frequency bands of interest and applying the inverse Fourier transform to obtain back the temporal signal band-limited to the selected frequency band.

The Fourier transform projects the signal to a set of orthogonal bases that represent sinusoids in various frequencies. Alternatively, different families of basis function have been used in the EEG analysis context: The wavelet basis, for example. By scaling and translating a prototypical basis function designed to capture a specific frequency band and time window the original signal can be

filtered obtaining time-frequency decomposition. All three methods have been considered in various paradigms in EEG analysis.

3.1.2 Spatial Filtering

Let $\mathbf{X} \in \mathbb{R}^{D \times T}$ denote an EEG observation. A spatially filtered signal $\mathbf{x}_{filter} \in \mathbb{R}^{T}$ is defined by a projection vector $\mathbf{u} \in \mathbb{R}^{D}$ as

$$\mathbf{x}_{filter} = \mathbf{u}^{\top}\mathbf{X} \tag{3.2}$$

The characteristics of the spatial filter will depend on what assumption one enforces on the spatial projection $\mathbf{u}$. The motivation for spatial filtering in EEG is based on the fact that electrodes measure the superposition of signals generated from various sources in the brain. Simulation studies [29], shows that only half of the contribution in one electrode come from sources within a 3cm radius. This makes it difficult to capture local activity from an electrode especially if the signal of interest is weak, thus contaminated by non-local sources. By properly selecting or estimating the vector $\mathbf{u}$ spatial filtering techniques produce localized signals.

One approach used for defining spatial filters is to manually design the vector $\mathbf{u}$ based on the required properties of specific applications. Some of these methods include *Bipolar Filters, Laplace Filters, Common Average Reference.* Bipolar filters rely on the assumption that neighboring electrodes measure non-local activity of equal intensity. Hence re-referencing each electrode to the nearest neighboring electrode cancels out non-local EEG activity. In the spatial filtering framework, bipolar filtering of an electrode i whose nearest electrode is j amounts to a projection vector $\mathbf{u}$

defined by

$$\mathbf{u}(k) = \begin{cases} 1, & k = i; \\ -1, & k = j; \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

The Laplace filter is again based on the same idea as the bipolar filters, aiming to remove signal content which does not originate near the recording electrode. The Laplace filter subtracts the average signal of surrounding electrodes. Let I_i be an index set of neighboring electrodes to electrode i. The laplace filter coefficients are defined as:

$$\mathbf{u}(k) = \begin{cases} 1, & k = i; \\ -1/|I_i|, & k \in I_i; \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

Another approach for spatial filter estimation is data driven. In EEG analysis three primary methods have been used, namely Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Common Spatial Pattern (CSP). We will briefly motivate the first two methods and discuss in greater detail the Common Spatial Patters, since it is a method with which our method.

The motivation for the use of PCA in EEG is based on the assumption that directions with the highest variance contain most of the information in the EEG signal. By considering every point in time (columns of the EEG matrix $\mathbf{X} \in \mathbb{R}^{D \times T}$) as an observation in the electrode space, PCA extracts a set of orthonormal vectors that span a subspace that captures most of the variance in the data. A set of spatial filter vectors $\mathbf{u}_1, .., \mathbf{u}_k$ are then defined as the first k principal components of the resulting subspace. On the other hand, ICA is motivated by the assumption that brain activity of interest comes from independent sources in the brain that are spatially separated. Independent

Component Analysis thus finds a set of spatial filters such that the resulting filtered signals are as independent as possible in the statistical sense. Unlike PCA, the solution of the ICA problem is not unique and it changes based on the criterion used for statistical independence and the algorithm implementation.

Both PCA and ICA are mostly used as artifact removal and outlier detection techniques. For example PCA has been used as ocular artifact removal method, by identifying components with highest variance during eye-blink calibration session. Similarly, ICA has been used to extract independent components during eye movement calibration session. The ICA resulting components are visually inspected, identified as artifacts and removed from the signal.

The Common Spatial Patters (CSP) algorithm is a supervised technique that learns the best spatial filters in terms of their discriminant ability. It has been proven useful in identifying ERD/ERS effects. Given two distributions in a high-dimensional space, corresponding to two classes, CSP seeks projection vectors that maximize the variance of one class while simultaneously minimizing the variance of the other class in the projected space. By filtering the EEG signals to a frequency band of interest, the high variance indicates strong rhythmic activity in that band while low variance indicates a weak rhythmic activity.

Let $\mathbf{\Sigma}^+, \mathbf{\Sigma}^- \in \mathbb{R}^{D\times D}$ be the estimates of the covariance matrix of the band-pass filtered EEG signal for class + and − respectively.

$$\mathbf{\Sigma^c} = \frac{1}{|I_c|}\sum_{i\in I_c}\mathbf{X}_i\mathbf{X}_i^\top, (c \in \{+,-\}) \tag{3.5}$$

where I_c is the index set corresponding to trials belonging to each class and $|I_c|$ the cardinality of the set. In the above expression $\mathbf{X}$ is filtered to the frequency band of interest, tis mean across

time subtracted and scaled by $\sqrt{T}$. Further define the $\mathbf{S}_d$ and $\mathbf{S}_c$ as follow:

$$\begin{aligned} \mathbf{S}_d &= \mathbf{\Sigma}^+ - \mathbf{\Sigma}^- \\ \mathbf{S}_c &= \mathbf{\Sigma}^+ + \mathbf{\Sigma}^- \end{aligned}$$

where $\mathbf{S}_d$ corresponds to the discriminant activity, i.e., the power difference in the selected frequency band between the two conditions while $\mathbf{S}_c$ correspond to the common activity between the two conditions that provide no information for discriminating the two classes. Then the CSP solution is obtained by solving the following optimization problem.

$$\arg_{\mathbf{u}} \max \frac{\mathbf{u}^\top \mathbf{S}_d \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_c \mathbf{u}} \tag{3.6}$$

Clearly the optimization above maximizes the difference in variance between the two classes in the projected space and minimizes the total variance.

Note that the objective function 3.6 is invariant w.r.t scaling of the vector $\mathbf{u}$, since

$$\frac{\mathbf{u}^\top \mathbf{S}_d \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_c \mathbf{u}} = \frac{(\alpha \mathbf{u})^\top \mathbf{S}_d (\mathbf{u}\alpha)}{(\alpha \mathbf{u})^\top \mathbf{S}_c (\mathbf{u}\alpha)} = \frac{\hat{\mathbf{u}}^\top \mathbf{S}_d \hat{\mathbf{u}}}{\hat{\mathbf{u}}^\top \mathbf{S}_c \hat{\mathbf{u}}} \tag{3.7}$$

Hence we can always choose $\mathbf{u}$ such that the denominator of $\mathbf{u}^\top \mathbf{S}_c \mathbf{u} = 1$. This enables us to write the maximization problem as a mathematical program

$$\arg_{\mathbf{u}} \min -\frac{1}{2} \mathbf{u}^\top \mathbf{S}_d \mathbf{u} \tag{3.8}$$

subject to the constraint

$$\mathbf{u}^\top \mathbf{S}_c \mathbf{u} = 1$$

A Lagrangian multiplier can be used in optimizing (3.8). We write the corresponding Lagrangian as

$$L(\mathbf{u}, \lambda) = -\frac{1}{2}\mathbf{u}^\top \mathbf{S}_d \mathbf{u} + \frac{1}{2}\lambda(\mathbf{u}^\top \mathbf{S}_c \mathbf{u} - 1) \tag{3.9}$$

Taking the derivative with respect to $\mathbf{u}$ and enforcing the Karush-Kuhn-Tucker (KKT) [8] conditions one obtains

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{u}} &= -\mathbf{S}_d \mathbf{u} + \lambda \mathbf{S}_c \mathbf{u} = 0 \\ \frac{\partial L}{\partial \lambda} &= (\mathbf{u}^\top \mathbf{S}_c \mathbf{u} - 1) = 0 \end{aligned}$$

This results in a generalized eigenvalue problem of the form

$$\mathbf{S}_d \mathbf{u} = \lambda \mathbf{S}_c \mathbf{u} \Rightarrow \mathbf{S}_c^{-1} \mathbf{S}_d \mathbf{u} = \lambda \mathbf{u} \tag{3.10}$$

Let $\hat{\mathbf{u}}$ and $\hat{\lambda}$ denote the solution of the eigenvalue problem (3.10) once plugged into the objective function (3.12) we obtain,

$$\frac{\hat{\mathbf{u}}^\top \mathbf{S}_d \hat{\mathbf{u}}}{\hat{\mathbf{u}}^\top \mathbf{S}_c \hat{\mathbf{u}}} = \frac{\hat{\mathbf{u}}^\top \mathbf{S}_c \hat{\lambda} \hat{\mathbf{u}}}{\hat{\mathbf{u}}^\top \mathbf{S}_c \hat{\mathbf{u}}} = \hat{\lambda} \tag{3.11}$$

Hence the optimal vector $\hat{\mathbf{u}}$, that maximizes the 3.6 is the largest eigenvalue of the generalized eigenvalue problem in (3.10).

3.2 Classification Methods

Several classification algorithms have been proposed to address the question of single-trial EEG classification problem. In this section we review some of those algorithms as they have been used in the context of EEG analysis and Brain Computer Interfaces.

3.2.1 Fisher Discriminant Analysis

The Fisher Linear Discriminant has long been proposed as a supervised method for dimensionality reduction and classification [14]. It searches for a subspace such that the projected data on that subspace is well separated. The objective of the Fisher Linear Discriminant is to find the subspace projection such that the class-means are further apart from one another when measured relative to the sum of the variances of the data assigned to a particular class. Formally, Fisher Discriminant Analysis estimates a projection vector $\mathbf{w}$ that maximizes the following objective function know as the Rayleigh coefficient:

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}} \tag{3.12}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ refers to the *between class scatter* and *within class scatter* defined as

$$\mathbf{S}_B = \sum_{c \in \{-1,1\}} N_c (\bar{\mathbf{x}}_c - \bar{\mathbf{x}})(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})^\top \tag{3.13}$$

$$\tag{3.14}$$

$$\mathbf{S}_W = \sum_{c \in \{-1,1\}} \sum_{n|y_n=c} (\mathbf{x}_n - \bar{\mathbf{x}}_c)(\mathbf{x}_n - \bar{\mathbf{x}}_c)^\top \tag{3.15}$$

where C denotes the set of all classes, N_c denotes the number of points that belong to class c, $\bar{\mathbf{x}}_c$ is the mean of class c defined as $1/N_c \sum_{\{n|y_n=c\}} \mathbf{x}_n$, and $\bar{\mathbf{x}}$ is the total mean of the data $1/N \sum_{\mathbf{x} \in \mathbf{X}} \mathbf{x}$.

3.2.2 Logistic Regression

Logistic regression is a well studied statistical model for classification[7][30][11]. It is based on modeling the *log-odds* of the probability of a class y, given the observation $\mathbf{x}$ as a linear function

of the parameters. This defines a linear classification boundary in the space. Formally, logistic regression models

$$log\left(\frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})}\right) = \mathbf{w}^\top \mathbf{x} + w_0 \tag{3.16}$$

Further, y is Bernoulli distributed, hence $E[y|\mathbf{x}] = P(y=1|\mathbf{x})$, which gives rise to a model of the $E[y|\mathbf{x}]$

$$E[y|\mathbf{x}] = P(y=1|\mathbf{x}) = \pi(\mathbf{x}) = \frac{1}{1+e^{-(\mathbf{w}^\top \mathbf{x}+w_0)}} \tag{3.17}$$

Since $y|\mathbf{x}$ is a binary random variable it follows a Bernoulli distribution, hence the probability of y is given by

$$P(y|\mathbf{x}) = \pi(\mathbf{x})^y(1-\pi(\mathbf{x}))^{1-y} \tag{3.18}$$

Note: In this section we assume that $y \in \{0,1\}$ instead of $\{-1,1\}$, since this is the traditional way logistic regression is presented. In the case $y \in \{-1,+1\}$ labeling notation one would define $p(y|\mathbf{x}) = \frac{1}{1+e^{-y(\mathbf{w}^\top \mathbf{x}+w_0)}}$. So given a new observation $\mathbf{x}_{new}$ under the logistic regression model, the predicted label is given by

$$y = f(\mathbf{x}_{new}) = \begin{cases} 1, & P(y|\mathbf{x}_{new}) > 0.5); \\ 0, & \text{otherwise.} \end{cases} \tag{3.19}$$

Hence, under the logistic regression model, given the parameters $\mathbf{w}$ and w_0 and an input $\mathbf{x}$ we can make a prediction about y.

The likelihood of the model is defined as

$$L(\mathbf{w}, w_0; X, \mathbf{y}) = p(\mathbf{y}|\mathbf{w}, w_0, X) \tag{3.20}$$

where $\mathbf{y} = [y_1, ..., y_N]^\top$ and X is the *design matrix* defined as $X = [\mathbf{x}_1, ..., \mathbf{x}_N]$.

Under the assumption that y_n is drawn $i.i.d$ from the Bernoulli distribution, then the likelihood of the model can be written as:

$$\begin{aligned} L(\mathbf{w}, w_0; X, \mathbf{y}) &= \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \mathbf{w}, w_0) \\ &= \prod_{n=1}^{N} \pi(\mathbf{x}_n)^{y_n}(1-\pi(\mathbf{x}_n))^{1-y_n} \\ &= \prod_{n=1}^{N} \pi(\mathbf{x}_n)^{y_n}(1-\pi(\mathbf{x}_n))(1-\pi(\mathbf{x}_n))^{-y_n} \\ &= \prod_{n=1}^{N} \frac{\pi(\mathbf{x}_n)^{y_n}}{(1-\pi(\mathbf{x}_n))^{y_n}}(1-\pi(\mathbf{x}_n)) \\ &= \prod_{n=1}^{N} e^{(\mathbf{w}^T\mathbf{x}_n+w_0)y_n}(1-\pi(\mathbf{x}_n)) \\ &= \prod_{n=1}^{N} e^{(\mathbf{w}^\top\mathbf{x}_n+w_0)y_n}(1+e^{\mathbf{w}^\top\mathbf{x}_n+w_0})^{-1} \end{aligned}$$

The log likelihood is then given as:

$$log(L(\mathbf{w}, w_0; X, \mathbf{y})) = l(\mathbf{w}, w_0; X, \mathbf{y}) = \sum_{n=1}^{N}(\mathbf{w}^\top\mathbf{x}_n + w_0)y_n - log(1 + e^{\mathbf{w}^\top\mathbf{x}_n+w_0}) \tag{3.21}$$

The *Gradient* and *Hessian* of the log likelihood can be obtained analytically and are determined by the equations

$$\frac{\partial l(\mathbf{w}, w_0; X, \mathbf{y})}{\partial w_0} = \sum_{n=1}^{N}(y_n - \pi(\mathbf{x}_n)) \tag{3.22}$$

$$\frac{\partial l(\mathbf{w}, w_0; X, \mathbf{y})}{\partial \mathbf{w}^\top} = \sum_{n=1}^{N}\mathbf{x}_n(y_n - \pi(\mathbf{x}_n)) \tag{3.23}$$

$$\frac{\partial^2 l(\mathbf{w}, w_0; X, \mathbf{y})}{\partial w_0 \partial w_0} = -\sum_{n=1}^{N}\pi(\mathbf{x}_n)(1-\pi(\mathbf{x}_n)) \tag{3.24}$$

$$\frac{\partial^2 l(\mathbf{w}, w_0; X, \mathbf{y})}{\partial \mathbf{w} \partial \mathbf{w}^\top} = -\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top \pi(\mathbf{x}_n)(1 - \pi(\mathbf{x}_n)) \tag{3.25}$$

$$\frac{\partial^2 l(\mathbf{w}, w_0; X, \mathbf{y})}{\partial w_0 \partial \mathbf{w}^\top} = -\sum_{n=1}^{N} \mathbf{x}_n \pi(\mathbf{x}_n)(1 - \pi(\mathbf{x}_n)) \tag{3.26}$$

The maximum likelihood estimation of the parameters $\mathbf{w}$ can be achieved using the 'Damped Newton' iterations method or the *Iterative Re-weighted Least Squares* (IRLS) [3].

However, when the training data is separable, the function $l(\mathbf{w})$ can be made arbitrarily large, so a prior distribution on $\mathbf{w}$ is crucial. This motivates the adoption of a maximum a posteriori (MAP) estimate with $p(\mathbf{w})$ being the prior distribution over $\mathbf{w}$. Then one optimizes the log posterior

$$\hat{\mathbf{w}}, \hat{w}_0 = \arg\max_{\mathbf{w}, w_0} [l(\mathbf{w}, w_0; X, \mathbf{y}) + log(p(\mathbf{w}))] \tag{3.27}$$

A Gaussian prior is usually used, in which case the 'Damped Newton' optimization or *iterative re-weighted least squares* methods can be used as before. A Laplacian prior which promotes sparse solutions can be used instead [23]. LASSO or LARS algorithm [18] can be used to solve the resulting optimization. Logistic regression has a natural extension to the multi class case by assuming a multinomial distribution of the class variable y [23].

3.3 EEG Methods to Compare

In the previous section we provided generic overview of a number of feature selection and classification methods that form the building blocks for most EEG analysis techniques. In this section we introduce the algorithms that will serve as benchmarks for comparison with our algorithm. These methods are considered to be the state-of-the-art in single-trial EEG analysis as they are used in their corresponding fields. Specifically, we will review three algorithms. First we present

Hierarchical Discriminant Component Analysis (HDCA) [20] and Bilinear Discriminant Component Analysis (BDCA)[17]. These two methods are used in EEG single-trial classification for Event Related Potential based experiments, such as Rapid Serial Visual Presentation (RSVP). We then present Matrix Logistic Regression (MLR)[40], which was proposed specifically for experiments that involve imaginary movements. These three methods along with the CSP algorithm presented in the previous chapter will be our performance baseline for the work presented in this dissertation.

3.3.1 Hierarchical Discriminant Component Analysis

The HDCA method combines Fisher Discriminant and Logistic Regression in a hierarchical order for single-trial classification of EEG [31, 20]. The algorithm was proposed as a way to combine both spatial and temporal information in EEG experiments in which ERP features are informative. It has been used successfully in analyzing signals obtained by the Rapid-Serial-Visual Presentation (RSVP) experiment [20], that is known to cause the P300 signal (i.e positivity in amplitude of the signal at around 300ms, in electrodes located over the visual cortex).

According the algorithm each EEG observation $\mathbf{X} \in \mathbb{R}^{D \times T}$, corresponding to half second of EEG signal, is segmented into ten non-overlapping temporal windows of 50ms $< \mathbf{X}_1, ..., \mathbf{X}_{10} >$. The underlining assumption is that the signals in each window are now stationary, i.e., the mean and covariance does not change over time. Under this assumption, the method considers the mean vector across time as an estimate of the signals amplitude in each window, thus transforming the entire observation in that window to a D-dimensional vector. Each observation $\mathbf{X} \in \mathbb{R}^{D \times T}$ is now represented by a tuple of ten vectors $< \mathbf{x}_1, ..., \mathbf{x}_{10} >$ where $\mathbf{x}_i \in \mathbb{R}^D$.

Given a training set of N observations from two classes, Fisher Discriminant is applied indepen-

dently to each window, where each temporal sample is treated as an observation. This results in ten different linear classifiers, one per window, that define the first level of the hierarchical classifier. Applying the first level classifiers to the input tuple of vectors $< \mathbf{x}_1, ..., \mathbf{x}_{10} >$ result in a tuple of scalars $\mathbf{t} =< x_1, ..., x_{10} >$ where the ith value is the classification output of the ith classifier. Each original observation has been reduces now to a 10-dimensional vector. The second level of the classifier involves training another linear model, namely logistic regression to project the resulting vector to a single scalar. This scalar is the final result of the classification.

By segmenting the signal into ten temporal windows, HDCA reduces the dimensions of the space dramatically from $D \times T$ to $10 \times D$. Further, using the hierarchical structure of multiple classifiers, it eliminates the problem of limited sample size, since for the comparatively small D-dimensional sample, samples sizes in the order of $2000 \times D$ are available. This arbitrary selection of the ten windows is also one of the limitations of the method, since it reduces the temporal resolution of the data, it misses any discriminant information embedded in-between the windows.

3.3.2 Bilinear Discriminant Component Analysis

In the early stages on my thesis work I contributed in the formulation of a new algorithm that first introduced the usages of multi-linear algebra in single-trial EEG analysis. We termed this algorithm Bilinear Discriminant Component Analysis (BDCA). Motivated by the need to utilize the spatio-temporal information in EEG and the ERP analysis features,we proposed an algorithm that uses a low rank representation of the data matrix [17].

Bilinear Discriminant Component Analysis defines the following linear discriminant model. For

a matrix $\mathbf{X}$, the log-odds ratio of class y is given by

$$\log \frac{P(y=+1|\mathbf{X})}{P(y=-1|\mathbf{X})} = \text{Trace}\, \mathbf{U}^\top \mathbf{X} \mathbf{V} \tag{3.28}$$

where $\mathbf{U} \in \mathbb{R}^{D\times K}, \mathbf{V} \in \mathbb{R}^{T\times K}$ are the model parameters, and K a user specified parameter (typical values: $K = 1$ or $K = 2$). The above linear discriminant model implies an expression for the probability of the class y given an observation $\mathbf{X}$ as follows:

$$P(y=+1|\mathbf{X}) = \frac{1}{1+e^{-y\,\text{Trace}\{\mathbf{U}^\top \mathbf{X}\mathbf{V}\}}} \tag{3.29}$$

The method proceeds in learning the parameter matrices $\mathbf{U}$ and $\mathbf{V}$ by maximizing the likelihood of the parameters given the data. To reduce the effect of overtraining, BDCA employ regularization techniques that enforce prior distribution on the parameters [17].

The main advantage of BDCA is that by enforcing a bilinear structure in obtaining a linear combination of the elements of $\mathbf{X}$, it reduces the dimension of the space from $D \times T$ to $K \times (D+T)$ without loosing any temporal resolution. As in the case of HDCA, BDCA has been used successfully, in RSVP experiments but fails in experiments involving non-phase lock information - i.e., movement imagination.

3.3.3 Matrix Logistic Regression

A Discriminative approach to capture ERD characteristics in EEG was proposed in [40]. We will refer to this method as Matrix Logistic Regression (MLR). Similar to the BDCA the authors of [40] model the log-odds of a class y given an observation $\mathbf{X}$ as a linear model. However unlike BDCA the linear model applies on the covariance of $\mathbf{X}$ rather than the raw observation. Their exact model is as follows:

$$\log \frac{P(y=+1|\mathbf{X})}{P(y=-1|\mathbf{X})} = \text{Trace}\,\{\mathbf{W}^\top \mathbf{X}\mathbf{X}^\top\} \tag{3.30}$$

where $\mathbf{W} \in \mathbb{R}^{D \times D}$ is the matrix of model parameters. The main difference between MLR and BDCA is that the former operates on the second order matrix $\mathbf{X}\mathbf{X}^{\top}$ rather than the raw EEG signals. By filtering the data $\mathbf{X}$ in some frequency of interest, MLR captures linear combination of the power of the signal in that frequency band. This corresponds to capturing ERD features in EEG trials.

The estimation procedure for the parameter matrix $\mathbf{W}$ involves maximizing the likelihood of the parameters, given some training data. Further, to reduce generalization error the method introduces a regularization term in the optimization that involves the matrix covariance [40].

Chapter 4

SECOND ORDER BILINEAR DISCRIMINANT ANALYSIS

4.1 Introduction

In EEG the signal-to-noise ratio (SNR) of individual channels is low, often at, or below -20dB. To overcome this limitation, all analysis methods perform some form of averaging, either across repeated trials, across time, or across electrodes. Traditional EEG analysis averages signals across many repeated trials for each individual electrode. Typical in this case is to average the measured potentials following stimulus presentation, thereby canceling uncorrelated noise that is not reproducible from one trial to the next. This averaged activity, called an event related potential (ERP), captures activity that is time-locked to the stimulus presentation but cancels induced oscillatory activity that is not locked in phase to the timing of the stimulus. Alternatively, many studies compute the oscillatory activity in specific frequency bands by filtering and squaring the signal before averaging. Induced changes in oscillatory activity are termed event related synchronization or desynchronization (ERS/ERD).

Surprisingly, Discriminant analysis methods developed thus far by the machine learning community have followed this dichotomy: First order methods in which the amplitude of the EEG signal is considered to be the feature of interest in classification – corresponding to ERP – and second order methods in which the power of the feature is considered to be of importance for classification – corresponding to ERS/ERD. First order methods include temporal filtering + thresholding [2], Fisher linear Discriminant [30, 4], hierarchical linear classifiers [20] and bilinear Discriminant anal-

ysis [17, 41]. Second order methods include logistic regression with a quadratic term [42] and the well known common spatial patterns method (CSP) [37] and its variants: common spatio-spectral patterns (CSSP)[24], and common sparse spectral spatial patterns (CSSSP)[12].

In the past the process for choosing features for classification has been *ad hoc* and driven primarily by prior knowledge and/or assumptions regarding the underlying neurophysiology and task. From a machine-learning point of view, it seems limiting to commit *a priori* to only one type of feature. Instead, it would be desirable for the analysis method to extract the relevant neurophysiological activity *de novo* with minimal prior expectations.

In this chapter we present a new framework that combines both first and second order features in the analysis of EEG. Through a bilinear formulation, the method can simultaneously identify spatial linear components as well as temporal modulation of activity. These spatio-temporal components are identified such that their first and second order statistics are maximally different between two conditions. Further, through the bilinear formulation, the method exploits the spatio-temporal nature of the EEG signals and provides a reduced parameterization of the high dimensional data space. We show that a broad set of state-of-the-art EEG analysis methods can be characterized as special cases under this bilinear framework. Simulated EEG data are then used to evaluate performance of the different methods under varying signal strengths. We conclude this paper with a performance comparison on human EEG. In all instances the performance of the present method is comparable or superior to the existing state-of-the-art methods.

4.2 Second Order Bilinear Disctiminant

To introduce the new method we start by formally defining the classification problem in EEG. We then present the bilinear model, discuss interpretation in the context of EEG, and establish a link to

current analysis methods. The section concludes with the optimization criterion and regularization approaches. As the title of this section suggests, we termed our method Second Order Bilinear Discriminant Analysis (SOBDA).

4.2.1 Problem setting

Suppose that we are given examples of brain activity as a set of trials $\{\mathbf{X}_n, y_n\}_{n=1}^N, \mathbf{X}_n \in \mathbb{R}^{D\times T}, y_n \in \{-1, 1\}$, where for each trial n, the matrix $\mathbf{X}_n$ corresponds to the EEG signal with D channels and T time samples while y_n indicates the class to which this trial corresponds. The class label can indicated one of two conditions, i.e. imagined right or left hand movement, stimulus or non-stimulus control, etc. Given these examples the task is then to predict the class label y for a new trial with data $\mathbf{X}$.

4.2.2 Second order bilinear Model

To solve this problem we propose the following discriminant function

$$f(\mathbf{X};\theta) = C\,\mathrm{Trace}(\mathbf{U}^\top\mathbf{X}\mathbf{V}) + (1-C)\,\mathrm{Trace}(\Lambda\mathbf{A}^\top(\mathbf{X}\mathbf{B})(\mathbf{X}\mathbf{B})^\top\mathbf{A}) + w_o\,, \tag{4.1}$$

where $\theta = \{\mathbf{U} \in \mathbb{R}^{D\times R}, \mathbf{V} \in \mathbb{R}^{T\times R}, \mathbf{A} \in \mathbb{R}^{D\times K}\,\mathbf{B} \in \mathbb{R}^{T\times T'}\}$, $w_o \in \mathbb{R}$ are the parameters and $\Lambda \in diag(K)|\lambda_{ii} \in \{-1, +1\}, 0 \leq C \leq 1$ is specified by the user. The goal will be to use the N examples to optimize these parameters such that the discriminant function takes on positive values for examples with $y_n = +1$ and negative values for $y_n = -1$. To accomplish this we will use a standard probabilistic formalism – logistic regression – which will permit us to incorporate regularization criteria as prior probabilities on the parameter as will be explained in sections 4.6 and 4.7.

4.3 Interpretation and rationale of the model

The discriminant criterion defined in (4.1) is the sum of a linear and a quadratic term, each combining bilinear components of the EEG signal. The first term can be interpreted as a spatio-temporal projection that captures the first order statistics of the signal. Specifically, the columns $\mathbf{u}_r$ of $\mathbf{U}$ represent R linear projections in space (rows of $\mathbf{X}$). Similarly, each of the R columns of $\mathbf{v}_r$ in matrix $\mathbf{V}$ represent linear projections in time (columns of $\mathbf{X}$). By re-writing the term as:

$$\text{Trace}(\mathbf{U}^\top \mathbf{X}\mathbf{V}) = \text{Trace}(\mathbf{V}\mathbf{U}^\top \mathbf{X}) = \text{Trace}(\mathbf{W}^\top \mathbf{X})\,, \tag{4.2}$$

where we defined, $\mathbf{W} = \mathbf{U}\mathbf{V}^\top$, it is easy to see that the bilinear projection is a linear combination of elements of $\mathbf{X}$ with a rank-R constraint on $\mathbf{W}$. This expression is linear in $\mathbf{X}$ and thus captures directly the amplitude of the signal. In particular, the polarity of the signal (positive evoked response versus negative evoked response) will contribute to discrimination if it is consistent across trials. This term, therefore, captures phase locked event related potential in the EEG signal. This bilinear projection reduces the number of model parameters of $\mathbf{W}$ from $D \times T$ dimensions to $R \times (D + T)$ which is a significant reduction in alleviating the problem of over-fitting. This projection assumes that the discriminant information in an observation can be captures using low-rank matrices. This holds true in EEG data, where an electrical current source which is spatially static in the brain will give a rank-one contribution to the spatio-temporal $\mathbf{X}$ [17].

The second term of equation (4.1) is the power of spatially and temporally weighted signals and thus captures the second order statistics of the signal. As before, each column of matrix $\mathbf{A}$ and $\mathbf{B}$ represent components that project the data in space and time respectively. Depending on the structure one enforces in matrix $\mathbf{B}$, different interpretations of the model can be achieved. In the

general case where no structure on $\mathbf{B}$ is assumed, the model captures a linear combination of the elements of a rank-T' second order matrix of the signal $\mathbf{XB}(\mathbf{XB})^\top$. In the case where Toeplitz structure is enforced on $\mathbf{B}$ (see Section 4.6.1), then $\mathbf{B}$ defines a temporal filter on the signal and the model captures powers of the filtered signal. Further, by allowing $\mathbf{B}$ to be learned from the data, we may be able to identify new frequency bands that have so far not been identified in novel experimental paradigms. The spatial weights $\mathbf{A}$ together with the Trace operation ensure that the power is measured, not in individual electrodes, but in some component space that may reflect activity distributed across several electrodes. The diagonal matrix $\mathbf{\Lambda}$ implements a linear combination of powers allowing thus to capture also differences in power between components. Finally the parameter C defines a convex combination of the first order term and the second order term. $C = 1$ indicates that the discriminant activity is dominated by the first order features, $C = 0$ indicates that the activity is dominated by second order features, and any value in between denotes the importance of one component vs the other.

4.4 Incorporating prior knowledge into the model

Realizing that the parameters of the SOBDA model have a physical meaning (i.e. $\mathbf{u}_r$ and $\mathbf{a}_r$ map the sensor signal to a current-source space, $\mathbf{v}_r$ are temporal weight on a source signal and $\mathbf{b}_r$ can be arranged to represent a temporal filter) it becomes intuitive for the experimenter to incorporate prior knowledge of an experimental setup in the model. If the signal of interest is known to be in a specific frequency band, one can fix matrix $\mathbf{B}$ to capture only the desired frequency band. For example, $\mathbf{B}$ can be fixed to a Toeplitz matrix with coefficients corresponding to an 8Hz-12Hz band-pass filter, then the second-order term is able to extract power in the alpha-band which is known to be modulated during motor related tasks. It is often the case that experimenters have a

hypothesis about the temporal profile of the signal of interest, for example the P300 signal or the N170 are known EEG responses with a positive peak at 300ms or negative peak at 170ms and are associated with surprise or processing of faces respectively. In such a scenario the experimenter can fix the temporal profile parameter $\mathbf{V}$ to emphasize time samples around the expected location of the peak and optimize over the rest of the parameters. The model also provides the ability to integrate information from fMRI studies. fMRI has high spatial resolution and can provide locations within the brain that may be known to participate in the processing during a particular experimentation paradigm. This location information can be incorporated into the present model by fixing the spatial parameters $\mathbf{u}_r$ and $\mathbf{a}$ to reflect a localized source (often approximated as a current dipole). The remaining temporal parameters of the model can then be optimized.

4.5 SOBDA as a generalized EEG analysis framework

The present model provides a generic framework that encompasses a number of popular EEG analysis techniques. The following list identifies some of the algorithms and how they relate to the model used in the SOBDA framework:

- Set $C = 1$, $R = 1$ and choose temporal component $\mathbf{v}$ to select a time window of interest (i.e. set $\mathbf{v}(j) = 1$ if j is inside the window of interest,$\mathbf{v}(j) = 0$ otherwise). Learn the spatial filters $\mathbf{u}$. This exactly corresponds to averaging over time and classifying in the sensor space as in [20]

- Set $C = 1$ and select some $R > 1$ and choose the component vectors $\mathbf{v}_r$ to select multiple time windows of interest as 1. Learn for each temporal window the corresponding spatial vector $\mathbf{u}_r$ from examples separately and then combine these components by learning a linear

combination of the elements. This corresponds to the multiple window hierarchical classifier as in [31]

- Set $C = 1$, $R = D$ while constraining $\mathbf{U}$ to be a diagonal matrix and select, separately for each channel, the time window $\mathbf{v}_r$ which is most discriminative. Then train the diagonal terms of $\mathbf{U}$ resulting in a latency dependent spatial filter [25]. Alternatively, in the first step, use feature selection to find the right set of time windows $\mathbf{v}_r$ simultaneously for all channels [26].

- Set $C = 1$,$R = 1$ and learn the spatial and temporal components $\mathbf{u}, \mathbf{v}$ simultaneously. This reduces to the rank-one bilinear discriminant as in [16]

- Select $C = 1$ and some $R > 1$ and learn all columns of the spatial and temporal projection matrix $\mathbf{U}$ and $\mathbf{V}$ simultaneously. This results in the *Bilinear Discriminant Component Analysis (BDCA)* [17].

- Set $C = 0$, $K = 2$ and fix $\mathbf{B}$ to a Toeplitz structure encoding a specific frequency band and set the diagonal of $\mathbf{\Lambda}$ to be [1 -1]. Then learn the spatial component $\mathbf{A}$. This is then reduced to the logistic regression with a quadratic term [42]. This formulation has been shown to be related to the Common Spatial Patters, (CSP) algorithm [37], see details in [42].

- Define $\hat{\mathbf{X}}$ to be the concatenation of $\mathbf{X}$ with itself delayed by τ samples, where τ is specified by the user, fix $\mathbf{B}$ to a Toeplitz structure, $C = 0$, and $\mathbf{A} \in \mathbb{R}^{2D \times 2}$, learn the matrix $\mathbf{A}$. This configuration can be related to the Common Spatio-Spectral Pattern [24].

4.6 Logistic regression

To optimize the model parameters $\mathbf{U}, \mathbf{V}, \mathbf{A}$ and $\mathbf{B}$ we use a Logistic Regression (LR) formalism. The probabilistic formalism is particularly convenient when imposing additional statistical properties on the coefficients such as smoothness or sparseness. In addition, in our experience, linear LR performs well in strongly overlapping high-dimensional datasets and is insensitive to outliers, the later being of particular concern when including quadratic features.

Under the Logistic Regression model the probability that a trial belongs to class y after seeing data $\mathbf{X}$ is given by the class posterior probability

$$P(y|\mathbf{X};\theta) = \frac{1}{1+e^{-y(f(\mathbf{X};\theta)+w_o)}}\,. \tag{4.3}$$

With this definition, the discriminant criterion given by the log-odds ratio of the posterior class probability

$$\log\frac{P(y=+1|\mathbf{X})}{P(y=-1|\mathbf{X})} = f(\mathbf{X};\theta)\,, \tag{4.4}$$

is simply the discriminant function which we chose to define in (4.1) as a sum of linear and quadratic terms. The log-likelihood of observing N independent samples under this model is then given by

$$L(\theta) = -\sum_{n=1}^{N}\log(1+e^{-y(f(\mathbf{X}_n;\theta)+w_o)})\,. \tag{4.5}$$

Training consists of maximizing this likelihood using a gradient assent algorithm. Analytic gradients

of the log likelihood (4.5) with respect to the various parameters are given by:

$$\frac{\partial L(\theta)}{\partial \mathbf{u}_r} = C\sum_{n=1}^{N} y_n \pi(\mathbf{X}_n)\mathbf{X}_n\mathbf{v}_r\,. \tag{4.6}$$

$$\frac{\partial L(\theta)}{\partial \mathbf{v}_r} = C\sum_{n=1}^{N} y_n \pi(\mathbf{X}_n)\mathbf{u}_r\mathbf{X}_n\,. \tag{4.7}$$

$$\frac{\partial L(\theta)}{\partial \mathbf{a}_r} = 2(1-C)\lambda_r\sum_{n=1}^{N} y_n \pi(\mathbf{X}_n)(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{a}_r\,. \tag{4.8}$$

$$\frac{\partial L(\theta)}{\partial \mathbf{b}_t} = 2(1-C)\sum_{n=1}^{N} y_n \pi(\mathbf{X}_n)\mathbf{X}^\top\mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top\mathbf{X}\mathbf{b}_t\,. \tag{4.9}$$

where we define

$$\pi(\mathbf{X}_n) = 1 - P(y|\mathbf{X}) = \frac{e^{-y(f(\mathbf{X}_n;\theta)+w_o)}}{1+e^{-y(f(\mathbf{X}_n;\theta)+w_o)}}\,. \tag{4.10}$$

and $\mathbf{u}_i, \mathbf{v}_i, \mathbf{a}_i$ and $\mathbf{b}_i$ correspond to the i_{th} columns of $\mathbf{U}, \mathbf{V}, \mathbf{A}$ and $\mathbf{B}$ respectively.

4.6.1 Enforcing structure on B

If matrix $\mathbf{B}$ is constrained to have a circular Toeplitz structure then it can be represented as $\mathbf{B} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}$, where $\mathbf{F}^{-1}$ denotes the inverse Fourier matrix, and $\mathbf{D}$ is a diagonal complex-valued matrix of Fourier coefficients. In such a case we can re-write equations (4.8) and (4.9) as

$$\frac{\partial L(\theta)}{\partial \mathbf{a}_r} = 2(1-C)\sum_{n=1}^{N} y_n \pi(\mathbf{X}_n)(\mathbf{X}_n\mathbf{F}^{-1}\hat{\mathbf{D}}\mathbf{F}^{-\mathrm{T}}\mathbf{X}_n^\top)\mathbf{a}_r\,. \tag{4.11}$$

$$\frac{\partial L(\theta)}{\partial d_i} = 2(1-C)\sum_{n=1}^{N} y_n \pi(\mathbf{X}_n)(\mathbf{F}^{-\mathrm{T}}\mathbf{X}_n^\top\mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top\mathbf{X}_n\mathbf{F}^{-1})_{i,i}d_i\,. \tag{4.12}$$

$$\tag{4.13}$$

where $\hat{\mathbf{D}} = \mathbf{D}\mathbf{D}^{\mathrm{H}}$, and the parameters are now optimized with respect to Fourier coefficients $d_i = \hat{\mathbf{D}_{i,i}}$.

This way of modeling $\mathbf{B}$ opens up a new perspective on the capabilities of the model. By allowing the columns of $\mathbf{F}$ to represent an orthogonal basis of an appropriate transformation in the

temporal space, we can incorporate this transformation in the model without any change to the estimation procedure. For example, the columns of F can represent a set of wavelet basis vectors. We note that a wavelet basis can be thought of as time-frequency representation of the signal; hence, proper selection of a wavelet basis allows for the method to not only capture the stationary power of the signal, but also the local changes in power within the T samples of matrix $\mathbf{X}$.

4.7 Regularization

Due to the high dimensional space in which the model lies and the limited samples available during training (typically in the order of 100), a maximum likelihood estimate of the parameters will be over-trained and the classifier will have poor generalization performance. To ensure good generalization performance additional regularization criteria need to be invoked. The maximum likelihood formalism of the Logistic Regression makes this particularly convenient as one can formulate standard prior probabilities on the parameter space to generate maximum a posteriori (MAP) estimates. We choose Gaussian process priors [38] on the various parameters of the model and ensure smoothness by choosing the proper covariance matrices. Specifically, the spatial components of the model (i.e. columns of $\mathbf{U}$ and $\mathbf{A}$) follow a normal distribution with $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_u)$ and $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_a)$. The covariance matrix K_u is constructed using the Matérn function given by:

$$\mathrm{k}_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{\nu} \mathrm{B}\left(\frac{\sqrt{2\nu} r}{1} \right) , \tag{4.14}$$

where l is a length-scale parameter, and ν is a shape parameter. Parameter l can be thought of as the distance within which points are significantly correlated [38]. The parameter ν defines the degree of ripple. The covariance matrix $\mathbf{K}$ is then built by evaluating the covariance function

$$(\mathbf{K})_{ij} = \sigma^2 \, \mathrm{k}_{\text{Matérn}}(r_{ij}) \tag{4.15}$$

where $r_{i,j}$ denotes the physical distance of sensor-i from sensor-j, and σ^2 defines the overall scale parameter. Similarly, the Gaussian prior can be used on the columns of the temporal matrix $\mathbf{V}$ (i.e $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_v)$).

Regularizing logistic regression amounts to minimizing the negative log likelihood and adding a log-prior on each of the columns of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{A}$ and parameters of $\mathbf{B}$, which is written as:

$$\arg \min_{\mathbf{U},\mathbf{V},\mathbf{A},\mathbf{B},w_o} -L(\theta) - \sum_{r=1}^{R}(\log p(\mathbf{u}_r) + \log p(\mathbf{v}_r)) - \sum_{k=1}^{K} \log p(\mathbf{a}_k) - \sum_{t=1}^{T'} \log(p(\mathbf{b}_t))\,, \tag{4.16}$$

where the log-prior is given for each of the parameters as $\log p(\mathbf{u}_k) = \mathbf{u}_k^\top \mathbf{K}^{(u)} \mathbf{u}_k$, $\log p(\mathbf{v}_k) = \mathbf{u}_k^\top \mathbf{K}^{(v)} \mathbf{u}_k$, $\log p(\mathbf{a}_k) = \mathbf{a}_k^\top \mathbf{K}^{(a)} \mathbf{a}_k$ and $\log p(\mathbf{b}_k) = \mathbf{b}_k^\top \mathbf{K}^{(b)} \mathbf{b}_k$. $\mathbf{K}^{(u)} \in \mathbb{R}^{D\times D}, \mathbf{K}^{(v)} \in \mathbb{R}^{T\times T}, \mathbf{K}^{(a)} \in \mathbb{R}^{D\times D}, \mathbf{K}^{(b)} \in \mathbb{R}^{T\times T}$ are kernel matrices that control the smoothness of the parameter space. Details on the regularization procedure can be found in [17].

4.8 Optimization

Optimization (4.16) is achieved using a coordinate decent type algorithm. We obtain analytic expressions for both the gradient and the Hessian of the function, however, in the optimization only the gradient information is used.The reason for discarding the Hessian information is that the computational cost is prohibitively high to calculate it at each iteration and further, the optimization function itself is non-convex. Hence, at each iteration an approximation of the Hessian would have been needed to convert it to positive definite. The numerical results reported in the next section were obtain using this type of optimization. We note that this method only finds local minima. This limitation is overcome by carefully selecting the regularization matrix and proper initialization. As noted in the regularization section above, prior knowledge can be encoded in the covariance matrix

of the regularizer, hence avoiding sub-optimal local minima. Further, a good practice is to initialize the corresponding parameter vector with the solutions from traditional methods.

4.9 Performance evaluation

We evaluated our algorithm on 3300 simulated datasets as well as 6 real EEG recordings, including two datasets used in Brain Computer Interface Competitions I and II. The simulation aims to quantify the algorithm's performance on a broad spectrum of conditions and various noise levels, as well as to compare the extracted spatial, temporal and frequency components with ground truth. Real dataset evaluation compares the cross-validation performance of the proposed method with three popular methods used in EEG analysis and BCI. Results show that our method outperformed these methods on the real datasets, decreasing classification error rates by 25%-30%. We further report results on an independent benchmark test dataset from the Brain Computer Interface (BCI) Competition and compare the performance against the competition's results.

4.9.1 Simulated EEG data

To evaluate the algorithm as a function of signal quality we generated several data sets of simulated data for a two-class problem was generated using standard EEG simulation software (BESA software). This software can generate electrode measurements under the assumption of dipolar current sources in the brain. We used 3 dipoles at three different locations, with one dipole used to generate evoked response activity, one dipole to generate induced oscillatory activity, and one dipole to generate unrelated noise/interference. The first dipole component simulates a P300 evoked response potential signal. We used a half-sinusoid lasting 125ms with the peak positioned at 300ms after trial-onset and a trial-to-trial Gaussian temporal jitter with standard deviation of 10ms. The

second dipolar component simulates ERS/ERD) in the frequency band of 8Hz to 30Hz. A variable signal in this frequency band was generated by bandpass filtering an uncorrelated Gaussian noise. The third dipole was used to generate noise in the source space representing brain activity that is not related to the evoked/induced activity. Electric potentials at $D = 31$ electrode locations were generated corresponding to 500ms of EEG signal sampled at 100Hz ($T = 50$ samples). In addition to this rank-one noise we added noise to each sensor representing other sources of noise (muscle activity, skin potentials, inductive noise, amplifier noise, etc.). All noise sources were white. Trials belonging to the first class ($y_n = +1$) contained the ERP and IS source signals scaled appropriately to achieve a specified SNR for each dataset. The second class was generated by only including the noise with no ERP or IS activity. A dataset is specified by indicating the SNR for the ERP component and the SNR for the IS component. A total of 500 trials for each class were generated for each classification problem. The SNR of the ERP component is in the range of -33dB to -13dB, and in the range of -22dB to -10dB for the oscillatory component. This is a very broad range in terms of SNR. We note that -20dB translates to the signal being 10 times smaller than the noise. ERP signals are known to be as low as $-20dB$ so this evaluation captures some extreme cases of SNR. We generated 35 datasets for each combination of SNR resulting to a total of 3300 datasets.

Cross-validation performance on these datasets was compared against the performance of three popular algorithms in EEG analysis, namely: Bilinear Discriminant Component Analysis (BDCA) [17], Common Spatial Patterns (CSP)[37], and Matrix Logistic Regression (MLR)[42]. Since both CSP and MLR require the data to be band-pass filtered to the frequency of interest, datasets where filtered in the range of 8Hz-30Hz for these two methods. This configuration gives an advantage to these two methods over our algorithm which instead is designed to identify the relevant frequency

band directly from the data.

4.9.2 Performance results on simulated data

The simulation results are summarized in Figure 4.1 which compares in each panel at the top row the performance of the SOBDA algorithm with each of three existing methods. Each point in the scatter plot represents the performance of the two algorithms for a given dataset. The diagonal line indicates methods performing equally well on that dataset, points above the line indicate better performance for SOBDA algorithm. The fraction of datasets falling above and below that line is given in percent. It is clear that in all three comparisons, SOBDA performs better than the other methods in a larger fraction of the total datasets. Further, when SOBDA performs better it does so with a greater increase in performance while when SOBDA is outperformed it is within a small margin. Figure 4.1 also shows the performance of each of the methods as a function of the SNR. The contours of the classification performance for each method as a function of the SNR of the first order and the second order components are shown. It is clear that BDCA performance is only affected by the noise in the linear term while CSP and MLR performance only changes as a function of the second order component's SNR. SOBDA however, utilizes both first and second order terms, hence performs well in datasets where at least one of the components has reasonable SNR. This finding confirms that SOBDA performs well in a broader range of SNRs than the other three competitive methods.

As a decomposition method, SOBDA extracts spatial, temporal and frequency components. The advantage of simulated data is that we can now compare the extracted information to ground truth. The component recovered from a dataset with ERP SNR at $-22dB$ and IS SNR at $-15dB$

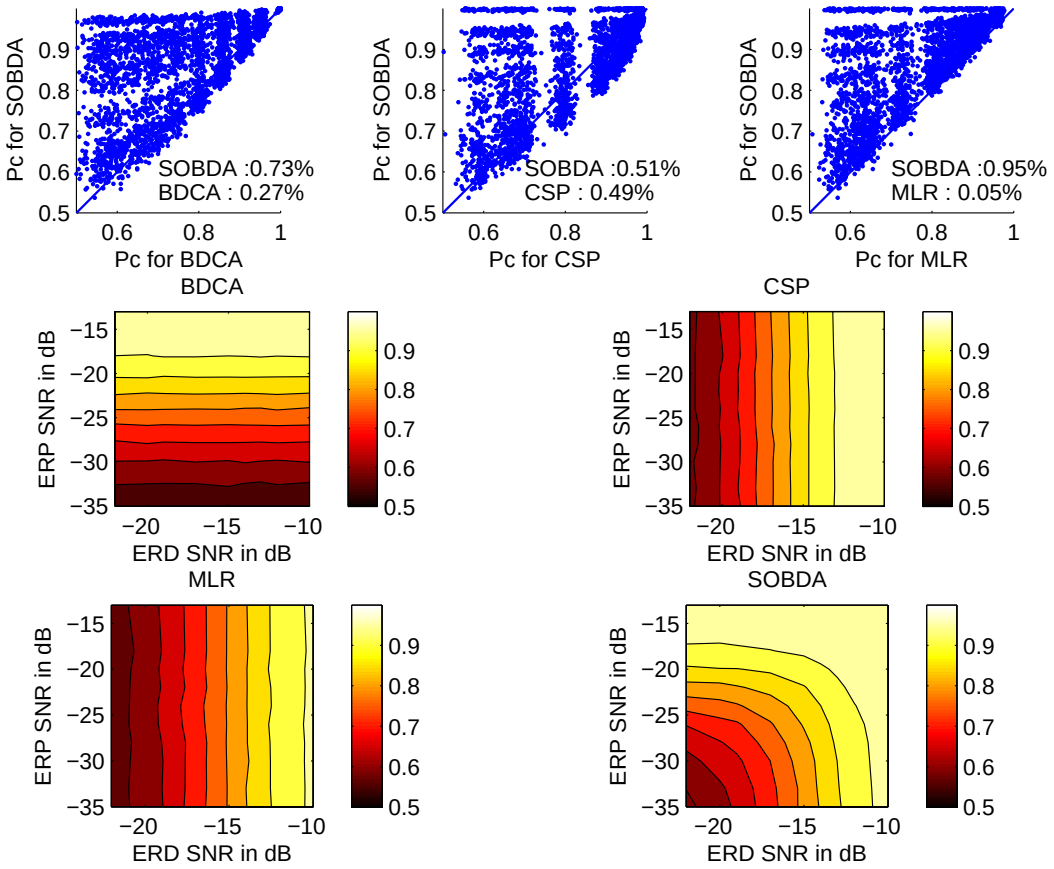

Figure 4.1: Performance results on simulated data. *Top row:* Each scatter plot compares probability of correct identification (Pc) achieved by BDCA, MLR and CSP based classifier vs SOBDA. Each point represents Pc for one of the simulated dataset. The portion of datasets lying above/below the diagonal line is given in %. *Second and third row:* Pc as a function of the component's SNR. SOBDA equi-performance contours span larger area in the SNR space than any of the other three algorithms.

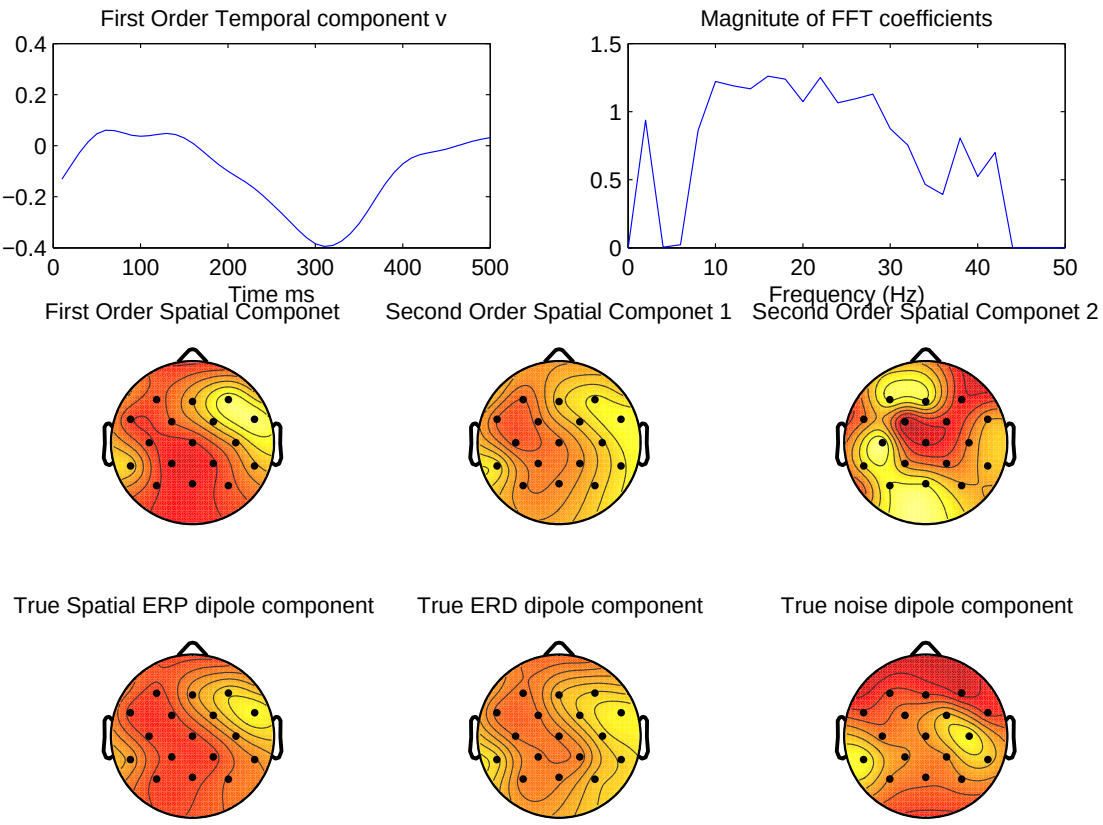

Figure 4.2: Extracted components on simulated dataset with first order SNR at $-22dB$ and second order SNR at $-15dB$. *Top row:* Extracted temporal weight of linear term (left) and frequency weights of quadratic term (right). *Center row:* Extracted spatial weights. *Bottom row:* Distribution of electric potentials corresponding to the three dipoles used during stimulus generation.

is shown in figure 4.2. The first row shows the extracted temporal component **U** and the frequency component **d** . We can see that the method extracted a temporal component with a peak at 300ms which is exactly the signal used in the simulation data design and also the frequency band extracted shows a higher amplitude in the range of 8Hz-30Hz which is exactly the band used to generate the oscillatory component. The spatial components extracted and the corresponding dipole used for data generation are shown in rows two and three in the figure. It is clear that the topography of the extracted components is similar for the first order component and the second order components. The last column of the figure captures the second power oscillatory component and the dipole of the rank one noise. We note that the ability to visually inspect the components allows the researcher to give interpretations to the signals and match them to existing neurological findings. Further, the results can be utilized as input to a source localization method to identify the exact location in the brain of the signal of interest, or used as a guide to reduce the number of electrodes necessary for a particular brain computer interface.

4.9.3 Human subject EEG - Anticipate Hand Movement Task

To evaluate the performance of the proposed method on real data we first applied the algorithm to an EEG data set that was made available through The BCI Competition 2003 ([6], Data Set IV). EEG was recorded on 28 channels for a single subject performing self-paced key typing, that is, pressing with the index and little fingers corresponding keys in a self-chosen order and timing. Key-presses occurred at an average speed of 1 key per second. Trial matrices were extracted by epoching the data starting 630ms before each key-press. A total of 416 epochs were recorded, each of length 500ms. For the competition, the first 316 epochs were used for classifier training, while

the remaining 100 epochs were used as a test set. Data was recorded at 1000 Hz with a pass-band between 0.05 and 200 Hz, then down sampled to 100Hz sampling rate.

For this experiment, the matrix $\mathbf{B}$ was fixed to a Toeplitz structure that encodes a 10Hz-33Hz bandpass filter and only the parameters $\mathbf{U}, \mathbf{V}, \mathbf{A}$ and w_0 were trained. The number of columns of $\mathbf{U}$ and $\mathbf{V}$ were set to 1, where two columns were used for $\mathbf{A}$. The temporal filter was selected based on prior knowledge of the relevant frequency band. This demonstrates the flexibility of our approach to either incorporate prior knowledge when available or extract it from data otherwise. Regularization parameters where chosen via a five fold-cross validation procedure (details can be found in [17]).

Benchmark performance was measured on the test set which had not been used during either training or cross validation. The number of misclassified trials in the test set was 13 which places our method in a new first place ranking, based on the results of the competition ([6]). Hence, our method works as a classifier producing a state-of-the art result on a realistic data set. The receiver-operator characteristic curve (ROC) for cross-validation and for the independent test set are shown in figure (4.3). The Figure also shows the contribution of the linear and quadratic terms for every trial for the two types of key-presses.

To further validate our method we obtained five more EEG recordings that follow the same experimental paradigm as that of the BCI competition's dataset described above. For each dataset and each algorithm we performed 20 repetitions of a five-fold cross-validation procedure. Each repetition uses a different partitioning of the data used in the cross-validation. The mean performance and standard deviation of each dataset and algorithm are summarized in table 4.9.3. In the mean, SOBDA outperforms competitive methods in five out of the six datasets, while achieving a

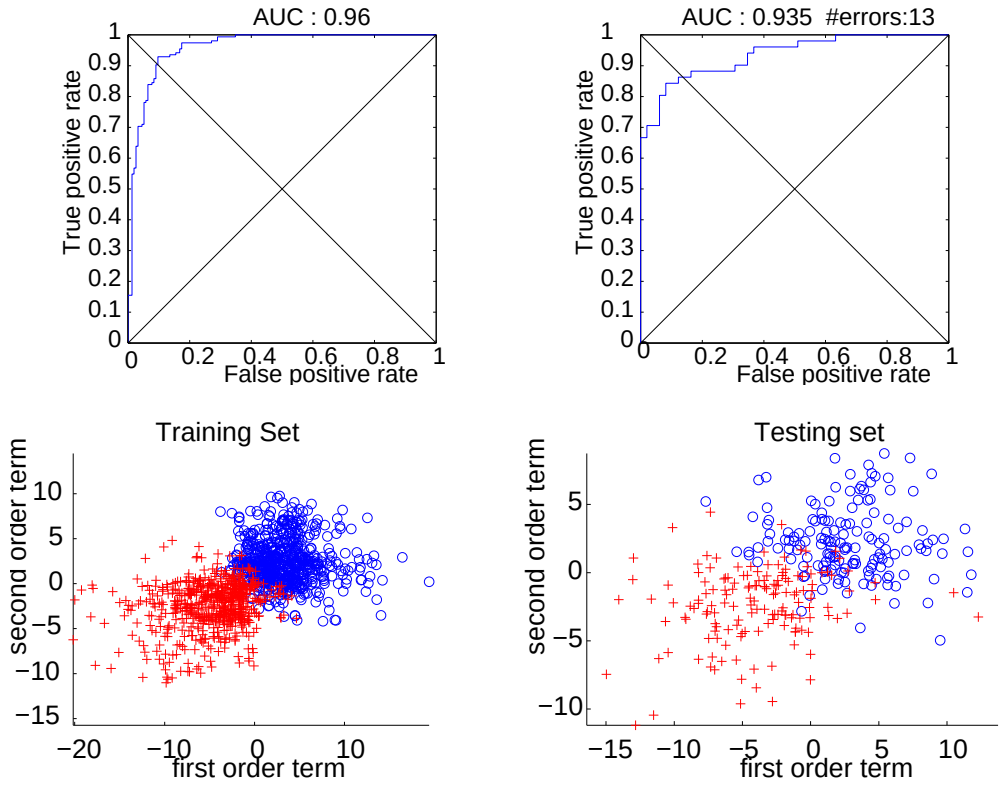

Figure 4.3: Results on human EEG for BCI. *Top row:* ROC curve with area under the curve 0.96 for the cross-validation on the benchmark dataset (left). ROC curve with area under the curve 0.93, on the independent test set, for the benchmark dataset. The were a total of 13 errors on unseen data, which is less than any of the results previously reported, placing this method on first place in the benchmark ranking. *Bottom row:* Scatter plot of the first order term vs second order term of the model, on the training and testing set for the benchmark dataset ('+' left key, and 'o' right key). It is clear that the two types of features contain independent information that can help improve the classification performance.

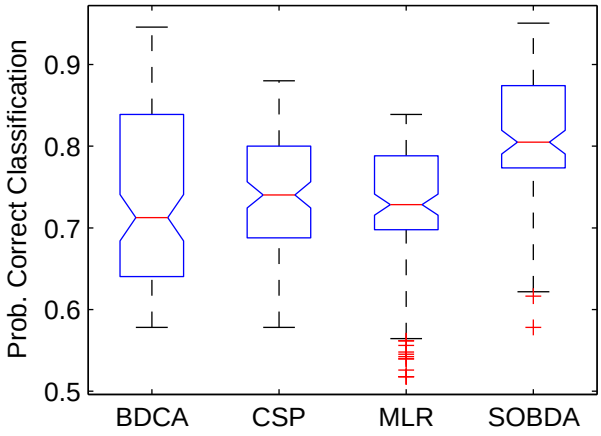

Figure 4.4: Estimate of the spread of the probability of correct identification from multiple cross-validation repetitions. Lines show lower quartile, median, and upper quartile values for each of the methods on all datasets. + symbols represent outliers.

comparable performance on dataset 2. Figure (4.4) shows performance distribution across these bootstrap repetitions using a standard boxplot. The average decrease in the probability of error is 25%-30%. A one-way ANOVA shows that the difference in performance across algorithms shown in this figure is highly significant ($p < 10^{-11}$).

Figure 4.5 shows the extracted components for 3 of the 6 datasets. We note that in all three cases the extracted components follow the general shape of the pre-motor or readiness potential (a.k.a. Bereitschafts potential) which is well-know signal in the EEG which precedes a voluntary muscle movment. In addition, for two of the datasets, the frequency weightings suggests that alpha band activity also provides discriminant information for this task. This finding is consistent with the changes in the μ rhythm–i.e. alpha band activity localized over the motor cortex and associated with motor planning and execution. This demonstrates the ability of our method to learn first and second order features that are consistent with and can be linked to existing knowledge of the underlying neuronal signal generators.

Experiment	BDCA	CSP	MLR	SOBDA
1	0.84 ± 0.011	0.80 ± 0.017	0.82 ± 0.011	**0.88 ± 0.013**
2	0.69 ± 0.037	**0.84 ± 0.017**	0.77 ± 0.028	0.83 ± 0.021
3	**0.63 ± 0.018**	0.62 ± 0.016	0.55 ± 0.020	**0.63 ± 0.017**
4	0.72 ± 0.021	0.78 ± 0.015	0.77 ± 0.015	**0.79 ± 0.018**
5	0.64 ± 0.018	0.70 ± 0.022	0.70 ± 0.011	**0.78 ± 0.013**
6	0.93 ± 0.010	0.70 ± 0.016	0.72 ± 0.010	**0.94 ± 0.008**
Mean	0.7412	0.7388	0.7213	0.8068
Error decrease	25%	26%	30%	

Table 4.1: Probability of correct identification for the six EEG datasets obtained by each of the four methods. The last row indicates the percentage of decrease in the classification error achieved by SOBDA compared to each one of the methods. ± range indicates one standard deviation for results of multiple cross-validation repetitions.

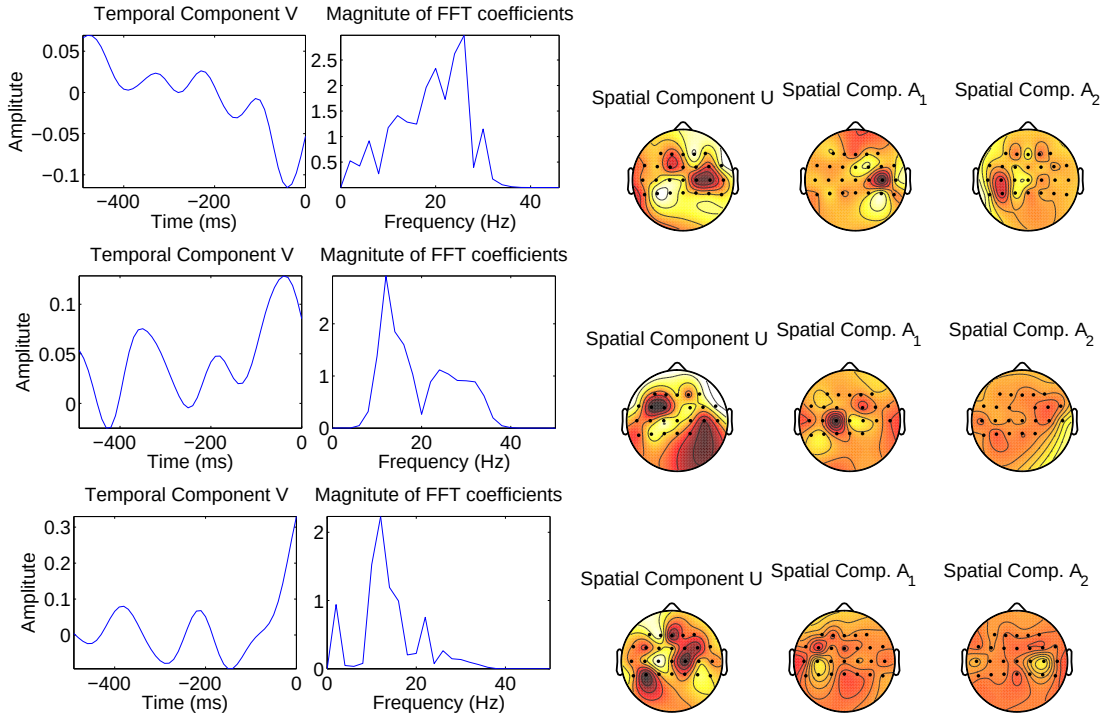

Figure 4.5: Extracted components in EEG for datasets 6, 4, and 3. *Left:* Temporal weights of linear component (first column) and and frequency weights of quadratic component (second column). *Right:* Spatial weights of linear component (third column) and two spatial weights for second order spatial components (fourth and fifth column).

Chapter 5

BILINEAR FEATURE-BASED DISCRIMINATION

5.1 Motivation

In the two previous chapters we presented two algorithms that address the single-trial classification of EEG, based on the bilinear model. Initially, Bilinear Discriminant Component Analysis (BDCA) algorithm introduced the bilinear decomposition that use the spatio-temporal structure of EEG to reduce the dimensionality of the data. BDCA is applied to raw EEG signals and extract ERP characteristics (i.e. differences in the signal amplitude between the two conditions). This restricts the application of BDCA to specific type of EEG experiments where Evoke Potentials are present. Second Order Bilinear Discriminant(SOBDA) attempts to address the limitations of BDCA by extending the model to capture oscillatory type of features using again bilinear projections. We showed that SOBDA can be though of as a model that generalizes several single-trial classification algorithms. However, due to the non-convexity of the optimization problem, SOBDA requires careful seeding of the parameters and an appropriate regularization procedure. Further, even though SOBDA incorporate prior knowledge by specifying time-frequency basis of interest (i.e. Fourier matrix, wavelet basis etc.), the selection of these bases remains an *ad-hock* process. Furthermore, the number of samples one has to train SOBDA is bounded by the number of trials that are obtained from single experimental sessions. This is due to the inter-subject and inter-session variability in EEG. All, basis selection, parameters seeding and regularization procedure are challenging tasks that the user needs to address when using SOBDA. This makes the applicability of SOBDA to a

broad range of experimental paradigms difficult.

To address these limitations we develop now a new concept we term Bilinear Feature Based Discriminant (BFBD). It uses the bilinear formulation for the spatio-temporal decomposition of EEG that capture both event related potentials as well as oscillatory features. The following hypothesis motivates our approach:

> **Hypothesis**: Spatio-temporal characteristics of the EEG vary greatly among subjects or repetitions of the EEG experiment. However, for any specific experimental paradigm the underlying neural rhythmic activity (the cause of oscillatory and evoke related features in EEG) associated with a task of interest remains invariant.

This hypothesis is implicitly assumed in most current BCI experiments. For example, in the BCI task where a subject is asked to imagine a movement of the left or the right hand, a preprocessing step involves band-pass filtering the data in the α-band (8Hz - 13Hz). The selection of this band was motivated by neurological studies and is tied to this specific experimental paradigm independent of the subject performing the task. Another example is the case of Rapid Serial Visual Presentation (RSVP) experiment were a series of images is shown to a subject. The aim is for the subject to identify images that contain a targets of interest. In such experimental paradigms one look for evoke potential at around 300ms after the stimulus (called P300 signal). Again this signal is assumed to occur in this specific experimental paradigm independent of the subject.

The goal of the BFBD method is to take advantage of the fact that certain characteristics of the EEG signals are subject-invariant and session-invariant, while some others are subject-specific. By distinguishing between subject-invariant parameters and subject specific parameters, we can use EEG recordings from many subjects to estimate the subject invariant-parameters, while using

a recoding from an individual subject to fine-tune the subject specific parameters. The advantage of this approach is of course the fact that while we have a limitation in the number of trials we can obtain from a single subject during a single session (i.e. subject gets tired after couple of hours) there is no limitation in repeating the experiment on multiple subjects/ multiple session. Hence part of the model can be trained with many more trials, reducing the problem of over training.

5.2 Problem definition

As before we represent an EEG trial by a matrix $\mathbf{X} \in \mathbb{R}^{D \times T}$ where D denotes the number of sensors and T the number of temporal samples of the EEG. Multiple trials from a single experiment are represented as a set of such matrices. Each EEG trial is associated with some underlying mental state, that constitutes the label of the trial. Further, we have now multiple experiments, denoted by a superset of sets. We can then define the following classification problem.

> **Classification Problem** Given a set of training examples $\mathcal{D}_1 = \{\mathbf{X}_n, y_n\}_{n=1}^{N}, \mathbf{X} \in \mathbb{R}^{D \times T}, y \in \{-1, 1\}$, where $\mathbf{X}_n$ corresponds to the EEG signal of D channels and T sample points and y_n indicates one of two conditions or classes (e.g. right or left hand imaginary movement, stimulus versus control conditions, etc.), and given M more such datasets $\{\mathcal{D}_m\}_{m=2}^{M+1}$, the task is then to predict the class label y for an unobserved trial $\mathbf{X}$ sampled from the same distribution as the samples from session $\mathcal{D}_1$.

In terms of classification this is the same problem as before, however, in this formulation we have available information from multiple datasets. In the sections that follow, we formulate a new classification algorithm that uses this extra information.

5.3 Bilinear Feature Pool

Before we present the new algorithm we define the concept of a Bilinear Feature Pool. Let Φ be a given set of functions with elements $\phi_k : \mathbb{R}^{D \times T} \rightarrow \mathbb{R}^{D_k \times T_k}$. Note that ϕ_k denotes a function from a matrix space to a matrix space. For EEG we design this set of functions to be relevant to the features of interest. Below we list some possible families of functions and their relevance to EEG:

Hilbert Transform Hilbert transform is useful for analyzing the instantaneous power content of an oscillatory signal as a function of time. The Hilbert transform of a real-valued signal is a complex signal where the real part can be though of as the same as the original signal while the imaginary part is the original signal with $\pi/2$ phase shift - i.e. every sinusoidal component of the signal shifted. The magnitude squared of its complex coefficients defines the instantaneous power of the signal over time. Filtering the signal for a specific frequency band before applying the Hilbert transform we can obtain the instantaneous power over that specific band. In terms of the EEG signal, this change in power is useful since it captures modulation of oscillatory features of interest.

Multitaper Spectrum Estimation Multitaper is a spectrum estimation method that minimizes the problem of spectral leakage [36]. It uses Slepian Sequences, a set of temporal basis functions designed to concentrate the maximum energy in a narrow frequency band around zero; under the assumption of signals with unit variance. By multiplying the basis with sinusoid of some center frequency f the basis can capture the power in the same bandwidth centered on f. Time varying windows are used to capture the power of the signal in a narrow frequency band at various center frequency and temporal resolutions.

Temporal downsampling EEG is usually recorded using high temporal resolution, with sampling rates ranging from 1024Hz to 2048Hz. However, this high sampling rate increases unnecessarily the dimensionality of the parameter space. Temporal downsampling helps in reducing the number of dimensions without compromising the signal quality.

Spatial smoothing Spatial smoothing include channel subset selection, whereby a small number of channels is selected. This reduces the number of dimensions to learn. An alternative way is the grouping and averaging of channels based on their location. Spatial filtering has been used extensively in EEG, some of the methods we reviewed in Chapter 3.

Identity Transform The Identity function applied on a matrix return the exact same matrix. We include the identity function in our pool of features for uniformity in notation.

For each such function $\phi_k \in \Phi$, we can define the bilinear projection of the resulting transformation as follows :

$$t_k = \text{Trace}\{(\mathbf{U}_k^\top \phi_k(\mathbf{X})\mathbf{V}_k)\}$$

where $\mathbf{U}_k \in \mathbb{R}^{\hat{D}\times R}, \mathbf{V}_k \in \mathbb{R}^{\hat{T}\times R}$ are parameters and $\mathbf{X} \in \mathbb{R}^{D\times T}$ the EEG signal. The resulting scalar t_k is a linear combination of the elements of the transformed matrix $\mathbf{X}$. Note that the parameters are tide with the selection of the kth function ϕ_k.

5.3.1 Multi-taper log spectral estimation

Since multi-taper spectral estimation is the family of transformations we will be using in the evaluation of our method, we present the calculation of the transformation in some detail here.

The important elements in the multi-taper method are the Slepian sequences. A list of n-dimensional orthonormal vectors $< \mathbf{b}_1, ..., \mathbf{b}_K >$ are called Slepian sequences if they maximize the energy $E = \int_{-w}^{w} |B_k(f)|^2 df$ in the frequency band $[-wq, wq], 0 < w < 1/2$, where $B_k(f)$ is the fourier transform of the sequence $\mathbf{b}$ and q is the nyquist frequency and w is a parameter for the basis. For a given frequency of interest $\hat{f}$ the list of vectors $< \hat{\mathbf{b}}_1^{\hat{f}}, ..., \hat{\mathbf{b}}_K^{\hat{f}} >$ obtain by

$$\hat{\mathbf{b}}_k^{\hat{f}}(t) = \mathbf{b}_k(t) e^{i2\pi \hat{f} t} \tag{5.1}$$

concentrate the maximum energy in the frequency band $[\hat{f} - wq, \hat{f} + wq]$. The log-power estimate for a data segment $\mathbf{x} \in \mathbb{R}^n$ is then obtained using

$$p(\mathbf{x}; \hat{f}) = log(\mathbf{x}^\top \hat{\mathbf{B}} \hat{\mathbf{B}}^\top \mathbf{x}) \tag{5.2}$$

where we defined $\hat{\mathbf{B}}$ a matrix whose columns are the list of vectors $\hat{\mathbf{b}}_1^{\hat{f}}, ..., \hat{\mathbf{b}}_K^{\hat{f}}$.

The family of multi-taper spectral estimation transformations is defined by a set of parameters $< n, \hat{f}, n_o >$. The transformation is obtained by first fragmenting each channel of the EEG signal in windows of length n, with overlapping n_o, and then for each such data segment obtain the log-power using the equation 5.2.

5.4 Discriminative Model

Consider a set of functions $\{\phi_1, ...\phi_K\} \subset \Phi$. For each $\phi_k \in \Phi$ we define the following discriminant model

$$log \frac{P(y = +1 | \mathbf{X}, \mathbf{U}_k, \mathbf{V}_k, \phi_k)}{P(y = -1 | \mathbf{X}, \mathbf{U}_k, \mathbf{V}_k, \phi_k)} = g_k(\mathbf{X}; \mathbf{U}_k, \mathbf{V}_k, \phi_k) = \text{Trace}\{\mathbf{U}_k^\top \phi_k(\mathbf{X}) \mathbf{V}_k\} \tag{5.3}$$

where $\mathbf{U}_k, \mathbf{V}_k$ are the model parameters. Let $Q = \{(\mathbf{U}_k, \mathbf{V}_k)\}_{k=1}^K$ the list of K model parameters each associated with the kth feature function. Define the parameter index set $I \subset \{1, .., K\}$, whose

elements are in ascending order. We denote I_i and Q_i to be the ith element of sets I and Q respectively. We use $Q(I)$ to denote the subset of Q whose elements are determined by the index set I, i.e., $\{Q_{I_1}, ..., Q_{I_{|I|}}\}$. We can express the feature vector $\mathbf{t}(\mathbf{X}; Q, I)$

$$\mathbf{t}(\mathbf{X}; Q, I) = \begin{pmatrix} g_{I_1}(\mathbf{X}; Q_{I_1}, \phi_{I_1}) \\ ... \\ g_{I_{|I|}}(\mathbf{X}; Q_{I_{|I|}}, \phi_{I_{|I|}}) \end{pmatrix}$$

note that the dimensions of the space $\mathbf{t}(\mathbf{X}; Q, I) \in \mathbb{R}^{|I|}$ depends on the cardinality of the index set $|I|$. We identify I as the the subject-invariant parameter of our model. Finally we define the classification of an observation $\mathbf{X}$, as:

$$f(\mathbf{X}) = sign(\mathbf{w}^\top \mathbf{t}(\mathbf{X}; Q, I)) \tag{5.4}$$

where the parameter vector $\mathbf{w} \in \mathbb{R}^{|I|}$ defines a linear combination of the elements of $\mathbf{t}(\mathbf{X}; Q, I)$. We identify the parameters $Q(I)$ and $\mathbf{w}$ as the subject-specific parameters of our model. In the following sections we give an interpretation of the model and define the optimization problem to estimate the model parameters $I, Q(I)$, and $\mathbf{w}$.

5.4.1 Model interpretation

As explained in the previous section, the functions in the feature set Φ transforms an input EEG matrix $\mathbf{X}$ to a potentially more informative matrix $\phi(\mathbf{X})$. The motivation is that some of these functions can possibly decrease the dimensions of the space while increase the signal-to-noise ratio of the observation. Note that we have no restriction on the type of functions in the set, it can be either linear or non-linear transformations. The function g_k is associated with a function ϕ_k

and can be thought of as a parametric feature extractor. It implements a bilinear combination of the elements of $\phi_k(\mathbf{X})$. The output of g_k is considered as a single feature obtained by the EEG observation. The effectiveness of each feature depends on its parameters $\mathbf{U}_k$ and $\mathbf{V}_k$ as well as the selection of the corresponding ϕ_k. We determine the proper values of the parameters $\mathbf{U}_k$ and $\mathbf{V}_k$ by means of an optimization procedure, that we present in the next section. The index set I specifies a selection of features from the feature pool Φ. Depending on the experimental paradigm, different functions in Φ might be informative (i.e., increase the signal-to-noise ratio). We identify I as the subject-invariant parameter of our model because we obtain it using recordings from multiple datasets as opposed to $Q(I)$ and $\mathbf{w}$ which are determined using a single-subject/single-session recording. Finally, parameter $\mathbf{w}$ defines a linear discriminant in the feature space.

5.5 Optimization

The model involves a number of parameters that need to be optimized. Specifically, the index set I, the vector $\mathbf{w}$ and the parameters $Q(I)$. We will first formulate the optimization of the subject-specific parameters $Q(I)$ and $\mathbf{w}$ for a fixed index set I. Then we will formulate a combinatorial optimization problem and present an algorithm to find an approximate solution for the index set I.

Given a dataset $\{\mathbf{X}_n, y_n\}_{n=1}^N$, fix an index set I. We can then express the log-likelihood of the parameters $(\mathbf{U}, \mathbf{V}) \in Q(I)$ given the corresponding ϕ by:

$$L((\mathbf{U}, \mathbf{V}); \{\mathbf{X}_n, y_n\}_{n=1}^N, \phi) = -\sum_{n=1}^{N} log(1 + \exp^{-y_n g(\mathbf{X}_n;(\mathbf{U},\mathbf{V}),\phi)}) \tag{5.5}$$

We further introduce regularization terms for each of the parameters $\mathbf{U}, \mathbf{V}$ that we denote here as $reg(\mathbf{U}, \mathbf{V})$. We motivate the use of regularization in the following section and provide implemen-

tation details. Then the parameters $Q(I)$ are the solutions of the following optimization problems.

$$\begin{aligned}
\hat{\mathbf{U}}_{I_1}, \hat{\mathbf{V}}_{I_1} &= \arg\max L((\mathbf{U}_{I_1}, \mathbf{V}_{I_1}); \{\mathbf{X}_n, y_n\}_{n=1}^N, \phi_{I_1}) + reg(\mathbf{U}_{I_1}, \mathbf{V}_{I_1}) \\
\hat{\mathbf{U}}_{I_2}, \hat{\mathbf{V}}_{I_2} &= \arg\max L((\mathbf{U}_{I_2}, \mathbf{V}_{I_2}); \{\mathbf{X}_n, y_n\}_{n=1}^N, \phi_{I_2}) + reg(\mathbf{U}_{I_2}, \mathbf{V}_{I_2}) \\
&... \\
\hat{\mathbf{U}}_{I_{|I|}}, \hat{\mathbf{V}}_{I_{|I|}} &= \arg\max L((\mathbf{U}_{I_{|I|}}, \mathbf{V}_{I_{|I|}}); \{\mathbf{X}_n, y_n\}_{n=1}^N, \phi_{I_{|I|}}) + reg(\mathbf{U}_{I_{|I|}}, \mathbf{V}_{I_{|I|}})
\end{aligned}$$

We obtained analytic formulas for the gradient function of L and reg, with respect to the parameters $(\mathbf{U}, \mathbf{V})$ (see appendix). We solve the above optimization using the gradient decent algorithm.

5.5.1 Regularization

The Maximum likelihood [1] estimate of the parameters will over-train the data and have poor generalization performance due to the high dimensional space in which the model lies and the limited samples available during training (typically in the order of 100), As in the case of BDCA and SOBDA, to ensure good generalization performance additional regularization criteria need to be invoked. The maximum likelihood formalism of the Logistic Regression makes it particularly convenient as one can formulate standard prior probabilities on the parameter space to generate maximum a posteriori (MAP) estimates. We choose Gaussian process priors [38] on the various parameters of the model and ensure smoothness by choosing the proper covariance matrices. Specifically, the spatial components of the model (i.e. columns of $\{\mathbf{U}_k\}_{k=1}^K$) follow a normal distribution

[1] Maximum likelihood estimate in our model corresponds to maximizing the likelihood function $L(\mathbf{U}, \mathbf{V})$ without the addition of the regularization term reg.

with $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_u)$.

The covariance matrix $\mathbf{K}_u$ depends on the Matérn function

$$\mathrm{k}_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{\nu} \mathrm{B}\left(\frac{\sqrt{2\nu} r}{1} \right) , \tag{5.6}$$

where l is a length-scale parameter, and ν is a shape parameter. Parameter l can be roughly though of as the distance within which points are significantly correlated [38]. The parameter ν defines the degree of ripple. The covariance matrix $\mathbf{K}_u$ is then built by evaluating the covariance function

$$(\mathbf{K}_u)_{ij} = \sigma^2 \, \mathrm{k}_{\text{Matérn}}(r_{ij}) \tag{5.7}$$

where $r_{i,j}$ denotes the physical distance of sensor-i from sensor-j, and σ^2 defines the overall scale parameter. Similarly, the Gaussian prior can be used on the columns of the temporal matrix $\mathbf{V}$ (i.e $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_v)$). In the case of $\mathbf{K}_v$ the length-scale parameter l can be thought of as the temporal correlation of the samples.

We can now define the regularization terms $reg(\mathbf{U}, \mathbf{V})$ as:

$$reg(\mathbf{U}, \mathbf{V}) = \sum_{r=1}^{R} \mathbf{u}_r^{\top} \mathbf{K}_u \mathbf{u}_r + \mathbf{v}_r^{\top} \mathbf{K}_v \mathbf{v}_r \tag{5.8}$$

5.5.2 Estimating the parameter vector $\mathbf{w}$

For an optimal set of parameters $\hat{Q} = \{\hat{\mathbf{U}}_k, \hat{\mathbf{V}}_k\}_{k=1}^{|I|}$, obtained by the optimization in 5.6 we transform the input data-set $\{\mathbf{X}_n, y_n\}_{n=1}^{N}$ to $\{\mathbf{t}(\mathbf{X}_n; \hat{Q}, I), y_n\}_{n=1}^{N}$. For simplicity in notation, we denote $\mathbf{t}_n = \mathbf{t}(\mathbf{X}_n; \hat{Q}, I)$. In this new space we assume a Normal class-conditional distribution on $\mathbf{t}|y$ with

equal covariance matrix with

$$\begin{aligned} p(\mathbf{t}|y=+1) &= N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\ p(\mathbf{t}|y=-1) &= N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \end{aligned}$$

where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Under this normality assumption, the Fisher Discriminant gives the Bayes optimal classifier with an analytic solution for the parameter $\mathbf{w}$ given as:

$$\hat{\mathbf{w}} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \tag{5.9}$$

where $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$ are estimates of the mean for the two classes, and $\hat{\boldsymbol{\Sigma}}$ is the estimated pool covariance matrix.

A new observation $\mathbf{X}_{new}$ can then be classified by evaluating the function:

$$f(\mathbf{X}_{new}) = sign(\hat{\mathbf{w}}^\top \mathbf{t}(\mathbf{X}_{new}, \hat{Q}, I)) \tag{5.10}$$

5.5.3 Estimation of subject-invariant parameter I

In the section above we defined the discriminant model and expressed the corresponding optimization problems to estimate the subject specific parameters $Q(I)$ and $\mathbf{w}$ for a fixed I. In this section we define the optimization of the subject-invariant parameter set I and suggest a heuristic algorithm to find satisfactory solutions.

According to the problem definition in section 5.2, a superset of data-sets $\{\mathcal{D}_2, ..., \mathcal{D}_M\}$ is given, where each $\mathcal{D}_m = \{(\mathbf{X}_n, y_n)\}_{n=1}^{N^m}$ corresponds to observations from a specific subject and a specific EEG session. The subject specific parameters of the model are trained on individual data-sets. Our

goal in estimating the subject-invariant parameter set I is to take advantage of the information provided by all data-sets combined.

We proceed in formulating the optimization problem by defining the following functions. Let $cv(I; \mathcal{D}_m)$ be the function that calculates the five-fold cross-validation performance [2] in terms of area under the ROC curve, evaluated for a fixed parameter set I and on data-set $\mathcal{D}_m$, using the optimization procedures described in the previous section. We define the average performance of index set I on the given data-set as

$$mean\_cv(I; \{\mathcal{D}_m\}_{m=1}^{M}) = \frac{1}{M-1} \sum_{m=2}^{M} cv(I; \mathcal{D}_m) \tag{5.11}$$

We can define the optimization procedure now as

$$\hat{I} = \arg_I \max mean\_cv(I; \{\mathcal{D}_m\}_{m=2}^{M}) \tag{5.12}$$

The optimization problem specified in 5.12 fall into the category of combinatorial problems. There are many algorithms proposed in the literature to solve this type of optimization. They either employ different heuristics to search the parameter lattice for solutions or exploit different properties of the optimization function using branch and bound techniques [1]. Since in our formulation it is desired for I to have small cardinality (i.e., $|I| < 4$) we employ a stochastic search strategy that promotes sparseness. In the code listing 1 we provide the pseudo-code for our proposed algorithm that approximates the solution to the combinatorial problem. Alternatively, we can use a generic algorithm for feature selection was proposed by [22].

---

[2] Five fold cross validation is a standard technique in patter recognition where a training set is partition into five blocks with approximately equal number of observation in each block. Four of the blocks are used for training a classifier, and the fifth applying the classifier. Rotate blocks five times until the classifier is applying to all data.

Algorithm 1 Combinatorial optimization algorithm

Input: $\{\mathcal{D}\}_{m=1}^{M}$, $\Phi = \{\phi_1, ..., \phi_k\}$,$J = \{1, .., K\}$

Choose $MaxNumOfDimensions \in \{1, ..4\}$

Choose $RepetitionsPerDimension \leq K$

Initialize $bestFeaturePool = \{\}$

for k =1:$MaxNumOfDimensions$ **do**

Initialize $currentFeaturePool = \{\}$

for i =1:$RepetitionsPerDimension$ **do**

I_{cur1} ←randomly sample from $bestFeaturePool$ a tuple $< Az, I >$, assign I to I_{cur1}

I_{cur2} ←randomly sample from J.

$I_{cur} \leftarrow I_{cur1} \cup I_{cur2}$

$Az \leftarrow mean\_cv(I_{cur}, \{\mathcal{D}\}_{m=1}^{M})$

Add $< Az, I_{cur} >$ to $currentFeaturePool$

end for

$bestFeaturePool \leftarrow$ Assign five of $< Az, I > \in \{currentFeaturePool \cup bestFeaturePool\}$ with highest Az value.

end for

return $< Az, I > \in bestFeaturePool$ with highest Az value.

5.6 Evaluation

We evaluated our method on 20 real EEG data-sets obtained from 10 subjects/individuals on a Perceptual Categorization task [19]. We further report results on 14 real EEG data-sets from a Target Detection task. In the case of the Perception categorization, for each subject s, the data-sets corresponding to all subjects except the sth one (18 data-sets) were used in the estimation procedure of the subject-invariant set $\hat{I}$. Having obtained $\hat{I}$, the set of subject-specific parameters are estimated using the data-sets associated with the sth subject (2 data-sets). Performance on each data-set is measured in terms of area under the ROC curve (we refer to as Az values) using five-fold cross validation within each data-set. In the rest of this section we briefly introduce the Perception Categorization Task and present the results of our method.

5.6.1 EEG Based Perceptual Categorization Task

A total of 10 subjects (5 females and 5 males, age range 21-37 years) performed a simple categorization task where they had to discriminate between images of faces and cars. The data-sets were obtained in a study presented by [34], where the experiment is described as follows:

> A set of 12 faces (Max Plank Institute face database) and 12 car gray-scale images (image size 512x512 pixels,8-bit/pixcel) were used. Each image was processed to have different phase coherence values (20%,25%,30%,35%,40%,45%) by manipulating their corresponding phase spectra [34]. Coherence levels adjust the recognizability of the image by the human subject. Images with high coherence are easier to recognize than images with lower coherence. Subjects reported their decision regarding the type of image by pressing one of two mouse buttons - left for faces and right for cars - using

their right index and middle fingers respectively. A block trial consisted of 24 trials of both face and car images at each of 6 different coherence levels, a total of 144 trials. There are a total of 4 blocks in each experiment. At the beginning of a block of trials, subjects fixated at the center of the screen. Images were presented for 30ms followed by an inter-stimulus-interval (ISI) which was randomized on the range of 1500-2000ms. Subjects were instructed to respond when they identified the type of image and, before the next image was presented. A schematic representation of the behavioral paradigm is given in figure 5.1. Trials where subjects failed to respond within the ISI were marked as no-choice trials and were discarded from further analysis.

EEG data where acquired during the perceptual categorization task in an electrostatically shielded room (ETS-Lindgren, Glendale Heights, IL) using a high input impedance Sensorium EPA-6 (Charlotte, VT) electrophysiological amplifier (1G) with a gain of 10K from 60 Ag/AgCl electrodes attached to an EEG cap. The electrodes were positioned according to the International 10-20 system of electrode placement. All channels were referenced to the left mastoid with chin ground. Data were sampled at 1000 Hz with an analog pass band of 0.01-300 Hz using 12 dB/octave high pass and 8th order elliptic low pass filters. EEG data were recorded via National Instruments (Austin, TX) data acquisition cards, on a separate computer, each of which can sample 64 channels at a maximum rate of 2000 Hz.

Once EEG data were collected digital 0.5 Hz high pass filter (4th order Butterworth) was used to remove DC drifts and 60 Hz and 120 Hz (harmonic) notch filters were applied to minimize line noise artifacts. These filters were designed to be linear-phase

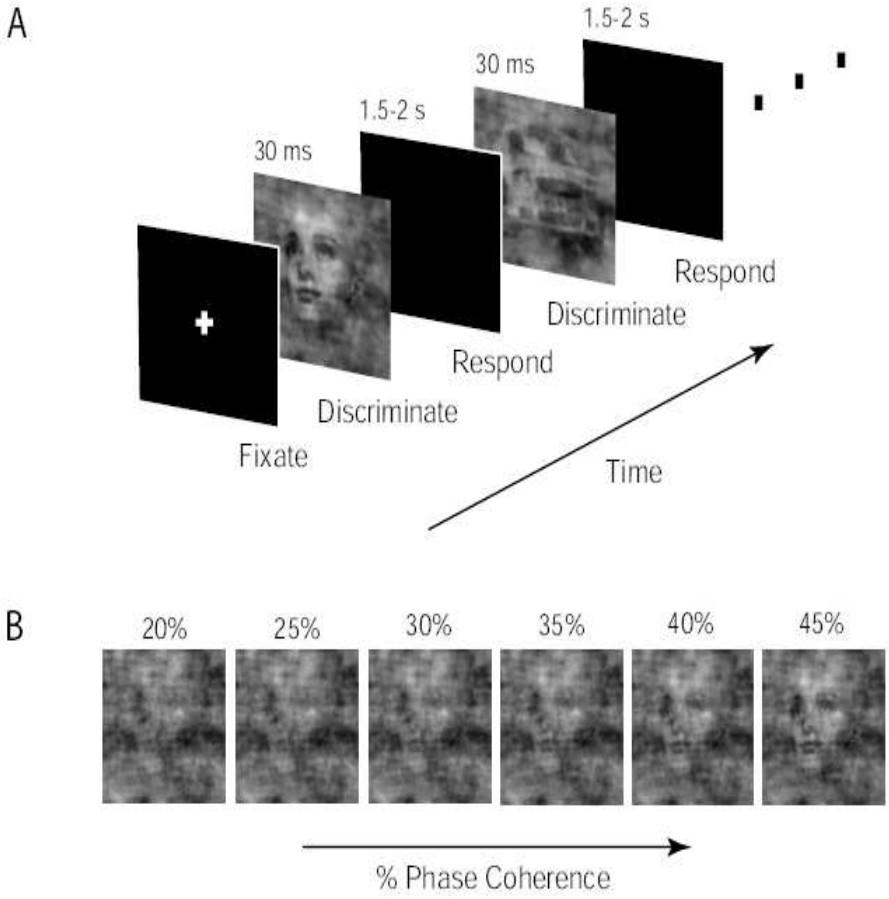

Figure 5.1: Schematic representation of the behavioral paradigm. (A) Within a block of trials subjects were instructed to fixate on the center of the screen and were subsequently presented, in random order, with a series of different face and car images at one of the six phase coherence levels shown in (B). Each image was presented for 30 ms followed by an inter-stimulus-interval lasting between 1500- 2000 ms during which subjects were required to discriminate among the two types of images and respond by pressing a button. A block of trials was completed once all face and car images at all six phase coherence levels have been presented. (B) A sample face image at 6 different phase coherence levels (20, 25, 30, 35, 40, 45

to minimize delay distortions. Motor response and stimulus events recorded on separate channels were delayed to match latencies introduced by digitally filtering the EEG.

For the evaluation of our algorithm we formulated binary classification problems from the data described above. For each of the 10 subject, two data-sets were generated. First, a data-set of EEG trails corresponding to stimulus of images shown in the coherence levels $\{40\%, 45\%\}$. The second data-set includes EEG trials that correspond to coherence levels $\{35\%, 40\%\}$. Each trial measures EEG from 58 channels for 500ms starting from the stimulus onset. A total of 110 trials (55 trials per condition) were used. Each trial is labeled according the stimulus type (i.e. face or car).

We compare our method against Bilinear Discriminant Component Analysis (BDCA) [17], and Hierarchical Fisher Discriminant Analysis (HDCA)[31] as introduced in chapter 3. Both methods represent the current state-of-the-art in EEG analysis applied on Rapid Serial Visual Presentation task. We use the area under the ROC curve as a measure of performance, denoted by Az values. The performance of each algorithm compared with one another is summarized using the scatter-plots in figure 5.2 for the Perception Categorization task, and in figure 5.6 for the Target Detection task. The x-axis represents the Az performance of one algorithm while y-axis represents the Az performance of an algorithm to compare with. Each point in a plot represents one of the datasets, 20 for the Perception Categorization task and 14 for the Target Detection task. The diagonal line shows the equal performance manifold (i.e. both algorithms perform equally well on a particular dataset.) We see from the first scatterplot in figure 5.2 that BFBD achieves higher Az values compared to BDCA on 70% of the datasets. We note that in the 30% of the datasets where BDCA performs better the difference between the performance is small, since most of those points are very close to the equi-performance manifold. The middle scatterplot in figure 5.2 compares

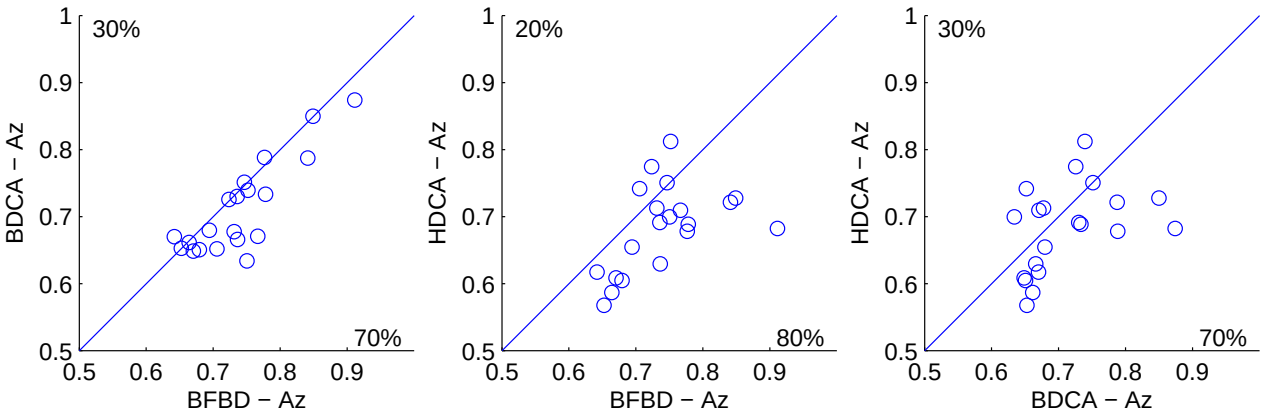

Figure 5.2: Each scatter plot compares area under the ROC curve (Az) achieved by BDCA, HDCA classifier vs BFBD. Each point represents Az for one of the twenty real EEG datasets. The portion of datasets lying above/below the diagonal line is given in %

BFBD against Hierarchical Fisher Discriminant Analysis in which case for 80% of the dataset BFBD achieves higher performance than HDCA. Finally the last scatter-plot compares the BDCA with HDCA in which case BDCA performance better for 70% of the datasets. In the scatterplot associated with the target detection task figure 5.6 BDBD outperforms the compiting method on 100% of the data-sets. In figure 5.3 and 5.4 the ROC curves obtain by BFBD and BDCA are shown for eight datasets. We can see from the ROC curves an improvement of BFBD in the true positive rate over BDCA is achieved at low false positive rates. We also note that the ROC curves obtained by BFBD are almost uniformly better for most datasets.

Figure 5.5 shows the performance distributions across datasets for the three methods on the Perception Categorization task. In the mean BFBD is shown to outperform the other two methods. The average decrease in error *i.e.*, $1 - Az$ is 10% and 18% against BDCA and HDCA respectively.

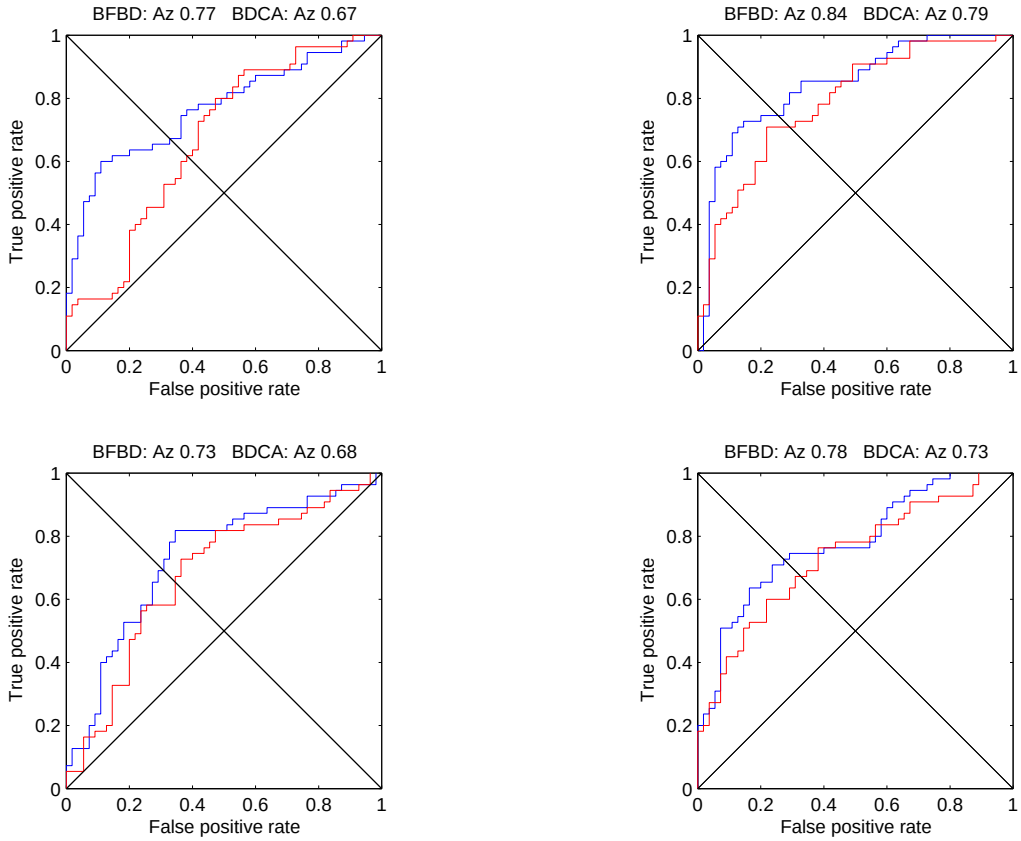

Figure 5.3: Receiver Operator Characteristic (ROC) obtained by BFBD (blue line) and BDCA (red line) on four of the datasets.*Note* the uniformity of red curve being above the blue curve as false positive rate increases. *Also note* that the increase in Az performance is already visible at the lowest of the false positive rates.

A one-way repeated measures ANOVA shows that the difference in performance across the three algorithms is highly significant ($p < 7 \times 10^{-4}$). Figure 5.7 shows the performance distributions across data-sets for the three methods on the Target Detection task. The average decrease in error *i.e.*, $1 - Az$ is 34% and 48% against BDCA and HDCA respectively.

For this evaluation we considered a pool of $K = 45$ transformations based on the Slepian basis with various parameter values resulting in 2^{45} possible feature spaces. For each subject the 18

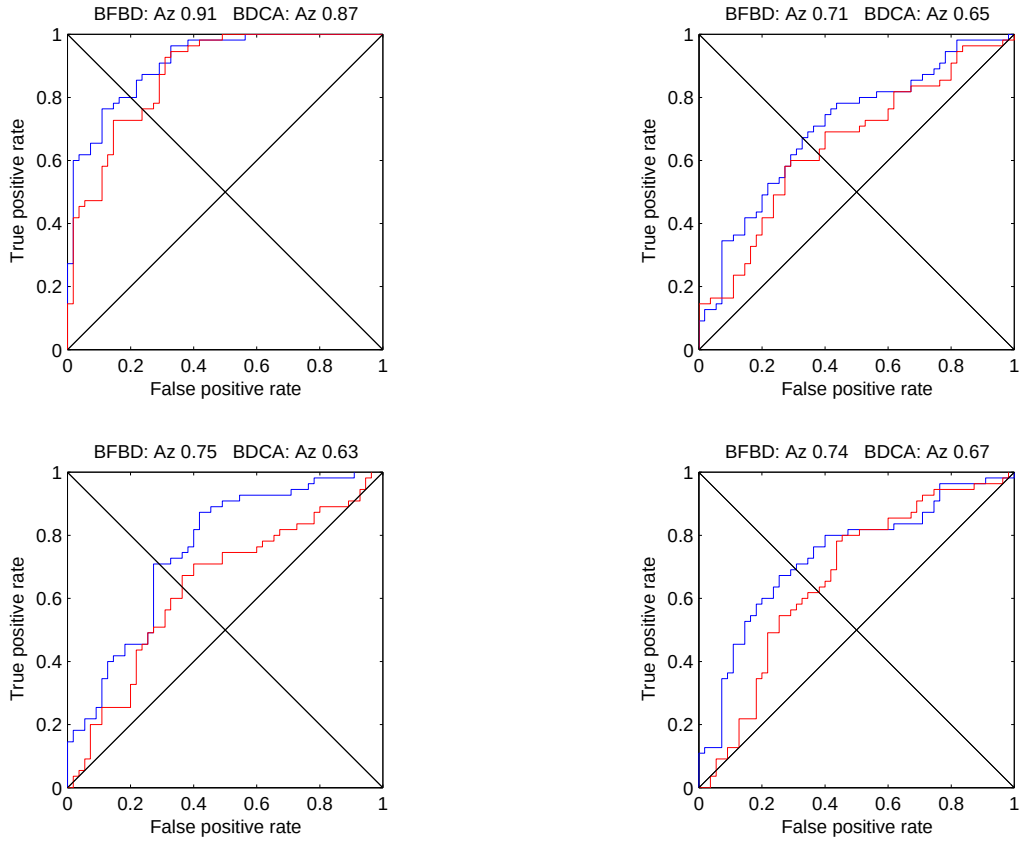

Figure 5.4: Receiver Operator Characteristic (ROC) obtained by BFBD (blue line) and BDCA (red line) on four more datasets.*Note* the uniformity of red curve being above the blue curve as false positive rate increases. *Also note* that the increase in Az performance is already visible at the lowest of the false positive rates..

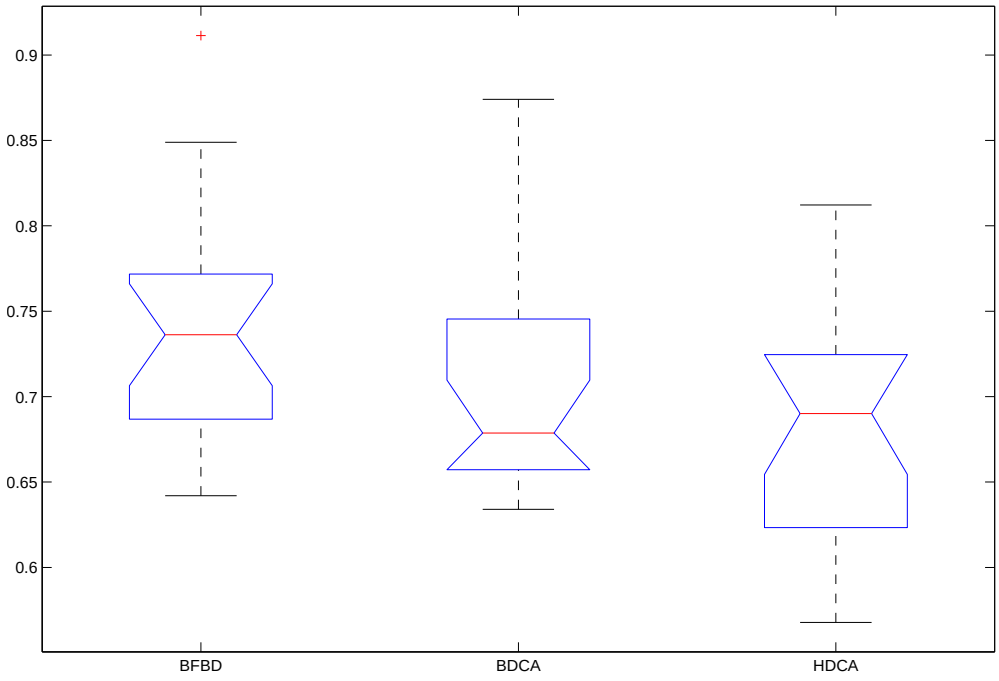

Figure 5.5: Distribution of the Az values for the three methods on the twenty EEG datasets. The boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers.

Feature	Description
ϕ_1	Raw EEG signal
ϕ_2	Slepian basis with center frequency 2Hz, overlap window of 200ms with overlap 50ms.
ϕ_3	Slepian basis with center frequency 5Hz, overlap window of 100ms with overlap 50ms.

Table 5.1: Description of the feature vector obtained by the optimization over I. A three dimensional vector has been extracted as the optimum feature.

data-sets[3] obtained by the other nine subjects are used to estimate the subject-invariant parameter I. We further fixed the first feature to be the identity transformation, hence operating on the raw EEG. For all subjects the procedure resulted to the exact same parameter I, which boosts the confidence to the estimation procedure. The selected feature is a three dimensional vector and the details of each dimension are summarized in table 5.6.1. The selected Slepian sequences are shown in figures 5.8 and 5.9.

The bilinear formulation of the problem not only allow us to obtain a classification performance but also to inspect visually the resulting space-time-frequency components of the model. For each of the three features the resulting spatial and temporal components are shown in figures 5.6.1 and 5.6.1 respectively for one of the datasets. The spatial projections, presented as scalp-plots show the weight of each electrode assigned by the classifier. This interpretation of the spatial components can help identify sensors that contribute to the discriminant ability and therefor provide a selection criteria for reducing the number of sensors used in brain computer interface applications. Further, the same scalp-plots can be used as input to source localization methods, that allows for the isolation

[3]Nine remaining subject each defining two datasets

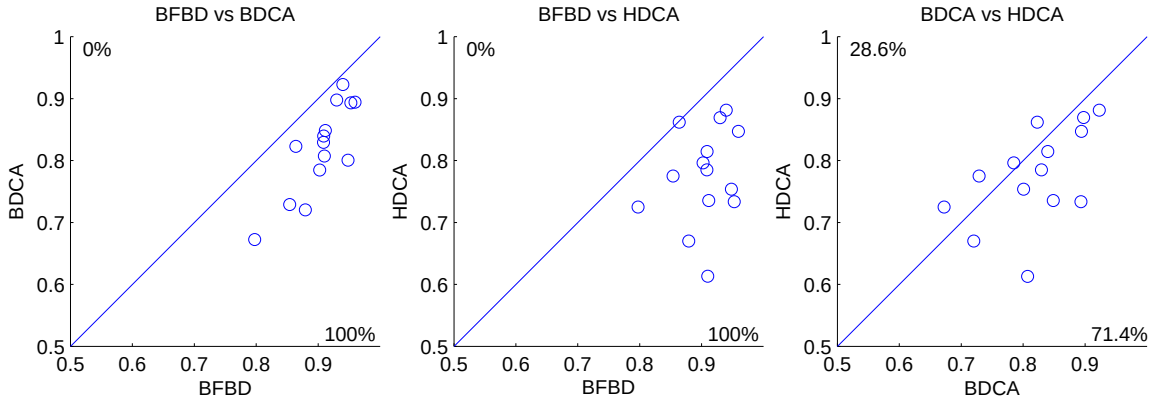

Figure 5.6: Each scatter plot compares area under the ROC curve (Az) achieved by BDCA, HDCA classifier vs BFBD. Each point represents Az for one of the twenty real EEG datasets. The portion of datasets lying above/below the diagonal line is given in %

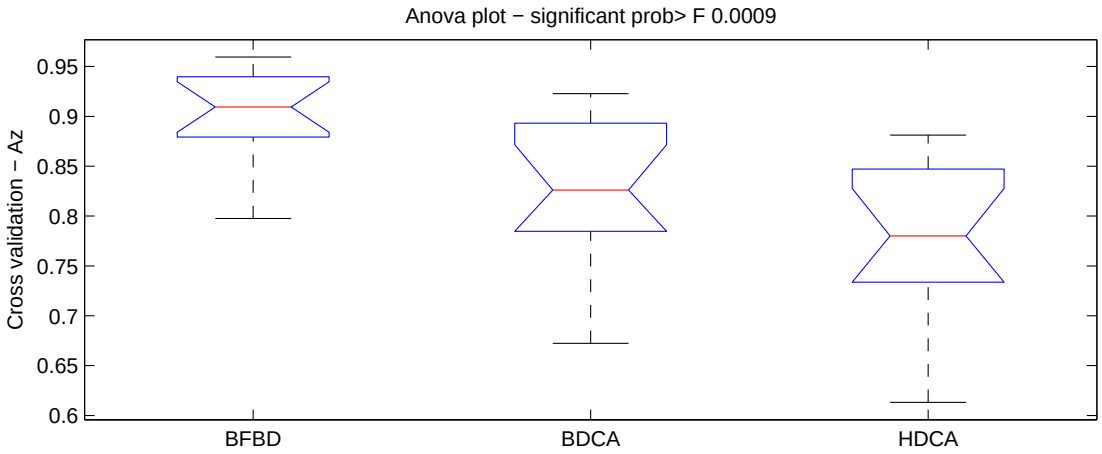

Figure 5.7: Distribution of the Az values for the three methods on the twenty EEG datasets. The boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers.

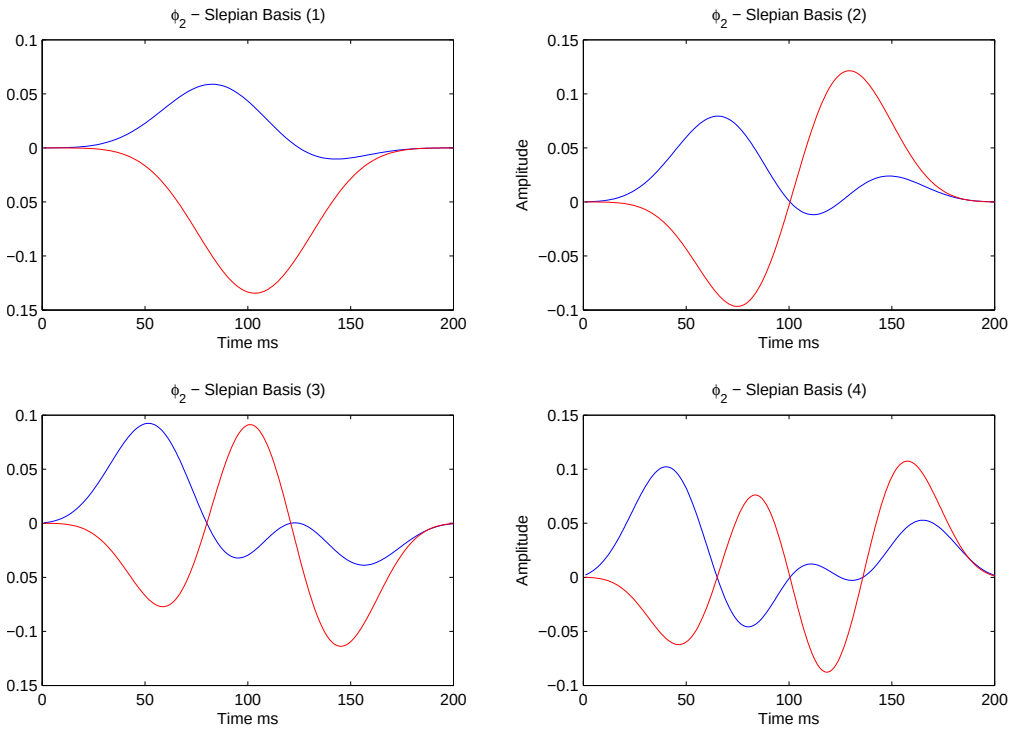

Figure 5.8: The first four basis used for the second feature transformation. The blue line indicates the real part of the basis while the red line plots the imaginary component of the basis

of the areas of neurons which generate the signal of interest. The temporal components show the timing of the signal of interest relative to the event onset. In the case of the first feature we observe that a signal of interest shows a negativity in the area between 200ms-250ms following by a positive signal around 400ms. This gives an indication as of where the signal of interest lies. Similarly, for feature 2, it seems to remain almost constant while feature 3 seems to decrease over time.

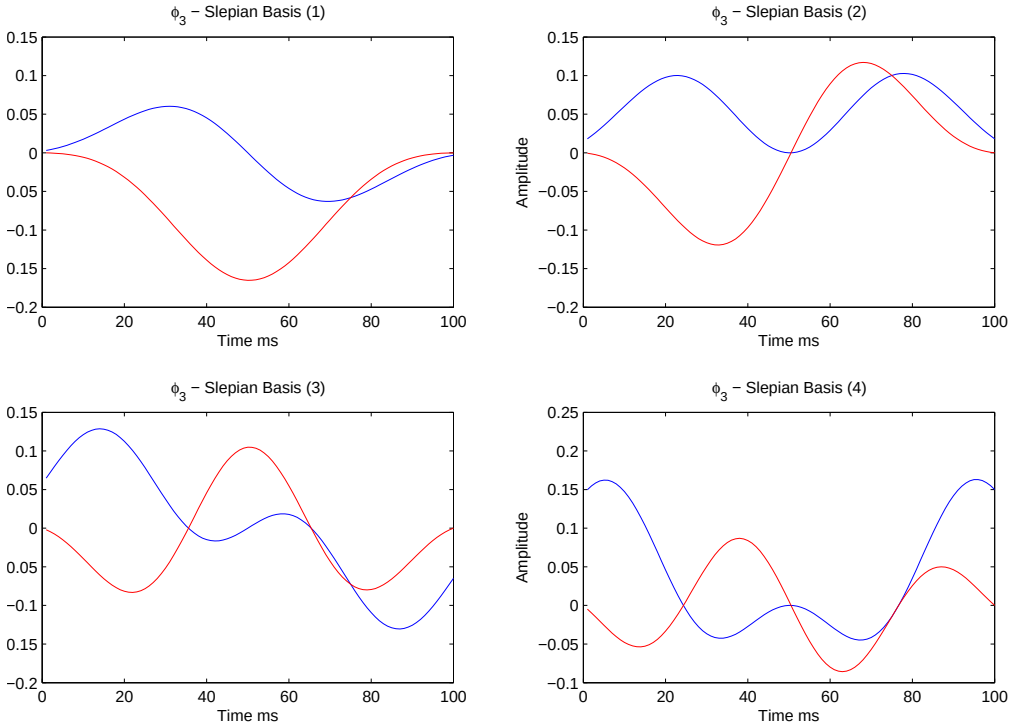

Figure 5.9: The first four basis used for the third feature transformation. The blue line indicates the real part of the basis while the red line plots the imaginary component of the basis

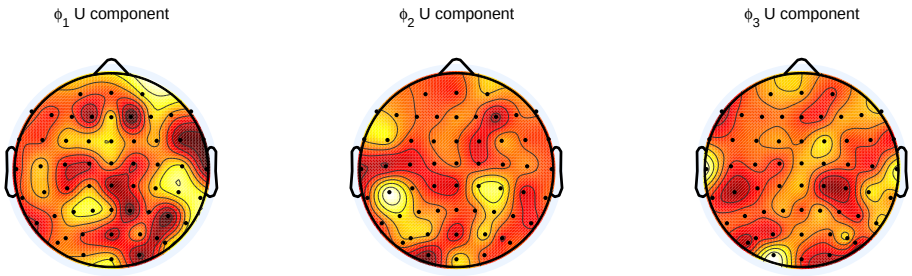

Figure 5.10: Resulting spatial components for the three features. *left:* The spatial component for the first feature (raw EEG feature) *center:* spatial component for the second feature (slepian basis centered at 2Hz) *right:* spatial component for third (slepian basis centered at 5Hz)

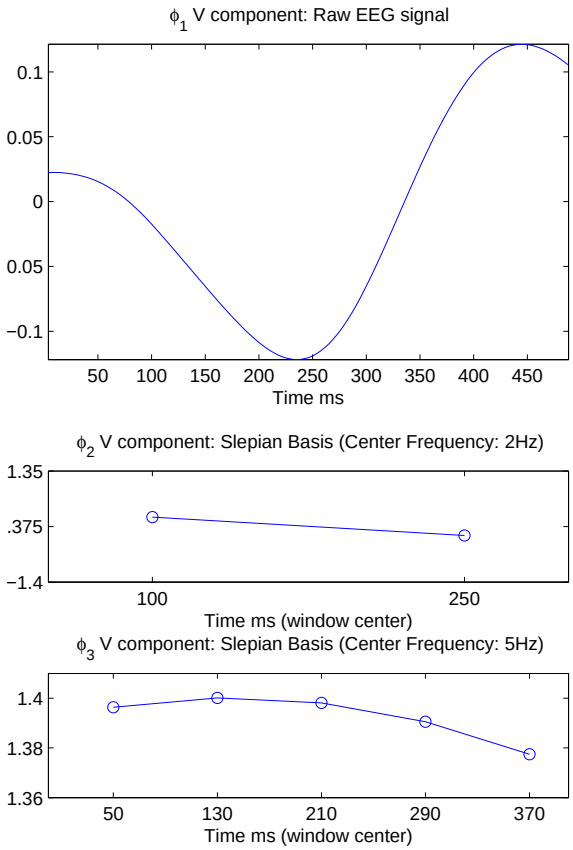

Figure 5.11: Resulting temporal components for the three features. *top:* The spatial component for the first feature (raw EEG feature) *center:* spatial component for the second feature (slepian basis centered at 2Hz) *bottom:* spatial component for third (slepian basis centered at 5Hz)

Chapter 6

CONCLUSIONS AND FUTURE RESEARCH

6.1 Discussion of Main Contributions

Our results in this dissertation demonstrate that Spatio-Spectro-Temporal representation is beneficial for single-trial classification of EEG signals. The Second Order Bilinear Discriminant Analysis framework we presented in chapter 4 uses spatial temporal and spectral components to reduce the dimensions of the space in a single discriminant model. We show that this representation outperforms state-of-the-art methods in terms of classification performance on real EEG data-sets. Similarly, the second algorithm we presented, namely Bilinear Feature-Based Discrimination, also uses spatial-temporal and spectral representation of the EGG in a hierarchical discriminant model.

It is interesting to note that the spatio-spectro-temporal formulation we presented in this dissertation, allows for one to visualize as well as interpret the results since each of the model parameters is associated with has a physical and conceptual meaning. The spatial components of our model are associated with the corresponding electrodes, hence one can visualize and identify areas on the scalp, where the signal of interest is concentrated. This provides the tools to perform tasks such as source localization and electrode reduction. Further, the temporal components of our model, generate a time-profile of the response. These profiles allow the experimentalist to hypothesize as to what the response time of the brain is to different stimuli, different experiment and different difficulty levels. Finally, spectral components in our model identify the frequency band of interest in various tasks, and compare or validate the results with existing literature.

A number of studies in single-trial EEG classification focus on using either Event Related Potential analysis or Event Related Synchronization. The motivation for choosing one vs. the other depend on the specific experimental paradigm. In this dissertation we presented a new method called SOBDA for analyzing EEG signals on a single-trial basis. The method combines linear and quadratic features. We evaluated the SOBDA algorithm in both simulated and real human EEG data-sets. We show a significant reduction in the classification error on human EEG when comparing our method to the state-of-the-art methods. The results on the EEG suggest that, in contrast to the common practice in the EEG analysis field, combining the two types of EEG characteristics increases performance significantly.

The results on simulated data of the same algorithm characterize the operational range of these algorithms in terms of SNR and show that the proposed algorithm operates well where other methods fail.

The parameterization of the discriminant criterion of SOBDA is intuitive, allowing one to incorporate prior knowledge as well as to derive spatial, temporal, and spectral information about the underlying neurological activity. This intuitive description of the model is particularly interesting, since there is a plethora of knowledge obtained in neural analysis from neurological experiments.

More than just a discriminant model, SOBDA provides a generic framework that encompasses a number of popular EEG analysis techniques. Specifically, by fixing some of the parameters of the SOBDA model, corresponds to introducing the various assumption of different methods about the characteristics of the underlining process, which enables us to get a deeper understanding of the underlining assumptions of the various methods. Our method, not only provides this unified view of EEG analysis but also suggests ways of future research and new approaches. More than

a neat mathematical formulation, our work provides the conceptual framework one can think of EEG analysis. A number of new ideas for understanding EEG analysis can be inspired by this formulation.

We further presented a second method we termed BFBF that performs single-trial analysis of EEG using more than one experimental session. The method allows the extraction of first order, second order but also arbitrary type of features through choice of a parametric feature bilinear functions. We evaluated BFBF on 20 real EEG datasets. We show a significant reduction in the classification error on real human EEG compared to two popular algorithms in EEG analysis.

In the evaluation of Bilinear Feature-Based Discrimination in chapter 5, we considered a perception categorization task. Previous studies [35],[34] that analyze EEG in this task focus on the extraction of ERP characteristics in the EEG. In this study however, we show that BDBD extracts power features, in addition to the event related potentials and increases the classification performance significantly over methods that only use one or the other characteristic.

Further, in the formulation of BFBD, we utilize the recording from multiple sessions to address the inter-subject and inter-session variability of EEG. By defining the model to distinguish between subject-specific and session-specific parameters vs subject-session invariant parameters our model can use more data during training and avoid over-fitting.

6.2 Extension to Future Work and Applications

A key component of this dissertation is the utilization of the bilinear combination of elements of a matrix. This formulation allowed us to reduce the number of dimensions in the various optimization problems we presented in an intuitive way. It further allowed for generating components that can be associated directly with physical objects and physical phenomena such as sensors, sensor positions,

neurons firing rate and response time to stimuli. We used Bilinear formulation to capture the spatial,temporal and spectral characteristics of the EEG, as a framework to unite a number of popular methods in EEG analysis , as a generic feature extraction function, and as a tool to address the inter-subject and inter-session variability problem that affects the EEG analysis. This formulation rightfully earned its place to the title of this work, which we call "The Bilinear Brain."

Building on this broad applicability of the Bilinear combination formulation, as part of our future work we will investigate the following extensions: Multi-class Bilinear Discriminant and Time-shift invariant bilinear discriminant. We briefly explain and motivate these extensions:

6.2.1 Multi-class Bilinear Discriminant

The formulation of the classification problems we considered in this dissertation are binary classification problems. An obvious extension to the bilinear formulation is its application to a multi-class problem. In the Brain Computer Interface scenario, a multi-class single-trial classifier allows for a higher information transfer rate per decision. A typical measure of information transfer rate in BCI is the bit-rate (i.e, the number of bit that are communicated per decision, given the performance of the classifier). It is defined as follows

$$bitrate = log_2(K) + Plog_2(P) + (1-P)log_2(\frac{1-P}{K-1}) \tag{6.1}$$

K is the number of different types of classes and P is the accuracy of the classifier in terms of probability of correct identification. Figure shows the bit transfer rate as a function of accuracy for two and three class classifier.

Figure (6.1) shows the information bit-rate of a 2-class and 3-class classifier, as a function of their accuracy. At perfect binary classifier with accuracy of 1.0 achieves a bit transfer rate of 1 bit,

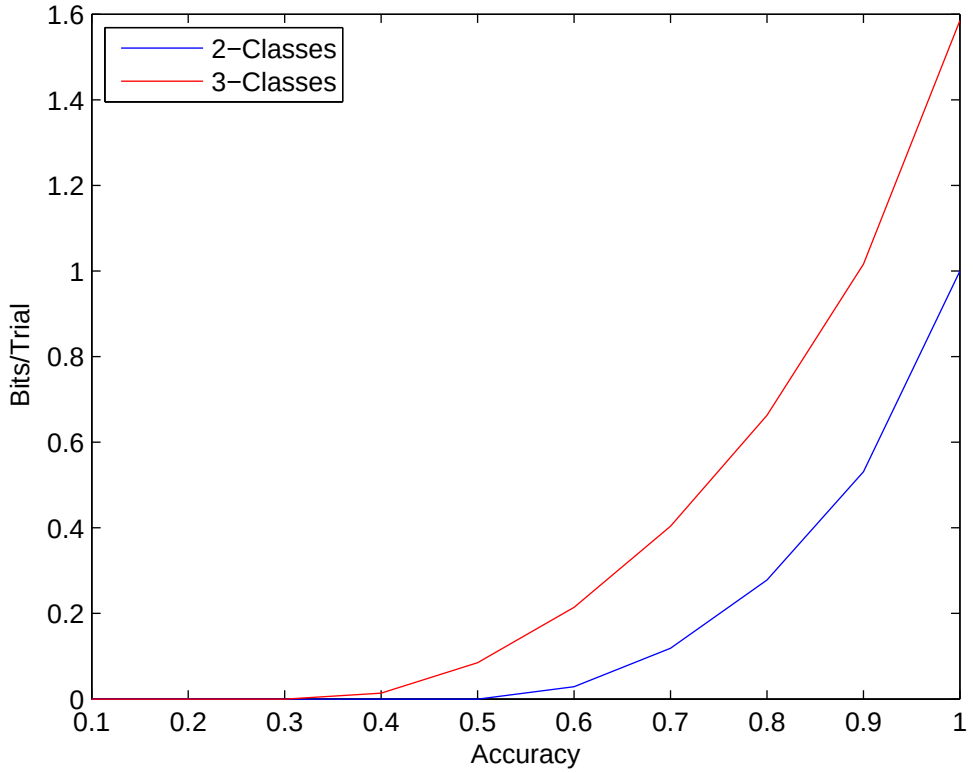

Figure 6.1: Information transfer rate curve as a function of accuracy, compares the improvement of multiple classes over the for a BCI system.

a similar three class classifier achieve 1.6 bit transfer rate. Hence more that two class classifier with reasonable performance potentially allows for higher bit rates. It is clear that a BCI that relying on a mulli-class classifiers has the potential to achieve higher information transfer rates.

In a multi-class classification problem, data are provided as a set of matrices, $\{\mathbf{X}_n \in \mathbb{R}^{D\times T}\}_{n=1}^{N}$, where each row of the matrix represents a channel/sensor and each column represents a time point in the time series. The entire matrix $\mathbf{X}_n$ represents an observation of the EEG activity, and it is associated with one out of possible K mental tasks denoted by a vector $\mathbf{y}_n \in \{e_1, ..., e_k\}$ where e_i is the $i-th$ columns of the identity matrix. Given such a data-set the aim of a the multi-class

classifier is to assign on of the label $\{e_1, ..., e_k\}$ to new observation $\mathbf{X}_{new}$

Since we define the bilinear discriminant based on logistic regression, the extension of the BDCA to multiple classes is straight forward. We achieve this by extending the probability of the class given an observation from a binomial distribution to a multi-nomial distribution and introduced the necessary parameters for each class. Septically, we model the probability of a class vector $\mathbf{y}$ given an observation $\mathbf{X}$ as

$$p(\mathbf{y} = e_k|\mathbf{X}, \mathbf{U}, \mathbf{V}) = \prod_{k'=1}^{K} \pi'_k(\mathbf{X})^{y^{(k')}} \tag{6.2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are matrices whose columns are the vectors $\mathbf{u}_1, ..., \mathbf{u}_K$, and $\mathbf{v}_1, ..., \mathbf{v}_K$ respectively,K corresponds to the number of classes, $y^{(k')}$ corresponds to the $k'th$ element of vector $\mathbf{y}$ and we defined $\pi'_k(\mathbf{X}) = \frac{e^{\mathbf{u}_l^\top \mathbf{X}\mathbf{v}_l}}{\sum_{k=1}^{N} e^{\mathbf{u}_k^\top \mathbf{X}\mathbf{v}_k}}$. For binary problems $K = 2$, this is known as a logistic regression model; for $K > 2$ the usual designation is multi-nomial logistic regression. Because of the normalization condition

$$\sum_{k=1}^{K} p(\mathbf{y} = e_k|\mathbf{X}, \mathbf{U}, \mathbf{V}) = 1 \tag{6.3}$$

the weight vector of one of the components need not to be estimated. Without loss of generality, we set the $\mathbf{u}_K = 0$ and $\mathbf{v}_K = 0$ and only the only parameters to be learned are $\mathbf{u}_1, ...\mathbf{u}_{K-1}$ and $\mathbf{v}_1, ...\mathbf{v}_{K-1}$. The log-likelihood of the model is defined as

$$L(\mathbf{U}, \mathbf{V}|\mathcal{D}) = \sum_{n=1}^{N} \sum_{k=1}^{K} y_n^{(k)} (\mathbf{u}_k^\top \mathbf{X}_n \mathbf{v}_k) - log \sum_{k=1}^{K} e^{\mathbf{u}_k^\top \mathbf{X}_n \mathbf{v}_k} e^{-\mathbf{u}_{k'}^\top \mathbf{K}_u \mathbf{u}_{k'}} \tag{6.4}$$

In a Maximum Likelihood (ML) estimate of the model parameters $\mathbf{U}$,$\mathbf{V}$ is given be the solution

of the following optimization problem

$$\hat{\mathbf{U}}, \hat{\mathbf{V}} = \arg_{\mathbf{UV}} \min -L(\mathbf{U}, \mathbf{V}|\mathcal{D}) \tag{6.5}$$

as in the case of BDCA one include regularization terms to avoid over-fitting the parameters. We obtained some preliminary result on simulated data, for a 3-class classifier. We measured the cross validation performance of our algorithm using three different measurements, namely, the probability of correct identification, bit-rate and cohen kappa values. We tried five different training sample sizes (i.e., 100,200,300,400,500 training trials per class). Figures (6.2),(6.3),(6.4) show the performance as a function on signal-to-noise ratio and for various sample sizes. It is interesting to note in figure (6.2) that for an SNR at $-19db$ and training sample of 500, the 3-class classier archives higher bit-rate than a binary classifier with 100% classification accuracy.

In future research we would like to identify multi class real EEG experimental paradigms and investigate the applicability of this formulation.

6.2.2 Time-Shift invariant bilinear Discriminant.

Inter-subject and inter-session variability was addressed in this report under the bilinear model. Another source of variability in EEG is known as the inter-trial variability that is cause for a number of reason. The temporal response of the brain to a stimulus may vary slightly from trial to trial. Another cause of this variability can be system related. In those cases the time-mark of a trial varies due to hardware limitations, hence the exact time of the beginning of the trial is only know approximately.

In our preliminary work we formulated a time-shift invariant version of the bilinear Discriminant model to address the trial-to-trial temporal variability in EEG. Specifically, we model the bilinear

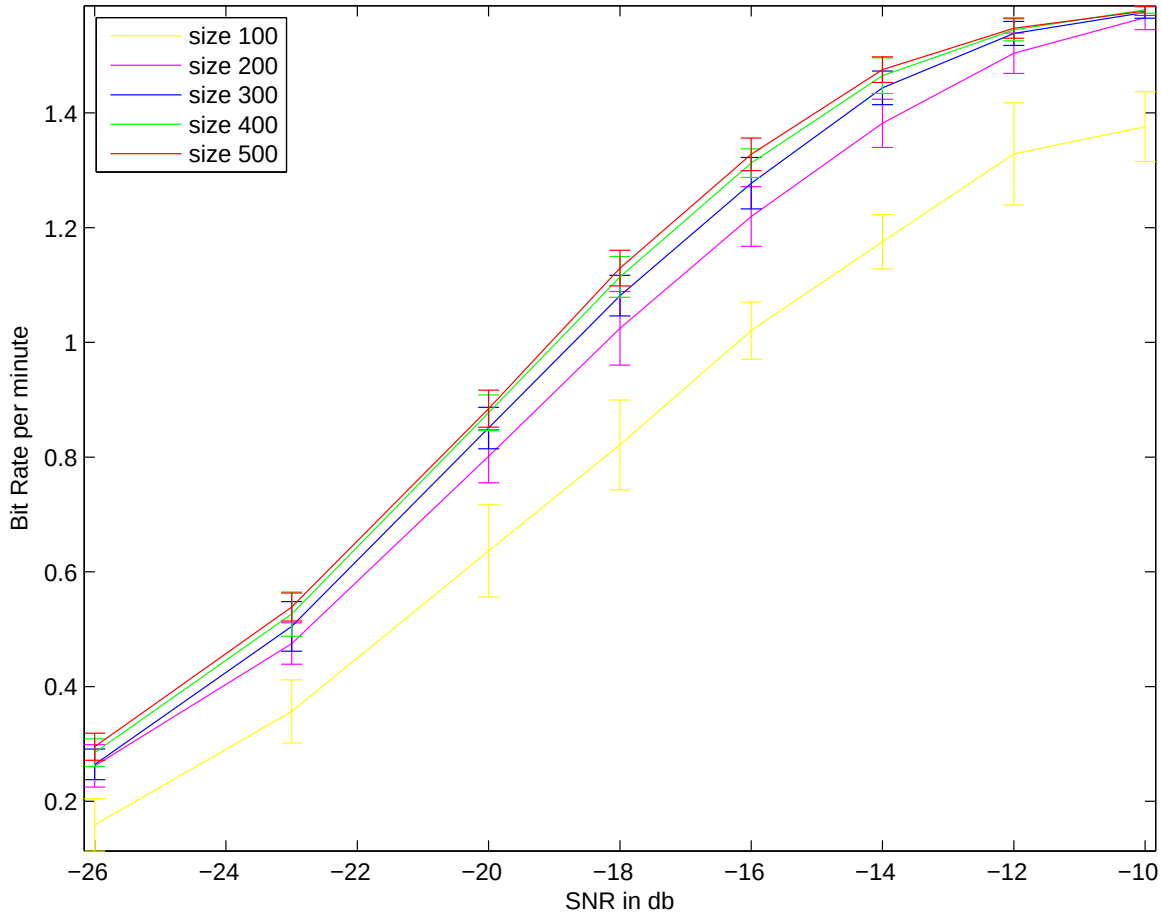

Figure 6.2: Simulated results of the 3-class classifier. Performance is measured in terms of bit-rate, and is given as a function of the components signal-to-noise ratio. Each curve corresponds to the results of a classifier training on a different training sample size.

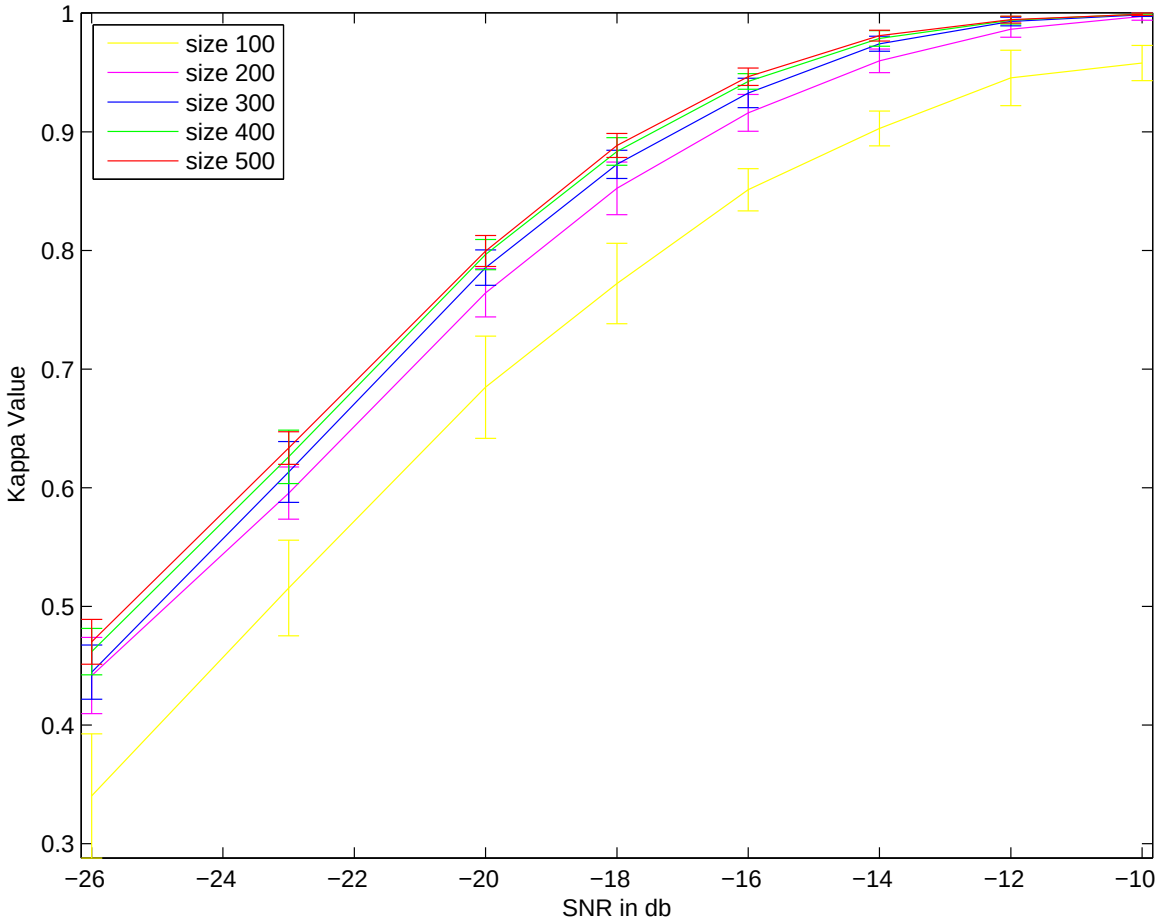

Figure 6.3: Simulated results of the 3-class classifier. Performance is measured in terms of Kappa values, and is given as a function of the components signal-to-noise ratio. Each curve corresponds to the results of a classifier training on a different training sample size.

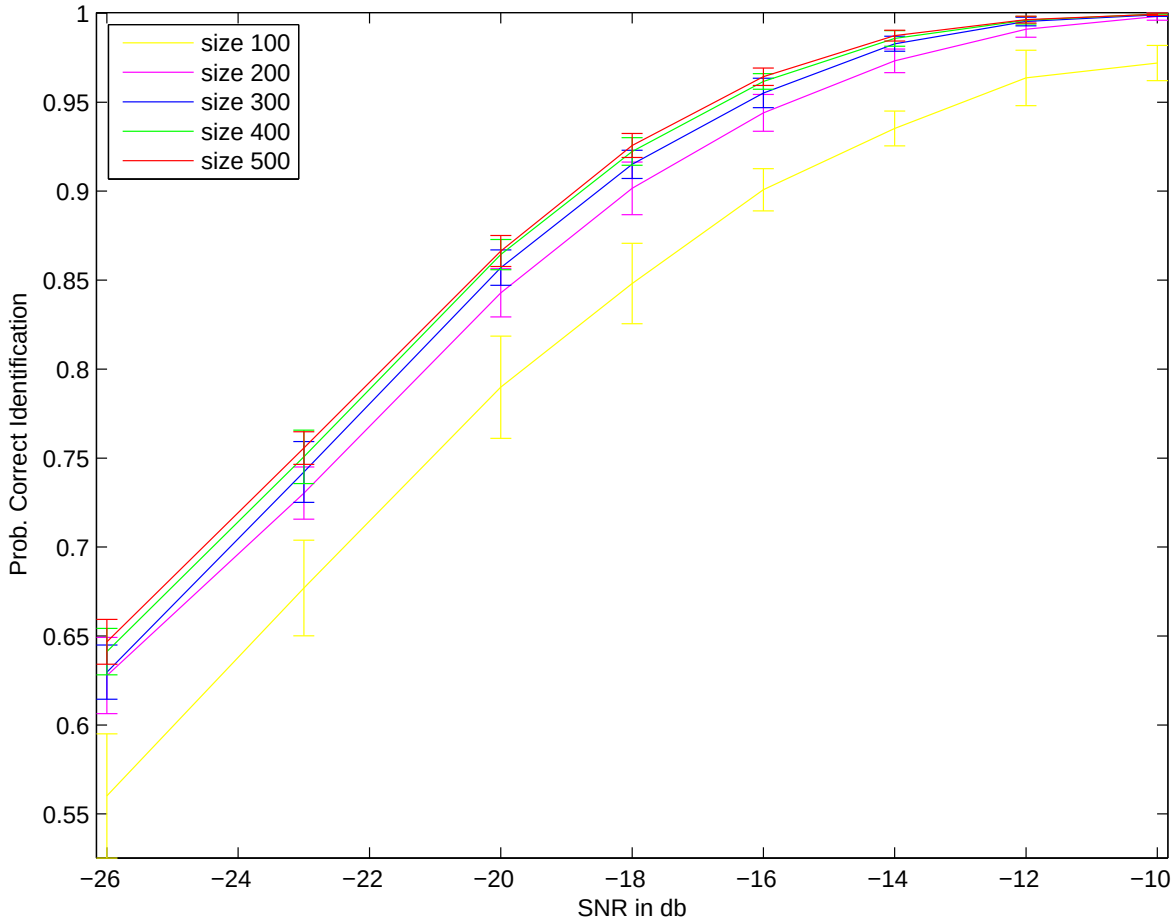

Figure 6.4: Simulated results of the 3-class classifier. Performance is measured in terms of probability of correct identification, and is given as a function of the components signal-to-noise ratio. Each curve corresponds to the results of a classifier training on a different training sample size.

combination of the nth training observation matrix as follows

$$g(\mathbf{X}_n) = \text{Trace}\left\{\mathbf{U}^\top \mathbf{X}_n \boldsymbol{\Delta}(\boldsymbol{s_n})\mathbf{V}\right\} \tag{6.6}$$

where $\mathbf{U} \in \mathbb{R}^{D\times R}, \mathbf{V} \in mathbbR^{T\times R}$ are the trial invariant parameters of the model, and the matrix delta $\boldsymbol{\Delta} \in \{0,1\}^{T\times T}$ is a Toeplitz matrix whose first column contains s_n zeros and $T - s_n$ ones. The model parameters are then $\mathbf{U}, \mathbf{V}$ and the trial specific parameters $\{s_n\}$ for each trial. Note that the matrix $\hat{X}_n = \mathbf{X}_n\boldsymbol{\Delta}(\boldsymbol{s_n})$ is the matrix $\mathbf{X}_n$ shifted temporally by s_n samples.

In future research we would like to investigate the convergence properties of this model, and how it compares to non-time-shift invariant bilinear formulation in terms of classification performance.

Appendix A

APPENDIX

In this section we derive the analytic gradient formulas of the negative log-likelihood function defined in (4.5). In general the gradient with respect to any of the variables can be expressed as:

$$\frac{\partial L(\theta)}{\partial \theta} = -\sum_{n=1}^{N} \frac{\partial \log(1+e^{-yf(\mathbf{X}_n;\theta)})}{\partial \theta} \tag{A.1}$$

$$= -\sum_{n=1}^{N} \frac{1}{1+e^{-yf(\mathbf{X};\theta)}} \frac{\partial\{1+e^{-yf(\mathbf{X}_n;\theta)}\}}{\partial \theta} \tag{A.2}$$

$$= \sum_{n=1}^{N} y \frac{e^{-yf(\mathbf{X};\theta)}}{1+e^{-yf(\mathbf{X};\theta)}} \frac{\partial f(\mathbf{X}_n;\theta)}{\partial \theta}\,, \tag{A.3}$$

Now one has to take the specific derivatives with respect to each of the variables in θ.

The gradient with respect to $\mathbf{u}_r$, the rth column of $\mathbf{U}$.

$$\begin{aligned}
\frac{\partial\{f(\mathbf{X}_n;\theta)+w0\}}{\partial \mathbf{u}_r} &= C\frac{\partial\{\text{Trace}\,\mathbf{U}^\top \mathbf{X}_n \mathbf{V}\}}{\partial \mathbf{u}_r} \\
&= C\frac{\partial\{\sum_{r'=1}^{R} \mathbf{u}_{r'}^\top \mathbf{X}_n \mathbf{v}_{r'}\}}{\partial \mathbf{u}_r} \\
&= C\mathbf{X}_n \mathbf{v}_r\,,
\end{aligned}$$

The gradient with respect to $\mathbf{v}_r$, the rth column of $\mathbf{V}$.

$$\begin{aligned}
\frac{\partial\{f(\mathbf{X}_n;\theta)+w0\}}{\partial \mathbf{v}_r} &= C\frac{\partial\{\text{Trace}\,\mathbf{U}^\top \mathbf{X}_n \mathbf{V}\}}{\partial \mathbf{v}_r} \\
&= C\frac{\partial\{\sum_{r'=1}^{R} \mathbf{u}_{r'}^\top \mathbf{X}_n \mathbf{v}_{r'}\}}{\partial \mathbf{v}_r} \\
&= C\mathbf{u}_r^\top \mathbf{X}_n\,,
\end{aligned}$$

The gradient with respect to $\mathbf{a}_r$, the rth column of $\mathbf{A}$.

$$\begin{aligned}
\frac{\partial\{f(\mathbf{X}_n;\theta)+w0\}}{\partial\mathbf{a}_r} &= (1-C)\frac{\partial\{\text{Trace}\,\mathbf{A}^\top(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{A}\}}{\partial\mathbf{a}_r} \\
&= (1-C)\frac{\partial\{\sum_{r'=1}^{K}\lambda_{r'}\mathbf{a}_{r'}^\top(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{a}_{r'}\}}{\partial\mathbf{a}_r} \\
&= 2(1-C)\lambda_r(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{a}_r\,,
\end{aligned}$$

The gradient with respect to $\mathbf{b}_r$, the rth column of $\mathbf{B}$.

$$\begin{aligned}
\frac{\partial\{f(\mathbf{X}_n;\theta)+w0\}}{\partial\mathbf{b}_r} &= (1-C)\frac{\partial\{\text{Trace}\,\mathbf{\Lambda}\mathbf{A}^\top(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{A}\}}{\partial\mathbf{b}_r} \\
&= (1-C)\frac{\partial\{\text{Trace}\,\mathbf{B}^\top\mathbf{X}_n^\top\mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top\mathbf{X}_n\mathbf{B}\}}{\partial\mathbf{b}_r} \\
&= (1-C)\frac{\partial\{\sum_{r'=1}^{K}\mathbf{b}_{r'}^\top\mathbf{X}_n^\top\mathbf{A}^\top\mathbf{\Lambda}\mathbf{A}^\top\mathbf{X}_n\mathbf{b}_{r'}\}}{\partial\mathbf{b}_r} \\
&= 2(1-C)(\mathbf{X}_n^\top\mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top\mathbf{X}_n)\mathbf{b}_r\,,
\end{aligned}$$

Appendix B

REGULARIZATION

The two models we present in this dissertation utilize regularization techniques to improve generalization performance of the classifier. In this appendix we provide details on the regularization procedure and the selection of its various parameters.

Specifically, we choose Gaussian process priors [38] on the various parameters of the two models and ensure smoothness by choosing proper proper covariance matrices. Spatial and temporal smoothness is typically a valid assumption in EEG [32]. Specifically, the spatial components of the model (i.e. columns of $\mathbf{U}$, and $\mathbf{A}$) follow a normal distribution with $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_u)$, $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_a)$ where the covariance matrices $\mathbf{K}_u$ and $\mathbf{K}_a$ define the degree and form of the smoothness of $\mathbf{u}$ and $\mathbf{a}$. This is done through choice of covariance function: Let r be a spatial or temporal measure in context of $\mathbf{X}$. For instance r is a measure of spatial distance between data acquisition sensors, or a measure of time difference between two samples in the data. Then a covariance function $k(r)$ expresses the degree of correlation between any two points with that given distance. For example, a class of covariance functions that has been suggested for modeling smoothness in physical processes, the Matérn class, is given by:

$$\mathrm{k}_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{\nu} \mathrm{B}\left(\frac{\sqrt{2\nu} r}{1} \right) , \tag{B.1}$$

where l is a length-scale parameter, and ν is a shape parameter. Parameter l can be roughly though of as the distance within which points are significantly correlated [38]. The parameter ν defines the

degree of ripple. The covariance matrix $\mathbf{K}$ is then built by evaluating the covariance function

$$(\mathbf{K})_{ij} = \sigma^2 \, \mathrm{k}_{\text{Matérn}}(r_{ij}) \tag{B.2}$$

where $r_{i,j}$ denotes the physical distance of sensor-i from sensor-j, and σ^2 defines the overall scale parameter. Similarly, the Gaussian prior can be used on the columns of the temporal matrix $\mathbf{V}$ (i.e $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_v)$). The Matérn function was preferred because it allows for a low parametrization of the covariance matrix (two parameters define the entire covariance), but also because of the physical and intuitive interpretation of its parameters. Specifically the parameter l is associated with the physical concept of distance between measurements (either in space or time). This understanding of the parameters is useful since it allows for an educated search strategy in setting the proper values for these parameters.

Following [38] the shape parameter was chosen to be $\nu = 100$ for the spatial components and $\nu = 2.5$ for the temporal components. Reasonable choices for the length-scale parameter l may be 25ms, 50ms or 100ms and in space 1cm, 2cm, and 3cm. Cross-validation was used to select among these choices. The overall scale parameters σ were chose to be the same for space and time components, but allowed to take on separate values for the first and second order component in the case of the SOBDA algorithm. We used a line-search procedure in combination with cross-validation to select appropriate values for σ.

Experiment	(σ, l, ν) of $(\mathbf{U})$	(σ, l, ν) of $(\mathbf{V})$	(σ, l, ν) of $(\mathbf{A})$	σ of $(\mathbf{B})$
1	(0.7,0.01,100)	(0.7,20,2.5)	(14,0.7,100)	10
2	(0.01,0.015,100)	(0.01,5,2.5)	(0.2,0.7,100)	5
3	(3.5,0.01,100)	(3.5,15,2.5)	(14,0.7,100)	10
4	(0.1,0.01,100)	(0.1,15,2.5)	(14,0.7,100)	10
5	(0.7,0.01,100)	(0.7,10,2.5)	(14,0.7,100)	10
6	(1.5,0.01,100)	(1.5,20,2.5)	(14,0.7,100)	10
Simulated	(3.5,0.01,100)	(3.5,20,2.5)	(14,0.7,100)	10

Figure B.1: Regularization parameters used for the evaluation of the SOBDA algorithm for each of the six Human subject EEG data-sets. Last row corresponds to the regularization parameters on all 3300 simulated EEG data-sets.

BIBLIOGRAPHY

[1] G. Ausiello, P. Crescenzi, V. Kann, Marchetti-Sp, Giorgio Gambosi, and Alberto M. Spaccamela. *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Springer, January 2000.

[2] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kubler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–8, Mar FebruaryMay 1999.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[4] B. Blankertz, G. Curio, and K. Müller. Classifying single trial EEG: Towards brain computer interfacing. In T. G. Diettrich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, 2002., 2002.

[5] B. Blankertz, G. Dornhege, C. Schfer, R. Krepki, J. Kohlmorgen, K. Müller, V. Kunzmann, F. Losch, and G. Curio. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Trans. Neural Sys. Rehab. Eng.*, 11(2):127–131, 2003.

[6] B. Blankertz, K.-R. Müller, G. Curio, T.M. Vaughan, G. Schalk, J.R. Wolpaw, A. Schlogl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schroder, and N. Birbaumer. The bci competition 2003: progress and perspectives in detection and discrimination of EEG single trials. *Biomedical Engineering, IEEE Transactions on*, 51(6):1044–1051, 2004.

[7] Dankmar Bohning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics (Historical Archive)*, 44:197–200, March 1992.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[9] Christoforos Christoforou, Robert Haralick, Paul Sajda, and Lucas C. Parra. Second order bilinear discriminant analysis. *J. Mach. Learn. Res.*, submitted, 2008.

[10] Christoforos Christoforou, Paul Sajda, and Lucas C. Parra. Second order bilinear discriminant analysis for single trial eeg analysis. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 313–320. MIT Press, Cambridge, MA, 2008.

[11] Mary Kathryn Cowles. Generalized, linear, and mixed models. charles e. mcculloch and shayle r. searle. *Journal of the American Statistical Association*, 101:1724–1724, December 2006.

[12] G. Dornhege, Blankertz B, and K.R. Krauledat M. Losch F. Curio G.Müller. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE Trans. Biomed. Eng. 2006*, 2006.

[13] G. Dornhege, B. Blankertz, G. Curio, and K. Müller. Combining features for BCI. *Advances in Neural Inf. Proc. Systems (NIPS 02)*, 15, 2003.

[14] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[15] Richard O. Duda, Peter E. Hart, and David G. Storck. *Pattern Classification*.

[16] M. Dyrholm and L.C. Parra. Smooth bilinear classification of EEG. In *In Proc. 28th Annu. Int Conf. IEEE Engineering in Medicine and Biology Society*, 2006.

[17] Mads Dyrholm, Christoforos Christoforou, and Lucas C. Parra. Bilinear discriminant component analysis. *J. Mach. Learn. Res.*, 8:1097–1111, 2007.

[18] Bradley Efron, Trevor Hastie, Lain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[19] Philiastides Marios G., Ratcliff Roger, and Sajda Paul. Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *Journal of Neuroscience*, 26(35):8965–8975, August 2006.

[20] Adam D. Gerson, Lucas C. Parra, and Paul Sajda. Cortically-coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14:174–179, June 2006.

[21] R.M. Haralick. Propagating covariance in computer vision. *ICPR-A*, 94:493–498.

[22] A.K. Jain and D. Zongker. Feature-selection: Evaluation, application, and small sample performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:153–158, 1997.

[23] Balaji Krishnapuram and Alexander J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):957–968, 2005. Fellow-Lawrence Carin and Senior Member-Mario A. T. Figueiredo.

[24] S. Lemm, B. Blankertz, G. Curio, and K. Müller. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Trans Biomed Eng., 52(9):1541–8, 2005.*, 2005.

[25] A. Luo and P. Sajda. Learning discrimination trajectories in EEG sensor space: Application to inferring task difficulty. *J. Neural Eng.*, 3:L1–L6, 2006.

[26] A. Luo and P. Sajda. Using single-trial EEG to estimate the timing of target onset during rapid serial visual presentation. In *Proc. Engineering in Medicine and Biology Society(EMBC2006)*, 2006.

[27] E. McDermott, E.A Woudenberg, and S. Katagiri. A telephone-based directory assistance system adaptively trained using minimum classification error/generalized probabilistic descent. In *ICASSP '96: Proceedings of the acoustics, Speech, and Signal Processing*, pages 3346–3349, 1996.

[28] M. Mørup. Analysis of brain data - using multi-way array models on the EEG. Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, 2005. Supervised by Prof. Lars Kai Hansen.

[29] P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch. EEG coherency I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and clinical Neurophysiology*, 103:499–515, 1997.

[30] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda. Recipes for the linear analysis of EEG. *Neuroimage*, 28(2):326–341, November 2005.

[31] Lucas C. Parra, Christoforos Christoforou, Adam D. Gerson, Mads Dyrholm, An Luo, Mark Wagner, Marios G. Philiastides, and Paul Sajda. Spatio-temporal linear decoding of brain state: Application to performance augmentation in high-throughput tasks. *Signal Processing Magazine, Special Issue on Brain Computer Interfaces*, 2007 (to appear).

[32] W. D. Penny, N. J. Trujillo-Barreto, and K. J. Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24:350362, 2005.

[33] G. Pfurtscheller and F. H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol*, 110(11):1842–1857, November 1999.

[34] M.G. Philiastides, R. Ratcliff, and P. Sajda. Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *Journal of Neuroscience*, 26(35):8965–8975, August 2006.

[35] M.G. Philiastides and P. Sajda. Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex*, 16(4), April 2006.

[36] J.W. Pitton. Time-frequency spectrum estimation: an adaptive multitaper method. *Time-Frequency and Time-Scale Analysis, 1998. Proceedings of the IEEE-SP International Symposium on*, pages 665–668, Oct 1998.

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

[38] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[39] Pradeep Shenoy, Matthias Krauledat, Benjamin Blankertz, Rajesh P. N. Rao, and Klaus-Robert Müller. Towards adaptive classification for bci. 2006.

[40] R. Tomioka, K. Aihara, and K. Müller. Logistic regression for single trial EEG classification. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1377–1384. MIT Press, Cambridge, MA, 2007.

[41] Ryota Tomioka and Kazuyuki Aihara. Classifying matrices with a spectral regularization. In *24th International Conference on Machine Learning*, 2007.

[42] Ryota Tomioka, Kazuyuki Aihara, and Klaus-Robert Müller. Logistic regression for single trial EEG classification. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1377–1384. MIT Press, Cambridge, MA, 2007.

[43] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clin Neurophysiol*, 113(6):767–791, June 2002.

[44] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2):301–320, 2005.