Separable Spatio-Spectral Patterns in EEG signals During Motor-Imagery Tasks

by

Amirhossein Shokouh Aghaei

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Electrical and Computer Engineering
University of Toronto

Abstract

Separable Spatio-Spectral Patterns in EEG signals During Motor-Imagery Tasks

Amirhossein Shokouh Aghaei

Doctor of Philosophy

Graduate Department of Electrical and Computer Engineering

University of Toronto

2014

Brain-Computer Interface (BCI) systems aim to provide a non-muscular channel for the brain to control external devices using electrical activities of the brain. These BCI systems can be used in various applications, such as controlling a wheelchair, neuroprosthesis, or speech synthesizer for disabled individuals, navigation in virtual environment, and assisting healthy individuals in performing highly demanding tasks. Motor-imagery BCI systems in particular are based on decoding imagination of motor tasks, e.g., to control the movement of a wheelchair or a mouse curser on the computer screen and move it to the right or left directions by imagining right/left hand movement. During the past decade, there has been a growing interest in utilization of electroencephalogram (EEG) signals for non-invasive motor-imagery BCI systems, due to their low cost, ease of use, and widespread availability.

During motor-imagery tasks, multichannel EEG signals exhibit task-specific features in both spatial domain and spectral (or frequency) domain. This thesis studies the statistical characteristics of the multichannel EEG signals in these two domains and proposes a new approach for spatio-spectral feature extraction in motor-imagery BCI systems. This approach is based on the fact that due to the multichannel structure of the EEG data, its spatio-spectral features have a matrix-variate structure. This structure, which has been overlooked in the literate, can be exploited to design more efficient feature extraction methods for motor-imagery BCIs.

Towards this end, this research work adopts a matrix-variate Gaussian model for the spatio-spectral features, which assumes a separable Kronecker product structure for the covariance of these features. This separable structure, together with the general properties of the Gaussian model, enables us to design new feature extraction schemes which can operate on the data in its inherent matrix-variate structure to reduce the computational cost of the BCI system while improving its performance. Throughout this thesis, the proposed matrix-variate model and its implications will be studied in various different feature extraction scenarios.

Acknowledgements

I would like to express my sincere gratitude and appreciation to my supervisor Professor Konstantinos N. Plataniotis for his invaluable advice, guidance, and encouragement. Without his kind supports during my PhD studies, this research would have been impossible. I am deeply grateful for all his helps during this period.

Furthermore, I would like to offer my sincere thanks to Professor Subbarayan Pasupathy for his insightful guidance and helpful suggestions throughout this work. His comments and guidelines have always been a source of inspiration for me. I would like to also thank my PhD committee members and my examination committee: Prof. Tom Chau, Prof. Dimitrios Hatzinakos, Prof. Deepa Kundur, Prof. Stergios Stergiopoulos, Prof. Shahrokh Valaee, and Prof. Z. Jane Wang for reviewing my work and offering their constructive comments and feedback.

I greatly appreciate the kind helps and supports of my colleagues and friends during my PhD program. Specially, I should thank my best friends: Houman Hosseinpour, Mahdi Ramezani, Mohammad Shahin Mahanta, Siavash Fazeli, Hossein Seyedmahdi, Ali Kalantarian, Mandana Einolghozati, Fahimeh Salimi, Saeid Rezaei, Mohsen Soleimani, Kaveh Ghasemloo, and Mohammad Norouzi; as well as my colleagues: Mahdi Hosseini, Kianoush Hosseini, Sanam Sadr, Gokul Sridharan, and Mahdi Marsousi. Moreover, I would like to thank Chris Aimone, Lou Pino, Om Bhatt, Sina Mackay, and Richard Gao for hours of technical discussion and helpful feedbacks.

Last, but certainly not least, my deepest gratitude goes to my family for their never-ending support, encouragement, and unconditional love. Despite the fact that we have been living far apart during my graduate studies, their endless support has always been with me during all the ups and downs of this journey. This thesis is dedicated to them.

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description	Page Number
2DLDA	Two-Dimensional Linear Discriminant Analysis	9
AAR	Adaptive Autoregressive	23
AR	Autoregressive	23
ARCH	Autoregressive Conditional Heteroscedastic	125
BCI	Brain Computer Interface	ii
CCR	Correct Classification Rate	58
CS	Channel Selection	6
CSP	Common Spatial Patterns	6
DA-FE	Domain-Agnostic Feature Extraction	6
DS-FE	Domain-Specific Feature Extraction	6
DTF	Directed Transfer Function	28
EEG	Electroencephalogram	ii
EOG	Electrooclugram	17
ERD	Event Related Desynchronization	19
ERS	Event Related Synchronization	19
FBCSP	Filter Bank Common Spatial Patterns	6
FE	Feature Extraction	6
FIR	Finite Impulse Response	34
ICA	Independent Component Analysis	6
IIR	Infinite Impulse Response	23
Lin	Linear	44
LDA	Linear Discriminant Analysis	7
MA	Moving Average	34
MI	Motor Imagery	6
MI-BCI	Motor Imagery Brain Computer Interface	6

Abbreviation	Description	Page Number
MVLDA	Matrix-to-Vector Linear Discriminant Analysis	9
NBPW	Nave Bayesian Parzen Window	44
OVR	One Versus Rest	57
PCA	Principle Component Analysis	7
PLV	Phase Locking Value	28
SCSSP	Separable Common Spatio-Spectral Patterns	10
SL	Surface Laplacian	6
STFT	Short-Time Fourier Transform	34
WT	Wavelet Transform	34

List of Important Symbols

Symbol	Description	Page Number
N_f	Number of Spectral Features	6
N_{ch}	Number of EEG channels	32
N_s	Number of Spatial Features	6
Ω_i	Class i	6
C	Number of Classes	6
d	Number of Features in Chapters 4 and 5	79
d_{csp}	Number of CSP Features	57
d_{scssp}	Number of SCSSP Features	92
N	Total Number of Training Samples in Chapters 4 and 5	53
N_i	Number of Training Samples for Class i	53
$\mathbf{\Sigma}$	Covariance Matrix of Vectorial Data	47
$\mathbf{\Phi}_i$	Spectral Covariance Matrix of Class i	46
$\mathbf{\Psi}_i$	Spatial Covariance Matrix of Class i	46
$\mathbf{\Psi}_i^f$	Spatial Covariance Matrix of Class i for The f^{th} Rhythm	91
$\mathbf{M}_i$	Mean of Class i	46
$\mathbf{S}_{BL}$	Between-class Spectral Scatter Matrix	54
$\mathbf{S}_{BR}$	Between-class Spatial Scatter Matrix	54
$\mathbf{S}_{WL}$	Within-class Spectral Scatter Matrix	54
$\mathbf{S}_{WR}$	Within-class Spatial Scatter Matrix	54
$\mathbf{I}_K$	Identity Matrix of Size K	89
$\mathbf{X}$	Matrix-Variate Data in Chapters 4 and 5	6
$\mathbf{Z}$	Complex-Valued Matrix-Variate Data in Chapter 6	113

Symbol	Description	Page Number
$\mathbf{x}, \mathbf{z}, \ldots$	Vector-Variate Data	42, 90
$\mathbf{z}_n$	n^{th} Column of $\mathbf{Z}$ in Chapter 6	115
$s(t, c\|\Omega_i)$	EEG Signal at Frequency f, Channel c, and time t, during task Ω_i	112
$z(t, f, c\|\Omega_i)$	Complex-Valued Feature Obtained at Frequency f, Channel c, and time t, during task Ω_i	112
$\mathbb{R}$	Real Domain	6
$\mathbb{C}$	Complex Domain	113
$\text{vec}(\cdot)$	Column-wise Vectorization Operator	47
$E\{\cdot\}$	Statistical Expectation	31
$\exp\{\cdot\}$	Exponential Function	47
$\log(\cdot)$	Natural Logarithm	90
$\text{var}\,(\cdot)$	Variance	90
$\otimes$	Kronecker Product Operator	47

Chapter 1

Introduction

Since the first studies of electrical activities of the brain about a century ago, there has been a great interest in analyzing and decoding these activities for clinical, diagnostic, and rehabilitation applications. Several studies have shown that certain characteristics of electrical signals emitted from the brain are unique to each brain activity and each individual person. As a result, these signals have been used in areas such as:

- Clinical applications: Brain signals are widely studied for diagnosis and treatment of various mental disorders such as *dementia* [3,4] and *epileptic seizure* [5–7]. Moreover, it has even been shown that brain signals can be used for early diagnosis of many psychiatric disorders, such as: *dyslexia* [8], which is a developmental reading disorder; and *autistic disorder* [9], which is related to impaired social interaction and communication.

- Biometric systems: Brain signals provide a universal biometric for identification of individuals, which cannot be easily forged or stolen [10–14]. Although it may not be suitable for many commercial applications, brain signal has the potential to be used as a biometric in highly secure environments. Moreover, brain signals can be used in conjunction with other biometric modalities to improve the reliability of the identification system.

- Brain-computer interfaces (BCI): Brain signals can provide a non-muscular channel for interaction with computers and the external world [15, 16]. A BCI, also known as *direct neural interface* or *brain-machine interface*, is basically an interface between the brain and the world outside, which translates the electrical activity of the brain into signals that control external devices. Early BCIs were mainly designed to help paralyzed or disabled patients to control assistive devices such as

wheelchairs, neuroprosthesis, and speech synthesizers [17–19]. However, new commercial applications have recently emerged for BCIs. Some of the commercial applications include: assisting healthy individuals in performing highly demanding mental tasks [20–23] and brain-controlled navigation in virtual environments [24, 25].

In order to record the electrical activities of the brain, the following three approaches have been used in the BCI literature:

- ***Invasive***: An array of sensors are implanted directly into the grey matter of the brain.
- ***Partially-invasive***: Sensors are implanted inside the skull but outside the brain.
- ***Non-invasive***: Sensors are located outside the skull, and there is no need for surgical implantation.

Invasive and partially-invasive methods require surgery to implant the sensors, most of which last for only a few years and hence need to be replaced by new sensors every couple of years. As a result, the use of invasive solutions for brain-computer interaction is currently very limited and is restricted to clinical trials. In contrast, non-invasive methods are of special interest in BCI applications due to their ease of use for both commercial and medical applications.

Non-invasive methods that are used in the literature for brain-computer interfaces include:

- functional Magnetic Resonance Imaging (fMRI): This method measures brain activities in different parts of the brain by detecting the associated blood flow changes. This measurement is based on the fact that active neurones require more oxygenated blood flow during their activity.
- functional Near-Infrared Spectroscopy (fNIR): In this method, near-infrared electromagnetic waves are used to measure the concentration of oxygenated and deoxygenated hemoglobin in different parts of the brain cortex. These measurements will in turn determine active parts of the cortex, similar to fMRI. An important difference between fNIR and fMRI is that the fNIR method has a very limited penetration depth, in the order of a few centimetres, whereas the fMRI method can measure the brain activities at any depth.
- Magnetoencephalography (MEG): This method directly measures the magnetic fields generated by the neural activities of the brain, using an array of highly sensitive magnetometers. MEG mostly records magnetic fields originated from tangential current sources, which are usually located on sulcal walls in the cortex [26]. One of the main advantages of using MEG for source localization

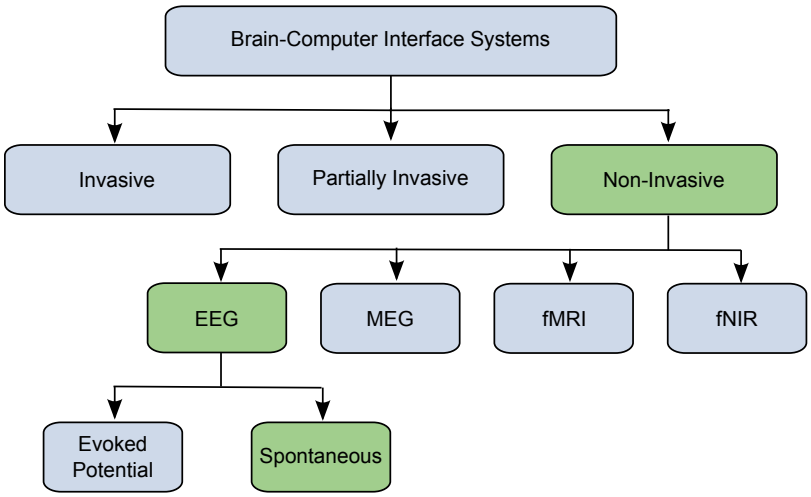


Figure 1.1: Commonly used approaches for brain-computer interfacing.

is that the skull and other tissues are almost transparent to the magnetic field, and hence they do not cause any attenuation or distortion on the MEG recordings.

- Electroencephalography (EEG): In this method the electrical fields generated by the neural assemblies across the brain are measured, using several small electrodes on the scalp. The EEG is mostly sensitive to electric fields that are generated by the radial current sources, which are usually located on the gyral surfaces in the cortex.

Among these methods, fMRI and MEG methods provide relatively higher spatial resolution compared to fNIR and EEG methods. However, fMRI and MEG require highly expensive equipments and controlled environments for their operation. Furthermore, fMRI and MEG are not portable and cannot be used for continuous daily usage, as required in most BCI applications. fNIR is a portable solution, however, it suffers from low temporal resolution (in the order of few seconds) which is dictated by the slow vascular response. As a result, EEG is the most widely used method for monitoring the brain activities in BCI application, and hence it will be the focus of our studies in this thesis.

It should be noted that EEG has two major limitations, which have to be taken into account in the design of any BCI system: (a) limited spatial resolution and (b) limited depth of penetration. In order to address these limitations, recent works have suggested to develop multimodal BCI systems that take advantage of different recording modalities to enhance the performance of the BCI system [27]. Considering the crucial importance of *portability* in most BCI applications, the best candidate for multimodal BCI systems is the combination of EEG and fNIR signals, ref. [27–29]. Although our focus in this thesis is on EEG-based BCI systems, the results of this research can be utilized in multimodal BCI systems as well.

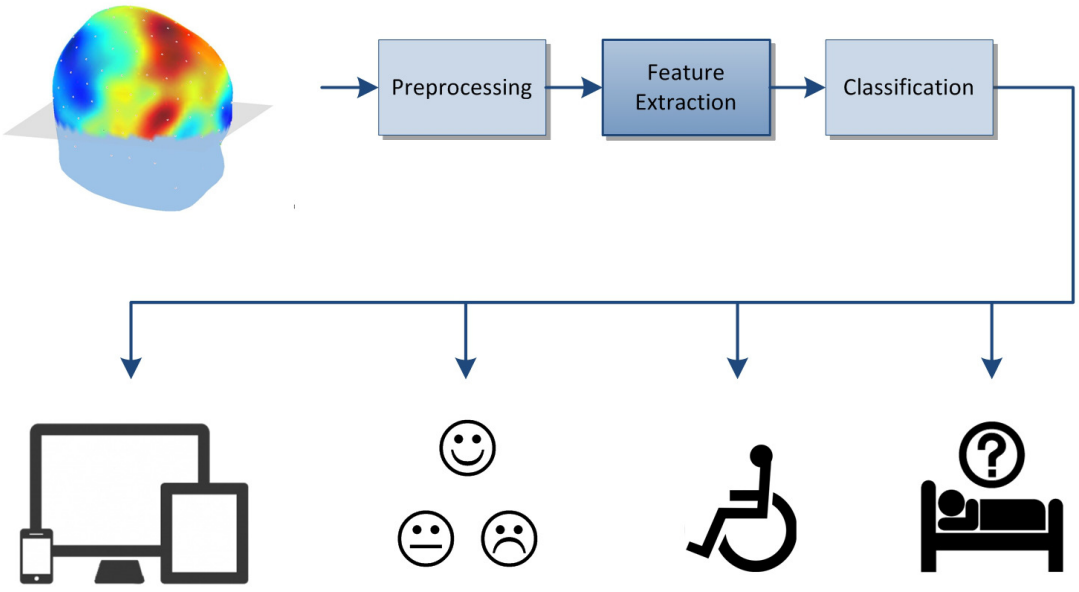


Figure 1.2: Processing pipeline and different applications of brain-computer interfaces, including human-computer interaction, emotion recognition, rehabilitation, and clinical diagnosis of brain disorders.

1.1 Motivation

During the past two decades, various EEG-based BCI systems have been developed to help disabled individuals. These systems have also recently been used in many commercial applications, such as navigation in virtual environments, neuromarketing, adaptive human machine interfacing, and cortically coupled computer vision.

A large portion of the currently existing BCI systems are based on *evoked* potentials, where the BCI works based on the EEG signal generated in response to a stereotyped sensory stimulation. As a case in point, assume that the user wants to spell out a word using a BCI system. One solution is to provided him/her with a screen which displays letters that are flashing with different frequencies. When the patient gazes at a desired letter, analysis of the resulting brain signals, called evoked potentials, can reveal which letter he/she is looking at. Although these evoked BCI systems are very accurate, they are not suitable for long term usage since the user will be constantly confronted with stimuli, which in turn can become exhaustive or even cause physiological problems for the user.

In order to alleviate this problem, recently there has been a growing interest in utilization of *spontaneous* BCI systems, which are based on detection of mental imaginations and do not require any external stimulation. Most of the spontaneous BCIs are based on motor imagination tasks, such as hand movement and foot movement. As a simple example, the user can control a cursor on the computer

screen, using the following motor imagery tasks: (a) Imagination of right hand movement, for moving the cursor to the right; (b) Imagination of the left hand movement, for moving the cursor to the left; (c) Imagination of the right foot movement, for moving the cursor up; and (d) Imagination of the left foot movement, for moving the cursor down. Similar commands can be used for moving a wheelchair to different directions (right, left, front, back).

One of the main benefits of using motor tasks in BCI systems is that they can be easily imagined and do not require any specific training. Particularly, in the applications where the BCI system is used for movement control, the imagined motor tasks can be naturally associated with the desired movement tasks. Moreover, the EEG signals generated by different users during motor tasks are relatively consistent, compared to other mental imagery tasks such as imagination of an object or an abstract concept. In general, motor imagination activates similar neural assemblies as motor execution (see [30] and references therein). As a result, motor-imagery BCI systems can be used by a wide range of motor-disabled individuals if their motor cortex has not been re-assigned to other tasks (ref. Section 2.1.2). As a result, several works have reported successful use of motor-imagery BCI systems for individuals with different levels of myopathy, spinal cord injury, tetraplegia, amputation, spino-cerebellar ataxia or multiple sclerosis (e.g., see [31, 32]). However, it should be noted that motor-imagery BCI may not be suitable for certain users, such as people with congenital motor impairment, patients in the complete locked-in state (CLIS), and motor-disabled patients who have lost their motor function many years ago (ref. [33, 34]).

During motor imagery tasks, EEG signals exhibit task-specific characteristics in both spatial domain and spectral (or frequency) domain [35–38]. These characteristics can be exploited in a BCI system to detect the user's intention. Towards this end, various *feature extraction* algorithms have been studied in the literature to extract EEG's discriminant information through spatial and spectral processing of the data. The main purpose of the feature extraction is to map the EEG data from its original measurement domain into another domain in which the motor imagery tasks are easily separable, according to a desired measure of separability (e.g., a linear or quadratic separability). Depending on the properties of the EEG data, this mapping may involve linear or nonlinear transformations in the spatial and/or spectral domains. The result of these transformations will be a multivariate (or univariate) representation of the data, where each variable is called a *feature* [39]. Accordingly, the multivariate space spanned by these variables is called the *feature space*. The extracted features are expected to provide an alternative representation in which the discriminant information of the data is preserved and at the same time the effect of the noise or interference is minimized. Therefore, one of the most important challenges in developing BCI systems is to consider both spatial and spectral characteristics of the signal during

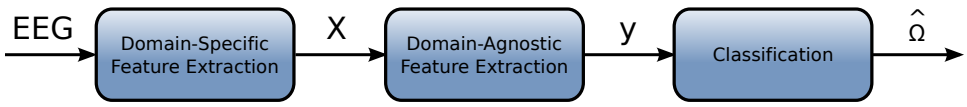


Figure 1.3: The processing pipeline for spatio-spectral feature extraction in MI-BCI systems.

the feature extraction and to take into account the inherent properties of the extracted spatio-spectral features in designing the classification algorithms, as will be described in the next section.

1.2 Problem Definition

Several combinations of spatial and spectral feature extraction (FE) techniques have been deployed for BCI systems in the literature to extract the most discriminant spatio-spectral features during motor-imagery (MI) tasks. Some of these FE methods are designed based on the existing knowledge about the neurophysiological characteristics of the EEG signals, while other methods are generic solutions that do not depend on such information. We call the former group *domain-specific feature extraction* (DS-FE) methods and the latter one *domain-agnostic feature extraction* (DA-FE) methods.

Consider a multichannel EEG signal that is recorded during the MI task Ω_i, $1 \leq i \leq C$, where C is the number of possible MI tasks. The goal of a MI-BCI is to classify the imagined motor task through analysis of the recorded EEG signal, and detect the imagined task. As illustrated in Figure 1.3, we divide this process into three major steps: (a) domain-specific feature extraction (DS-FE), (b) domain-agnostic feature extraction (DA-FE), and (c) classification.

DS-FE methods involve spatial processing, spectral processing, and in some cases joint spatio-spectral processing methods. Some examples include:

- channel selection (CS), independent component analysis (ICA), surface Laplacian (SL) filtering, and common spatial patterns (CSP) algorithms for spatial FE [40–42];
- parametric/nonparametric spectrum estimation and bandpass filtering for spectral FE [43–48];
- coherence analysis, directed transfer function modelling, filter-bank CSP (FBCSP), and common spatio-spectral patterns methods for joint spatio-spectral FE [49–52].

In general, the output of DS-FE stage is a *spatio-spectral feature matrix* of the form $\mathbf{X} \in \mathbb{R}^{N_f \times N_s}$, where N_f and N_s respectively represent the dimensionality of the spectral and spatial domains.

The common practice in MI-BCI systems is to vectorize the matrix $\mathbf{X}$, through concatenation of its columns (or rows), and pass it to a classifier either directly or through a DA-FE module. The DA-FE is usually used prior to classification to reduce the dimensionality of the feature space and

remove possible redundancies in the extracted features. The DA-FE stage may include any generic dimensionality reduction algorithm, such as principle component analysis (PCA), linear discriminant analysis (LDA), and methods based on mutual information or correlation (ref. [53, 54] and references therein).

We argue that the common approach for DA-FE, which requires vectorization of the matrix $\mathbf{X}$ by breaking it along the columns (or rows), introduces unnecessary degrees of freedom in the DA-FE stage by ignoring the inherent structure of the data along the broken dimension. In other words, vectorization of $\mathbf{X}$ removes the inherent spatio-spectral structure of the data, which could otherwise be exploited by the feature extractor.

The main problem that we address in this thesis is to design feature extraction techniques for motor-imagery BCI systems that takes into account the aforementioned inherent matrix-variate structure of the spatio-spectral features in order to (a) improve the overall performance of the MI-BCI system, and (b) reduce the computational cost of the feature extraction stage. Towards this end, we propose to utilize *matrix-variate* (or bilinear) algorithms for extraction of the most discriminant spatio-spectral EEG features.

In this thesis, we study how matrix-variate schemes can be used in the design of both domain-specific FE and domain-agnostic FE algorithms in motor-imagery BCI systems. We will then examine the benefits, challenges, and possible limitations of such schemes in two different MI-BCI experiments. The EEG data for these two experiments are obtained from two publicly available datasets that are widely used in the BCI literature for performance evaluation purposes. The first experiment represents a typical motor-imagery BCI scenario where enough training data is available to the algorithms. The second experiment represents the extreme case where the amount of training data is very limited. The latter case does not happen in most motor-imagery BCI systems, since these BCIs are generally designed for longterm utilization by the user. Nevertheless, the second experiment is included in this thesis to study the performance of different algorithms in extreme cases.

1.3 Technical Challenges

In the literature there exist numerous heuristic feature extraction solutions that aim to treat the matrix-variate data in their inherent structure through bilinear transformation techniques. One of the well known examples is the wide range of two-dimensional extensions of the LDA algorithm [55–63], all of which aim to extend the linear feature extraction procedure of LDA into a bilinear procedure that can be directly applied to matrix-variate data. As it will be discussed in Section 4.2, due to their

heuristic approach, most of these methods lead to unnecessary information loss and cannot capture all the discriminant information of the data, even in ideal Gaussian scenarios. Therefore, the most important challenge in matrix-variate analysis of the spatio-spectral features is to provide a solution which preserves the information content of the data.

The second important factor in designing matrix-variate solutions is the computational cost of the resulting algorithm. To clarify this point, consider the heuristic methods that are proposed in the BCI literature for extending the common spatial patterns (CSP) method to matrix-variate data [49,51,64–67] (ref. Section 3.1.1). One of the most successful extensions of CSP is called the filterbank CSP (FBCSP) method, which can be considered the state of the art solution in this area and outperforms most of the other solutions. Despite its high performance, this method has a relatively high computational cost and leads to a highly redundant feature space at its output, which in turn increases the computational cost of the classifier. Therefore, the second challenge in matrix-variate analysis of the spatio-spectral features is to design computationally efficient yet accurate algorithms.

The third challenge in matrix-variate analysis of the EEG features is the complex-valued nature of the spatio-spectral features obtained from domain-specific FE methods such as Fourier transformation. The common approach in the literature is to ignore the phase content of these features and only analyze their magnitude (or power). However, it has been recently shown in the literature that relevant information about the mental activities is conveyed by the phase of the EEG signal [68–71]. Therefore, it is of crucial importance to analyze such features in their inherent complex-valued format to be able to capture all the discriminant information of the data.

1.4 Thesis Contributions and Outline

In order to address the aforementioned technical challenges, we adopt a *matrix-variate Gaussian* distribution for modelling the spatio-spectral EEG features. This model lays the mathematical foundation for most of the theoretical designs and statistical studies in this thesis. This foundation enables us to theoretically derive computationally efficient yet accurate bilinear methods for spatio-spectral feature extraction in BCI systems.

The matrix-variate Gaussian model is a subset of the commonly used *multivariate Gaussian*. Beside the general assumptions of multivariate Gaussianity, the matrix-variate Gaussian model requires a certain Kronecker product structure for the covariance of the data, as will be described in Section 3.4. This extra condition on the covariance of the data is the key point that distinguishes the matrix-variate Gaussian model from the multivariate model which is conventionally used in various applications in the literature,

including the BCI systems. This condition allows us to present the data in a matrix-variate structure and process it using bilinear operations.

This thesis provides a general framework for spatio-spectral feature extraction from motor imagery EEG signals, which emphasizes the distinction between domain-specific feature extraction (DS-FE) and domain-agnostic feature extraction (DA-FE) in MI-BCI systems. This general framework not only does encompass existing feature extraction methods, but also suggests new alternative approaches for spatio-spectral feature extraction. We use the proposed framework to introduce a matrix-variate Gaussian model for the spatio-spectral EEG features. Based on this model, we design two new approaches for spatio-spectral feature extraction in motor-imagery BCI systems. Therefore, the main contributions of this work can be categorized as follows:

Domain-Agnostic Bilinear Feature Extraction for MI-BCI [72, 73]: In Chapter 4, we consider the homoscedastic matrix-variate structure of spatio-spectral features at the input of domain-agnostic FE stage. We propose to deploy matrix-variate feature extractors, instead of the conventional vector-variate DA-FE methods. Considering the fact that the Bayes optimal feature extraction strategy for homoscedastic vector-variate data is the linear discriminant analysis (LDA) method, we suggest to utilize a bilinear extension of LDA for DA-FE in motor-imagery BCI systems. The Bayes optimality of the FE method guarantees that the extracted features encapsulate all the discriminant features of the data, and there will be no performance loss caused by the dimensionality reduction procedure in the feature extractor.

Towards this end, we study the following two possible methods for bilinear extension of the LDA method: (a) An iterative two-sided extension of the LDA, called 2DLDA in this thesis, which has been proved to be highly successful in other applications in the pattern recognition literature. (b) A non-iterative method, called *matrix-to-vector linear discriminant analysis* (MVLDA), which directly takes advantage of the properties of the matrix-variate Gaussian model to extract the most discriminant features of the data. Both methods directly operate on the matrix-variate data, without any need for vectorization. They simultaneously take into account both spatial and spectral characteristics of the data, and have significantly less computational complexity compared to the conventional vector-variate LDA method.

To study the effectiveness of the proposed bilinear domain-agnostic FE schemes, we deploy the 2DLDA and MVLDA methods in conjunction with a widely used domain-specific FE method, called filterbank common spatial patterns. The experimental results show that the combination of FBCSP and MVLDA methods provide a significant performance improvement over the state

of the art solutions. Furthermore, we provide a comprehensive study of the effect of utilization of the surface Laplacian filtering and the channel selection, at the DS-FE level, on the overall performance of the proposed system.

Domain-Specific Bilinear Feature Extraction for MI-BCI [74, 75]: In Chapter 5, we consider the matrix-variate structure of the features generated during the domain-specific FE stage. We propose a novel DS-FE method which takes this structure into account during the extraction of the most discriminant features. The proposed method, called separable *common spatio-spectral patterns* (SCSSP) method, has low computational cost compared to the state of the art FBCSP method. The SCSSP method uses a heteroscedastic matrix-variate Gaussian model for the multi-band EEG rhythms, which allows it to efficiently rank the extracted features according to their discriminant power. As a result, the features generated by SCSSP method can be directly passed to a classifier, without any need for a separate domain-agnostic FE stage.

The proposed SCSSP method has two major differences with the FBCSP method. First, FBCSP ignores the spectral correlations between different EEG bands and independently extracts the spatial features of each band; whereas the SCSSP method simultaneously considers both spectral and spatial correlations of the data. Second, the FBCSP method assumes a unique spatial covariance for each EEG rhythm; whereas the SCSSP method considers a common structure for the spatial covariance matrices of different rhythms. These differences allow the SCSSP method to improve the performance and reduce the computational cost of the DS-FE stage, provided that enough training data is available to the algorithm.

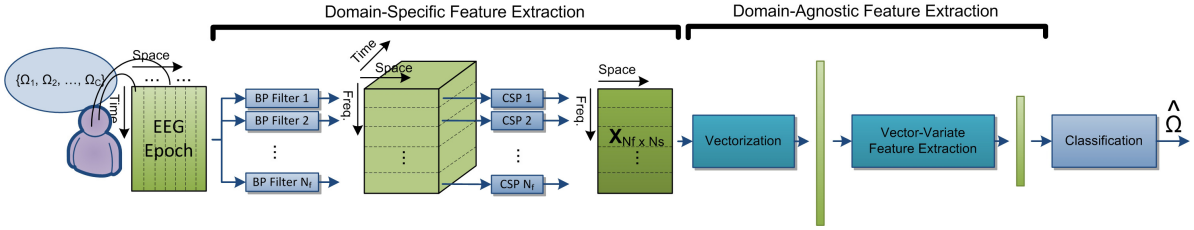
We study the performance of the SCSSP when combined with two different simple classifiers, namely the naive Bayes Parzen Window (NBPW) and the linear minimum distance classifier. The experimental results show that the linear classifier is the best match for the SCSSP method. We also perform a comprehensive experimental study on the effect of surface Laplacian filtering and channel selection on the overall performance of the BCI system, when they are used in conjunction with the SCSSP algorithm.

Statistical Characterization of Spatio-Spectral EEG Features in The Fourier Domain [76, 77]: The results of our experimental evaluations in Chapter 4 and Chapter 5 show a significant performance improvement by the proposed matrix-variate schemes compared to the state of the art solutions, which highly suggests that the matrix-variate Gaussian distribution provides a reasonable model for the statistical properties of the spatio-spectral EEG features. Motivated by these results, we provide an in-depth statistical study of the complex-valued spatio-spectral EEG fea-

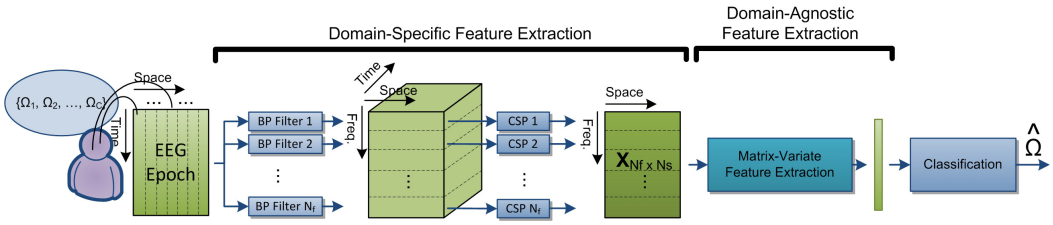
tures in Chapter 6. The results of the previous chapters highly suggest that the multiband EEG rhythms follow a matrix-variate Gaussian distribution. As a result, the Fourier domain representation of the data is also expected to exhibit similar properties (ref. Appendix A.4). One of the major benefits of focusing on the Fourier domain analysis of the data is to provide a model for the spatio-spectral features which can also take into account the information in the phase of the EEG data, as mentioned in the previous section.

In Chapter 6, we propose a complex-valued Gaussian model for the Fourier domain representation of the spatio-spectral EEG features and will study the link between this model and the matrix-variate Gaussian model that was explored in the previous sections. The validity of this model will be examined through several statistical tests. In the proposed complex-valued model, the second order characterization of the data requires the knowledge of both the covariance and the pseudo-covariance of the data. In case that the complex-valued features do not convey information in their phase, the pseudo-covariance of the data will be zero, and all the second order statistics of the data will be conveyed by its covariance matrix. This property provides us with a statistical tool to study whether any relevant information is conveyed in the complex-valued spatio-spectral features of the EEG signals. Our statistical tests highly confirm the hypothesis that the pseudo-covariance of these features is not zero, which in turn confirms that relevant information is conveyed in the phase of these complex-valued features. This finding agrees with the recent neurophysiological studies on the phase information of the EEG signals [68–71].

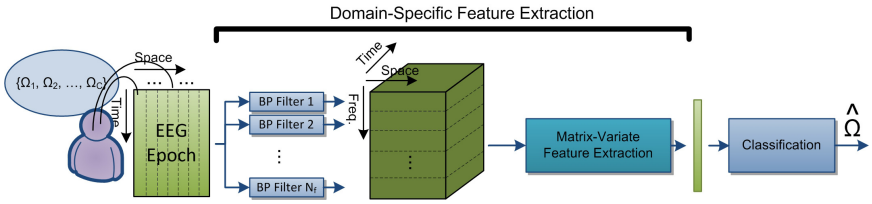
The rest of this thesis is organized as follows. Chapter 2 provides the required background information and preliminary knowledge about the EEG signals and their properties during the motor imagery tasks. In Chapter 3, we introduce our proposed framework for spatio-spectral feature extraction in motor-imagery BCI systems, and explain how various methods in the literature fit into this framework. Based on this framework, the matrix-variate Gaussian model for the spatio-spectral EEG patterns will be defined in this chapter. Chapters 4, 5, and 6 include the three main contributions of the thesis as explained above. Finally, the thesis summary and concluding remarks are presented in Chapter 7.



(a) Conventional Filterbank CSP Approach



(b) Proposed Bilinear Approach for Domain-Agnostic FE



(c) Proposed Bilinear Approach for Domain-Specific FE

Figure 1.4: Illustrative comparison of the proposed schemes with the state of the art filterbank CSP solution.

Chapter 2

Preliminaries

This chapter provides a brief review of the brain structure and the relationship between brain activities and EEG signals. We also provide a short description of the EEG signal acquisition techniques, and the artifacts that usually contaminate the EEG signals.

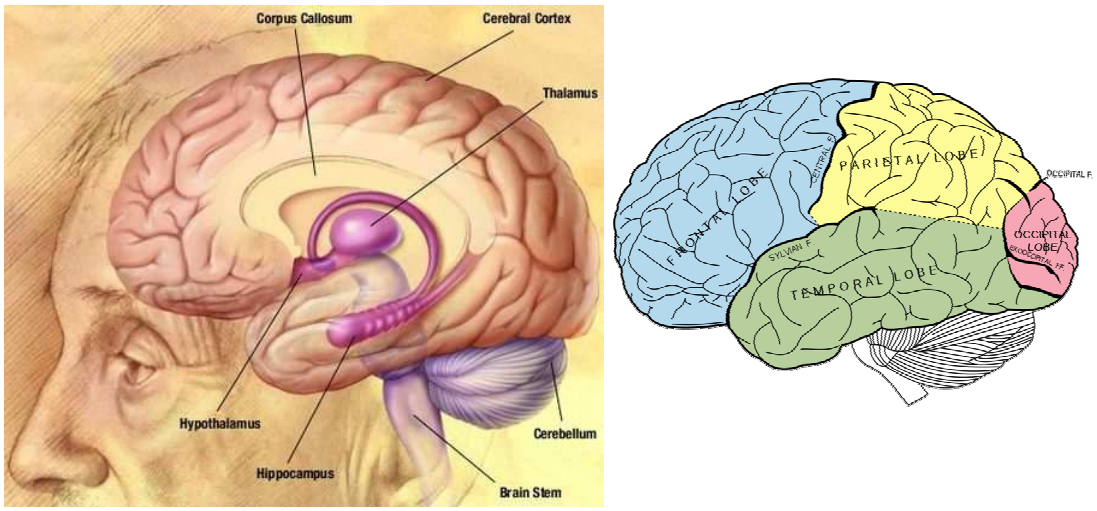
2.1 Structure of The Brain

The human brain can be divided into three major parts: the cerebrum, the cerebellum, and the brain stem. The cerebrum, which is the largest part of the brain, is divided into two hemispheres and contains the basal ganglia, the limbic system (hippocampus, hypothalamus, thalamus, etc.), and the cerebral cortex. The cerebral cortex is the outer layer of the cerebrum and is divided into four topographical major lobes: frontal, parietal, temporal, and occipital (see Figure 2.1.b[1]). This cortex plays an important role in high-level tasks in the brain such as processing of the sensory information, planning and controlling voluntary movements, and understanding the language. As it is illustrated in Figure 2.2.b, each of these tasks are performed in a different part of the cerebral cortex.

2.1.1 Motor Control in The Brain

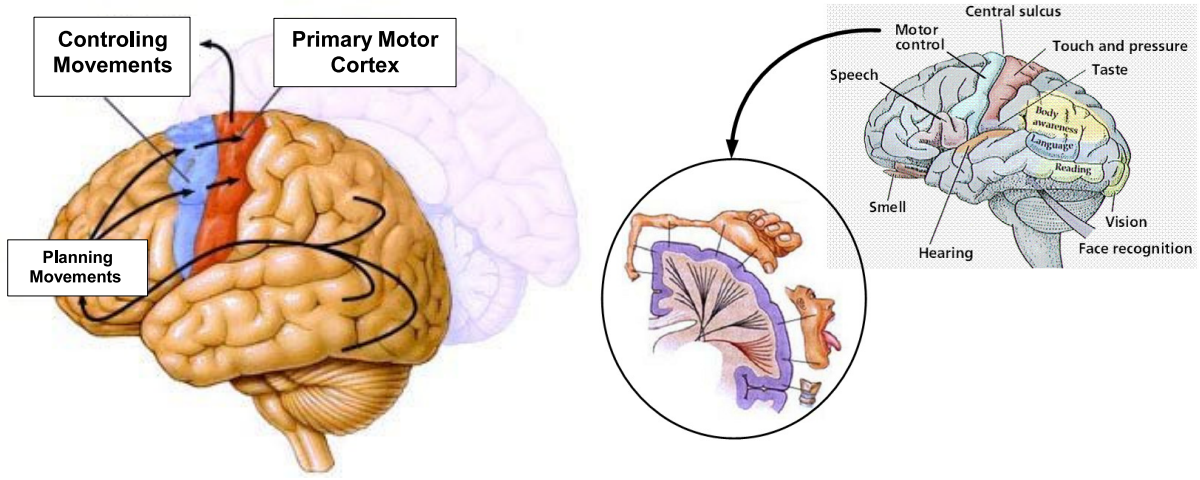
A part of the cerebral cortex which is mostly involved in controlling voluntary movements is called the motor cortex. As it is shown in the magnified part of Figure 2.2.b, different regions of the motor cortex control the movement of different parts of the body. In this figure, those parts of the body that are shown larger are the ones which occupy more space in the motor cortex and are responsible for finest movements. Although motor cortex is the main part of the brain responsible for voluntary movements,

[1]The figures adopted from other sources in this chapter are not copyright-protected.



(a) (b)

Figure 2.1: (a) Major parts of the human brain; (b) Topographical regions of the cerebral cortex.



(a) (b)

Figure 2.2: (a) The procedure of planning a movement in the brain; (b) Different parts of the cerebral cortex and their corresponding tasks. (Adopted from [1])

several other regions of the cerebral cortex are also involved in controlling these movements. Figure 2.2.a illustrates the process of planning for a voluntary movement in the brain. The planning process is done mainly in the forward portion of the frontal lobe, which receives information about the individual's current position from several other parts. Then, the required commands are issued to the first area on the motor cortex (known as area 6). This part of the motor cortex decides which set of muscles to contract to achieve the required movement, then issues the corresponding orders to the primary motor cortex. This area in turn activates specific muscles or groups of muscles via the motor neurones in the spinal cord. Therefore, in order to process EEG signals generated during motor imagery, both spatial

and temporal characteristics of the EEG signal should be considered.

2.1.2 Brain Plasticity

The mapping of tasks shown in Figure 2.2.b is a general mapping which may differ between different individuals and also may change for each individual over time. Indeed, the brain has the ability to change its structure based on the daily life experiences and needs. The following are some of the major cases where such changes may occur:

- If a particular part of the brain is exhaustively used over a long period of time, this part of the brain may expand its boundaries and grow in size to be able to meet the demands.
- When a certain part of the brain is damaged or injured, the other parts may try to adapt their structure to be able to compensate for some of the lost functions or take on some of the responsibilities of the damaged cells.
- When a specific part of the brain is not used for a long period of time, this part may be reassigned to perform other tasks in the brain. As an example, when someone goes blind and the input to the visual cortex is blocked, the corresponding visual parts of the cortex gradually changes its function and receives other sensory inputs, such as tactile or auditory inputs.

These functional changes should be taken into account in designing the BCI systems. In the case of healthy individuals, the brain plasticity results in inter subject variations in spatio-temporal characteristics of the brain signals. Consequently, BCI systems usually perform a subject specific training phase for each individual to take into account possible changes in the characteristics of the motor related areas in the brain.

Furthermore, one may argue that due to brain plasticity, the motor cortex of disabled individuals will be reorganized and they may not be able to perform motor imagery tasks required for BCI systems. In the case of paralyzed people with spinal cord injury, who did not have any damage in their motor cortex, several research works have studied this issue (e.g., see [78–80]). These works have shown that since the motor cortex is no longer used in these people, the reorganization of the motor cortex occurs over time; however, this reorganization does not significantly affect the motor-cortical activities corresponding to motor imagery tasks. These studies have shown that movement attempts in these individuals result in a set of brain activities in motor-related areas (including the primary motor area) that are closely similar to what is normally observed during the preparatory stages of movement execution in healthy individuals.

2.2 Electroencephalogram (EEG) Signals

In Chapter 1, it was mentioned that we study non-invasive EEG recordings of the brain activities. This section provides a brief review of the EEG signals and their characteristics which can be utilized in BCI systems.

2.2.1 EEG Signal Acquisition

EEG signals are usually recorded using several electrodes on the scalp, which aggregate the electric voltage fields from millions of neurones across the brain. The EEG recordings at the scalp surface are mostly generated by electrical current sources in the brain which are coherent over an area of at least a few square centimetres[2].

It has been shown in the literature that the skull tissue acts as a spatial lowpass filter, which highly attenuates the electric potentials generated by localized cortical sources while having little effect on the sources that are distributed on larger cortex areas [26]. In other words, the skull tissue acts as a lowpass spatial filter on the EEG signals. As a result, it can be considered as a natural anti-aliasing spatial filter which attenuates the high-frequency components of the EEG signal in the spatial domain. This anti-aliasing filter is of particular importance since we need to spatially sample the EEG signal with a limited number of electrodes.

Beside the low-pass filtering effect of the skull, the contact area of each electrode also plays an important role in avoiding spatial aliasing. Indeed, the conductive surface area of each electrode acts a lowpass spatial filter which eliminates the signal components with wavelengths approximately shorter than the electrode diameter. The combined effect of the skull issue and the electrodes' conductive surface enables us to spatially sample the EEG signal without aliasing[3].

The electrode locations are usually determined from the international $10-20$ standard system. In this system, 21 electrodes are located at the locations shown in Figures 2.3.a and 2.3.b. These locations are determined based on the following two anatomical reference points: Nasion, which is located between the eyes at the top of the nose; and Inion, which is located at the lower rear part of the skull. As it is illustrated in Figure 2.3, the distance between two adjacent electrodes is %10 or %20 of the total distance between the nasion and the inion. The naming for these electrodes follows the following convention: The letters F, T, P, O, and C respectively stand for frontal, temporal, parietal, occipital, and central parts of the scalp. The odd electrode numbers refer to the left hemisphere and the even numbers refer to the

[2]About 6 cm^2 of cortical gyri tissue must be synchronously active to produce a scalp potential which is recordable by conventional EEG sensors. This area, corresponds to approximately $600,000$ cortical microcolumns or $60,000,000$ neurons.

[3]To completely avoided spatial aliasing, the electrode diameter needs to be chosen to be equal to the edge-to-edge distance between the electrodes' conductors or spreads of gel layer

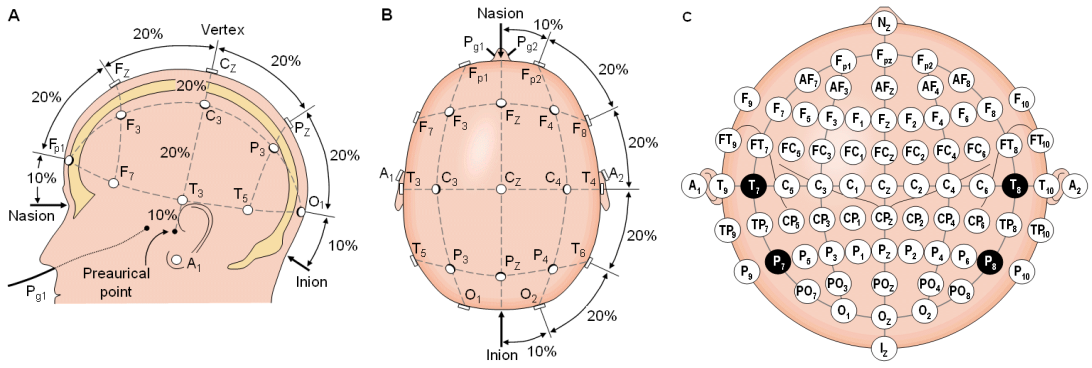


Figure 2.3: EEG electrode locations in (a) $10-20$ system, side view; (b) $10-20$ system, top view; (c) $10-10$ system, top view (Adopted from [2])

right hemisphere.

In order to use more electrodes on the scalp, this standard has been extended to $10-10$ and $10-5$ systems (see [81] and references therein). Figure 2.3.c illustrates the electrode locations for a $10-10$ system. In this research we will use motor imagery EEG databases available at [82], which are collected using the $10-10$ and $10-5$ systems.

The electrical signals collected from these electrodes are passed through a differential amplifier. There exist different standard methods for connecting the electrodes to the amplifiers, such as common reference, average reference, and bipolar. The databases used in this research are recorded using a common reference method, where the difference between the output signal of each scalp electrode and the output signal of a fixed *reference electrode* (usually ear electrode) is amplified by the differential amplifier. Each amplifier output forms an EEG channel; hence, in a $10-20$ system we will get 21 EEG channels.

Amplified signals are then highpass filtered (to prevent aliasing during sampling), uniformly sampled, and converted to digital signals. Databases used in this research include EEG signals which are sampled with a sampling frequency of 250 Hz or 1000 Hz. The resulting digitized multichannel EEG signal is usually recorded in a matrix of size $N_t \times N_{ch}$, where N_{ch} is the number of EEG channels, and N_t denotes the number of time samples for each EEG channel.

2.2.2 EEG Artifacts

The electrical voltages recorded by EEG electrodes are usually in the range of $10\mu V$ to $100\mu V$. Consequently, EEG recordings are very sensitive to interfering signals, also called *artifacts*, that are not generated by the brain. In general, artifacts can be categorized into two groups:

- Biological Artifacts, such as signals generated by eye movements (EOG) or eye blinks, electrical

activity of the heart (ECG) and muscle activation signals (EMG).

- Environmental Artifacts, such as powerline artifacts (50/60 Hz) and signals generated by cardiac pacemakers. Also, momentary movements of scalp electrodes can cause abrupt changes in the impedance of these electrodes and result an artifact in the EEG recording.

Some of these artifacts can be easily removed by appropriate filtering of the EEG signals, e.g., notch filtering at 50/60 Hz for powerline artifacts. Other artifacts, such as EOG/ECG/EMG, are usually removed using source decomposition techniques such as independent component analysis (ICA) method [83, 84].

2.2.3 EEG Rhythms

Early studies on EEG signals (by H. Berger in 1929) revealed that EEG signals can be expressed in terms of a number of rhythmic activities, each of which oscillates within a different frequency band. These rhythms are generated by numerous excitatory/inhibitory postsynaptic potentials (ESPS/IPSP) in the cerebral cortex. In order to study these rhythmic activities, the frequency spectrum of EEG signals is usually divided into the following frequency bands:

Frequency band:	< 4 Hz	$4 - 8$ Hz	$8 - 12$ Hz	$12 - 30$ Hz	30 Hz $<$
Rhythmic wave:	Delta	Theta	Alpha	Beta	Gamma

Several studies have shown that different rhythmic activities within these frequency bands correspond to different brain activities or states. As an example, delta waves are associated with deep sleep and alpha waves correspond to relaxed awareness without concentration. However, it should be noted that due to the complicated structure of the brain, there exist numerous different tasks which result in rhythmic signals within the same band [85]. In this research, we are specially interested in alpha and beta bands since they have been shown to be associated with motor activities (see Section 2.2.4).

It is worth mentioning that spatial and temporal properties of the EEG sources are coupled to a certain extent. It is known in the literature that small patches of cortical sources tend to generate higher temporal frequencies, whereas larger patches tend to generate lower temporal frequencies [26, 85]. Therefore, the number of active sources contributing to the EEG power for lower frequencies is much higher than the sources contributing to the EEG power in high frequencies. As a result, the EEG power spectrum has a general trend of the form $1/f$.

2.2.4 Event-Related Dynamics of Cortical Rhythms During Motor Imagery Tasks

It has been shown in many studies that when no cognitive or motor task is performed in the brain, large populations of firing neurones are synchronized to each other, which in turn result in steady rhythmic activities in the brain. During the execution of motor tasks, however, the synchronization of these populations decreases (or increases), which results in a decrease (or increase) in the power of corresponding oscillatory rhythms. This phenomenon is called *event related desynchronization*, ERD, (or *event-related synchronization*, ERS) [35, 86, 87].

One of the earliest works that has illustrated the ERD and ERS effects during imagination of hand movement is the experimental work in [86]. In this experiment, the participants were asked to imagine the right hand or left hand movement. Considering the fact that right hand movements are controlled in the left hemisphere (and vice versa), the results of [86] reveal a significant desynchronization (ERD) in the alpha band ($8-12$Hz) in the contralateral hemisphere[4], which corresponds to the motor imagery activities in this hemisphere. It also reveals a significant synchronization (ERS) in the ipsilateral hemisphere, which corresponds to the idle state of the motor cortex in this hemisphere. These ERD and ERS features are usually used in BCI systems for classification of right vs. left hand movement imagery task.

It should be noted that ERD/ERS effects are not constant over all frequency bands. The work in [38] has studied the power spectrum changes on the surface of the cortex during the hand movement task for a wide frequency range of $0 - 150$ Hz. This study shows that the motor task results in a power decrease (ERD) in low frequency rhythms ($f < 50$ Hz), while causing a significant power increase (ERS) in the high frequency rhythms.[5]

In order to study the spatial characteristics of ERD/ERS, we can perform similar analysis for all EEG channels. Figure 2.4 illustrates the spatial patterns of ERD/ERS for left and right hand movement imagery tasks. The value of ERD/ERS in the topographical maps is expressed as the relative power decrease (ERD) or power increase (ERS) with respect to the 0.5 sec interval before start of motor imagery task. As a result, in the colour bar represented in the topographical maps, the value -1 dB corresponds to the ERD effect, while $+1$ dB corresponds to the ERS effect. These topographical maps reveal the fact that a large number of EEG channels convey relevant information about the motor imagery tasks; hence, it is crucial to take into account the spatial characteristics of the EEG channels in classification of

[4]For right hand motor imagery tasks, the right hemisphere of the brain is the *ipsilateral* hemisphere and the left hemisphere is the *contralateral* hemisphere, and vice versa.

[5]It should be noted that the signals in [38] are recorded using partially-invasive electrocorticogram (ECoG) electrodes. In case of non-invasive EEG signals, the skull significantly dampens the high frequency rhythms. This will result in a down shift in high frequency components. Nevertheless, we can still observe a relative power increase in high frequency rhythms of EEG signals. A similar study of these high frequency oscillations in EEG signals is performed in [88].

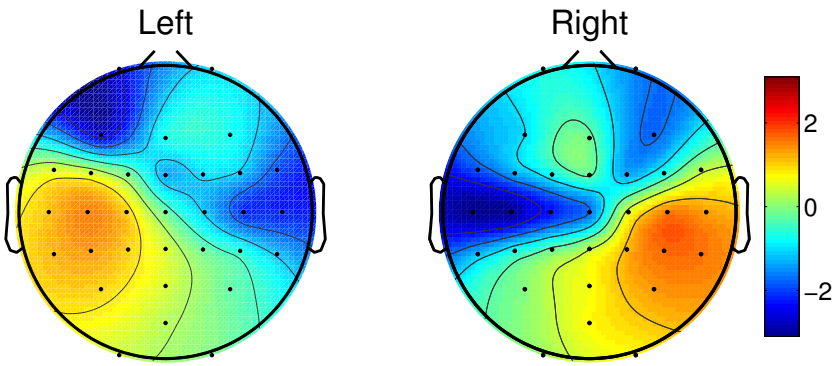


Figure 2.4: Spectral characteristics of EEG signal during left hand movement and right hand movement imagery tasks. The topographic maps are obtained by passing the EEG signal through a bandpass filter ($8 - 28$ Hz) and then averaging the signal over the time interval between $0.5 - 3.5$ seconds after the onset of motor imagination. The plotted values represent the power change relative to the time before the task onset (in dB scale).

these tasks. In general, we can conclude that in order to design a BCI system, one should simultaneously consider spectral, temporal, and spatial characteristics of the EEG signals.

2.3 Algorithms for pre-emphasizing localized sources in EEG

In Section 2.2.1, it was mentioned that the EEG signal recorded at the scalp surface is mostly due to the sources with low spatial frequencies. Consequently, the effects of localized sources in the EEG are usually dominated by widely spread sources. In other words, the raw EEG signal has a relatively low spatial resolution. In order to alleviate this problem, the following two methods have been proposed in the literature: (a) Dura Imaging method, also known as Spatial Deconvolution method; and (b) Surface Laplacian method.

The dura imaging method aims to estimate the electrical potentials on the inner surface of the skull, called dura potential, using a volume conductor model for the head. This method requires an accurate head model to determine the geometry of the cortical surface, inner and outer skull surfaces, and the scalp surface. This accurate model is usually obtained using magnetic resonance imaging (MRI) method. Although the dura imaging method is very accurate, it cannot be used in many BCI applications, due to the high cost and inconvenience of MRI scanners and in many cases lack of access to such scanners.

The surface Laplacian method aims to estimate the local radial current flux which passes through the skull at each point. This local current is closely related to the dura potential (i.e., the potential on the inner surface of the skull) generated by localized sources. Unlike the Dura Imaging method, surface Laplacian only requires the electrode locations and does not require the person's head model. Therefore,

it can be used in most BCI applications. We will briefly explain the surface Laplacian method in this section, as it will be used later in the thesis.

2.3.1 Surface Laplacian Method

Let V_s, J_s, V_c, and J_k respectively represent the electric potential at the outer surface of the skull, the current on the outer surface of the skull, the electric potential at the inner surface of the skull, and the radial current density passing through the skull. Note that according to the Ohm's law, J_k is proportional to $V_c - V_s$. Due to the lowpass spatial filtering property of the skull, $V_c \gg V_s$ for localized sources, and hence $J_k \propto V_c$ for localized sources. Based on this result, the surface Laplacian method tries to calculate the value of J_k in order to provide an estimate of V_c.

Since all the radial current J_k spreads on the skull surface once it reaches the outer surface of the skull, we can conclude that

$$J_k = \nabla_s \cdot J_s, \tag{2.1}$$

where ∇_s denotes the spatial derivative, or divergence, operator along the two surface coordinates. Assuming that the skull surface is spherical, ∇_s can be defined as follows:

$$\nabla_s = \frac{1}{r \sin\theta} \frac{\partial}{\partial \theta} (\sin\theta J_\theta) + \frac{1}{r \sin\theta} \frac{\partial J_\phi}{\partial \phi} \tag{2.2}$$

where (r, θ, ϕ) represents the spherical coordiantes: radios, polar angle, and azimuthal angle. Using the Ohm's law, the surface current in Equation 2.1 can be linked to the surface potential, as follows:

$$\begin{aligned} J_k &= \nabla_s \cdot (\sigma_s \nabla V_s) \\ &= \sigma_s \nabla_s^2 V_s \end{aligned} \tag{2.3}$$

where σ_s represents the scalp's conductivity and ∇_s^2 denotes the second spatial derivative along the two surface coordinates. For spherical surfaces, ∇_s^2 operator is defined as follows:

$$\nabla_s^2 = \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial J_\theta}{\partial \theta} \right) + \frac{1}{r^2 \sin^2\theta} \frac{\partial^2 J_\phi}{\partial \phi^2} \tag{2.4}$$

Therefore, the electric potential at the inner surface of the skull is approximately proportional to the surface Laplacian of the electric potential at the outer surface of the skull, i.e.,

$$V_c \propto \nabla_s^2 V_s, \tag{2.5}$$

It is worth mentioning that since we are mostly interested in the value of V_c at the electrode locations, in practice the surface Laplacian of V_s is only calculated at the electrode locations. Using this method, the surface Laplacian output will have the same spatial resolution as the original EEG signal.

It is also noteworthy that since the surface Laplacian method is based on the second order spatial derivative operator, it provides a reference free measurement, which is independent of the choice of reference electrode used for EEG recording.

2.3.2 Surface Laplacian Calculation From Spatially Sampled EEG Data

In order to apply the surface Laplacian operator to V_s, we need to have a continuous measurement of the V_s over the scalp surface. However, the EEG recording provides a spatially discrete signal which only contains information about V_s at the electrode locations. There are two approaches to estimate the surface Laplacian from the spatially discrete EEG signal.

The first approach is to use the finite difference approximation of the ∇_s^2 operator. In this approach, the value of $\nabla_s^2 V_s$ at each electrode location will be approximated by a linear combination of the V_s measured at that electrode and its neighbouring electrodes. As a case in point, assume that the value of EEG recording at a certain electrode is V_0, and it has N neighbouring electrodes with EEG recordings of $V_n, 1 \leq n \leq N$ which are equally distributed on a circle of radios d_0 around this electrode. Then the first order approximation of the surface Laplacian at this electrode location, can be calculated as follows:

$$\nabla_s^2 V_s \simeq \frac{1}{d_0^2}\left(V_0 - \frac{1}{N}\sum_{n=1}^{N} V_n\right) \tag{2.6}$$

This first order approximation simply removes signal components with low spatial frequency which are commonly sensed by all the neighbouring electrodes. Such transformation amplifies the effect of localized sources while attenuating the effect of distributed or distant sources.

The second approach is to use spline interpolation in the spatial domain to estimate the value of V_s over all the points on the scalp surface, based on which the surface Laplacian can be calculated. Depending on the type of geometry assumed for the scalp surface, the following three methods have been used in the literature:

- 2D Spline: This method projects the electrode locations onto a two-dimensional flat plane, and calculates the spline interpolations in that plane [89].

- Spherical Spline: In this method, the electrode locations will be projected onto a sphere, which approximately represents the scalp surface, and hence the spherical splines will be used for inter-

polation [90].

- 3D Spline: This method is the most accurate method in which the electrode potentials will be interpolated in the three-dimensional space, regardless of the scalp's surface geometry [91].

In this thesis, we use the spherical spline method since its spherical assumption for the scalp surface is more accurate than the 2D spline method and at the same time it is more computationally efficient compared to the 3D spline method. To implement this method, we have used the publicly available toolbox for MATLAB, called CSD-Toolbox [92–94]

2.4 Linear Prediction Models for EEG Signals

Various different methods have been suggested in the literature for modelling the EEG signals. These methods include but are not limited to: (a) Proney's method [95], which can be used for modelling evoked potentials; (b) Neural mass modelling [96], which is mainly used for modelling steady-state behaviors of neural systems; (c) Nonlinear chaotic modelling [97], which has been used to model EEG abnormalities such as epilepsy or psychiatric disease as well as normal EEG rhythms; (d) Linear prediction modelling, which has widely been used in various applications such as nonparametric spectrum estimation (Section 3.1.2) and coherence analysis (Section 3.1.1). Considering the wide range of applications in which linear models have been used for spontaneous BCI systems, we particularly focus on linear models in this section. Among existing linear models, autoregressive (AR) model is usually used for EEG signals; therefore, we briefly overview the AR model and its modified versions.

2.4.1 Autoregressive (AR) and Adaptive Autoregressive (AAR) Models

Using AR linear predictive model, we can express the EEG signals as the output of a linear system driven by a white noise at its input, as follows. Let $y_i(n)$ be the output of channel i at time instance n. Then, $y_i(n)$ can be defined in terms of previous outputs of this channel, using the following equation:

$$y_i(n) = -\sum_{k=1}^{p_i} a_{i,k} y_i(n-k) + x_i(n), \tag{2.7}$$

where $x_i(n)$ is the random white noise, $a_{i,k}$ are the AR model parameters and p_i is modelling order for channel i. Such AR model represents an all-pole system which has infinite impulse response (IIR system). It should be noted that in this approach each EEG channel is separately modelled, and the parameters $a_{i,k}$ and p_i are separately determined for each channel.

In this modelling approach, appropriate selection of the parameter p_i is of great importance. Overestimation of p generates false peaks in the estimated EEG spectrum, while underestimation of p_i results in an over-smoothed spectrum. One of the conventional methods used for finding the model order is the utilization of Akaike information criterion (AIC), defined as follows:

$$AIC(p_i) = N \ln(\sigma^2_{p_i}) + 2p_i, \tag{2.8}$$

where N is the number of samples and $\sigma^2_{p_i}$ is the prediction error using the model order p_i. This criterion can be viewed as a trade off between the complexity of the algorithm (the second term in the above equation) and the precision of fitting (the first term). It is worth mentioning that AIC has a strong bias when the sample size (N) is limited [98]. This bias can lead to unwanted overfitting for the model order. Therefore, it is highly recommended in the literature to utilize the corrected versions of AIC in limited sample size scenarios (ref. [98–100]).

Once the model order is defined, the model coefficients ($a_{i,k}$) can be determined using various different methods, such as Yule-Walker method, covariance method, Burg algorithm, least squares method, and maximum likelihood method. A comparative analysis of these methods for modelling EEG signals is presented in [101].

The AR model presented above generates a stationary signal ($y_i(n)$). In practice, however, EEG signals are nonstationary and their statistics change over time. In order to take this nonstationarity into account in AR models, two solutions have been suggested in the literature. The first solution is to divide EEG signals into small time segments over which the signal can be considered as a stationary signal, and update the model parameters ($a_{i,k}$) for each segment [102]. In this approach, as the length of these time segments decreases, the temporal resolution of the AR model increases while the estimation error of the AR model increases. In general, these time segments can have a fixed or variable length. Fixed-length segmentation algorithms use a fixed length for all the segments, where this fixed length should be short enough to guarantee stationarity over each individual time segment. Variable-length segmentation algorithms, each segment is identified such that it can capture an entire length of stationary state in the EEG signal; hence, segment boundaries are defined as time instances where the characteristics of the EEG signals change. A comparative review of different fixed and variable segmentation algorithms is presented in [103].

The second solution is to adaptively change the AR parameters at each time instance. In this model, known as adaptive AR (AAR) model, the AR parameters are adaptively updated for each time

instant, as follows:

$$y_i(n) = -\sum_{k=1}^{p_i} a_{i,k}(n) y_i(n-k) + x_i(n), \tag{2.9}$$

where $a_{i,k}(n)$ are the time-variant AR parameters. In this model, similar to the AR model in (2.7), different channels are modelled separately. The Akaike method is again used for determining the appropriate p_i, and least squares or recursive least squares methods are usually used for determining the AAR parameters $a_{i,k}(n)$. The main advantage of AAR model is that it does not require segmentation of the EEG data, and the model parameters are updated for each time instance. Thus, AAR models are more suitable for analysis of fast transitions of the brain state. This advantage come at the cost of high computational complexity of updating AAR parameters at each time instance. The AAR model can be used for spectral estimation of EEG signals.

2.4.2 Multivariate AR (MVAR) and Adaptive Multivariate AR (AMVAR) Models

One of the main disadvantages of both aforementioned AR and AAR models is that the output of each channel is independent from the outputs of other channels. This results in a poor signal modelling that does not agree with the characteristics of actual EEG signal, where outputs of different EEG channels are highly correlated to each other. In order to provide a more realistic model which considers the spatial characteristics of EEG signals, multivariate autoregressive (MVAR) models have been used in the literature. In MVAR model, output of the i^{th} channel is determined as follows [104, 105]:

$$y_i(n) \quad = \quad -\sum_{j=1}^{N_{ch}} \sum_{k=1}^{p} a_{jk}\, y_j(n-k) + x_i(n) \tag{2.10}$$

where N_{ch} is the total number of channels p is the model order. Without loss of generality, p is assumed to be the same for all the channels. In this model, the output of each channel not only does depend on the previous outputs of that channel, but also depends on the previous outputs of other channels. The model order p can be determined by minimization of the following Akaike criterion:

$$AIC(p) = N \ln\big(\det(\Sigma_p)\big) + 2pN_{ch}^2, \tag{2.11}$$

where Σ_p is the prediction error covariance matrix. The model parameters a_{jk} can be determined by means of solving a multivariate version of Yule-Walker equation [104].

In order to consider the time-varying properties of EEG signals, adaptive multivariate AR (AMVAR) models are recently used in the literature. In AMVAR model, similar to AAR, the model coefficients are updated for each time instance, and the channel outputs are defined as follows:

$$y_i(n) \quad = \quad -\sum_{j=1}^{N_{ch}} \sum_{k=1}^{p} a_{jk}(n)\, y_j(n-k) + x_i(n) \tag{2.12}$$

For estimation of time-varying coefficients $a_{jk}(n)$, the recursive least squares method is commonly used in the literature. The AMVAR model is widely used for coherence analysis of EEG (see Section 3.1.1).

2.5 Summary and Concluding Remarks

In this chapter, the background information regarding the neurophysiological properties of the EEG signals were reviewed. In particular, the event related dynamics of the EEG signals during the motor-imagery tasks was reviewed. The event-related synchronization/desynchronization (ERD/ERS) effects that were discussed in this chapter are the main properties of the EEG signals that are mostly used for motor-imagery BCI systems.

It is worth mentioning that most of the *spontaneous* BCI systems using spectral features that have successfully been implemented in practice are not *directly decoding* the brain tasks. In these systems, individuals ***learn*** *how to control* certain aspects of the electrophysiological signals emitted by their brains. As an example, consider the BCI system explained in [106] that is designed to move a cursor to *up* and *down* directions. This system does not really detect the imagination of moving the user's hand to up/down direction. Instead, the user controls the cursor by voluntarily increasing or decreasing the amplitude of the mu rhythm[6] (8-12 Hz) or beta rhythm (18-26 Hz) signals generated by the sensorimotor cortex of the brain. In other words, the users of these BCI systems develop a new skill to properly control their brain signals such that they can successfully operate the BCI device.[7] To solve this problem, some research groups are trying to minimize the role of subject training and impose the major learning load on the computer [108, 109], while others are proposing solutions to *directly decode* the brain tasks without any need for subject adaptation or training. The work in this thesis can be categorized in the latter group. Our BCI design is based on open-loop approach, where the subject is not provided with any type of neuro-feedback. As a result, the user has no information on whether or not the BCI system has been able to successfully decode the brain task; and hence he/she can perform the regular brain activity without receiving any reward/penalty from the BCI system.

[6] The alpha rhythm ($8-12$ Hz) which is recorded over the sensorimotor cortex is usually called mu (μ) rhythm.
[7] For a more detailed discussion on this issue and specific examples, see [15, Section 2.2] and [107, Page 524].

Chapter 3

General Framework for Spatio-Spectral Feature Extraction in MI-BCI

In the previous Chapter, it was mentioned that several studies on EEG signals have shown that during motor-imagery (MI) tasks, EEG exhibits event related desynchronization (ERD) or synchronization (ERS) over the alpha band and beta band [35,38,86,87]. For each motor-imagery task, these ERD/ERS effects vary across different cortical areas. As a result, a specific spatio-spectral pattern corresponds to each task, which can be used to classify it. Based on these properties, several methods have been proposed in the literature to extract the task-related spatio-spectral features from EEG signals.

In this chapter, we provide a general framework which categorizes the spatio-spectral feature extraction (FE) algorithms into domain-specific FE methods (DS-FE) and domain-agnostic FE methods (DA-FE), as illustrated in Figure 3.1. The former group consists of FE methods that are designed and used based on the prior knowledge about the neurophysiological characteristics of the EEG signals, whereas the latter group mostly consists of methods that are generic solutions for feature extraction or dimensionality reduction which are widely used in the pattern recognition literature. In Sections 3.1-3.3, we elaborate more on this framework and how existing solutions for motor imagery BCI fit into it.

Based on the proposed framework, we argue that the spatio-spectral features that are extracted at the domain-specific FE step construct a *matrix-variate* structure, which has been ignored in all the existing motor-imagery BCI solutions. Therefore, as part of our proposed framework, we suggest to

utilize feature extraction methods that can exploit this matrix-variate structure. Towards this end, we propose to model the spatio-spectral EEG features using the matrix-variate Gaussian distribution, as described in Section 3.4.

The first three sections of this chapter include a brief overview of the proposed framework, and the following existing solutions in the literature fit into this framework:

1. Domain-Specific Feature Extraction (DS-FE)
   - 1.1. Spatial FE
     - Surface Laplacian (SL) Filtering*
     - Independent Component Analysis (ICA)
     - Phase Locking Value (PLV)
     - Common Spatial Patterns (CSP)*
   - 1.2. Spectral FE
     - Bandpass Filtering*
     - Nonparametric Spectrum Estimation
       - Short-time Fourier Transform*
       - Wavelet Transform
     - Parametric Spectrum Estimation
       - Autoregressive (AR)
       - Adaptive Autoregressive (AAR)
   - 1.3. Spatio-Spectral FE
     - Spectral Coherence
     - Directed Transfer Function (DTF)
     - Spectrally-Filtered Extension of CSP*
2. Domain-Agnostic Feature Extraction (DA-FE)
   - Principle Component Analysis (PCA)
   - Linear Discriminant Analysis (LDA)*
3. Classification
   - Linear*
   - Naive Bayesian Parzen Window*

Among these methods, the ones that will be used in the later chapters in this thesis are marked by asterisk (*) in the above list and will be discussed with more detail in this chapter[1]. It should be noted that the main purpose for overviewing these methods in this chapter is to illustrate how the

[1]Note that surface Laplacian (SL) filtering was covered in Section 2.3 in the previous chapter.

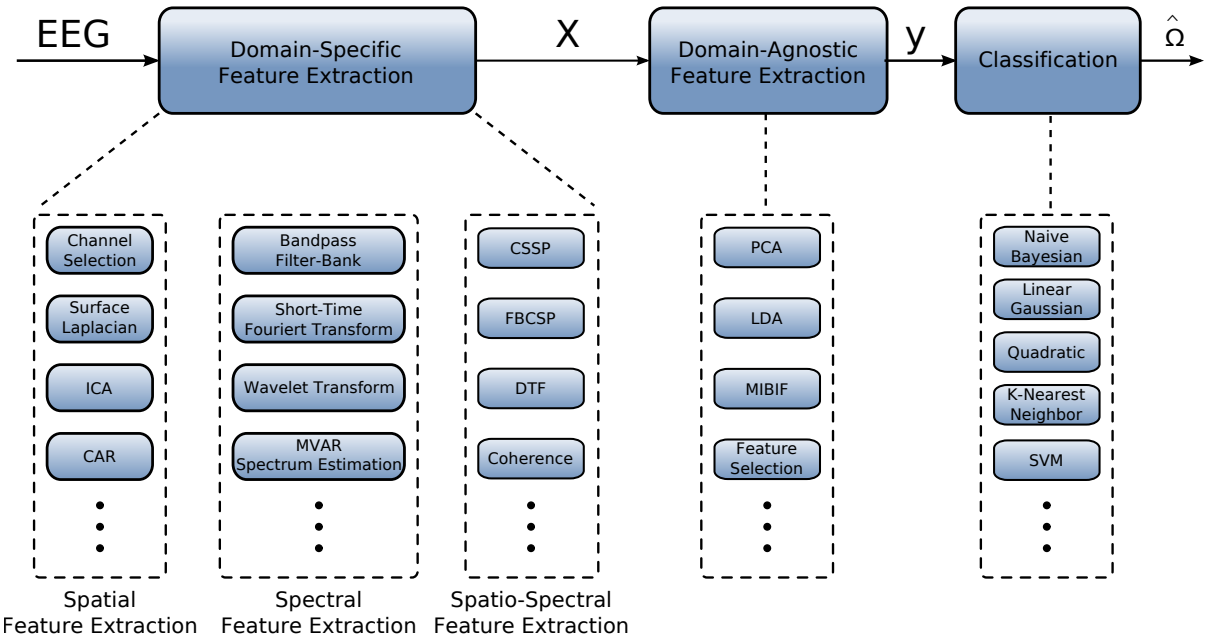


Figure 3.1: The general framework for spatio-spectral feature extraction in motor imagery BCI systems.

aforementioned feature-matrix is extracted/classified in the existing solutions. A comprehensive analysis of these solutions, however, is outside the scope of this thesis and the reader is referred to [16, 110] for further information about these methods.

3.1 Domain-Specific Feature Extraction (DS-FE)

The goal of the domain-specific FE is to use our knowledge about the characteristics of the EEG signal in spatial or spectral domains to transform the raw EEG data from its original representation space into a feature space in which the MI tasks are more separable, based on a desired measure of separability. From this perspective, many of the spatial and spectral transformation/filtering methods used in MI-BCIs can be categorized as domain-specific FE methods. Figure 3.2 demonstrates some of the most common domain-specific FE methods for MI-BICs. Note that in all these methods, the extracted spatial/spectral features are directly related to inherent neurophysiological characteristics of the EEG signal. Moreover, note that in many BCIs, the domain-specific FE step includes a number of spatial FE and spectral FE methods that are combined together. The resulting spatio-spectral feature matrix is denoted by $\mathbf{X} \in \mathbb{R}^{N_f \times N_s}$, where N_f and N_s represent the dimensionality in the spectral and spatial domains.

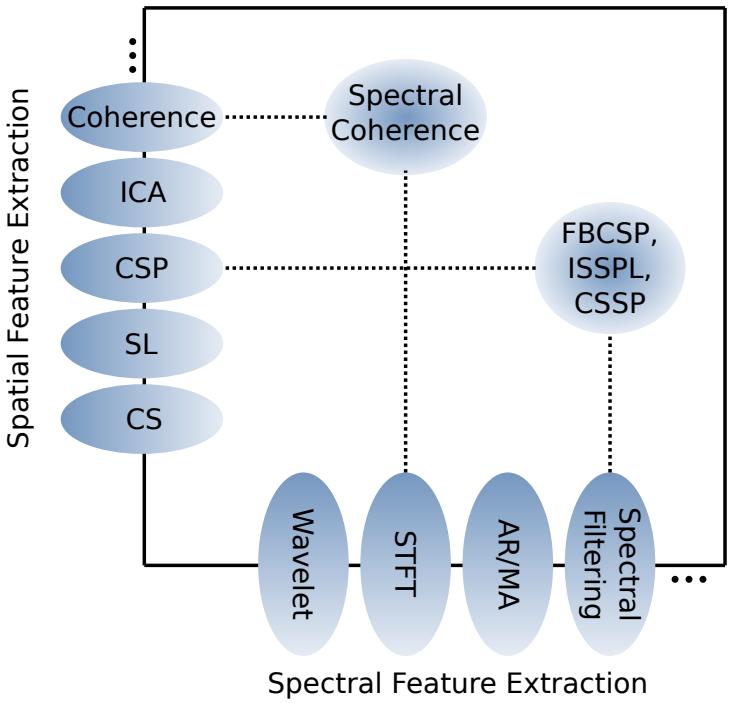


Figure 3.2: Domain-specific methods for extraction of spatio-spectral features.

3.1.1 Spatial FE

In Chapter 2, it was mentioned that spatial characteristics of EEG signals change depending on the type of brain activity. Spatial processing methods such as surface Laplacian (SL) filtering, beamforming, independent component analysis (ICA), phase locking value (PLV) common spatial patterns (CSP), and channel selection (CS) are among the most commonly used spatial FE algorithms in MI-BCIs. Various combinations of these methods can also be used in a MI-BCI (e.g., SL together with CS). A quick review of these methods is provided below.

As mentioned in Section 2.3, the surface laplacian method can be viewed as a highpass spatial filter that removes the non-localized signal components as well as the interference from neighbouring areas caused by volume conduction [26,42]. Independent component analysis is an unsupervised method which is widely used to decompose the EEG signal into independent sources. ICA is used both for removing the artifacts and for extracting the discriminant features from the EEG [41]. Channel selection (CS) is a strategy to reduce the dimensionality of the original EEG signal by only selecting the most informative EEG channels. CS can be performed either using automated machine learning algorithms or based on our prior neurophysiological knowledge about the cortical areas that will be mainly activated during a certain motor-imagery task [40].

Phase Locking Value (PLV)

During the last five years, a few studies have suggested to measure the phase coupling (or phase locking) of oscillatory activities from different parts of the brain, and use these measurements as discriminative features for BCI applications [111–113]. These algorithms make use of the fact that different neural assemblies in the brain are temporarily synchronized during performing perceptual, cognitive, and motor functions [114, 115].

The *phase locking value* is defined as follows. Assume that $x_1(t)$ and $x_2(t)$ are the signals corresponding to two electrodes which are recording EEG signals. $\phi_1(t)$ and $\phi_2(t)$ are defined as the corresponding instantaneous phases of these two signals.[2] Now, these two electrodes are called phase locked if $\Delta\phi(t) = \phi_1(t) - \phi_2(t) = constant$. By definition, the magnitude of the average value of $e^{j\Delta\phi(t)}$ over a short time interval will be considered as the *phase locking value*, i.e., $PLV = |E\{e^{j\Delta\phi(t)}\}|$.

Due to the low signal to noise/interference ratio in EEG signals, the PLV measurements are highly sensitive to the choice of reference electrode during data recording. In some cases, the synchrony reported in some studies have been proved to be a result of an exaggeration of the common contribution of the reference electrode. Nevertheless, it is still believed that the phase synchrony, if properly measured, conveys valuable information about the cognitive tasks in the brain (see [115] and the discussion therein).

Beamforming

Beamforming is a well-known approach in the array processing literature which has recently been also used for analysis of brain signals in the context of BCI systems. Beamforming was first deployed for analysis of MEG signals about two decades ago (e.g., see [116–118]), and with a long delay it has recently been used for EEG-based BCI systems (see [119, 120] and references therein). The main goal of beamforming is to linearly combine the EEG recordings from different sensors to emphasize the signal contribution from sources located in a certain part of the brain while suppressing the effect of all other sources. This technique can be used to find the location, magnitude, and direction of current sources inside the brain during different brain tasks.

Similar to the surface Laplacian filtering method, beamforming is an unsupervised method which does not require any labeled EEG data for training, but instead requires the knowledge of exact sensor locations. While surface laplacian only focuses on the radial current sources that are located on the surface of the cortex, beamforming is more flexible and can detect other types of sources as well. It is also worth mentioning that beamforming is mostly efficient when EEG signal is collected using a high

[2] $\phi_i(t)$ can be determined as follows: $\phi_i(t) = \arctan(\widetilde{x}_i(t)/x_i(t))$, where $\widetilde{x}_i(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x_i(\tau)}{t-\tau}d\tau$ is the Hilbert transform of $x_i(t)$.

density array of sensors (i.e., when N_{ch} is in the order of a hundred sensors or more).

There are two major approaches for utilizing beamforming in BCI systems. In the first approach, beamformers are used to scan the regions of interest in the brain on a voxel-by-voxel basis and locate the corresponding current sources. In the second, and most recent, approach the beamforming technique is used to selectively filter out the effect of sources located in the brain regions that are less likely to be active during the studied brain task. As an example, during motor-imagery tasks, the beamformer can be used to filter out any signal component that is generated by sources outside the motor cortex, and hence remove the effect of artifacts and interference from other parts of the brain.

Common Spatial Patterns (CSP)

CSP is a supervised method that was originally proposed for spatial FE in a binary classification scenario [36,39,121]. CSP method was first used in BCI systems with two-class problem, such as left hand movement vs. right hand movement. Given a set of training data, this algorithm tries to find spatial filters that maximize the variance for one class while minimizing the variance of the other class. In the case of ERD/ERS effects of left/right hand movement, this criterion completely matches the characteristics of EEG signals, since during the hand movement imagination, the power of ipsilateral channels is maximized (ERS) while the power of contralateral channels is minimized (ERD).

Let $\mathbf{S} \in N_t \times N_{ch}$ denote the EEG signal with N_{ch} channels and N_t temporal samples per channel, and let $\mathbf{\Sigma}_1$ be the spatial covariance matrix of the EEG data recorded during the left hand movement imagery task, i.e., $\Sigma_1 = E\{\mathbf{S}^T\mathbf{S}|\Omega_1\}$. Similarly, we can define $\mathbf{\Sigma}_2$ for the right hand movement imagery task. Provided that these two covariance matrices are correctly estimated during training period, the CSP algorithm finds a mapping matrix $\mathbf{W}$ such that:

$$\mathbf{W}^T\mathbf{\Sigma}_1\mathbf{W} = \mathbf{\Lambda}_1 \tag{3.1}$$

$$\mathbf{W}^T\mathbf{\Sigma}_2\mathbf{W} = \mathbf{\Lambda}_2 \tag{3.2}$$

$$\mathbf{\Lambda}_1 + \mathbf{\Lambda}_2 = \mathbf{I} \tag{3.3}$$

where $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are diagonal generalized eigenvalue matrices, and $\mathbf{I}$ is the identity matrix. Each column of matrix $\mathbf{W}$ can be considered as a spatial projection vector or a *spatial filter*.

The condition of Equation 3.3 on the eigenvalue matrices is illustrated in Figure 3.3. In this figure, each circle represents an eigenvalue, and the diameter of the circle is proportional to the magnitude of the eigenvalue. Due to the condition on eigenvalue matrices in Equation 3.3, if we sort the eigenvalues in $\mathbf{\Lambda}_1$ in descending order, the corresponding eigenvalues in $\mathbf{\Lambda}_2$ will be sorted in ascending order. Therefore,

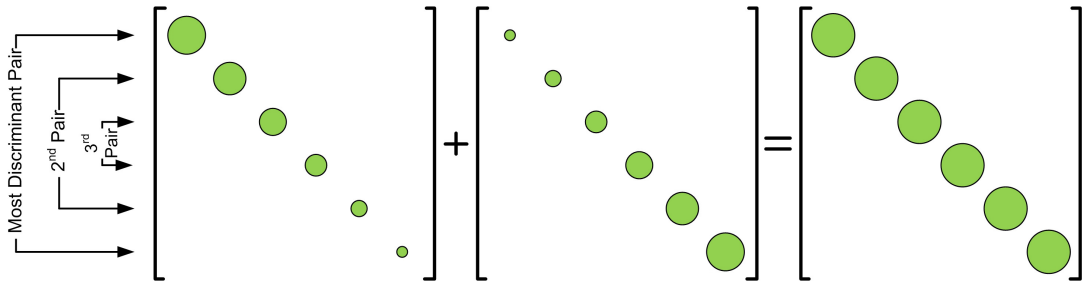


Figure 3.3: Illustration of sorted eigenvalues in the diagonal generalized eigenvalue matrices $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, when $\mathbf{\Lambda}_1 + \mathbf{\Lambda}_2 = \mathbf{I}$

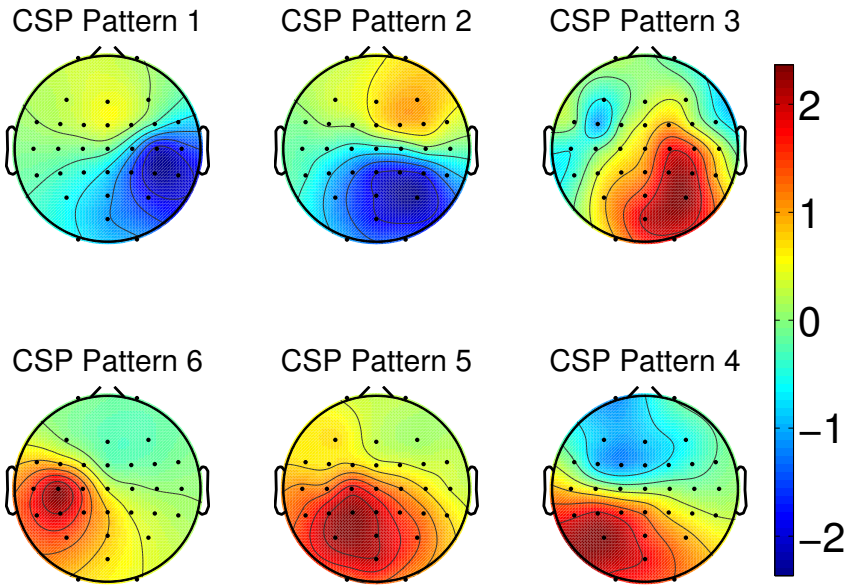


Figure 3.4: Spatial pattern pairs extracted by the CSP algorithm for the left/right hand movement imagery task. Patterns 1 and 6 represent the most discriminant pairs, whereas patterns 3 and 4 represent the least discriminant pair.

the projection vector in $\mathbf{W}$ that maximizes variance of the first class, minimizes the variance of the second class and vice versa. As a result, the first and last columns of $\mathbf{W}$, which correspond to the aforementioned projection vectors, form the most discriminant spatial filters. Similarly, the first and last columns of the matrix $\mathbf{W}^{-1}$ can be considered as the pair of *spatial patterns*, which have the most contribution to the left hand movement and right hand movement imagery tasks, respectively.

Figure 3.4 illustrates an example of the spatial patterns resulted from CSP algorithm for left/right hand movement tasks. These spatial patterns can be grouped as the following pairs: (1,6), (2,5), and (3,4), where the first pair represent the most discriminant pair of features and the last pair represent the least discriminant pair. These spatial patterns show how different cortical regions are activated or deactivated during the left/right hand movement task. Patterns 1-3 represent event-related desynchro-

nization (ERD) in the left hemisphere, where as Patterns 4-6 represent ERD in the right hemisphere, which agrees with our discussion in 2.2.4.

The CSP algorithm has also been generalized to find the spatial patterns for a multi-class BCI system, where the number of imagery tasks is more than 2 (see [122]). In the literature, CSP has been used together with time/frequency domain processing methods to improve the overall performance of the BCI (see [123] and the references in [53]). However, it should be noted that CSP has some disadvantages as well. One of the most important problems of CSP is its sensitivity to artifact, which becomes a critical problem when the size of training data is small. This problem has recently been addressed in [124], where the small sample size problem has been solved using generic learning algorithm. Another disadvantage of CSP is its sensitivity to the location of the sensors. This problem manifests itself when EEG data are collected in different sessions and the sensor locations may not be exactly the same for different sessions.

3.1.2 Spectral FE

There are three major approaches for extraction of spectral features in MI-BCIs: spectral filtering, nonparametric spectrum estimation, and parametric spectrum estimation. Spectral filters, such as bandpass filters, are mostly used to extract different EEG rhythms. Bandpass filters can be utilized for extracting bandpower features. They can also be deployed together with multiband extensions of CSP algorithm (ref. Section 3.1.3).

Nonparametric spectrum estimation methods include short-time Fourier-transformation (STFT), wavelet transformation (WT), and Fourier transformation of the windowed autocorrelation function. Parametric spectrum estimation methods include autoregressive/moving-average (AR/MA) methods and their variants such as adaptive AR (AAR) or multivariate AR (MVAR) methods (ref. [45] for a comparative study of different spectral FE methods). In many MI-BCIs, a combination of features obtained from both parametric and nonparametric methods is used [43, 44, 47]. Parametric spectrum estimation methods can also be deployed in conjunction with the directed transfer function (DTF) method which will be discussed in Section 3.1.3.

Bandpass filtering

In many MI-BCI systems, a bank of digital bandpass filters will be used to extract different EEG rhythms from the raw data. In general, both finite impulse response (FIR) and infinite impulse response (IIR) filters can be used for the EEG signals. The main advantages of FIR filters are the following: (a) FIR filters preserve the phase information of the signal, and hence do not cause any distortion; (b) FIR

filters allow for more control over the frequency response of the filter, as they have more degrees of freedom compared to the IIR filters. However, the computational cost of implementing an FIR filter in a realtime BCI system is prohibitive in most cases. In contrast, IIR filters are significantly more computationally efficient and at the same time introduce relatively low delay in the system. As a result, in most motor-imagery BCI systems, IIR filters are used for filtering the data.

Nonparametric spectrum estimation

A great number of spontaneous BCI systems are based on utilization of power spectral density of the EEG signals. As it was mentioned in Chapter 2, frequency components between $8-30$ Hz can be utilized as discriminative features for motor imagery brain activities [125]. However, due to the time-varying nature of EEG signals, the spectral analysis methods cannot perform well unless they are applied to EEG signals with short length. As a result, joint time-frequency analysis methods are usually used in BCI systems. In this section, we give a brief overview of *short-time Fourier transform* (STFT) and the *wavelet transform* (WT), both of which are nonparametric spectrum estimation methods, and adaptive autoregressive method which is a parametric spectrum estimation method.[3]

Short-Time Fourier Transform (STFT): STFT is a time dependent Fourier analysis which is applied to a windowed segment of the signal. The continuous-time STFT is defined as follows: $Y_i(t,\omega) = \int_{-\infty}^{\infty} y_i(\tau)w(\tau - t)e^{-j\omega\tau}d\tau$, where $y_i(t)$ is the output of the i^{th} EEG channel, and $w(t)$ is a window function, such as Tukey, Hamming, Hann, or Gaussian window, used to suppress the discontinuities at the interval edges. Since EEG signals are usually recorded in discrete-time format, the discrete-time STFT is usually used, which is defined as follows: $Y_i(n,\omega) = \sum_{k=-\infty}^{\infty} y_i(k)w(k-n)e^{-j\omega k}$. $Y_i(n,\omega)$ is in general a complex-valued two-dimensional signal; however, most of the studies in the literature only consider the power of these frequency components and ignore the information conveyed in their phases (ref. Chapter 6)

The STFT method suffers from a tradeoff between the time and frequency resolution. If the width of window is selected to be a small value, we will get a high temporal resolution but low spectral resolution, and vice versa. Nevertheless, due to its low complexity, STFT is widely used in the literature for analysis of EEG signals.

Wavelet Transform (WT): In order to solve the resolution tradeoff mentioned in the previous part,

[3] A comparative analysis of some of these spectral processing methods can be found in [45].

wavelet transforms (WT) can be used. A continuous-time WT is defined as follows:

$$Y_i(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} y_i(t) \psi^* \left(\frac{t-b}{a} \right) dt \tag{3.4}$$

where $\psi(t)$ is called the mother wavelet and should be continuous in both time and frequency domains. In this equation, a is a positive value, called the scale parameter, and b is called the position parameter. Both real-valued and complex-valued mother wavelets have been used in the literature. The wavelet transform has the advantage that it uses a short window for high frequency components and a long window for low frequency components, whereas STFT has a fixed window length for all frequency components. As a result the WT provides a high temporal resolution for rapidly changing high frequency components, while providing a high spectral resolution for long term low frequency components. This property of wavelet transform matches the characteristics of EEG signals, and makes the WT a useful tool for analysis of these signals [126, 127].

In the context of EEG analysis, usually discrete-WT is used in the literature since continuous-WT generates a highly correlated and redundant representation of the signal [127]. Recently, however, a few works have suggested that these redundancies in the continuous-WT can be exploited to improve the performance of the BCI system [128, 129].

Parametric spectrum estimation

Both STFT and WT methods explained in the previous sections are considered as nonparametric spectral estimation algorithms. As an alternative approach, we can make use of parametric methods. The main idea here is to fit a parametric model, such as linear predictive models of Section 2.4, to the EEG signal; and then use this model to estimate the signal's power spectrum.

Due to the time varying nature of EEG signals, adaptive AR models are commonly used in the literature for spectrum estimation of these signals. Let assume that the EEG signal is modelled with an AAR model given by Equation (2.9), and the model parameters have been determined as mentioned in Section 2.4. By taking the Fourier transform of both sides of the equation, we get:

$$|Y_i(\omega)|^2 = \frac{\sigma_x^2}{\left|1 + \sum_{k=1}^{p_i} a_{i,k} e^{-j\omega k}\right|^2} \tag{3.5}$$

One of the main advantages of AAR approach over nonparametric methods is that AAR modelling does not require any windowing of the observed data. This in turn results in a better spectral estimation specially when the length of the observed data is short. This property of the AAR method, is of great

interest since the nonstationary structure of EEG signals usually forces us to estimate the spectrum based on a short observation period.

3.1.3 Spatio-Spectral FE

For joint extraction of spatio-spectral features, three different approaches have been studied in the BCI literature:

(a) spectral coherence analysis,

(b) directed transfer function (DTF),

(c) spectrally-filtered extension of CSP.

The first two methods are based on the fact that several parts of the brain are involved during any mental activity, and the associated signals are communicated between these parts (ref. Section 2.1.1). Such a communication between different parts of the brain requires a type of temporary synchrony between these parts during the communication period. The goal of spectral coherence analysis and directed transfer function methods is to detect these transient synchronizations in order to study the corresponding mental tasks. In contrast to the first two approaches, the last approach is based on extending the main concept of common spatial patters such that it can also take into account the spectral characteristics of the EEG signals.

Spectral Coherence

One of the commonly used algorithms for analysis of synchronization between different EEG channels, is the measurement of coherence between individual frequency components of the signals in these channels [52]. By definition, the *spectral coherence* between channel i and channel j at frequency ω is defined as:

$$\mathrm{Coh}_{ij}^2 = \frac{\left|E\{C_{ij}(\omega)\}\right|^2}{E\{C_{ii}(\omega)\}\ E\{C_{jj}(\omega)\}} \tag{3.6}$$

where $C_{ij}(\omega) = Y_i(\omega)Y_j(\omega)$ can be viewed as the Fourier transform of the cross-correlation between the signals at channels i and j, i.e., $y_i(t)$ and $y_j(t)$ respectively. These spectral coherence measures can be used to study mental tasks which involve distant cortical areas.

Although Coh_{ij}^2 is a reasonable measure of coherence between two channels, the methods based on this measure suffer from the following problem. The spectral coherence does not provide any information regarding the timing and direction of coupling between two channels. In other words, when Coh_{ij}^2 has a

large value, it only represents the high amount of coupling between channel i and j; however, it cannot be determined whether or not this coupling has occurred at the same time instance or one of the channels has had a time lag with respect to the other.

Directed Transfer Function (DTF)

In order to solve the drawbacks of spectral coherency algorithm, the directed transfer function (DTF) algorithm has been proposed in [130], which utilizes a multivariate autoregressive (MVAR) model. Let $\mathbf{Y}(n) = [y_1(n), \ldots, y_{N_{ch}}(n)]^T$ be the vector of N_{ch} EEG channel outputs at time n. If we use an MVAR model of order p for $\mathbf{Y}(n)$, we will have

$$\mathbf{Y}(n) = -\sum_{k=1}^{p} \mathbf{A}_k \mathbf{Y}(n-k) + \mathbf{X}(n), \tag{3.7}$$

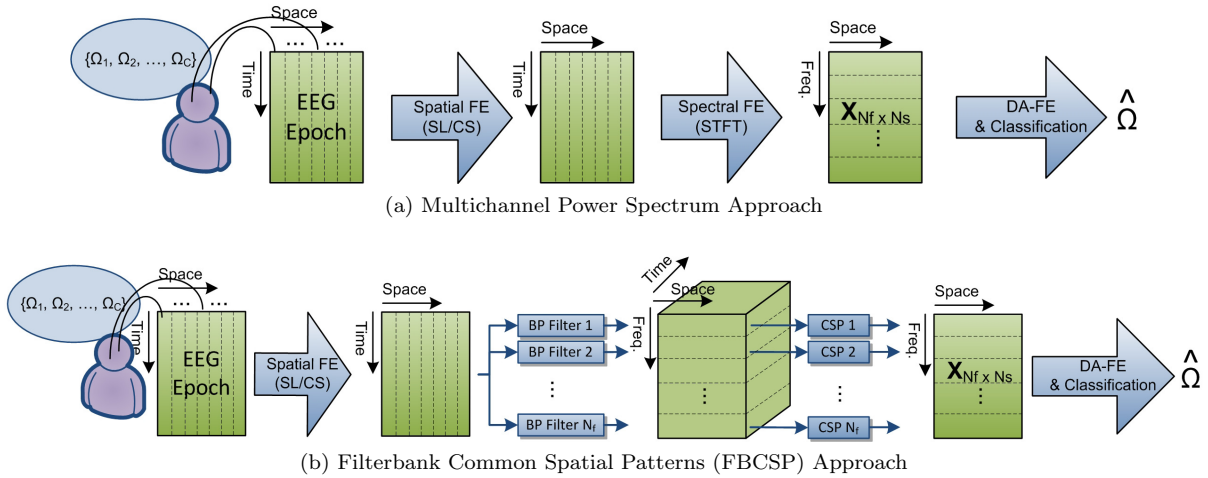
where $\mathbf{A}_k$ is an $N_{ch} \times p$ matrix of model coefficients and $\mathbf{X}(n)$ is a zero mean white noise. This MVAR model, can be represented in frequency domain as follows: $\mathbf{A}_f(\omega)\mathbf{Y}(\omega) = \mathbf{X}(\omega)$, where $\mathbf{A}_f(\omega) = \mathbf{I} + \sum_{k=1}^{p} \mathbf{A}_k e^{-j\omega k}$, and $\mathbf{X}(\omega) = \sigma_x^2 \mathbf{I}$. Thus, the transfer function of this MVAR system, can be expressed as: $\mathbf{H}(\omega) = \mathbf{A}_f^{-1}(\omega)$. Using this transfer matrix, the DTF value between channels i and j can be defined as follows [131]:

$$\Theta_{ij}^2 = \left|H_{ij}(\omega)\right|^2 \tag{3.8}$$

This value represents the *causal*[4] influence from channel j to channel i, which has been shown to be an important feature in detection of motor imagery tasks in BCI systems.

DTF algorithm is suitable for analysis of complicated motor tasks which require a more detailed analysis of the spatio-temporal characteristics of EEG signals. However, it should be noted that this benefit comes at the cost of computational complexity of this algorithm. Indeed, both spectral coherence and DTF methods are based on pairwise analysis of the EEG channels, and hence the dimensionality of their resulting spatio-spectral feature matrix is significantly high, compared to other domain-specific FE methods. This high dimensionality imposes challenging issues on the computational complexity and overall performances of the BCI system. As a consequence, currently the usage of spectral coherence and DTF methods is mostly restricted to analysis of the brain dynamics during MI tasks rather than classification purposes.

[4]By definition, $y_j(t)$ is causal to $y_i(t)$, if $y_i(t)$ can be causally predicted from $y_j(t)$



(a) Multichannel Power Spectrum Approach

(b) Filterbank Common Spatial Patterns (FBCSP) Approach

Figure 3.5: Most commonly used schemes for domain-specific spatio-spectral FE in motor-imagery BCI systems.

Spectrally-Filtered Extension of CSP

As mentioned in Section 3.1.1, the common spatial patterns (CSP) method is one of the most successful techniques for analysis of the motor-imagery tasks. However, it only relies on the spatial features and completely ignores the spectral characteristics of the EEG signal. To alleviate this problem, several works have suggested to utilize CSP together with spectral filters [51, 64–67]. Among these solutions, the work in [49], called filterbank CSP (FBCSP), is the most recent and most successful approach. Indeed, a large number of previous CSP extensions can be considered as a simplified version or special case of the work in [49].

FBCSP method is a multiband extension approach, in which a set of bandpass filters are used to extract different rhythmic activities of the brain from EEG signal. Each of these EEG rhythms is then passed to a separate CSP module to extract the spatio-spectral features corresponding to that frequency range. This scheme is illustrated in Figure 3.5. As illustrated in this figure, the FBCSP may also be preceded by simple spatial feature extraction methods such as surface Laplacian (SL) or channel selection (CS).

Figure 3.6 illustrates a simple example of the set of spatio-spectral patterns that FBCSP method generates for left/right hand motor imagery tasks. In this example we have used a set of seven bandpass filters, each of which with a passband of 4 Hz, to cover the range of $4-32$ Hz. For each frequency band, the two most discriminant pair of patterns are presented as Patterns (1,4) and Patterns (2,3). If we compare these patterns with the ones shown in Figure 3.4, it can be seen that the FBCSP provides more details regarding the spectral dependencies of the patterns, which is not available in the conventional CSP method.

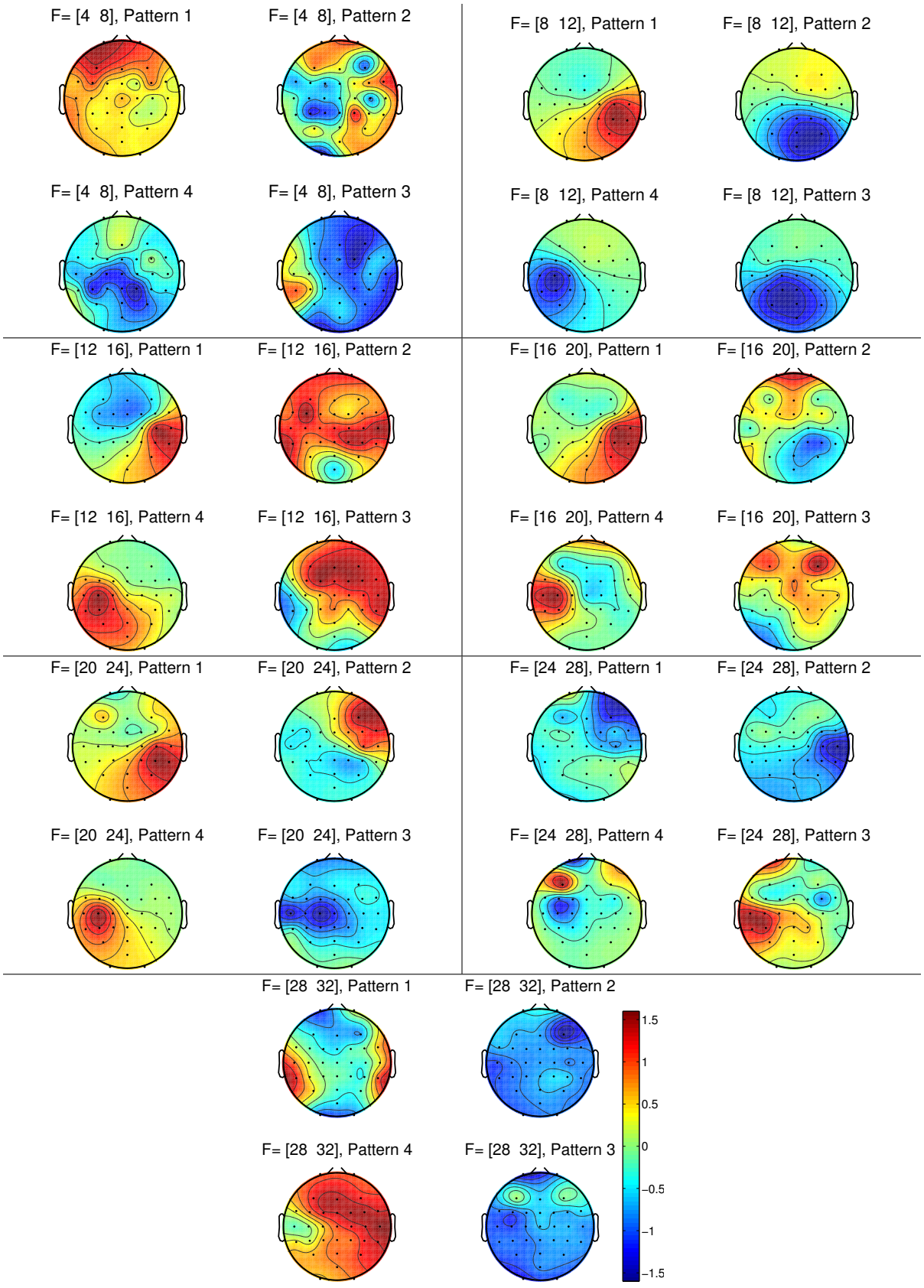


Figure 3.6: Spatio-spectral patterns obtained from FBCSP method for right hand motor imagery vs left hand motor imagery. For each frequency band, two most discriminant pattern-pairs are illustrated, which are presented as patterns (1,4) and (2,3).

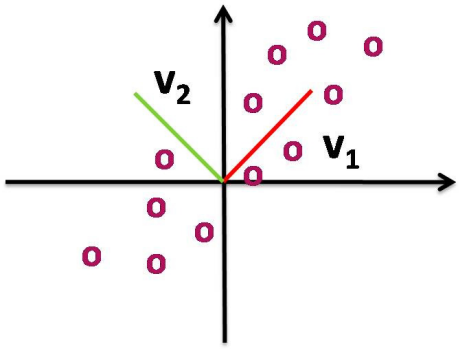


Figure 3.7: Using the PCA method to reduce the dimensionality of the data from two to one. The data will be mapped to v_1 direction (red direction) which provides a better representation of the data distribution, compared to the v_2 (green direction).

3.2 Domain-Agnostic Feature Extraction (DA-FE)

The feature space resulting from DS-FE in Figure 1.3 is usually a high-dimensional space which contains correlated or redundant components. This calls for the use of the domain-agnostic feature extraction to reduce the dimensionality of $\mathbf{X}$ prior to the classification step. We name this step *domain-agnostic* since, unlike the domain-specific FE step, the feature extractors used in this step do not depend on our knowledge about the neurophysiological characteristics of the EEG signals. Indeed, the DA-FE step usually consists of generic dimensionality reduction algorithms.

DA-FE methods can be categorized into two groups, as follows: (a) Methods such as principle component analysis (PCA) and linear discriminant analysis (LDA) that first transform the spatio-spectral features of $\mathbf{X}$ into a new feature space and then select the most discriminant components in the new feature space [132]. (b) Methods that do not require any transformation and directly select the most discriminant features from $\mathbf{X}$ based on a desired measure, such as mutual-information or correlation with the task labels [133–135].

Principle Component Analysis

The principle component analysis (PCA) method is an unsupervised algorithm, that tries to retain those directions in the feature space which convey most of the data variations, while discarding the directions that have little contribution to the data variations. Assuming that $\Sigma_{\mathbf{x}}$ represents the covariance of the spatio-spectral features at the output of domain-specific feature extraction step, the PCA retains feature directions that correspond to the most significant eigenvalues of $\Sigma_{\mathbf{x}}$. Figure 3.7 illustrates a simple example for a two-dimensional feature space. The v_1 and v_2 vectors in this figure represent the eigenvectors of $\Sigma_{\mathbf{x}}$, where v_1 corresponds to the larger eigenvalue.

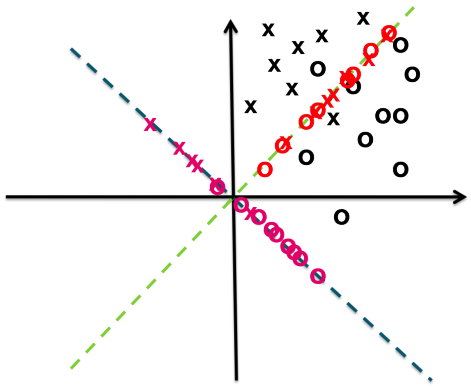


Figure 3.8: Comparison of PCA and LDA methods. Utilization of the PCA method results in projecting the data on the green direction due to the distribution of data points. However, LDA method lead will select an orthogonal direction, shown in black, which maximizes the class separability while minimizing the data variations within each class.

Linear Discriminant Analysis (LDA)

Unlike PCA, the linear discriminant analysis (LDA) method is a supervised approach for feature extraction, which takes advantage of labeled training data to find the desired set of features. LDA aims to linearly map the input data into a feature space for which the variations within each class is minimized while the distances between the means of different classes are maximized. Let $\mathbf{m}_1$, ..., $\mathbf{m}_C$ represent the mean vectors for classes $\Omega_1, \cdots, \Omega_C$, and $\mathbf{m}$ represent the total mean. Then, the within class scatter and the between class scatter matrices will be defined as follows:

$$\mathbf{S}_W = \sum_{i=1}^{C} E\{(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T|\Omega_i\} \tag{3.9}$$

$$\mathbf{S}_B = \sum_{i=1}^{C} (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T \tag{3.10}$$

In Equation (3.9), the term $E\{(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T|\Omega_i\}$ represents the scatter of the samples in class Ω_i around their corresponding mean, i.e., $\mathbf{m}_i$. In the LDA approach, it is assumed that the scatter of samples in all the classes are the same. Therefore, the matrix $\mathbf{S}_W$ represents the averaged scatter within different classes, hence it is called the *within class* scatter matrix. In Equation (3.10), the matrix $\mathbf{S}_B$ represents the scatter of different class means around the total mean $\mathbf{m}$, hence it is called the *between class* scatter matrix. It is worth mentioning that rank of $\mathbf{S}_B$ is at most equal to $C-1$; therefore, $\mathbf{S}_B$ will be a singular matrix when dimensionality of $\mathbf{x}$ is greater than $C-1$, which is usually the case in the motor-imagery BCI applications.

The LDA algorithm tries to find the transformation vector(s) $\mathbf{w}$ that maximizes the following measure:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{3.11}$$

In this equation, the numerator represents the within class scatter value of the transformed data, i.e., $\mathbf{w}^T\mathbf{x}$. Similarly, the denominator represents the between class scatter value of the transformed data. Therefore, the above measure, which is known as the generalized Rayleigh quotient, determines the ratio of the within class scatter over the within class scatter in the new feature space. The transformation $\mathbf{w}$ that maximizes this measure should satisfy (3.12).

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w} \tag{3.12}$$

Solving Equation 3.12 requires calculation of the generalized eigenvalues of $\mathbf{S}_B$ and $\mathbf{S}_W$ (ref. [136]). Since rank of $\mathbf{S}_B$ is at most $C-1$, the LDA method can provide up to $C-1$ orthogonal transformations for extraction of the most discriminant features.

In general, it can be shown that if the data $\mathbf{x}$ has the following two conditions, the $C-1$ transformations derived by LDA approach provide the minimum-dimension sufficient statistics for classification of $\mathbf{x}$ that conveys all the discriminant information of the data [39, 137]:

(a) $\mathbf{x}|\Omega_i$ has a normal distribution, i.e., the class conditional mean and covariance of the data completely describes the statistical characteristics of the data during each motor-imagery task.

(b) $\mathbf{x}$ is homoscedastic, i.e., the conditional covariances are the same for all the classes, and hence the only difference between distributions of different classes is the conditional mean of the data.

In other words, under the above two conditions the LDA method can reduce the dimensionality of the feature space to $C-1$, while guaranteeing that all the discriminant information of the data is preserved.

Figure 3.8 depicts an illustrative example to compare PCA and LDA methods, both trying to reduce the dimensionality of the feature space from two to one. The original data belonging to Class-1 and Class-2 are marked by black crosses and circles, respectively. The PCA method selects the direction shown by green dashed line for mapping the data, whereas LDA method selects the orthogonal direction shown in black dashed line. It can be seen that data points mapped by the LDA are more separable than the points mapped by the PCA method. Note that both PCA and LDA algorithms involve eigen decomposition and mapping the data along the eigenvectors corresponding to the largest eigenvalues. However, PCA is an unsupervised method which only has access to the total scatter of the data, whereas

LDA is a supervised method that takes advantage of the knowledge about individual means and scatters of different classes.

3.3 Classification

After extraction of the most discriminant features through domain-specific and domain-agnostic feature extraction steps, the resulting features will be passed to the classifier. There exist numerous classification methods in the machine learning literature, each of which is designed for a feature space with certain characteristics. Some of the classification methods that are commonly used in the BCI literature include: Naive Bayesian, linear Gaussian, quadratic, support vector machine (SVM), and k-nearest neighbours classifiers. Since the focus of this thesis is on feature extraction algorithms, we refer the reader to [53, 54, 138] for a comprehensive review of various classification methods used in MI-BICs. Throughout this thesis, we will mainly utilize the simple linear classifier (Lin) which classifies each sample based on its distances from the means of different classes, denoted by $\check{m}_i, i = 1, \cdots, C$. The sample would be assumed to belong to class Ω_i if $\check{m}_i$ is the closest class mean to the test sample.

We will also consider the Naive Bayesian Parzen Window (NBPW) classifier as a benchmark for FBCSP-based approaches since it has been shown to provide a competitive performance compared to other FBCSP-based solutions [49]. For any feature vector $\mathbf{x}$, the NBPW method uses the following classification rule to classify the data:

$$\hat{\Omega} = \arg\max_{\Omega_i} p(\Omega_i|\mathbf{x}) \tag{3.13}$$

where $p(\Omega_i|\mathbf{x})$ is determined using the Bayes rule, i.e,

$$p(\Omega_i|\mathbf{x}) \;=\; \frac{p(\mathbf{x}|\Omega_i)p(\Omega_i)}{p(\mathbf{x})} \;=\; \frac{p(\mathbf{x}|\Omega_i)p(\Omega_i)}{\sum_i p(\mathbf{x}|\Omega_i)p(\Omega_i)} \tag{3.14}$$

In order to estimate the conditional probability $p(\mathbf{x}|\Omega_i)$, in the NBPW method it will be naively assumed that the elements of the feature vector $\mathbf{x} = [x_1, \cdots, x_D]$ are conditionally independent, i.e., $p(\mathbf{x}|\Omega_i) = \prod_{d=1}^{D} p(x_d|\Omega_i)$. Finally, the conditional probability of each feature element, i.e., $p(x_d|\Omega_i)$ will be estimated using a Gaussian smoothing kernel function [139, 140], as follows:

$$\hat{p}(x_d|\Omega_i) = \frac{1}{n_i} \sum_{x_{d,j}^{(t)} \in \Omega_i} K\left(\frac{x_d - x_{d,j}^{(t)}}{h}\right) \tag{3.15}$$

where $\mathbf{x}_j^{(t)} = [x_{1,j}^{(t)}, \cdots, x_{D,j}^{(t)}] \in \Omega_i$ denotes the j^{th} sample from the set of *training* feature vectors from class Ω_i, and $K(\cdot)$ is the univariate Gaussian kernel function. The parameter h will be determined based on the standard deviation of x_d [140].

3.4 Matrix-Variate Gaussian Model for Spatio-Spectral Features

In the design of BCI systems, it is crucial to design both domain-specific and domain-agnostic feature extraction steps based on the characteristics of the corresponding spatio-spectral features. In the context of motor-imagery BCI systems, due to the multichannel structure of the EEG data, all of the domain specific FE methods that involve spectral feature extraction will generate a set of features which inherently form a matrix-variate structure.

Figure 3.5(a) illustrates a typical example where the BCI system uses the power spectral features of multichannel EEG signal for classification of the brain tasks. These features can be extracted using parametric techniques (e.g, auto-regressive/moving-average method) or non-parametric techniques (e.g., short-time Fourier transform or wavelet transform). Both cases generate a feature matrix, in which each row represents the set of spatial features that correspond to a certain frequency, and each column represents the set of spectral features that correspond to a certain EEG channel.

Figure 3.5(a) illustrates a similar matrix-variate structure when a joint spatio-spectral FE extraction method such as filterbank CSP is utilized. In this feature matrix, each row represents the spatial patterns corresponding to a certain frequency band, whereas each column represents the set of different spectral feature corresponding to a certain spatial pattern.

Using a similar analogy, it can be readily seen that for all of the domain-specific FE methods that were discussed in Sections 3.1.2 and 3.1.3, the resulting feature set forms a matrix-variate spatio-spectral structure. In this thesis, we argue that this matrix-variate structure conveys important information about the corresponding spatio-spectral features, which has been ignored in the BCI literature for both domain-agnostic FE and domain-specific FE.

Most of the BCI systems in the literature do not consider the joint characteristics of the spatial and spectral features at the domain-specific feature extraction stage. This problem manifests itself when the spatio-spectral features in different bands and/or different channels are extracted independently. One simple example of this case is the FBCSP method, in which the spectral feature extraction is performed using bandpass filters on each channel independent from the other channels, and subsequently the spatial

feature extraction is performed by applying the CSP method on each frequency band independent from the other bands.

Similarly, at the domain-agnostic feature extraction stage, most of the existing methods ignore the inherent structure of the spatio-spectral features that are passed to them by the domain-specific feature extraction methods. One simple example of this case is when generic vector-variate feature extraction methods, such as the LDA or PCA algorithms, are directly applied to the spatio-spectral features by concatenating all the spatio-spectral features into a single feature vector.

In this section we propose a new model for the spatio-spectral EEG features which provides a mathematical framework for both domain-specific and domain-agnostic feature extraction methods to take into account the joint characteristics of the spatial and spatial features. This model is based on the matrix-variate Gaussian assumption for the spatio-spectral EEG features. In order to introduce this model in this chapter, we use the following general notation for the spatio-spectral features. Let X_{ij} denotes the j^{th} spatial feature of the i^{th} spectral band. We construct a feature matrix, denoted by $\mathbf{X} \in \mathbb{R}^{N_f \times N_s}$, which contains all the spatio-spectral features at the output of the domain-specific FE step. Note that this notation can be used for any of the spatio-spectral feature extraction methods reviewed earlier in this chapter. In particular, we will use this notation in the next two chapters for representing the features in the FBCSP method. Nevertheless, the discussions and definitions presented in this chapter are general and are not restricted to the FBCSP algorithm.

3.4.1 Model Definition

Let $f(\mathbf{X}|\Omega_i)$ denote the conditional probability of matrix $\mathbf{X} \in \mathbb{R}^{N_f \times N_s}$ under class Ω_i, and let $P(\Omega_i)$ represent the prior probability of Ω_i. A matrix-variate Gaussian model [141] for the feature matrix $\mathbf{X}$ is denoted by:

$$\mathbf{X}|\Omega_i \sim \mathcal{N}(\mathbf{M}_i, \mathbf{\Phi}_i, \mathbf{\Psi}_i), \qquad 1 \leq i \leq C \tag{3.16}$$

Here, the matrices $\mathbf{M}_i, \mathbf{\Phi}_i, \mathbf{\Psi}_i$ denote the mean, spectral covariance, also called column-wise or left covariance, and the spatial covariance, also called row-wise or right covariance, of the class Ω_i. These

matrices are defined as follows:

$$\mathbf{M}_i = \mathrm{E}_{\mathbf{X}|\Omega_i}\left(\mathbf{X}\right), \tag{3.17}$$

$$\mathbf{\Phi}_i = \mathrm{tr}^{-1}(\mathbf{\Psi}_i) * \mathrm{E}_{\mathbf{X}|\Omega_i}\left((\mathbf{X}-\mathbf{M}_i)(\mathbf{X}-\mathbf{M}_i)^T\right), \tag{3.18}$$

$$\mathbf{\Psi}_i = \mathrm{tr}^{-1}(\mathbf{\Phi}_i) * \mathrm{E}_{\mathbf{X}|\Omega_i}\left((\mathbf{X}-\mathbf{M}_i)^T(\mathbf{X}-\mathbf{M}_i)\right). \tag{3.19}$$

Using this model, knowledge of the parameters $\mathbf{M}_i$, $\mathbf{\Phi}_i$, and $\mathbf{\Psi}_i$ will suffice to determine the conditional probability of $\mathbf{X}$ for different classes, as follows:

$$f(\mathbf{X}|\Omega_i) = \frac{\exp\left\{-\frac{1}{2}\,\mathrm{tr}\left[\mathbf{\Phi}_i^{-1}(\mathbf{X}-\mathbf{M}_i)^T\mathbf{\Psi}_i^{-1}(\mathbf{X}-\mathbf{M}_i)\right]\right\}}{(2\pi)^{\frac{N_fN_s}{2}}\det(\mathbf{\Phi}_i)^{\frac{N_f}{2}}\det(\mathbf{\Psi}_i)^{\frac{N_s}{2}}} \tag{3.20}$$

Vector-variate Gaussianity is a fairly common practical assumption for EEG signals as implied by utilization of relevant methods such as LDA [132]. However, the matrix-variate model in (3.16) corresponds to a specific structure for the covariance of the vectorized data, as follows. Assume a column concatenation operation vec(.) that operates on the matrix-variate data $\mathbf{X}$ and returns $\mathbf{x} = \mathrm{vec}(\mathbf{X})$. Then, the mean of $\mathbf{x}$ in Ω_i equals $\boldsymbol{\mu}_i = \mathrm{vec}(\mathbf{M}_i)$, and assuming that (3.16) holds, the class-conditional covariance of $\mathbf{x}$ equals

$$\mathbf{\Sigma}_i = \mathbf{\Psi}_i \otimes \mathbf{\Phi}_i, \tag{3.21}$$

where $\mathbf{\Sigma}_i \in \mathbb{R}^{N_fN_s \times N_fN_s}$, $\mathbf{\Psi}_i \in \mathbb{R}^{N_s \times N_s}$, $\mathbf{\Phi}_i \in \mathbb{R}^{N_f \times N_f}$, and the $\otimes$ symbol represents the Kronecker product operator (ref. Appendix A.2). Therefore, the matrix-variate Gaussianity implies a *separable* structure for the covariance matrix of the vectorized data as defined by (A.7). Let $m_{kj}^{(i)}$, $\phi_{kj}^{(i)}$, and $\psi_{kj}^{(i)}$ be the $(k,j)th$ elements in $\mathbf{M}_i$, $\mathbf{\Phi}_i$ and $\mathbf{\Psi}_i$ matrices, respectively. Then, Equation (A.7) implies that

$$\mathrm{E}_{\mathbf{X}|\Omega_i}\left((x_{k_1j_1} - m_{k_1j_1}^{(i)})(x_{k_2j_2} - m_{k_2j_2}^{(i)})\right) \;=\; \phi_{k_1k_2}^{(i)}\,\psi_{j_1j_2}^{(i)} \tag{3.22}$$

In other words, the covariance between any two spatio-spectral features can be decomposed into a spatial covariance term and a spectral covariance term. This separability is an important property which will be used in the algorithms proposed in the next two chapters for spatio-spectral feature extraction.

It is also worth mentioning that any bilinear transformation of the form $y = \mathbf{a}^T\mathbf{X}\mathbf{b}$ on the matrix-variate data $\mathbf{X} \in \mathbb{R}^{N_f \times N_s}$ is equivalent to a linear transformation on $\mathbf{x} = \mathrm{vec}(\mathbf{X})$, as follows: $y = \mathrm{vec}(\mathbf{b} \otimes \mathbf{a})^T\mathbf{x}$. In other words, bilinear spatio-spectral filtering of the matrix-variate data is equivalent to a certain class of vectorial filtering that has a Kronecker product structure.

3.4.2 Homoscedastic vs Heteroscedastic Models

The matrix-variate Gaussian model can be assumed to be either heteroscedastic or homoscedastic, depending on the properties of the spatio-spectral EEG features. If the spatial-covariance matrices are the same for all the classes and the spectral-covariance matrices are also the same, then the corresponding model will be called homoscedastic. Otherwise, the model will be called heteroscedastic. In other words, a homoscedastic model requires the following condition to be satisfied:

$$\mathbf{\Phi}_i = \mathbf{\Phi}, \qquad 1 \leq i \leq C \tag{3.23}$$

$$\mathbf{\Psi}_i = \mathbf{\Psi}, \qquad 1 \leq i \leq C \tag{3.24}$$

In the context of BCI systems, both homoscedastic and heteroscedastic assumptions have been used in different methods. As an example the CSP method explained in Section 3.1.1, and all its variants, are based on the heteroscedastic assumption; whereas the LDA method is based on the homoscedastic assumption.

3.5 Summary and Concluding Remarks

In this chapter, a general framework for feature extraction in motor-imagery BCI systems was introduced. The framework encompasses most of the existing solutions for MI-BCI in the literature. Based on this framework, it was shown that the feature sets extracted by most of the domain-specific FE methods form an inherent spatio-spectral feature matrix of the form $\mathbf{X} \in \mathbb{R}^{N_f \times N_s}$, where N_f and N_s represent the dimensionalities in the spectral and spatial domains, respectively. Based on this observation, we proposed to use a matrix-variate Gaussian distribution to model the statistical characteristics of $\mathbf{X}$.

The main difference between the matrix-variate Gaussian model and the conventional multivariate Gaussian model is the restrictive Kronecker structure assumption for the covariance of the data. This specific covariance structure can be exploited by the DS-FE/DA-FE methods to reduce the computation cost of the feature extraction stage. More importantly this assumption allows us to estimate the covariance of the data in lower dimensional spaces of $\mathbf{\Phi}$ and $\mathbf{\Psi}$, and improve the estimation accuracy of the second order statistics of the data, which in turn can improve the overall performance of the system. In the next two chapters, we will study how this matrix-variate Gaussian model can be used in the design of domain-agnostic FE and domain-specific FE methods.

In the next two chapters, we will mainly focus on spatio-spectral feature extraction methods, such as FBCSP, that are based on combination of bandpass filtering and the CSP algorithm. The main reason

for this focus is the fact that CSP has been proven to be one of the best DS-FE algorithms whose theoretical assumptions very well match the neurophysiological properties of the EEG signals during motor imagery tasks. However, the proposed matrix-variate Gaussian model can potentially be used in other spatio-spectral FE approaches that were reviewed in this chapter. In particular, our studies in Chpater 7 will examine the possibility of using the matrix-variate Gaussian model for the complex-valued spatio-spectral features that are generated through Fourier domain analysis of the EEG data.

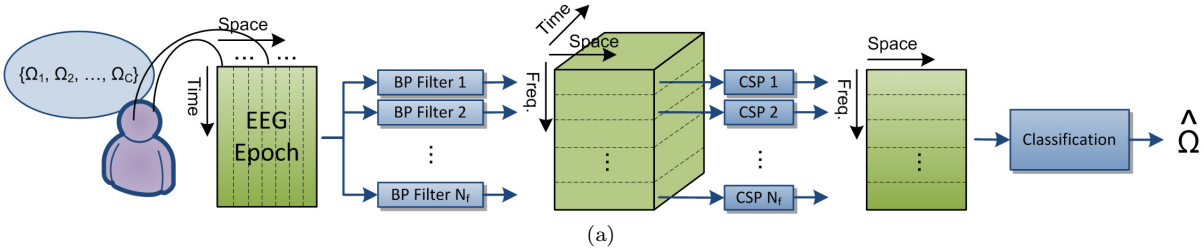
Finally, it is noteworthy that the surface Laplacian and channel selection methods for domain-specific feature extraction, which were briefly reviewed in Sections 2.3 and 3.1.1 are two of the most important yet simple methods which are widely used in the BCI literature in conjunction with other DS-FE algorithms. In this thesis, therefore, we will comprehensively study the effect of each of these methods on the overall performance of the newly proposed algorithms in Chapters 4 and 5.

Chapter 4

Domain-Agnostic FE Based on Matrix-Variate Model for FBCSP Features

In this chapter, we use the proposed framework of Chapter 3 to introduce a new domain-agnostic FE approach for extraction of the most discriminant features from the spatio-spectral matrix $\mathbf{X}$. We argue that the common approach in the BCI literature for domain-agnostic FE, which requires vectorization of the matrix $\mathbf{X}$ by breaking it along the columns (or rows), introduces unnecessary degrees of freedom by ignoring the inherent structure of the data along the broken dimension. In other words, vectorization of $\mathbf{X}$ removes the inherent spatio-spectral structure of the data. This inherent structure can potentially be exploited by the feature extractor to reduce the computational cost and/or improve the accuracy of the overall system.

In this section, we focus on the state of the art filterbank common spatial patterns (FBCSP) method, which is proved to be highly successful as a domain-specific FE algorithm for motor-imagery BCI systems. Following our general reasoning regarding the matrix-variate structure of the extracted spatio-spectral features, we propose to use the FBCSP method in conjunction with matrix-variate (or bilinear) feature extractors in the domain-agnostic FE stage. In particular, we will study the bilinear extensions of the linear discriminant analysis (LDA) method, which is the Bayes optimal strategy for features extraction in homoscedastic Gaussian scenarios. In order to emphasize the importance of the *bilinear* operations in the domain-agnostic FE stage, we will compare the proposed approach with the case where FBCSP



(a)

Figure 4.1: Filter-bank common spatial pattern (FBCSP) method for spatio-spectral feature extraction in a typical motor-imagery BCI system.

is used in conjunction with the conventional LDA method.

4.1 Matrix-Variate Gaussian Model for FBCSP Features

As mentioned in Section 3.1.1, the filterbank common spatial pattern (FBCSP) method is a highly successful multiband extension of the CSP method. In the FBCSP method, first different EEG rhythms are obtained by means of bandpass filtering the EEG signal, and then a bank of CSP modules is deployed to separately extract spatial features from each EEG rhythm; hence the name filter-bank CSP. The resulting features are then used for classification of the EEG data, as illustrated in Figure 4.1.

In this approach, each spectral band is processed independently by a separate CSP module, and hence possible correlations between different EEG rhythms are not considered by the bank of CSP filters. Therefore, the resulting spatio-spectral feature space is potentially redundant and relatively high dimensional. This redundancy in the feature space increases the computational cost of the classification step and can lead to potential performance loss.

Following our discussions in Section 3.4 about the matrix-variate structure of the FBCSP features, In this chapter, we introduce a new approach for domain-agnostic FE in motor-imagery BCIs. In this approach, the matrix-variate structure of the FBCSP features will be taken into account in the domain-agnostic FE step. Towards this end, we adopt a homoscedastic matrix-variate Gaussian model for the FBCSP features, which provides us with an efficient mathematical framework for developing the desired domain-agnostic FE algorithm.

Let $\mathbf{X} \in \mathbb{R}^{N_f \times N_s}$ represent the spatio-spectral feature matrix at the output of the domain-specific FE step, where X_{ij} denotes the j^{th} spatial feature of the i^{th} spectral band. Here, we have assumed that a total of N_f bandpass filters have been deployed, and for each frequency band, a total of N_s spatial

features have been extracted. A homoscedastic matrix-variate Gaussian model for $\mathbf{X}$ implies:

$$\mathbf{X}|\Omega_i \sim \mathcal{N}(\mathbf{M}_i, \mathbf{\Phi}, \mathbf{\Psi}), \qquad 1 \leq i \leq C \tag{4.1}$$

matrices $\mathbf{M}_i$ is the mean of the FBCSP feature matrix during task Ω_i, $\mathbf{\Phi}$ is the spectral covariance of the data,$\mathbf{\Psi}$ denotes the spatial covariance of the data. As mentioned in Section 3.4, the above matrix-variate Gaussian model is equivalent to the following vector-variate Gaussian model for the column-wise vectorized representation of $\mathbf{X}$, denoted by $\mathbf{x} = \text{vec}(\mathbf{X})$,

$$\mathbf{x}|\Omega_i \sim \mathcal{N}(\text{vec}(\mathbf{M}_i), \mathbf{\Psi} \otimes \mathbf{\Phi}), \qquad 1 \leq i \leq C \tag{4.2}$$

It has been shown in the literature that for vector-variate homoscedastic Gaussian data, the linear discriminant analysis (LDA) feature extractor followed by a linear classifier provides the Bayes optimal solution for classification of the data [39, 137]. This motivates us to focus on the LDA-based solutions for classification of the matrix-variate Gaussian data $\mathbf{X}$. A Naive approach is to vectorize the feature-matrix $\mathbf{X}$ and apply the LDA algorithm to the feature vector $\mathbf{x} = \text{vec}(\mathbf{X})$. Theoretically, this approach provides the Bayes optimal solution when the distribution parameters are known. In practice, however, this vectorization of the data unnecessarily increases the dimensionality of the feature space and imposes several challenges in estimation of the distribution parameters, which in turn can lead to a significant performance loss, as will be discussed in the experimental results (ref. Section 4.3). The alternative approach is to deploy a bilinear extension of LDA method as will be discussed in the next section.

4.2 Bilinear Domain-Agnostic FE for Matrix-Variate Gaussian Data

As mentioned in the previous section, the LDA algorithm provides a Bayes optimal solution for feature extraction from homoscedastic vector-variate Gaussian data. In the pattern recognition literature, there exist several works that have attempted to extend the LDA algorithm to be applicable to matrix-variate data [55–63].

The simplest approach is the work in [55–57], in which the LDA is applied only to the rows (or columns) of the matrix $\mathbf{X}$, and provides a one-sided solution for reducing the dimensionality of the $\mathbf{X}$ across either the rows or columns of $\mathbf{X}$. This one-sided approach is not suitable in the context of BCI systems, since it only deals with the feature matrix $\mathbf{X}$ in either the spectral domain or the spatial

domain, while ignoring the other domain.

The second approach is the approach used in [58–62], which involves intuitive two-sided (or bilinear) variations of the LDA method. Among these works, the solution proposed by [58] is one of the most promising methods, which has been shown to provide a competitive performance in the context of image processing. It is worth mentioning that despite its high performance, the work in [58] does not generally provide the Bayes optimal solution [142–144].

The third approach for bilinear extension of LDA is approach taken by the recent works in [63, 73], which have directly used the matrix-variate Gaussian assumption for the data in order to derive the optimal Bayesian strategy for this type of data. Unlike the previous two approaches, this third approach does not suffer from any unnecessary information loss.

In the rest of this chapter, we will study the possibility of deploying the following two particular bilinear LDA methods in conjunction with the FBCSP algorithm:

- The iterative bilinear extension of LDA as suggested by [58], which will be referred to as the 2DLDA method,

- The Bayes optimal bilinear extension of LDA as suggested by [73], which will be referred to as the matrix-to-vector LDA (MVLDA) method.

These two methods will be briefly reviewed in the next subsections.

4.2.1 Two-Dimensional Linear Discriminant Analysis (2DLDA)

The 2DLDA method proposed by [58], is a suboptimal bilinear extension of the LDA method. Let $\mathbf{X} \in \mathbb{R}^{N_f \times N_s}$ represent the spatio-spectral feature matrix at the output of the domain-specific FE step, and assume that N_i training samples $\mathbf{X}_{i,n}$, $1 \leq n \leq N_i$, are available for each class Ω_i, and the total number of training samples is $N = \sum_{i=1}^{C} N_i$. The 2DLDA method uses these training samples in an iterative approach to provide the bilinear operators $\mathbf{U} \in \mathbb{R}^{N_f \times D_1}$ and $\mathbf{V} \in \mathbb{R}^{N_s \times D_2}$. These bilinear operators, will then be applied to the matrix-variate data $\mathbf{X}$ to reduce its dimensionality from $N_f \times N_s$ to $D_1 \times D_2$, as follows:

$$\mathbf{Y} = \mathbf{U}^T \mathbf{X} \mathbf{V}, \tag{4.3}$$

In order to derive these operators, the 2DLDA method first estimates the class conditional means

and the total mean of the data as follows:

$$\hat{\mathbf{M}}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\mathbf{X}_{i,n} \tag{4.4}$$

$$\hat{\mathbf{M}} = \sum_{i=1}^{C}\hat{P}(\Omega_i)\hat{\mathbf{M}}_i \tag{4.5}$$

where $\hat{P}(\Omega_i) = \frac{N_i}{N}$. Then, at the first round of the iterative algorithm, the matrix $\mathbf{V}_0$ will be assumed to be a diagonal matrix, where the first D_2 diagonal entries are equal to one and the rest are zero. After applying this spatial transformation matrix $\mathbf{V}_0$ to the data, the following between-class spectral scatter and within-class spectral scatter matrices will be estimated:

$$\hat{\mathbf{S}}^0_{BL} = \sum_{i=1}^{C}\hat{P}(\Omega_i)(\hat{\mathbf{M}}_i - \hat{\mathbf{M}})\mathbf{V}_0\mathbf{V}_0^T(\hat{\mathbf{M}}_i - \hat{\mathbf{M}})^T, \tag{4.6}$$

$$\hat{\mathbf{S}}^0_{WL} = \frac{1}{N_s N}\sum_{i=1}^{C}\sum_{n=1}^{N_i}(\mathbf{X}_{i,n} - \hat{\mathbf{M}}_i)\mathbf{V}_0\mathbf{V}_0^T(\mathbf{X}_{i,n} - \hat{\mathbf{M}}_i)^T \tag{4.7}$$

Now, denote the eigenvectors of $(\hat{\mathbf{S}}^0_{WL})^{-1}\,\hat{\mathbf{S}}^0_{BL}$ by $\mathbf{u}^0_d$, where $1 \le d \le D_1$, and form the following spectral transformation matrix: $\mathbf{U}_0 = [\mathbf{u}^0_1, \cdots, \mathbf{u}^0_{D_1}]$. After applying this spectral transformation to the data, the following between-class spatial scatter and within-class spatial scatter matrices will be estimated:

$$\hat{\mathbf{S}}^0_{BR} = \sum_{i=1}^{C}\hat{P}(\Omega_i)(\hat{\mathbf{M}}_i - \hat{\mathbf{M}})^T\mathbf{U}_0^T\mathbf{U}_0(\hat{\mathbf{M}}_i - \hat{\mathbf{M}}), \tag{4.8}$$

$$\hat{\mathbf{S}}^0_{WR} = \frac{1}{N_f N}\sum_{i=1}^{C}\sum_{n=1}^{N_i}(\mathbf{X}_{i,n} - \hat{\mathbf{M}}_i)^T\mathbf{U}_0^T\mathbf{U}_0(\mathbf{X}_{i,n} - \hat{\mathbf{M}}_i) \tag{4.9}$$

Accordingly, the spatial transformation matrix $\mathbf{V}_1$ will be formed as $\mathbf{V}_1 = [\mathbf{v}^1_1, \cdots, \mathbf{v}^1_{D_2}]$, where $\mathbf{v}^0_d$, $1 \le d \le D_1$, denote the eigenvectors of $(\hat{\mathbf{S}}^0_{WR})^{-1}\,\hat{\mathbf{S}}^0_{BR}$.

The above procedure will be repeated by substituting $\mathbf{V}_0$ with $\mathbf{V}_1$ in Equations (4.6) and (4.8) and calculating the corresponding $\mathbf{U}_1$ matrix, which in turn will be used to calculate $\mathbf{V}_2$. This iterative procedure will be repeated for a few iterations to allow for spectral and spatial transformation matrices converge to the stable values $\mathbf{U}$ and $\mathbf{V}$, respectively. In our experimental analysis on motor-imagery EEG signals, we have observed that 10 iterations is large enough for the convergence of the transformation matrices in different scenarios, and hence we have fixed the number of iterations to 10 to provide a fair comparison across different scenarios.

4.2.2 Matrix-to-vector Linear Discriminant Analysis (MVLDA)

The MVLDA method is based on the matrix-variate Gaussian model described in Section 3.4. This model implies that the covariance between any two spatio-spectral features can be decomposed into a spatial covariance term and a spectral covariance term. The corresponding spatial/spectral covariance matrices can be estimated using the following equations[1] , assuming that N_i training samples $\mathbf{X}_{i,n}$, $1 \leq n \leq N_i$, are available for each class Ω_i:

$$\hat{\mathbf{\Psi}} = \frac{1}{N_f N} \sum_{i=1}^{C} \sum_{n=1}^{N_i} (\mathbf{X}_{i,n} - \hat{\mathbf{M}}_i)^T (\mathbf{X}_{i,n} - \hat{\mathbf{M}}_i), \tag{4.10}$$

$$\hat{\mathbf{\Phi}} = \frac{1}{N_s N} \sum_{i=1}^{C} \sum_{n=1}^{N_i} (\mathbf{X}_{i,n} - \hat{\mathbf{M}}_i)(\mathbf{X}_{i,n} - \hat{\mathbf{M}}_i)^T. \tag{4.11}$$

where $\hat{\mathbf{M}}_i = \frac{1}{N_i} \sum_{n=1}^{N_i} \mathbf{X}_{i,n}$.

Moreover, the MVLDA method also assumes a separable model for the between-class scatter matrix $\mathbf{S}_B = \mathbf{S}_{BR} \otimes \mathbf{S}_{BL}$, where

$$\hat{\mathbf{S}}_B = \sum_{i=1}^{C} \hat{P}(\Omega_i)(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}})^T \tag{4.12}$$

$$\hat{\mathbf{S}}_{BL} = \sum_{i=1}^{C} \hat{P}(\Omega_i)(\hat{\mathbf{M}}_i - \hat{\mathbf{M}})(\hat{\mathbf{M}}_i - \hat{\mathbf{M}})^T, \tag{4.13}$$

$$\hat{\mathbf{S}}_{BR} = \mathrm{tr}^{-1}(\mathbf{S}_{BL}) * \sum_{i=1}^{C} P(\Omega_i)(\hat{\mathbf{M}}_i - \hat{\mathbf{M}})^T (\hat{\mathbf{M}}_i - \hat{\mathbf{M}}). \tag{4.14}$$

Here, $\hat{\boldsymbol{\mu}} = \mathrm{vec}(\hat{\mathbf{M}})$, $\hat{\mathbf{M}} = \sum_{i=1}^{C} \hat{P}(\Omega_i)\hat{\mathbf{M}}_i$, and $\hat{P}(\Omega_i) = \frac{N_i}{N}$.

Under this set of assumptions, we denote the eigenvalues and eigenvectors of $\hat{\mathbf{\Phi}}^{-1}\hat{\mathbf{S}}_{BL}$ by λ_l and $\mathbf{u}_l$ respectively, where $1 \leq l \leq N_f$. Similarly, we denote the eigenvalues and eigenvectors of $\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{S}}_{BR}$ by γ_j and $\mathbf{v}_j$ respectively, where $1 \leq j \leq N_s$. Now, let λ_l and γ_j be sorted in descending order. Then, the Bayes optimal features for a matrix-variate Gaussian data with separable $\mathbf{\Sigma}$ and $\mathbf{S}_B$ matrices, can be obtained through a bilinear operation of the following form:

$$\mathbf{Y} = \mathbf{U}^T \mathbf{X} \mathbf{V}, \tag{4.15}$$

[1]Equations (4.10) and (4.11) provide moment estimates of the spatial and spectral covariances [63]. Alternatively, one can use the iterative approach of [145] which provides the maximum-likelihood (ML) estimates [146]. However, our studies have shown that for EEG signals, the above non-iterative estimators provide similar performance compared to the ML estimators.

Table 4.1: Pseudocode for training the MVLDA feature extractor.

Inputs:

- N_i training samples $\mathbf{X}_{i,n}$,$1 \leq n \leq N_i$ for each class Ω_i, $1 \leq i \leq C$. The total number of samples is N.
- The number of desired extracted features, denoted by N_f.

Outputs:

- The feature extraction operators $\mathbf{U}_{N_f \times N_f}$ and $\mathbf{V}_{N_s \times N_s}$.
- The corresponding λ_l and γ_j values which determine the priority in selecting the elements of the resulting feature matrix.

Procedure:

1. Estimate the class means $\mathbf{M}_i$, $1 \leq i \leq C$, the spatial covariance matrix $\mathbf{\Psi}$, and the spectral covariance matrix $\mathbf{\Phi}$, using (4.10), and (4.11).
2. Calculate $\mathbf{S}_{BL}$ and $\mathbf{S}_{BR}$ according to (4.13) and (4.14).
3. Calculate the eigenvalues λ_l and γ_j and the corresponding eigenvectors $\mathbf{u}_l$, $1 \leq l \leq N_f$, and $\mathbf{v}_j$, $1 \leq j \leq N_s$, for $\mathbf{\Phi}^{-1}\mathbf{S}_{BL}$ and $\mathbf{\Psi}^{-1}\mathbf{S}_{BR}$ respectively.
4. Construct $\mathbf{U}$ and $\mathbf{V}$ according to (4.16)

where

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{N_f}] \quad \text{and} \quad \mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_s}] \tag{4.16}$$

are spectral and spatial linear operators, respectively, whose columns are $\mathbf{u}_l$ and $\mathbf{v}_j$ vectors. This procedure projects $\mathbf{X}$ onto columns of $\mathbf{U}$ and $\mathbf{V}$ to get the feature matrix $\mathbf{Y}$. Finally, we select the y_{l_j} elements of $\mathbf{Y}$ which correspond to the N_f largest $\lambda_l \gamma_j$ values, and stack them in the $\mathbf{y}$ feature vector; hence it is called *matrix-to-vector* LDA. This is one of the most important advantages of MVLDA in comparison to 2DLDA method. Recall that the 2DLDA method is a *matrix-to-matrix* transformations and provides a matrix-variate set of features at its output, without any measure for sorting the features based on their discriminant power.

Table 4.1 outlines the pseudo-code for training the MVLDA method. The proposed MVLDA solution relies only on the N_f- and N_s-dimensional operations. Therefore, the computational complexity of the eigen decomposition step for MVLDA is broken down into $O({N_f}^3 + {N_s}^3)$, compared to vector-variate LDA's complexity of $O((N_f N_s)^3)$. Moreover, in MVLDA the two eigen decompositions of order $O({N_f}^3)$ and $O({N_s}^3)$ can be implemented in parallel, which is a significant advantage for implementation of this algorithm in real time. Finally, it is worth mentioning that the lower-dimensional covariances $\mathbf{\Phi}$ and $\mathbf{\Psi}$ can be estimated more reliably than the higher-dimensional covariance matrix $\mathbf{\Sigma}$ required by LDA.

4.3 Experimental Analysis

In this section, Data set V from BCI competition III [147] and Data set 2a from BCI competition IV [148] will be used to study the performance of MVLDA and 2DLDA methods as two candidates for matrix-variate domain-agnostic FE in MI-BCIs. We will also compare the performance of these two methods against the conventional vector-variate LDA, to emphasize the importance of utilizing matrix-variate solutions in the domain-agnostic FE step.

In order to study the interplay between different domain-specific and domain-agnostic feature extractors, and its effect on the overall performance of the BCI system, we consider the following scheme. The EEG data is first passed through an optional spatial FE module which contains surface Laplacian (SL) filtering and/or channel selection (CS). Then, a bank of bandpass filters is used to extract different rhythmic activities of the signal. Finally, the resulting rhythms are passed through a filter bank of CSP modules to extract spatio-spectral features [49]. To apply this scheme in a multiclass motor-imagery scenario, we use the one-versus-rest (OVR) multiclass extension of the FBCSP method [49], as explained in Section 4.3.3.

In our simulations, the effect of including the SL or CS are also studied separately. As a result, a total of 12 combinations for DS-FE scheme and DA-FE methods are considered: $2 \times 2 \times 3$ combinations of SL (Yes/No), CS (Yes/No), and LDA/2DLDA/MVLDA. Since our focus in this thesis is on the feature extraction steps, we mainly consider a simple linear Gaussian classifier for all combinations of these domain-specific FE and domain-agnostic FE methods.

For completeness of the results, we have also studied the case where no domain-agnostic FE is utilized and the FBCSP features are directly passed to the classifier. In this case, we have considered both the linear classifier and the naive Bayes Parzen window (NBPW) classifier that was discussed in Section 3.3. The NBPW classifier has been included in this study as a benchmark solution as suggested by the work in [49]. When no domain-agnostic FE method is present prior to the classifier, we take the following strategy to provide a fair performance comparison with other methods that are benefiting from a dimensionality reduction step prior to the classification step. Recall that in the FBCSP method, the spectral features in each frequency band are inherently sorted by the corresponding CSP module, though there is no sorting across different bands. In order to be able to manually adjust the number of features that are passed from the FBCSP to the linear or NBPW classifier, we naively select the "d_{csp}" most significant features from each band, which reduces the dimensionality of the feature matrix from $N_f \times N_s$ to $N_f \times d_{csp}$.

4.3.1 Experiment Setup

BCI competition III, Data set V (Exp. 1)

The goal of this competition is to design a BCI algorithm which can classify the following imagined mental tasks: left-hand movement (Ω_1), right-hand movement (Ω_2), and generation of words beginning with a random letter (Ω_3). This data set contains EEG recordings of three normal subjects recorded in four sessions. Each session consists of sequential 15-second trials of the three tasks. The first three sessions will be used for training purposes, whereas the last session is only used as unseen data for competition, i.e. testing phase. The signals are recorded using 32-electrode Biosemi system at 512Hz sampling rate, and the BCI algorithm is required to provide the estimated label $\widehat{\Omega}$ every 0.5 second, using only the last one second of EEG recording. The performance measure for this competition is the correct classification rate (CCR) of the overall system, defined as the ratio of number of successfully classified samples over the total number of samples. The chance of random classification in this experiment is $P_{rand} = 1/C = 0.33$, and the winning algorithm for this competition in the literature achieves a performance of %62.72 at the classifier output [149] [2].

BCI competition IV, Data set 2a (Exp. 2)

This competition aims to design a BCI algorithm which can classify the following motor imagery tasks: left hand (Ω_1), right hand (Ω_2), both feet (Ω_3), and tongue (Ω_4) movement. This data set contains EEG recordings of nine normal subjects recorded in two sessions. The signals are recorded using 22 Ag/AgCl electrodes at 250Hz sampling rate. Each session consists of 6 runs, each of which includes 48 trials of length 3 seconds, yielding a total of 288 trials per session. The first session will be used for training and the second session is only used as unseen data for testing phase. This data set also contains three electrooclugram (EOG) channel recordings that are provided for subsequent application of artifact processing methods and shall not be used for classification. The competition requires the BCI algorithms to provide a continuous classification output for each sample in the form of the estimated label $\widehat{\Omega}$. The performance measure for this competition is the kappa coefficient (κ) of the overall system [152], which is defined as follows: $\kappa = (CCR - P_{rand})/(1 - P_{rand})$. Here, P_{rand} is the probability of random classification, i.e., $P_{rand} = 1/C = 0.25$ for this experiment. Note that the measure κ is normalized such that $\kappa = 0$ for a random classifier, and its maximum value is $\kappa = 1$ for ideal classifier, as illustrated in Figure 4.2. The winning algorithm for this competition in the literature is the FBCSP-NBPW method,

[2] It is worth mentioning that after the original competition, the works in [150, 151] have outperformed the algorithm of [149] by deploying more complicated classifiers. However, all of these works are based on using short-time Fourier transformation for extraction of the spatio-spectral features from EEG signal.

Table 4.2: Parameters Used for Domain-Specific Feature Extraction Algorithms in Exp. 1 and Exp. 2

		Experiment 1	Experiment 2
Raw Data	Sampling Rate	512 Hz	250 Hz
	Number of Channels (N_{ch})	32	22
Surface Laplacian (SL)	Filter Order	2	2
	Regularization Factor (λ)	0.01	0.01
Channel Selection (CS)	Selected Centro-Parietal Channels	C3, Cz, C4, CP1, CP2, P3, Pz, P4	Channels 8-12, 14-21
	Number of Selected Channels (N_{ch})	8	13
Bandpass Filter Bank	Filter Type	Chebyshev Type-II	Chebyshev Type-II
	Filter Order	6	6
	Selected Bands	8-12, 12-16, $\cdots$, 28-32 Hz	4-8, 8-12, $\cdots$, 36-40 Hz
	Number of Bandpass Filters (N_f)	6	9
Common Spatial Patterns	Epoch Length	1 sec	2 sec
	Overlapping Factor	1/2	19/20

which has been faithfully re-implemented in our experimental studies.

Table 4.2 presents the parameters used to implement the processing steps of the DS-FE schemes and extract the spatio-spectral feature matrix $\mathbf{X}_{N_f \times N_s}$. It should be noted that

- The channel selection (CS) is performed by selecting the centro-parietal channels located over the motor cortex in each experiment. Hence, $N_{ch} \in \{8, 32\}$ for Exp. 1 and $N_{ch} \in \{13, 22\}$ for Exp. 2, depending on whether or not CS is used.

- In each experiment, the epoch length and frequency range used by the winning algorithm in the original competition are adopted in this chapter to provide a fair comparison between alternative solutions.

It is noteworthy that in both experiments, the EEG signals are collected under controlled conditions, where the subjects are asked to sit relaxed on a chair and minimize their body movements in order to minimize the amount of interfering artifacts. Furthermore, the dataset providers have asked EEG experts to visually inspect the recorded signals to mark the trials that are contaminated with artifacts. As recommended by the dataset providers, these artifact contaminated trials are excluded from our experimental analysis, and hence no automated artifact removal procedure is used.

A Comparative Note on The Datasets Used in Exp. 1 and Exp. 2

As mentioned in the description of each experiment, both databases in Exp. 1 and Exp. 2 contain multichannel EEG data which is collected during motor-imagery tasks. However, these two datasets are significantly different in terms of the availability of the training data. In Exp. 1, each trial is of length 15 seconds, which is significantly longer than the 3 second trial length in Exp. 2. In the context of motor-imagery tasks, the training trial length is of great importance. When the training trials are

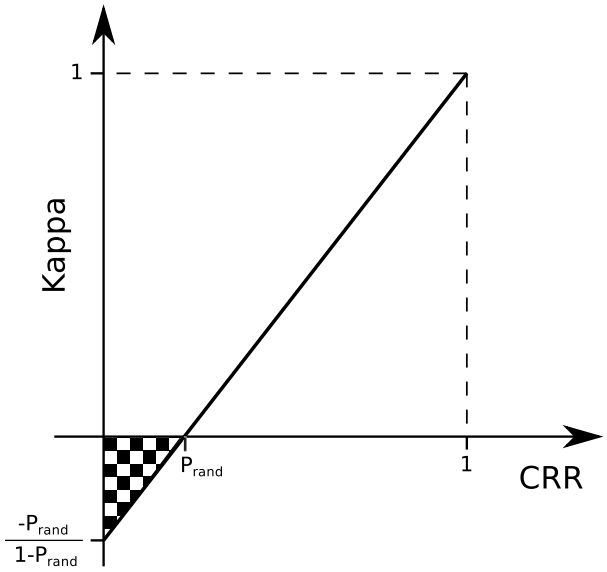


Figure 4.2: Kappa value defined as a normalized version of the correct classification rate (CCR): $\kappa = (CCR - P_{rand})/(1 - P_{rand})$. Note that $P_{rand} = 1/3$ for Exp. 1 and $P_{rand} = 1/4$ for Exp. 2. The shaded part of the graph illustrates the performance values (CRR and Kappa) that are not acceptable (i.e., random performance or worse).

longer, the subjects will have enough time to concentrate on the desired motor-imagery task and produce stable brain rhythms that can be reliably used for training.

Furthermore, the training data in Exp. 1 has been collected during three different sessions, whereas Exp. 2 only includes one session of EEG recording for training. It is well known in the context of motor-imagery BCIs that the EEG characteristics exhibit inter-session variations, which need to be taken into account while training the BCI algorithm.

Since motor-imagery BCI systems are mostly designed for longterm utilization by the user, it is usually assumed that the BCI algorithm has access to a training dataset with long enough trials which are collected over at least two different recording sessions. From this perspective, the training dataset in Exp. 1 can be considered as a typical dataset for motor-imagery applications, whereas the training set in Exp. 2 is an extreme case where only one recording session with very short trials is available is available for training the algorithms. Although this extreme case is unlikely to happen in the MI-BCI applications, we have included Exp. 2 in our analysis to study the robustness of different algorithms in the extreme conditions.

Finally, it should be mentioned that in order to apply the surface Laplacian filter to the EEG data, we need the exact locations of the EEG sensors. The dataset providers in Exp. 1 have provided the exact coordinates of the locations of the EEG sensors, using the standard 10-10 system. In contrast, the dataset in Exp. 2 only contains the approximate relative locations of the EEG sensors. In order to be

able to use the surface Laplacian filter in the Exp. 2, we have mapped these approximate locations, to the following closest standard locations: Fz, FC3, FC1, FCz, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP3, CP1, CPz, CP2, CP4, P1, Pz, P2, POz. The effect of this approximate mapping will be discussed latter in the experimental results.

4.3.2 Bandpass Filter Design

In this section, we will briefly discuss the design criteria which are of particular interest for motor-imagery BCI systems. The appropriate design of the bandpass filters has a great influence on the overall performance of the FBCSP method.

Selecting The Type of Filter

As mentioned in Section 3.1.2, in case of the motor-imagery BCI systems, infinite-impulse-response (IIR) digital filters are commonly used. In order to design the IIR filter, the following criteria are of particular importance for us: (a) Flat passband, (b) Small delay, (c) Sharp transitions band. Among the commonly used IIR filters, only the Chebyshev Type II and Butterworth filters have a flat passband, and both have a sharp transition band. Moreover, the Butterworth filter and the Chebyshev Type II filter introduce less distortion the signal, compared to other IIR filters such as Elliptic or Chebyshev Type I filters, since they have a flatter group delay response.

In order to provide a better insight into the differences between the characteristics of these two IIR filters, consider the following example. In order to extract the alpha rhythm from the EEG signal, we need a bandpass filter with the following criteria:

- Passband Frequency Range: $8 - 12$ Hz
- Stopband Frequency Range: $f < 6$ Hz and $f > 14$ Hz
- Sopband Attenuation: 60 dB

Based on these criteria, we have designed a Chebyshev Type II filter and a Butterworth filter[3], whose frequency response and impulse response are illustrated in Figure 4.3. In order to design these filters, the minimum filter order which satisfies the above passband/stopband requirements has been used; hence, the Butterworth filter is of order 28, and the Chebyshev Type II filter is of order 14. Figure 4.3 reveals that although these two filters have similar frequency responses in the passband and the transition band, the delay introduce by the Chebyshev Type II filter is half the delay introduced by the Butterworth filter.

[3]It is worth mentioning that these filters are designed using MATLAB's Filter Design and Analysis Tool (FDATool), which is part of the signal processing toolbox.

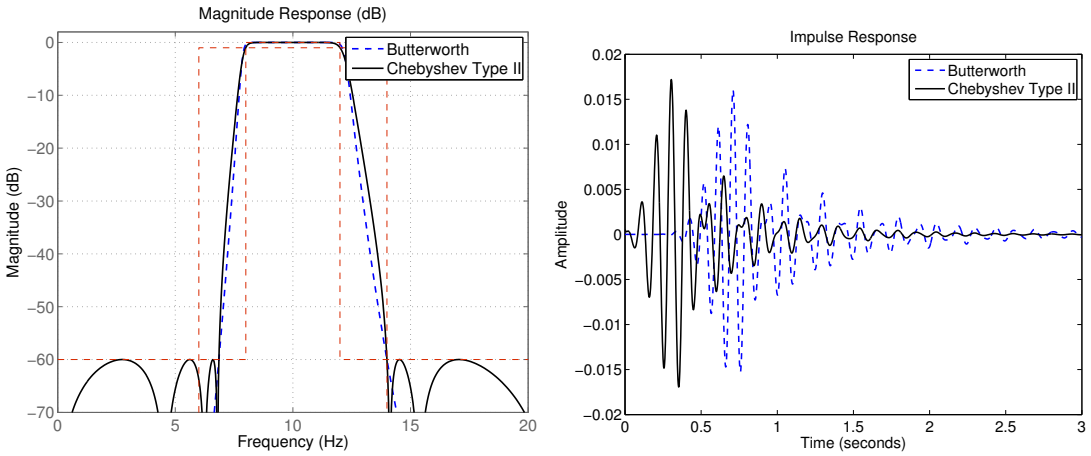


Figure 4.3: Comparison of the Frequency response and impulse response of the Butterworth and the Chebyshev Type II filters.

This is due to the fact that the order of the Butterworth filter is two times the order of the Chebyshev Type II filter.

Based on the above discussions, the Chebyshev Type-II filter will be used in Chapters 4 and 5 to implement the bandpass filterbank.

Implementation of high order IIR filters

It should be noted that any high order IIR filter can be implemented as a series of second-order sections. Throughout this thesis, we use the second-order implementation instead of the original high order transfer function to avoid the round-off errors. The effect of round-off errors for high order IIR filters is such detrimental that even with double-precision floating point arithmetics the resulting transfer function would be completely deferent form the desired one.

The effect of round-off error has been illustrated in an example in Figure 4.4. In this example, a Chebyshev Type II filter with the same characteristics has been implemented using (a) its original transfer function, and (b) its equivalent second-order sections. This figure reveals that the transfer function implementation leads to a completely incorrect frequency response.

4.3.3 Multiclass Extension of the FBCSP Method

Due to the fact that CSP modules utilized in FBCSP method are originally designed for binary classification scenarios, the FBCSP algorithm is also inherently suitable for binary classification cases. Similar to the CSP algorithm, however, there are several methods to extend FBCSP for multiclass scenarios. The work in [49] provides a comparative study of different multiclass extensions of the FBCSP method,

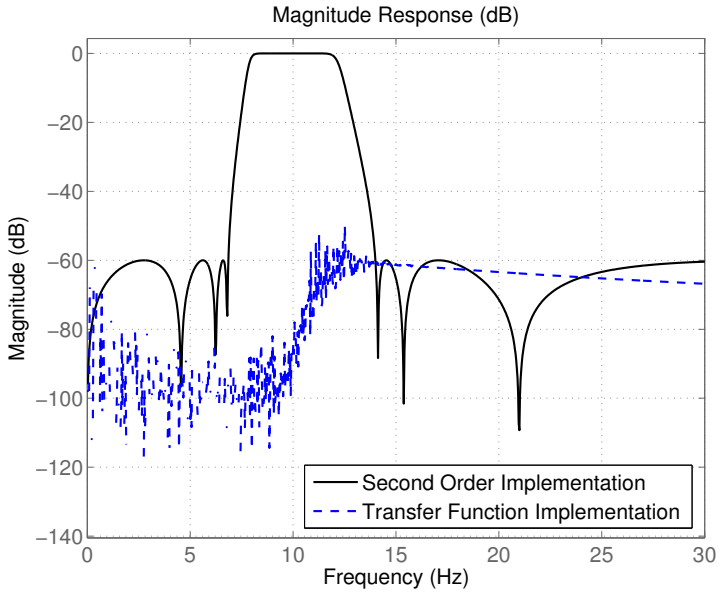


Figure 4.4: Comparison of the Frequency response of the Chebychev Type II bandpass filter implemented with (a) high order transfer function, (b) second-order sections. Both transfer functions have been calculated using MATLAB's Filter Visualization Tool (FVTool) with double-precision floating point arithmetics.

including divide-and-conquer strategy, pairwise classification, and the one-versus-rest strategy. It is shown in [49] that the one-versus-rest approach provides the best performance in comparison to other strategies; therefore, we adopt this approach in our experimental studies.

The one-versus-rest (OVR) approach for multiclass extension of FBCSP method works as follows. Let Ω_i' be the set of all motor-imagery task excluding the i^{th} task Ω_i. The CSP modules will first focus on the features that discriminate task Ω_1 versus the *rest* of tasks, i.e., Ω_1', by creating a pool of training samples from all other tasks and assigning them to Ω_1'. Accordingly, each CSP module will extract $d_{csp}\,(\leq N_{ch})$ features to discriminate Ω_1 from Ω_1'. This procedure will then be repeated for other classes by selecting one class at a time an comparing it against the rest of classes. This procedure results in a set of $d_{csp} * C$ features being generated for each CSP module, which eventually forms a feature matrix of size $N_f \times (d_{csp} * C)$. The resulting feature matrix will then be passed to the domain-agnostic feature extractor or directly the classifier.

It should be noted that the value of d_{csp} needs to be an even number due to the fact that CSP algorithm provides output features in paired groups (ref. Section 3.1.1). Moreover, it is assumed that the value of d_{csp} is fixed for all the choices of Ω_i versus Ω_i', and over all the frequency bands. Finally, it is worth mentioning that the extracted features at the output of each CSP module can be sorted, based on their discriminant power, into groups of size $2C$, where the first group includes the most discriminant

pair of CSP features for classification of Ω_i versus Ω'_i, $1 \leq i \leq C$, and so on. However, it is not possible to sort the features across different bands, since FBCSP deals every band independently.

4.3.4 Cross-validation Results

The performance of BCI algorithms highly depends on the dimensionality of the feature space at the classifier's input, denoted by d. To determine the optimal value of d, denoted by d_{opt}, for each feature extraction scheme, we perform cross-validation on the training data. In case of Exp. 1, since we have access to three different training sessions, a three fold cross-validation is performed to make sure that for each validation run the BCI system has access to two distinct sessions for training and one session for analyzing the performance. This strategy is very helpful in making sure that the inter-session variations of the EEG data are taken into account during the validation phase.

The dataset in Exp. 2, however, only contains one training session which prevents us from adopting the same cross-validation strategy as Exp. 1. Therefore, we chose to perform a 5×5-fold randomized cross validation strategy. In this strategy, the training data will be randomized five times. After each randomization, the data will be divided into five folds. In each validation run, four of these folds will be used for training the BCI algorithm and the remaining fold will be used for analyzing the resulting performance. This procedure results in five validation runs for each randomization, which leads to a total of $25 = 5 \times 5$ validation runs.

The complete results of this cross-validation for all the subjects in the two experiments are presented in Table 4.3 and Table 4.4. In Table 4.3, the first five rows in Table 4.3 and the first five paired rows in Table 4.4 provide the results for the following combinations of domain-specific FE, domain-agnostic FE, and Classification: FBCSP-NBPW (no DA-FE), FBCSP-Lin (no DA-FE), FBCSP-LDA-Lin, FBCSP-2DLDA-Lin, and FBCSP-MVLDA-Lin. In these first five rows, no surface Laplacian (SL) or channel selection (CS) has been applied to the data. Similarly, the next groups of five rows provide the results when different combinations of SL and CS feature extractors are used. For each subject, the optimal size of the feature space, i.e., d_{opt}, is chosen to be the one which maximizes the average performance over all the cross-validation runs. The corresponding average performance (and its standard error) are reported in these two tables. Note that the performance measure is the correct classification rate (CCR) for Exp. 1, and the Kappa coefficient (κ) for Exp. 2.

It is noteworthy that different methods have different limitations for possible values of d, as follows:

- The LDA algorithm can only provide up to $C - 1$ features, i.e., two features for Exp. 1 and three features for Exp. 2.

- 2DLDA provides dimensions of the form $d = m * n$ where $1 < m < D_1$, $1 < n < D_2$, $D_1 = rank(\mathbf{S}_{BL}) = \min(N_f, N_s * (C-1))$, and $D_2 = rank(\mathbf{S}_{BR}) = \min(N_s, N_f * (C-1))$.
- MVLDA method can provide any dimensionality in the range of $1 < d < D_1 * D_2$, where $D_1 = rank(\mathbf{S}_{BL})$ and $D_2 = rank(\mathbf{S}_{BR})$, as defined in the 2DLDA case.
- In case where FBCSP features are directly passed to the NBPW or the linear classifiers, we manually choose $N_f * N_s = N_f * d_{csp} * C$ features, where $d_{csp} \in \{2, 4, \cdots, N_{ch}\}$. (refer to the discussion at the beginning of Section 4.3)

It should also be mentioned that in Exp. 2, the LDA method fails to operate in most cases, since the dimensionality of the feature matrix at the output of FBCSP method is higher than the number of training samples available for calculation of the within-class scatter matrix in LDA method. To alleviate this problem, for the case of LDA, we manually decrease the dimensionality of the FBCSP matrix, by choosing the $d_{csp} = 4$ most significant features from each band, which results in a matrix of size $N_f \times (2 * C)$. Even in this case, the LDA cannot operate when only surface Laplacian filter is applied to the data, i.e., SL = Yes, CS = No, since the within class scatter matrix turns is singular for this case. This effect will be discussed in more detail in the discussions related to the effect of surface Laplacian filtering.

The results in Table 4.3 and Table 4.4 exhibit a large inter-subject variation in the performance results, which is expected in the context of motor-imagery BCI systems. This inter-subject variation is mostly due to the fact that motor-imagery tasks require person's concentration and engagement during the trials which is not necessarily the same for different subjects particularly since our experiments do not provide any neuro-feedback to the users. The other factor is the differences between characteristics of EEG signals from different subjects. Some subjects have better ability to control their mental states, specially those who are routinely involved in activities that require high levels of mindfulness.

Despite the aforementioned inter-subject variations, the general performance trends are similar in most of the subjects. Thus, to better illustrate the performance differences between different methods, Figures 4.5(a) and 4.6(a) provide the bar-plots of the validation results averaged over all the subjects. Similarly, Figures 4.5(b) and 4.6(b) provide the bar-plots of the testing results averaged over all the subjects, as will be discussed in Section 4.3.5.

The results in Figures 4.5 and 4.6 reveal that both MVLDA and 2DLDA, which are matrix-variate domain-agnostic FE methods, provide better performances in comparison to the vector-variate LDA algorithm. Particularly, the proposed MVLDA method outperforms all other DA-FE methods (including 2DLDA) in majority of DS-FE scenarios. The MVLDA method also outperforms the cases where NBPW

Table 4.3: Cross-validatoin performance results for different algorithms in Experiment-1. For each subject, the average correct classification rate (CCR) for the optimal dimension (d_{opt}) and its corresponding standard error are reported. The rightmost column represents the average performance over all the subjects. Note that the performance of a random classifier in this experiment is $\%CCR = \%33.3$.

DS-FE				**DA-FE**	**Clas.**	**Performance in the Cross-validation Phase**						
Spatial		**Spectral**	**Spatio-**			**Subject 1**		**Subject 2**		**Subject 3**		**Average**
SL	**CS**		**Spectral**			**%CCR**	d_{opt}	**%CCR**	d_{opt}	**%CCR**	d_{opt}	**%CCR**
No	No	BPF	FBCSP	-	NBPW	58.22±4.19	6*18*3	48.10±3.38	6*8*3	45.62±0.46	6*10*3	50.65±2.12
				-	Lin	63.25±1.99	6*12*3	50.42±3.37	6*12*3	48.73±1.00	6*10*3	54.13±2.03
				LDA	Lin	46.77±0.88	2	44.30±2.79	2	41.38±1.95	2	44.15±0.97
				2DLDA	Lin	53.50±1.90	5*87	50.42±1.37	4*92	45.97±1.44	2*96	49.97±1.25
				MVLDA	**Lin**	**69.50±2.81**	**2**	**53.94±0.99**	**2**	**46.75±1.20**	**27**	**56.73±1.56**
	Yes	BPF	FBCSP	-	NBPW	60.64±2.56	6*6*3	46.62±3.27	6*8*3	45.83±2.31	6*6*3	50.97±1.76
				-	Lin	65.99±2.28	6*6*3	49.79±1.06	6*8*3	43.50±2.16	6*8*3	53.09±1.63
				LDA	Lin	65.22±1.85	2	49.79±1.06	2	43.50±0.95	2	52.84±0.52
				2DLDA	Lin	67.33±4.63	3*23	51.69±1.11	6*12	44.92±2.02	6*19	54.64±2.54
				MVLDA	**Lin**	**71.67±2.66**	**17**	**51.55±2.41**	**26**	**48.16±0.46**	**6**	**57.13±1.35**
Yes	No	BPF	FBCSP	-	NBPW	58.71±4.01	6*18*3	48.10±2.86	6*14*3	46.40±1.71	6*16*3	51.07±2.11
				-	Lin	59.26±3.03	6*16*3	52.60±4.97	6*6*3	48.45±1.10	6*12*3	53.43±2.41
				LDA	Lin	52.45±1.73	2	44.30±1.67	2	41.67±0.98	2	46.14±0.39
				2DLDA	Lin	65.36±1.91	3*67	63.08±1.84	1*85	48.31±2.56	3*82	58.91±1.92
				MVLDA	**Lin**	**69.77±3.44**	**2**	**57.45±1.86**	**2**	**48.94±1.69**	**54**	**58.72±1.78**
	Yes	BPF	FBCSP	-	NBPW	61.99±1.90	6*8*3	44.66±5.13	6*8*3	47.53± 2.20	6*8*3	51.39±0.83
				-	Lin	65.14±1.53	6*8*3	55.20±2.37	6*8*3	48.52±1.16	6*6*3	56.28±0.82
				LDA	Lin	65.14±1.53	2	55.20±2.37	2	48.31±1.48	2	56.22±0.90
				2DLDA	Lin	69.98±2.41	4*24	55.20±0.86	6*14	49.22±0.85	6*16	58.14±0.75
				MVLDA	**Lin**	**74.06±1.59**	**3**	**62.45±2.72**	**2**	**53.60±0.44**	**16**	**63.37±1.49**

Table 4.4: Cross-validatoin performance results for different algorithms in Experiment-2. For each subject, the Kappa coefficient (κ) for the optimal dimension (d_{opt}) and its corresponding standard error are reported. The rightmost column represents the average performance over all the subjects. Note that the performance of a random classifier in this experiment is $\kappa = 0$.

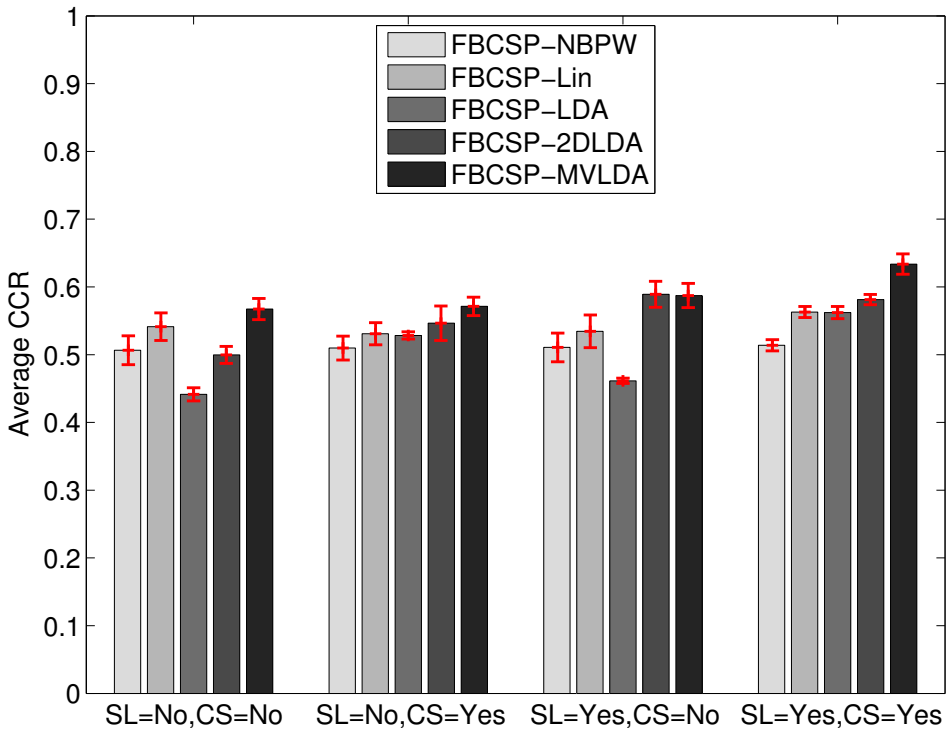
DS-FE																
SL	**Spatial CS**	**Spectral**	**Spatio-Spectral**	**DA-FE**	**Clas.**	**Performance in the Validation Stage (κ and d_{opt})**									
						Subj. 1	**Subj. 2**	**Subj. 3**	**Subj. 4**	**Subj. 5**	**Subj. 6**	**Subj. 7**	**Subj. 8**	**Subj. 9**	**Average**
No	No	BPF	FBCSP	-	NBPW	67.88±2.87 9*10*4	42.18±3.35 9*2*4	77.87±2.78 9*4*4	51.77±3.73 9*6*4	50.17±4.20 9*2*4	45.97±3.13 9*18*4	87.50±2.30 9*2*4	85.79±2.03 9*12*4	76.31±2.86 9*4*4	65.05±1.12
				-	Lin	68.76±2.56 9*8*4	29.21±3.02 9*12*4	72.07±2.67 9*2*4	49.24±2.45 9*20*4	52.54±3.82 9*2*4	48.91±3.26 9*2*4	84.63±2.40 9*6*4	85.96±2.41 9*8*4	76.46±2.89 9*2*4	63.09±0.90
				LDA	Lin	53.20±3.07 3	24.11±3.59 3	63.65±3.07 3	29.70±3.02 3	34.75±3.11 3	15.30±3.66 3	74.69±3.65 3	52.00±3.70 3	55.14±4.51 3	44.73±1.10
				2DLDA	Lin	63.45±2.81 3*9	29.86±1.27 7*6	65.54±4.06 6*4	26.08±3.23 7*13	42.91±2.73 8*3	24.59±2.67 8*13	68.26±3.96 8*5	68.11±3.18 8*3	71.82±2.75 9*3	51.18±1.38
				MVLDA	**Lin**	80.71±2.25 32	41.65±3.53 44	78.03±3.18 101	47.24±2.63 61	54.89±3.40 148	45.81±3.79 242	90.27±1.95 143	88.85±2.17 158	76.92±3.14 42	67.15±0.86
	Yes	BPF	FBCSP	-	NBPW	65.69±3.06 9*4*4	42.32±2.36 9*2*4	80.37±2.41 9*2*4	40.12±3.71 9*2*4	39.04±3.59 9*6*4	50.58±3.57 9*10*4	78.57±2.84 9*8*4	77.67±3.16 9*4*4	76.60±3.12 9*4*4	61.22±1.15
				-	Lin	66.96±2.99 9*4*4	33.32±3.14 9*6*4	80.89±2.69 9*2*4	49.14±2.47 9*6*4	42.29±3.30 9*4*4	46.30±3.79 9*6*4	73.01±2.90 9*8*4	83.04±2.48 9*12*4	76.17±3.09 9*2*4	61.24±0.98
				LDA	Lin	50.93±3.00 3	19.80±4.33 3	57.08±3.60 3	24.76±2.85 3	18.81±3.21 3	28.09±4.30 3	51.94±3.86 3	65.28±3.02 3	49.64±4.08 3	40.70±1.25
				2DLDA	Lin	65.18±2.57 3*7	28.45±1.26 2*6	68.17±1.87 3*8	27.93±3.35 2*5	28.07±2.15 7*15	26.05±4.05 3*9	62.44±2.09 4*3	69.23±2.69 5*5	71.59±3.27 4*8	49.68±1.33
				MVLDA	**Lin**	75.42±2.47 16	38.39±3.37 20	80.01±2.65 157	44.29±2.50 80	35.03±4.20 90	53.43±4.47 96	79.85±2.78 19	85.64±2.62 11	75.89±3.11 12	63.11±1.14
Yes	No	BPF	FBCSP	-	NBPW	71.45±1.65 9*4*4	40.10±1.78 9*4*4	71.00±1.72 9*4*4	39.30±2.05 9*4*4	57.56±1.86 9*2*4	30.37±2.26 9*6*4	81.43±1.23 9*2*4	77.98±0.94 9*4*4	73.93±1.71 9*6*4	60.35±0.62
				-	Lin	70.54±4.41 9*2*4	35.11±1.71 9*4*4	69.43±1.57 9*6*4	41.93±2.41 9*4*4	40.79±7.40 9*2*4	31.71±2.82 9*10*4	79.30±1.35 9*8*4	79.93±1.00 9*10*4	69.79±1.94 9*6*4	57.61±1.15
				LDA	Lin	- -	- -	- -	- -	- -	- -	- -	- -	- -	-
				2DLDA	Lin	46.76±15.66 5*9	22.66±7.78 3*9	51.54±15.77 3*7	30.76±4.28 4*6	18.97±14.38 8*5	21.70±6.24 7*6	67.99±3.11 7*5	53.79±16.01 4*8	73.56±2.38 3*6	43.08±6.27
				MVLDA	**Lin**	54.39±8.97 5	37.88±1.67 112	71.32±4.51 41	31.82±4.99 136	24.03±9.03 126	28.17±3.82 154	74.01±6.64 195	68.90±7.86 243	64.46±7.36 5	50.55±2.42
	Yes	BPF	FBCSP	-	NBPW	74.26±1.56 9*6*4	40.09±1.30 9*2*4	74.26±1.38 9*2*4	43.01±2.22 9*6*4	61.68±1.64 9*2*4	35.86±1.94 9*6*4	85.74±0.99 9*2*4	81.30±1.14 9*4*4	75.35±1.52 9*2*4	63.51±0.57
				-	Lin	76.37±1.52 9*4*4	39.70±1.63 9*6*4	78.13±1.26 9*4*4	46.67±2.22 9*6*4	59.49±1.72 9*4*4	35.39±2.29 9*6*4	84.74±1.50 9*6*4	82.38±1.14 9*4*4	71.56±1.77 9*2*4	63.82±0.49
				LDA	Lin	59.32±1.63 3	21.96±2.02 3	54.19±1.35 3	28.22±1.56 3	36.84±1.54 3	10.32±2.13 3	62.55±1.64 3	56.74±1.64 3	51.64±1.89 3	42.42±0.64
				2DLDA	Lin	66.60±2.71 3*8	30.01±4.19 6*2	70.65±2.54 3*5	32.58±3.07 3*7	45.80±3.84 8*3	24.20±4.65 5*9	75.17±1.44 5*7	74.68±2.87 8*4	73.40±3.73 4*5	54.79±1.84
				MVLDA	**Lin**	80.23±1.10 76	38.90±1.70 26	78.45±1.43 82	45.91±1.78 128	62.45±1.45 90	38.30±1.83 63	87.08±1.05 91	83.13±1.05 29	78.53±1.70 4	65.89±0.51

Table 4.5: Test performance results for different algorithms in Experiment-1. The correct classification rate (CCR) for each subject and the total average over all the subjects are reported. Note that the performance of a random classifier in this experiment is $\%CCR = \%33.3$.

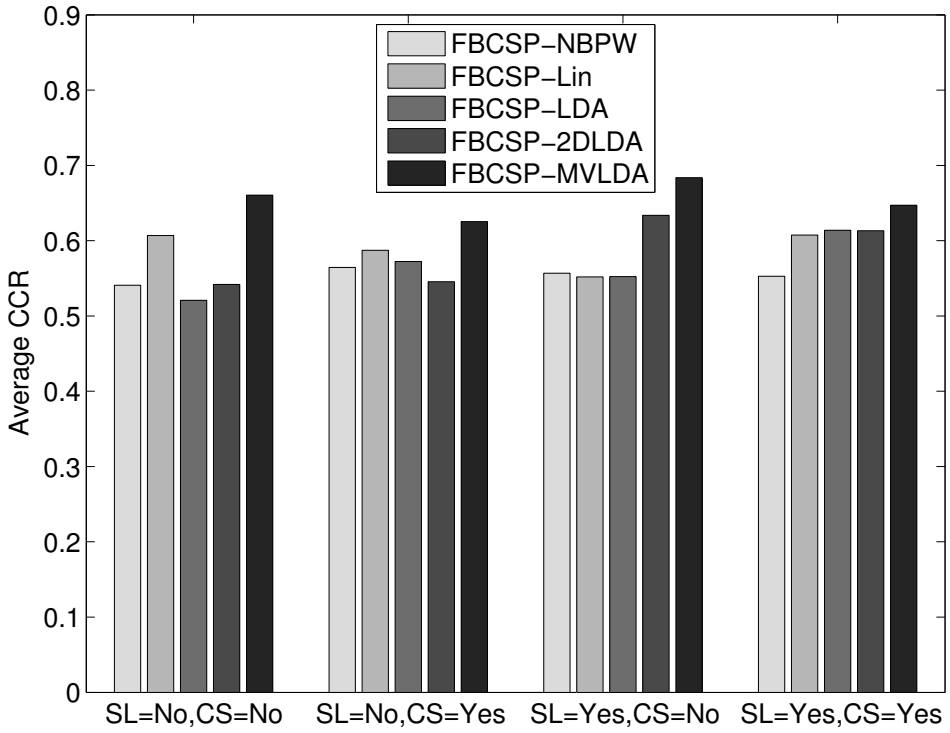
DS-FE									
Spatial		**Spectral**	**Spatio-**	**DA-FE**	**Classifier**	**Performance in the Test Stage (%CCR)**			
SL	**CS**		**Spectral**			**Subj. 1**	**Subj. 2**	**Subj. 3**	**Average**
No	No	BPF	FBCSP	-	NBPW	68.09	54.85	39.29	54.07
				-	Lin	71.06	62.66	48.32	60.68
				LDA	Linear	59.36	51.05	45.80	52.07
				2DLDA	Linear	62.34	56.54	43.70	54.19
				MVLDA	**Linear**	**80.64**	**68.99**	**48.53**	**66.05**
	Yes	BPF	FBCSP	-	NBPW	64.04	55.91	49.37	56.44
				-	Lin	74.04	53.38	48.74	58.72
				LDA	Linear	72.34	53.38	46.01	57.24
				2DLDA	Linear	64.26	52.53	46.85	54.55
				MVLDA	**Linear**	**79.57**	**55.91**	**52.10**	**62.53**
Yes	No	BPF	FBCSP	-	NBPW	66.17	60.55	40.34	55.68
				-	Lin	66.60	50.00	48.95	55.18
				LDA	Linear	60.85	54.85	50.00	55.23
				2DLDA	Linear	75.11	67.30	47.69	63.37
				MVLDA	**Linear**	**81.70**	**70.04**	**53.36**	**68.37**
	Yes	BPF	FBCSP	-	NBPW	69.79	50.42	45.59	55.27
				-	Lin	72.77	60.97	48.53	60.76
				LDA	Linear	72.77	60.97	50.42	61.39
				2DLDA	Linear	74.04	60.97	48.95	61.32
				MVLDA	**Linear**	**79.15**	**60.97**	**53.99**	**64.70**

Table 4.6: Test performance results for different algorithms in Experiment-2. The correct Kappa coefficient (κ)) for each subject and the total average over all the subjects are reported. Note that the performance of a random classifier in this experiment is $\kappa = 0$.

DS-FE															
SL	**Spatial CS**	**Spectral**	**Spatio-Spectral**	**DA-FE**	**Clas.**	**Performance in the Test Stage (κ)**									
						Subj. 1	**Subj. 2**	**Subj. 3**	**Subj. 4**	**Subj. 5**	**Subj. 6**	**Subj. 7**	**Subj. 8**	**Subj. 9**	**Average**
No	No	BPF	FBCSP	-	NBPW	63.67	31.20	60.10	36.73	23.33	19.43	64.20	57.07	61.39	46.35
				-	Lin	68.07	30.25	70.68	39.34	28.42	25.01	56.89	59.13	54.06	47.98
				LDA	Lin	60.85	15.59	55.11	22.76	17.57	9.61	53.23	40.28	37.01	34.67
				2DLDA	Lin	55.79	18.40	57.95	24.21	23.96	12.54	51.61	48.79	38.74	36.89
				MVLDA	**Lin**	65.94	23.84	68.41	39.54	41.77	21.00	68.18	63.43	57.76	49.98
	Yes	BPF	FBCSP	-	NBPW	60.72	25.06	56.65	34.97	9.61	18.10	55.56	53.92	53.33	40.88
				-	Lin	64.46	26.79	60.81	39.07	11.19	24.63	66.53	53.97	54.84	44.70
				LDA	Lin	50.66	16.49	54.92	20.56	2.73	15.07	41.36	42.73	38.44	31.44
				2DLDA	Lin	58.14	15.37	56.41	25.82	6.49	17.59	48.70	48.68	46.50	35.97
				MVLDA	**Lin**	70.28	23.27	61.20	33.23	7.68	22.55	61.57	60.97	51.45	43.58
Yes	No	BPF	FBCSP	-	NBPW	64.11	32.17	65.95	43.33	27.37	22.15	64.58	31.93	56.23	45.31
				-	Lin	69.31	36.63	69.52	48.49	14.20	27.82	61.17	-24.06	57.17	40.03
				LDA	Lin	-	-	-	-	-	-	-	-	-	-
				2DLDA	Lin	55.34	26.37	55.29	31.59	19.51	15.38	52.19	-23.14	43.80	30.70
				MVLDA	**Lin**	74.03	37.94	70.10	45.82	25.03	23.61	71.23	-25.65	54.76	41.88
	Yes	BPF	FBCSP	-	NBPW	64.87	41.45	60.56	40.46	21.82	19.85	65.74	54.56	50.38	46.63
				-	Lin	67.55	40.17	66.60	43.54	23.69	25.84	62.14	53.12	49.05	47.97
				LDA	Lin	61.31	17.73	60.33	27.30	0.22	6.50	50.63	33.79	37.48	32.81
				2DLDA	Lin	59.46	23.36	61.93	29.27	17.74	18.58	51.60	49.73	44.04	39.52
				MVLDA	**Lin**	71.45	35.81	69.78	41.32	34.54	26.03	60.08	60.51	50.20	49.97

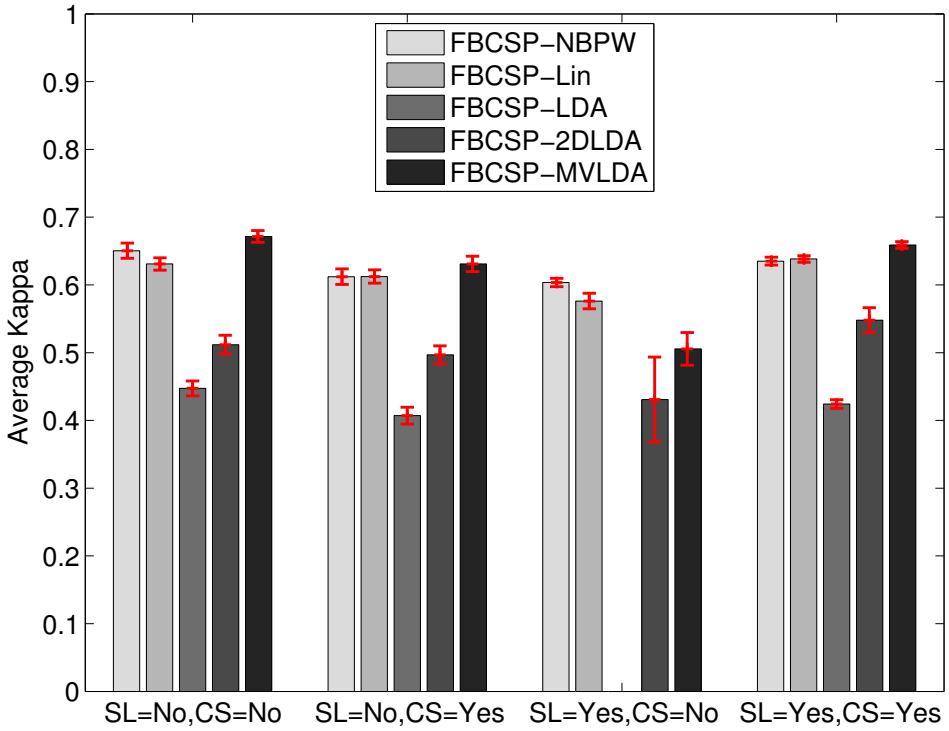


(a) Average validation results (correct classification rates) in Experiment-1.

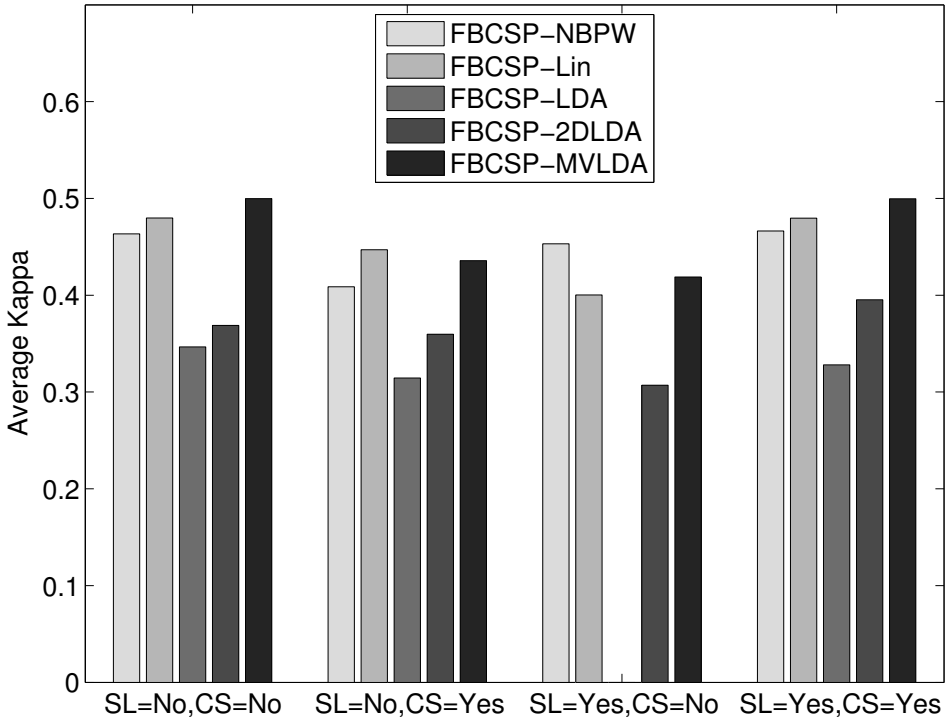


(b) Average test results (correct classification rates) in Experiment-1.

Figure 4.5: Performance results for different methods, averaged over all the subjects in (a) Validation phase of Experiment-1 and (b) Testing phase of Experiment-1. For each method, the averaged result over all the subjects are plotted. In case of validation results, standard error corresponding to performance variations over different validation runs is also presented. For more clarity, the results are illustrated in four groups, depending on whether or not the surface Laplacian (SL) and channel selection (CS) are applied in the domain-specific feature extraction step. Note that the performance measure in Experiment-1 is the Correct Classification Rate (CCR), and a random classifier results in $\%CCR = \%33.3$.



(a) Average validation results (kappa values) in Experiment-2.



(b) Average test results (kappa values) in Experiment-2.

Figure 4.6: Performance results for different methods, averaged over all the subjects in (a) Validation phase of Experiment-2 and (b) Testing phase of Experiment-2. For each method, the averaged result over all the subjects are plotted. In case of validation results, standard error corresponding to performance variations over different validation runs is also presented. For more clarity, the results are illustrated in four groups, depending on whether or not the surface Laplacian (SL) and channel selection (CS) are applied in the domain-specific feature extraction step. Note that the performance measure in Experiment-2 is the Kappa coefficient (κ), and a random classifier results in $\kappa = 0$.

or the Linear classifier are directly applied to the manually selected features, where as the 2DLDA provides an inferior performance compared to them. This can be attributed to the weakness of 2DLDA in extraction of highly discriminant features.

It is noteworthy that the classification performance for different brain tasks are not necessarily the same. In order to illustrate this fact, consider the confusion matrices of the FBCSP-BMLDA method in different scenarios during the validation phase of the second experiment, as shown in Table 4.7. In this table, the confusion matrices for different combinations of surface Laplacian (SL) and channel selection (CS) are presented. In these confusion matrices, the $(i, j)^{th}$ element represents the probability of classifying an EEG epoch belonging to task Ω_i as task Ω_j. Therefore, the diagonal elements of the confusion matrix represent the correct classification rate for each task, while the off-diagonal terms represent the miss-classification rates[4].

Table 4.7 shows that in general the third task (feet movement) and fourth task (tongue movement) have respectively the lowest and the highest correct classification rates, except for the case where both surface Laplacian and channel selection are applied to the data. This difference in the performances can be attributed to the location and the extent of the cortex area that are responsible for these motor tasks. Recall from Figure 2.2 that the motor cortex responsible for foot movement is relatively small (compared to hand and tongue movement) and is located in the area between the right and left hemispheres of the brain. In contrast the tongue movement involves a large area on the motor cortex. Moreover, it should be noted that when no surface Laplacian is applied to the EEG data, the probability of miss-classifying tongue movement as left-hand movement is almost twice the probability of miss-classifying it as right-hand movement. However, when surface Laplacian is applied to the EEG data, these two miss-classification probabilities are almost the same[5].

4.3.5 Test (Competition) Results

Table 4.5 and Table 4.6 outline the correct performance results of different methods when they applied to the unseen competition data in Exp. 1 and Exp. 2, respectively. Recall that three training sessions are available in Exp. 1, whereas only one training session is provided for Exp. 2. At this phase, the value of d_{opt} for each method and each subject is set based on the cross-validation results of Table 4.3 and

[4]The confusion matrix for an ideal classifier will be equal to identity matrix.

[5]This phenomenon requires further neurophysiological investigation based on specific information about the subjects, specially whether they are right-handed or left-handed, which is not available in the database descriptions. In more than %90 of right-handed people and more than %60 of left-handed people, it is expected that the left hemisphere of the brain is more active during vocal tasks [153, 154]. However, depending on the handedness of the subjects and how they are imagining the tongue movement, it is possible that the right hemisphere becomes more active during the tongue movement task; in which case the tongue movement will be more likely to be miss-classified as left-hand movement in comparison to right-hand movement.

Table 4.7: Normalized confusion matrices averaged over all the subjects for FBCSP-BMLDA method during the validation phase in Exp. 2. Note that the tasks are in the following order: left hand (Ω_1), right hand (Ω_2), both feet (Ω_3), and tongue (Ω_4) movement.

	CS=No	CS=Yes
SL=No	$\begin{bmatrix} 0.7631 & 0.0876 & 0.0580 & 0.0913 \\ 0.1180 & 0.7615 & 0.0561 & 0.0644 \\ 0.0737 & 0.1075 & 0.7002 & 0.1185 \\ 0.0575 & 0.0389 & 0.0752 & 0.8285 \end{bmatrix}$	$\begin{bmatrix} 0.7154 & 0.0845 & 0.0996 & 0.1005 \\ 0.1338 & 0.7420 & 0.0514 & 0.0728 \\ 0.1247 & 0.0848 & 0.6897 & 0.1009 \\ 0.0956 & 0.0415 & 0.0815 & 0.7814 \end{bmatrix}$
SL=Yes	$\begin{bmatrix} 0.6364 & 0.1373 & 0.0831 & 0.1432 \\ 0.1581 & 0.6305 & 0.0865 & 0.1250 \\ 0.1064 & 0.1221 & 0.6154 & 0.1561 \\ 0.1199 & 0.1203 & 0.1126 & 0.6472 \end{bmatrix}$	$\begin{bmatrix} 0.7625 & 0.1173 & 0.0590 & 0.0612 \\ 0.1207 & 0.7455 & 0.0637 & 0.0701 \\ 0.0709 & 0.0639 & 0.7461 & 0.1191 \\ 0.0749 & 0.0679 & 0.1178 & 0.7394 \end{bmatrix}$

Table 4.4. It should be noted that the winning method in the literature for Exp. 1 uses a combination of surface Laplacian + channel selection + short-time Fourier transformation + LDA method and achieves an average performance of %62.72 at the classifier output[6]. Also, the winning method in the literature for Exp. 2 uses the FBCSP-NBPW approach which does not use any surface laplacian or channel selection, which is provided as the benchmark solution in Table 4.6.

The averaged results over all the subjects in Exp. 1 and Exp. 2 are shown in Figures 4.5(b) and 4.6(b), respectively. If we compare these average results, with the average validation results in Figures 4.5(a) and 4.6(a), we can see that the general trends in testing phase are very similar to the trends in the validation phase. The minor differences in the performance trends can be attributed to the inter-session variation of the EEG characteristics, which has a more dominant effect in Exp. 2 since the it was not observable during the validation phase.

Figures 4.5 and 4.6 reveal that in both validation phase and test phase, the highest performance in both experiment is achieved when MVLDA method is utilized in the domain-agnostic FE step. Moreover, it can be seen that the vector-variate LDA method has a very poor and inconsistent performance, specially in Exp. 2. The 2DLDA method has a reasonable performance in Exp. 1, but fails to provide a consistent performance in Exp. 2 where the training data is very limited.

4.3.6 Bayes Optimality of the MVLDA

In Section 4.2, it was mentioned that both MVLDA and LDA are Bayes optimal for homoscedastic matrix-variate Gaussian data when the covariance matrices are known. However, when the covariance matrices need to be estimated from experimental data, MVLDA takes advantage of the reliable

[6]This winning algorithm also post processes the classifier outputs to correct for some misclassifications, which results in the final performance of %67.

matrix-variate estimation in lower-dimensional feature space and hence outperforms LDA. This reliable estimation improves the discriminance power of the extracted features, which in turn improves the performance of the BCI system, as shown by the experimental results.

It should be noted that both MVLDA and 2DLDA take advantage of reliable estimation in the lower dimensional space. Nevertheless, there is a significant performance gap between 2DLDA and MVLDA for many DS-FE cases since 2DLDA does not necessarily provide Bayes optimal features.

4.3.7 The Effect of surface Laplacian (SL) filtering

A closer look at the average performances in Figures 4.5 and 4.6 reveals that the surface Laplacian filtering improves the classification performance in all cases in Exp. 1, but it is not helpful in Exp. 2. The significant difference between these two experiments is caused by the fact that in Exp. 2 we do not have access to the exact locations of the EEG sensors, which in turn affects the accuracy of the surface Laplacian filtering. By comparing the results of Exp. 1 and Exp. 2, we can conclude that if the exact locations of the EEG sensors are known, then the use of surface Laplacian filtering is highly recommended. However, the use of surface Laplacian filter with approximate location information might corrupt the data and result in a significant performance loss.

Assuming that the surface Laplacian is accurately calculated, it will improve the performance of the subsequent feature extraction and classification methods, since it acts as a spatial highpass filter which emphasizes the effect of localized sources and increasing the spatial resolution of the EEG recordings. This effect can be clearly seen in Figures 4.5. Among different algorithms, the 2DLDA method benefits most from the surface laplacian filtering, specially when no channel selection is performed. As a result, the poor performance of the 2DLDA method on the raw EEG data (i.e., when no SL or CS has been performed) can be fixed by deployment of the surface Laplacian filter.

Finally, it should be mentioned that in Exp. 2, the LDA method cannot operate when surface Laplacian has been applied to the data and all channels are used for feature extraction. In this case, the LDA suffers from the fact that within-class scatter matrix of the FBCSP features is singular, which is caused by inaccurate calculation of the surface Laplacian transform. However, this problem can be resolved if surface Laplacian is combined with channel selection, which reduces the dimensionality of the data and results in a non-singular within-class covariance matrix.

4.3.8 The Effect of Channel Selection

In Section 3.1.1, it was mentioned that channel selection is considered as a simple strategy for dimensionality reduction. The results of Table 4.3 and Table 4.4 confirm that d_{opt} decreases for most methods when only the centro-parietal channels are used instead of all the channels. In case of FBCSP-LDA, d_{opt} is not affected by channel selection, mostly due to the fact that for LDA method, d can only take limited value up to $C-1$.

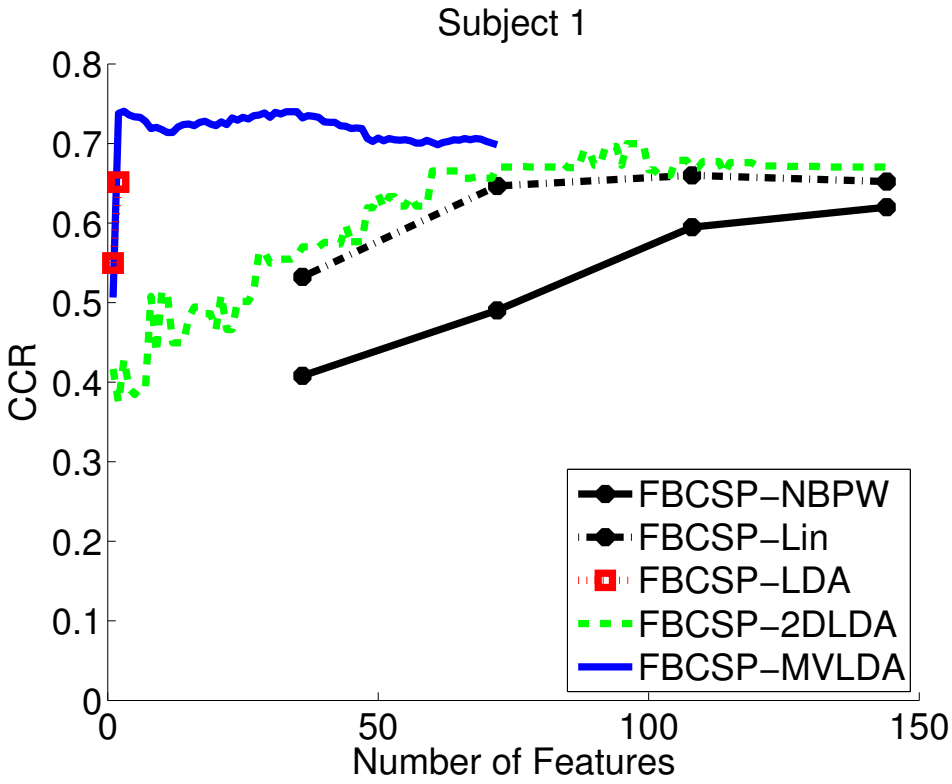
The effect of channel selection on the performance is not always consistent when applied to the raw EEG data. On one hand, channel selection helps to reduce the dimensionality of the data by only selecting the EEG channels which are located closer to the motor cortex area, which in turn can help in extraction of more relevant features at the next steps. On the other hand, channel selection completely ignores the data from discarded channels which may contain relevant information regarding the motor tasks.

However, when channel selection is applied to the EEG data which is already passed through surface Laplacian filtering, we can reasonably assume that each EEG channel mostly contains data from its neighbouring cortex area, and the information from EEG channels which are not close to the motor cortex can be safely discarded to improve the performance of the BCI system. The results from Figures 4.5 and 4.6 confirm this assumption. It should be noted that even in Exp. 2 the combination of channel selection and surface Laplacian results in high performances for all the methods. Therefore, we can conclude that channel selection is mostly effective when combined with surface Laplacian filtering.

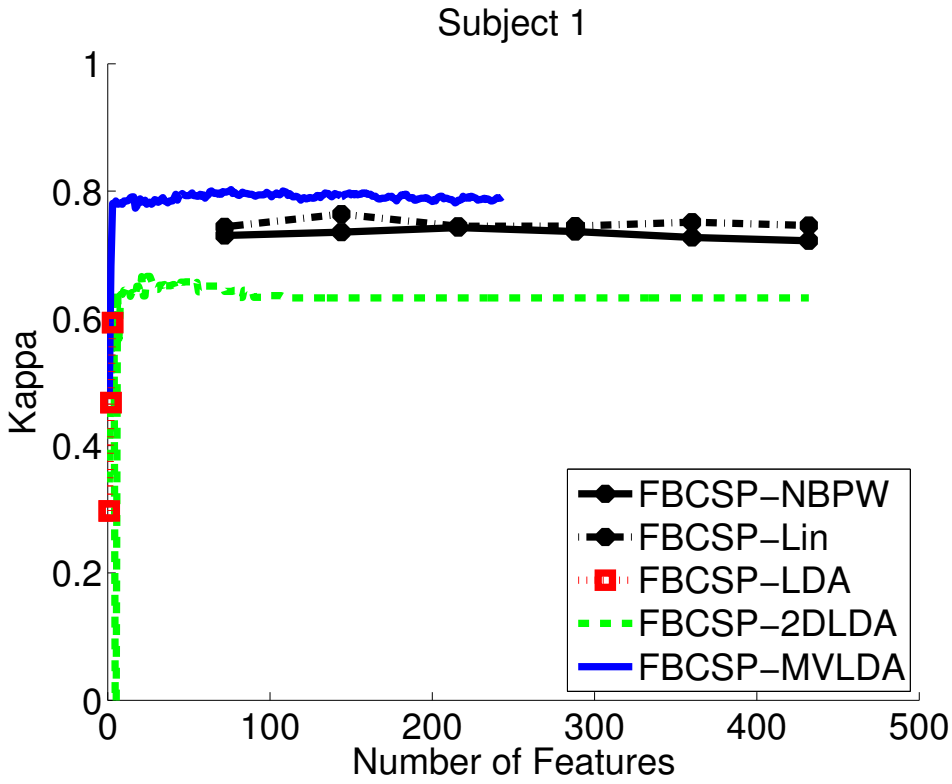
4.3.9 The Effect of Feature Space Dimensionality

As mentioned at the beginning of Section 4.3.4, the performance of BCI algorithms highly depends on the dimensionality of the feature space which is passed to the classifier. Figure 4.7 illustrates the effect of dimensionality on the performance of different methods for Subject 1 in Exp. 1 and Subject 1 in Exp. 2, when both surface Laplacian and channel selection are applied to the data during the validation phase. The performances reported in this figure are the average performances calculated over all the validation runs. Figures 4.8 and 4.9 illustrate similar results for the rest of subjects in Exp. 1 and Exp. 2, respectively. In order to clarify the inter-subject variations of the results, all these results are plotted in the same scale.

It can be seen from these figures that despite the inter-subject variations in the maximum performance of different methods, a similar trend exists for the relative performance of different methods in all the subject. In both experiments, for most of the subjects, the MVLDA method achieves the highest



(a) Correct Classification Rate (CCR) results for Subject 1 in Experiment-1.



(b) Kappa coefficient (κ) results for Subject 1 in Experiment-2.

Figure 4.7: Performance results for different methods versus the number of features in the validation phase for the first subject in (a) Experiment-1 and (b) Experiment-2.

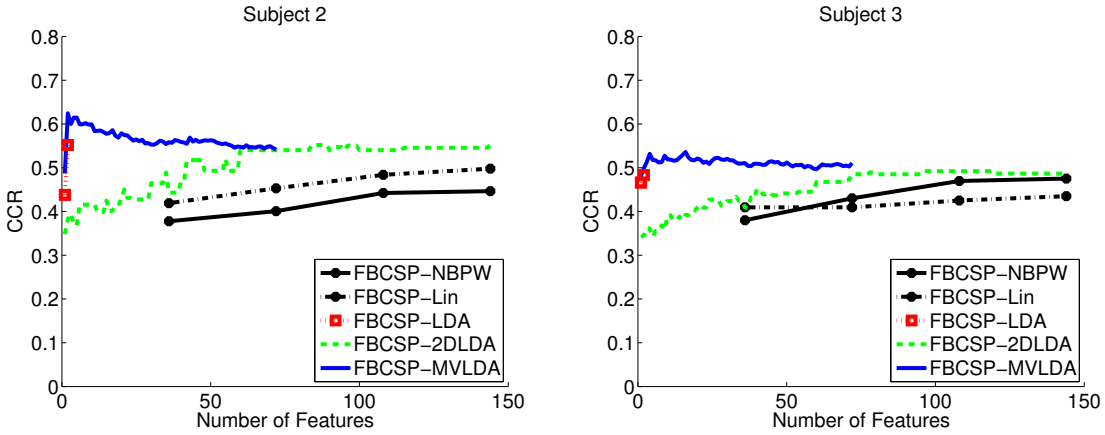


Figure 4.8: Correct Classification Rate (CCR) for different methods versus the number of features for all the subjects in the validation phase of Experiment-1. The illustrated results are for the case where both surface laplacian filtering and channel selection have been performed on the data. Note that FBCSP-LDA method provides at most $C-1$ features. Also, the minimum dimension for FBCSP-NB method is $N_f \times (2C)$.

performance among all the methods with a relatively small number of features. This behaviour is owing to the ability of MVLDA in extraction of highly discriminant features and more importantly sorting them according to their discriminance power.

In Exp. 1, the 2DLDA method achieves the second highest performance, after MVLDA, however its best performance at much higher number of features compared to the MVLDA method. This fact demonstrates the relative weakness of 2DLDA in dimensionality reduction and extraction of the most significant features. In Exp. 2 were the training data is very limited, the 2DLDA has a very poor performance and has the second worse performance, after the LDA method, which also suffers from the small number of training samples.

In cases where no domain-agnostic FE is deployed, i.e., FBCSP-NBPW and FBCSP-Lin, the naive Bayes classifier and the linear classifier have very close performance for most of the subjects in both experiments. In Exp. 1, where enough training samples are available, the performance of both classifiers tend to constantly increase as the number of features passed to the classifier increases. This trend shows that most of the features extracted by the FBCSP contain discriminant information. If we compare this trend with performance of FBCSP-MVLDA method, it can be seen that the MVLDA module has been highly successful in finding a very low dimensional subspace which contains all the discriminant information of the data.

In Exp. 2, where the training data is extremely limited, the performance of both FBCSP-NBPW and FBCSP-Lin methods is flat or decreasing for all the subjects. This trend suggests that most of the features extracted by the FBCSP algorithm do not contain discriminant information or they are highly

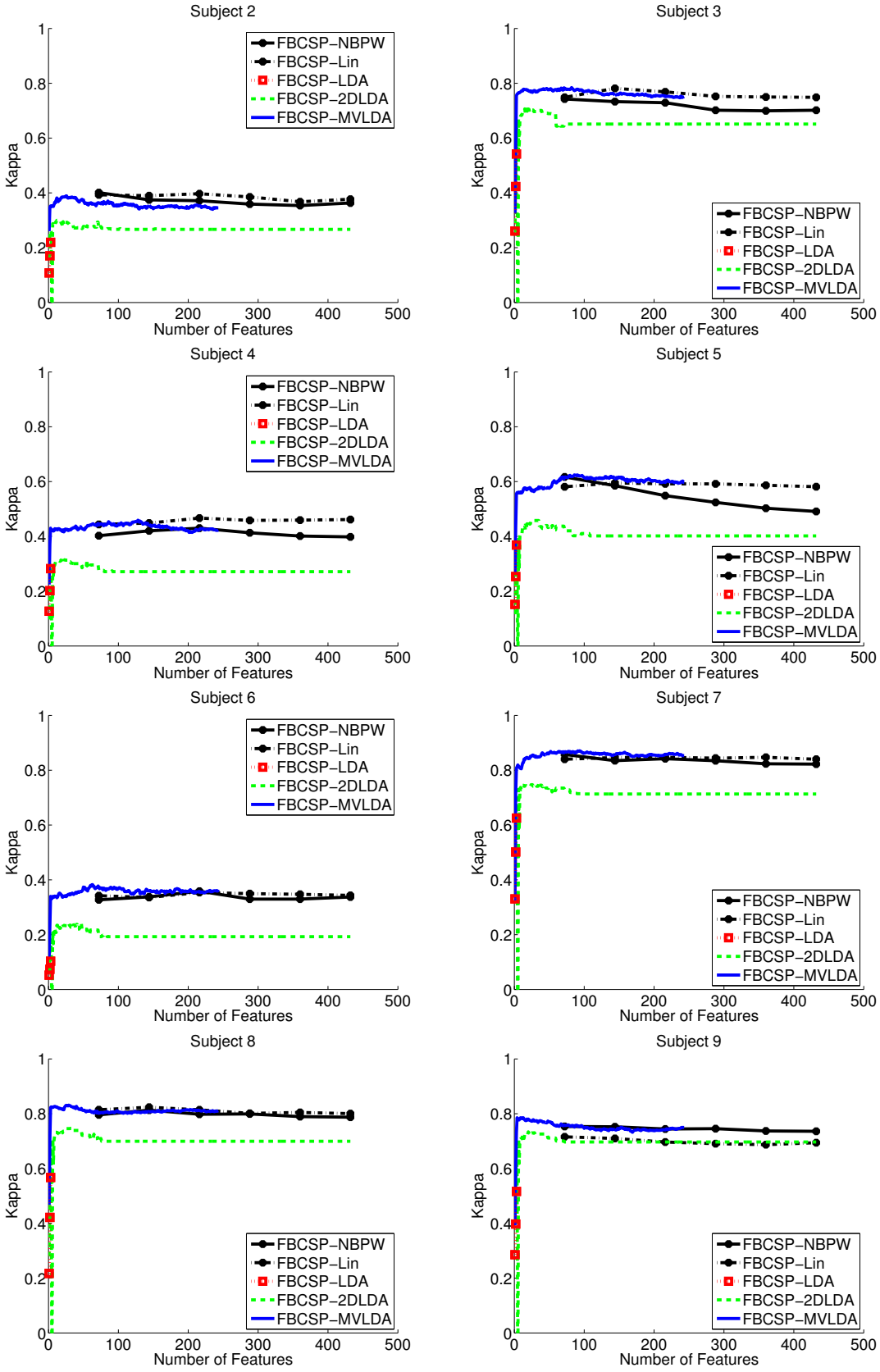


Figure 4.9: Kappa coefficient (κ) for different methods versus the number of features for all the subjects in the validation phase of Experiment-2. All these methods are applied to the raw EEG data, i.e., No surface laplacian or channel selection has been performed. Note that FBCSP-LDA method provides at most $C-1$ features. Also, the minimum dimension for FBCSP-NB method is $N_f \times (2C)$.

contaminated with noise. As a result the performance gap between FBCSP-MVLDA and FBCSP-Lin is not as pronounced as Exp. 1.

In order to Further study the effect of feature space dimensionality on the performance of each method when surface laplacian and/or channel selection are not applied to the data, we have provided the performances for Subjects 1 and 2 from Exp. 1 and Subjects 7 and 5 from Exp. 2 in Figures 4.10 - 4.14. For brevity of the results, we have selected one high performing subject and one low performing subject from each experiments. Similar trends can be seen in other subjects as well.

Figure 4.10 provides the comparative results for MVLDA in both experiments. It can be seen that the combination of surface Laplacian and channel selection significantly improves the performance of MVLDA regardless of the number of output features. The only exception is Subject 7 in Exp. 2, in which case MVLDA already achieves a very high performance of more than %90 using the raw data. It can also be seen that the use of surface Laplacian without channel selection has little positive effect on the performance in Exp. 1 while having a deteriorative effect in Exp. 2.

Similarly, Figure 4.11 provides the comparative results for 2DLDA method. Since 2DLDA provides the features in a matrix-variate structure, it can only support feature numbers of the form $d = m * n$. This limitation is the cause of the discontinuities in these plots. Moreover, it should be noted that for most values of d, there are several values of m and n that can result in a total of d features. As an example, for $d = 12$ features, the following are the possible cases for (m, n) values: $(1, 12), (2, 6), (3, 4), (4, 3), (6, 2), (12, 1)$. In order to have a fair comparison with other methods, for each value of d we have considered the (m, n) combination which provides the best performance. In Exp. 1, where enough training data is available, both surface Laplacian and channel selection highly improve the performance of the 2DLDA. However, in Exp. 2, where training samples are very limited, the performance of 2DLDA can only be improved when both surface Laplacian and channel selection are used together. The same trend can be observed for for LDA method in Figures

4.4 Summary and Concluding Remarks

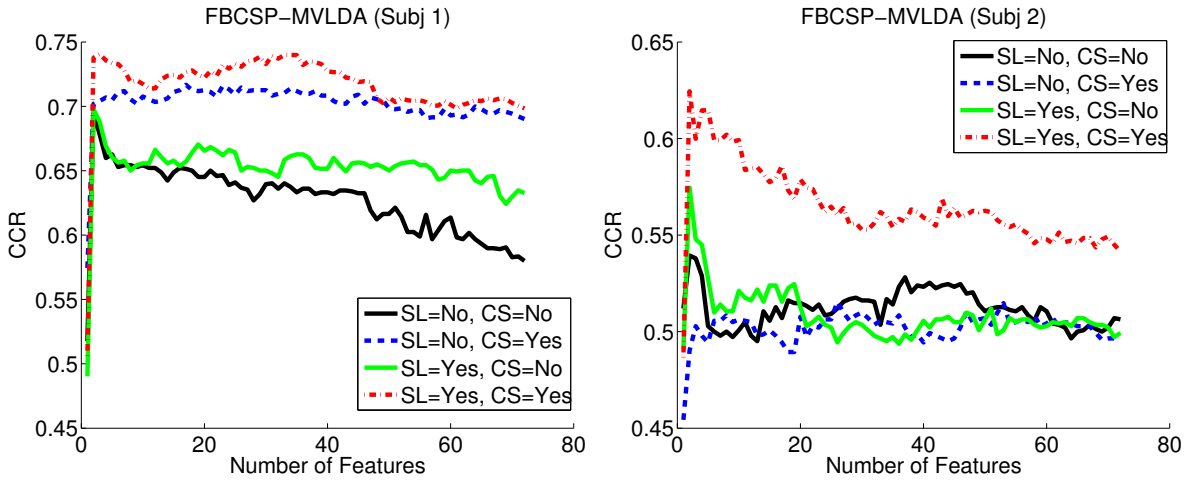
In this chapter, a new matrix-variate (or bilinear) approach was proposed for domain-agnostic FE in the MI-BCI systems. Based on a homoscedastic matrix-variate Gaussian model for the spatio-spectral features extracted by the FBCSP method, the 2DLDA and MVLDA methods were studied as two main candidates for matrix-variate extension of the LDA algorithm. Both 2DLDA and MVLDA methods directly operate on the matrix-variate data, using bilinear spectral and spatial operators.

Compared to LDA, MVLDA provides a reduced computational complexity, allows for possibility

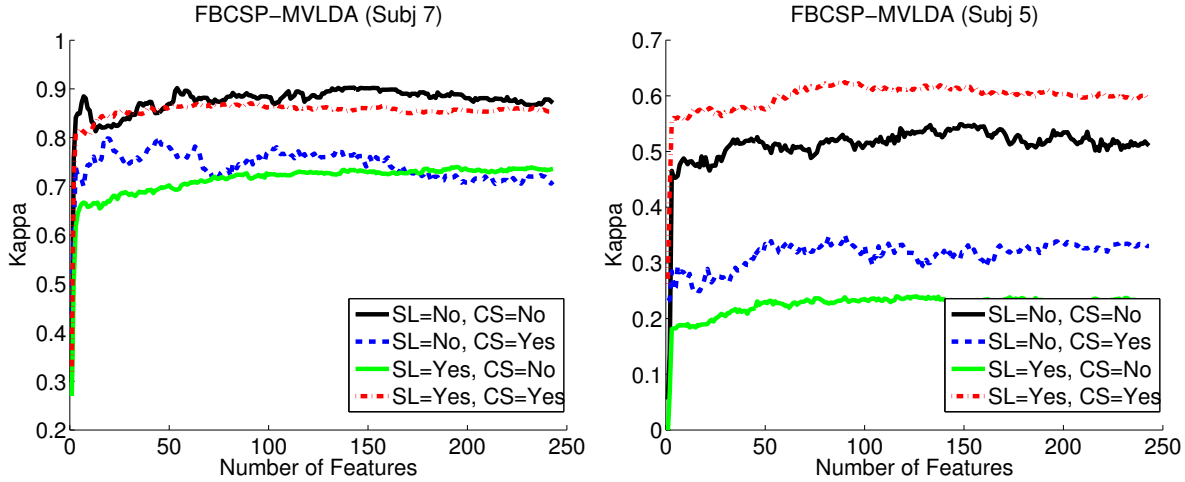
of parallel training of spatial/spectral operators, and most importantly, utilizes more reliable parameter estimates. Furthermore, compared to the 2DLDA method, the MVLDA method is non-iterative, and more importantly can determine the most discriminant features for an arbitrary reduction in the dimension. The performance of these schemes was evaluated in two different experiments. The first experiment represented a typical MI-BCI scenario where training data is collected over multiple sessions and each training trial lasts for 15 seconds. The second experiment represented an extremely restricted case where only one training session, with trials of length 3 seconds, is available. In both experiments, the MVLDA method outperformed the other algorithms, which shows that the assumed matrix-variate Gaussian distribution provides a reasonable model for the FBCSP features.

Finally, the effect of surface Laplacian (SL) and channel selection (CS) methods on the performance of the proposed methods was analyzed. The experimental results show that the channel selection is mostly beneficial when it is combined with surface Laplacian filtering. The surface Laplacian filtering assures that each EEG channel mostly conveys localized information regarding its neighbouring area on the brain cortex, which allows us to ignore the EEG channels which are not close to the motor cortex area and manually reduce the dimensionality of the input data.

It is worth mentioning that motor-imagery BCI systems generally exhibit high inter-subject variability, which can be attributed to various factors such as the individual difference in the level of concentration/engagement as well as the neuro-phisiological differences. In both cases the performance of the BCI can be significantly improved by increasing the amount of training time. In the former case, extra training with real-time feedback helps the user to improve his/her concentration level, which in turn improves the performance of the BCI system. In the latter case, the extra training helps the algorithm to have a better estimation of the signal parameters and avoid overfitting.

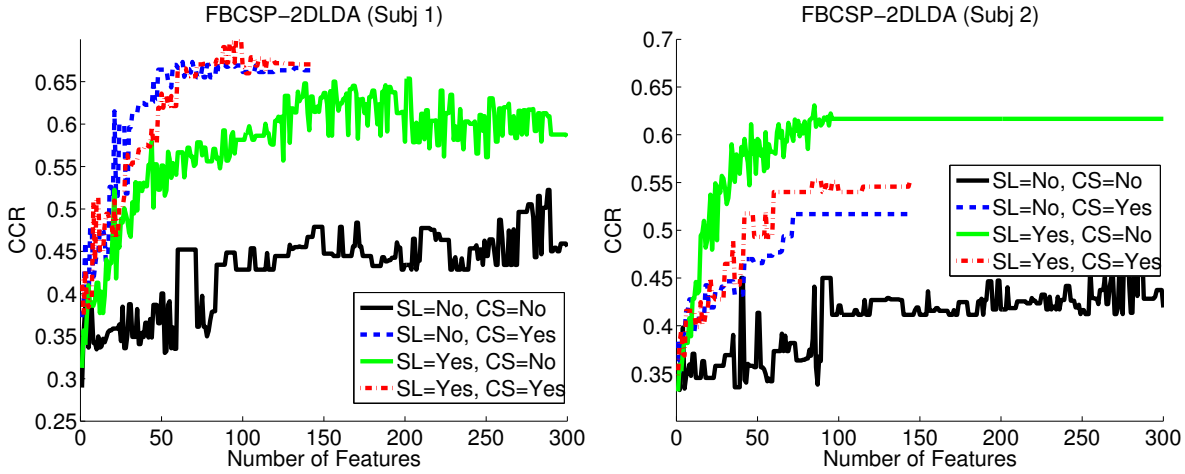


(a) Correct Classification Rate (CCR) Results, Experiment-1

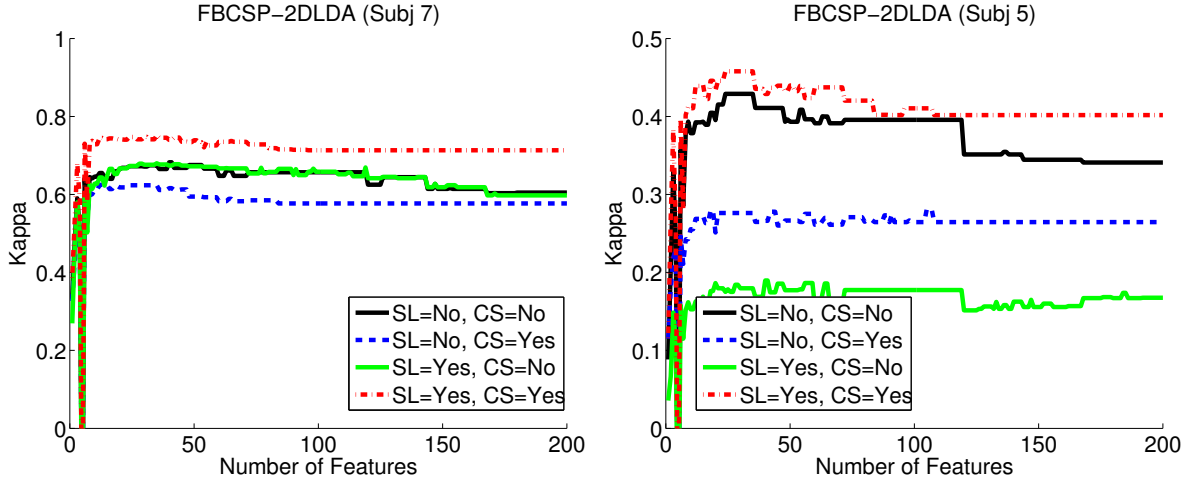


(b) Kappa coefficient (κ) Results, Experiment-2

Figure 4.10: The effect of surface Laplacian (SL) filtering and channel selection (CS) on the performance of FBCSP-MVLDA method versus the number of features that are used for classification. For brevity, only the results of two subjects from each experiment are presented to illustrate the general trends in one high-performing subject and one low-performing subject.

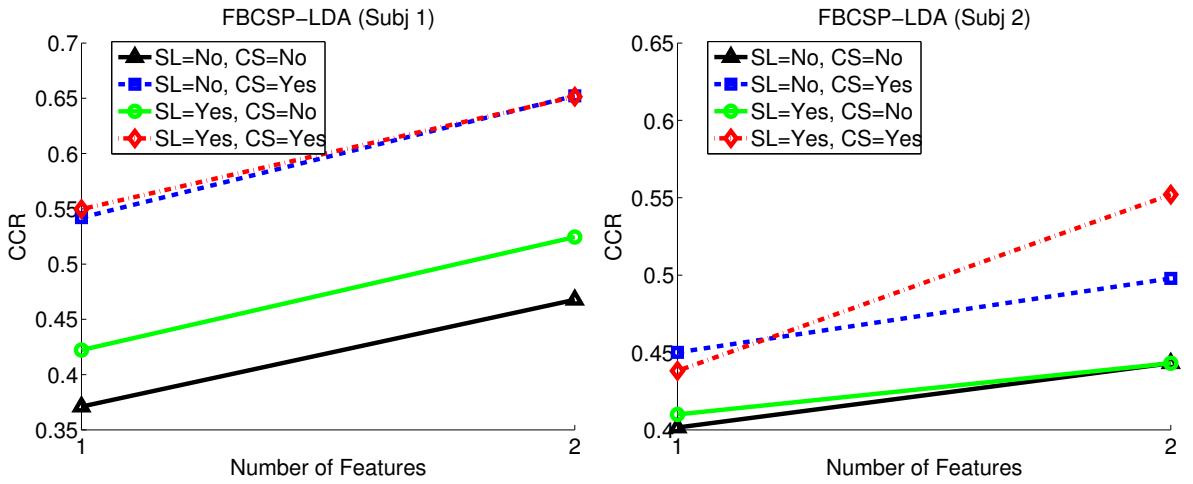


(a) Correct Classification Rate (CCR) Results, Experiment-1

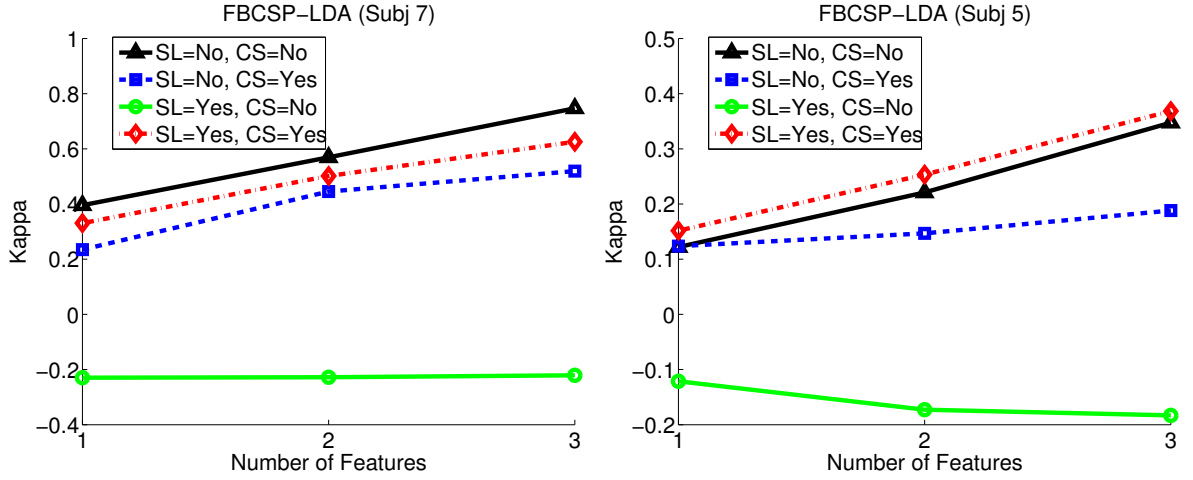


(b) Kappa coefficient (κ) Results, Experiment-2

Figure 4.11: The effect of surface Laplacian (SL) filtering and channel selection (CS) on the performance of FBCSP-2DLDA method versus the number of features that are used for classification. For brevity, only the results of two subjects from each experiment are presented to illustrate the general trends in one high-performing subject and one low-performing subject. To provide a more clear illustration, the graphs are zoomed in to the range of $0-300$ features for Exp. 1 and $0-200$ features for Exp. 2.

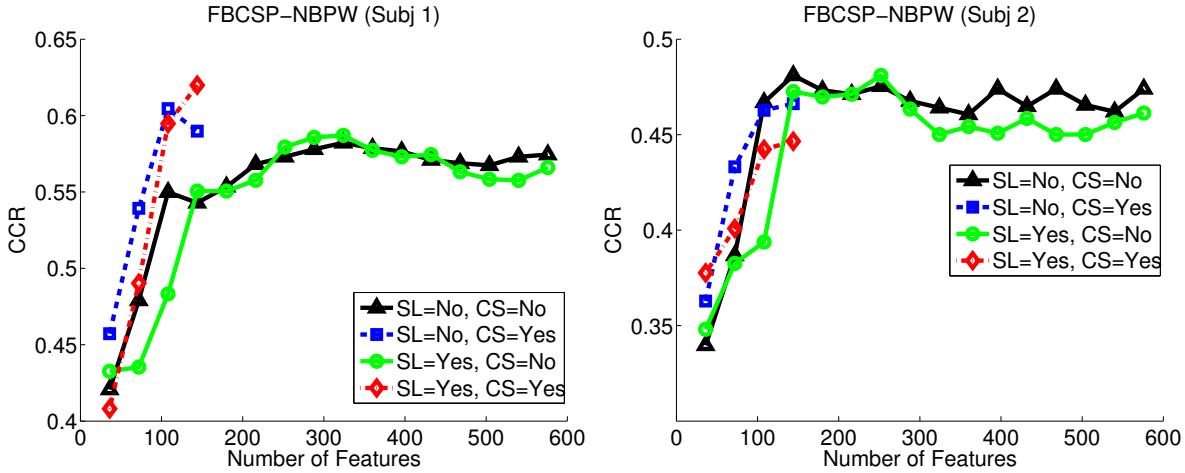


(a) Correct Classification Rate (CCR) Results, Experiment-1

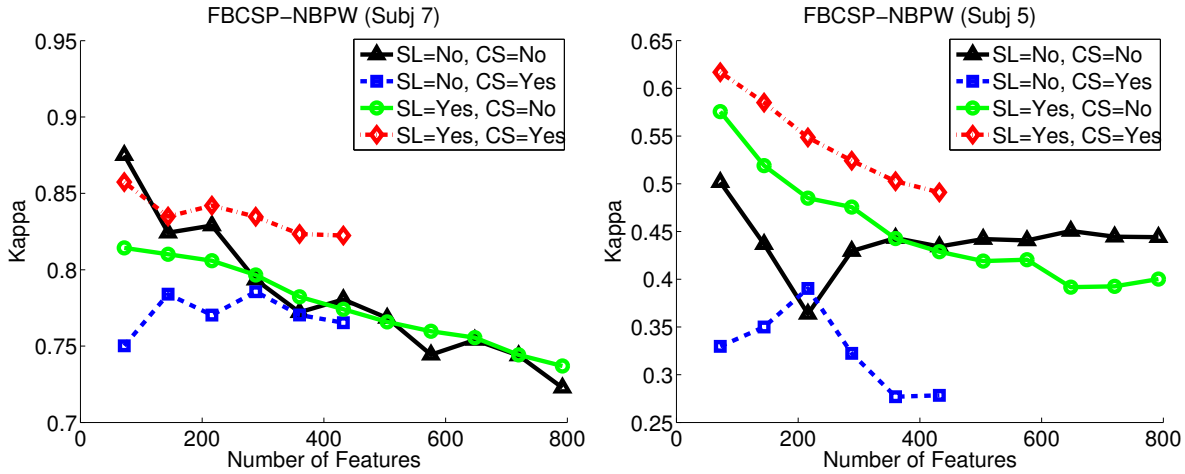


(b) Kappa coefficient (κ) Results, Experiment-2

Figure 4.12: The effect of surface Laplacian (SL) filtering and channel selection (CS) on the performance of FBCSP-LDA method versus the number of features that are used for classification. For brevity, only the results of two subjects from each experiment are presented to illustrate the general trends in one high-performing subject and one low-performing subject. Note that FBCSP-LDA method provides at most $C-1$ features, i.e, two features in Exp. 1 and three features in Exp. 2.

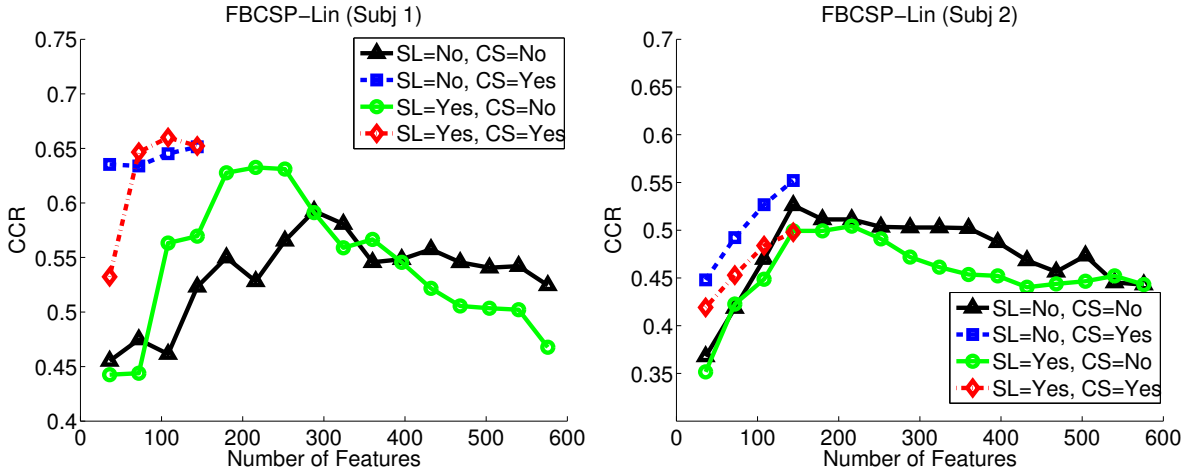


(a) Correct Classification Rate (CCR) Results, Experiment-1

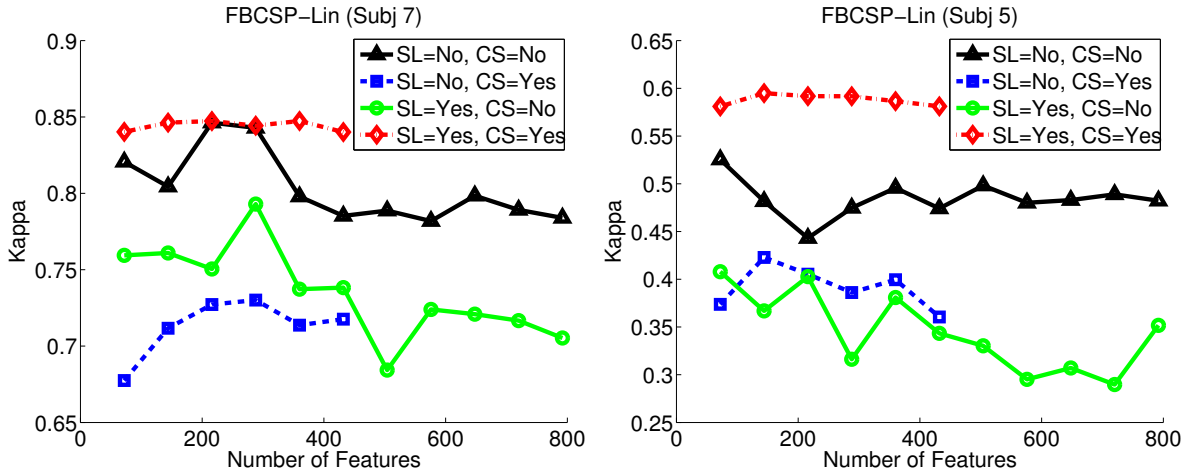


(b) Kappa coefficient (κ) Results, Experiment-2

Figure 4.13: The effect of surface Laplacian (SL) filtering and channel selection (CS) on the performance of FBCSP-NBPW method versus the number of features that are used for classification. For brevity, only the results of two subjects from each experiment are presented to illustrate the general trends in one high-performing subject and one low-performing subject. Note that FBCSP-LDA method provides at most $C-1$ features, i.e, two features in Exp. 1 and three features in Exp. 2.



(a) Correct Classification Rate (CCR) Results, Experiment-1



(b) Kappa coefficient (κ) Results, Experiment-2

Figure 4.14: The effect of surface Laplacian (SL) filtering and channel selection (CS) on the performance of FBCSP-Lin method versus the number of features that are used for classification. For brevity, only the results of two subjects from each experiment are presented to illustrate the general trends in one high-performing subject and one low-performing subject. Note that FBCSP-LDA method provides at most $C-1$ features, i.e, two features in Exp. 1 and three features in Exp. 2.

Chapter 5

Domain-Specific FE Based on Matrix-Variate Model for Multiband EEG Rhythms

In Chapter 4, it was shown that the spatio-spectral features that are generated by the filterbank common spatial pattern (FBCSP) method can be modelled as a matrix-variate Gaussian data, based on which efficient domain-agnostic FE schemes can be developed to improve the performance of the overall BCI system. The results of the previous chapter motivates us to have a closer look at the FBCSP method and examine if the assumption of matrix-variate Gaussianity can be directly used at the domain-specific FE stage to improve the efficiency of the system.

Despite its high performance, the FBCSP method suffers from a number of shortcomings as listed below:

- FBCSP suffers from high computational cost at the training phase since it requires a separate feature extractor for each spectral band, each of which requires calculation of generalized eigenvectors for covariance matrices of size $N_{ch} \times N_{ch}$, where N_{ch} denotes the number of EEG channels.
- Since each spectral band is treated independently, possible correlations between different EEG rhythms are completely ignored by the FBCSP method, which in turn causes redundancy in the extracted feature set.
- FBCSP does not provide any measure for comparing discriminant power of the features obtained from different spectral bands. Although the CSP features within each band are sorted based on

their discriminant power, it is not possible to sort the features across different bands.

In this chapter, we propose a novel algorithm which simultaneously processes the EEG rhythmic activities in both spatial and spectral domains, and extracts the most discriminant spatio-spectral features across all the frequency bands. The proposed method, called *separable common spatio-spectral patterns (SCSSP)*, is based on a matrix-variate Gaussian model for spatio-spectral EEG patterns which allows us to develop a bilinear feature extractor. Compared to the FBCSP method, our algorithm has the following main advantages: First, it involves only two CSP-type modules, regardless of the number of frequency bands (N_f). As a result, the computational cost of training SCSSP algorithm in a practical BCI is less than FBCSP. Second, the features are extracted based on joint analysis of both spatial and spectral characteristics of the signal. Therefore, correlations between different spectral bands can be exploited for feature extraction. Third, a measure is provided to rank the discriminatory power of extracted spatio-spectral features, which eliminates the need for a subsequent feature selection stage.

5.1 System Model

Figure 5.1(a) illustrates the processing pipeline of our proposed algorithm and how it compares with the FBCSP method (Figure 5.1(b)). Consider an EEG epoch with N_t samples from N_{ch} channels. After passing the EEG epoch through a set of N_f bandpass filters, we get N_t matrices of size $N_f \times N_{ch}$, each of which representing a spatio-spectral EEG pattern. The ultimate goal is to extract the most discriminant features from these matrix-variate patterns.

Let $\mathbf{X} \in \mathbb{R}^{N_f \times N_{ch}}$ denote the matrix-variate EEG pattern at the output of the bandpass filterbank. Each motor-imagery task, denoted by class Ω_i, is characterized by the likelihood density $f(\mathbf{X}|\Omega_i)$. We adopt the heteroscedastic matrix-variate Gaussian model of Section 3.4.1 for these likelihoods, i.e.,

$$\mathbf{X}|\Omega_i \sim \mathcal{N}(\mathbf{M}_i, \mathbf{\Phi}_i, \mathbf{\Psi}_i), \qquad 1 \leq i \leq C \tag{5.1}$$

where, $\mathbf{M}_i$ denotes the class mean, $\mathbf{\Phi}_i$ is the spectral covariance, also called column-wise or left covariance, and $\mathbf{\Psi}_i$ is the spatial covariance, also called row-wise or right covariance. Since $\mathbf{X}$ is obtained from bandpass filtering of the EEG signal, all classes have zero mean, i.e., $\mathbf{M}_i = \mathbf{0}$ for $1 \leq i \leq C$. Therefore, the discriminant information are contained in the second order statistics of the data.

In the proposed method, we directly focus on the matrix-variate structure of the multiband EEG rhythms at the output of the bandpass filterbank, and use the statistical model of (5.1) to develop a bilinear domain-specific FE method for $\mathbf{X}$.

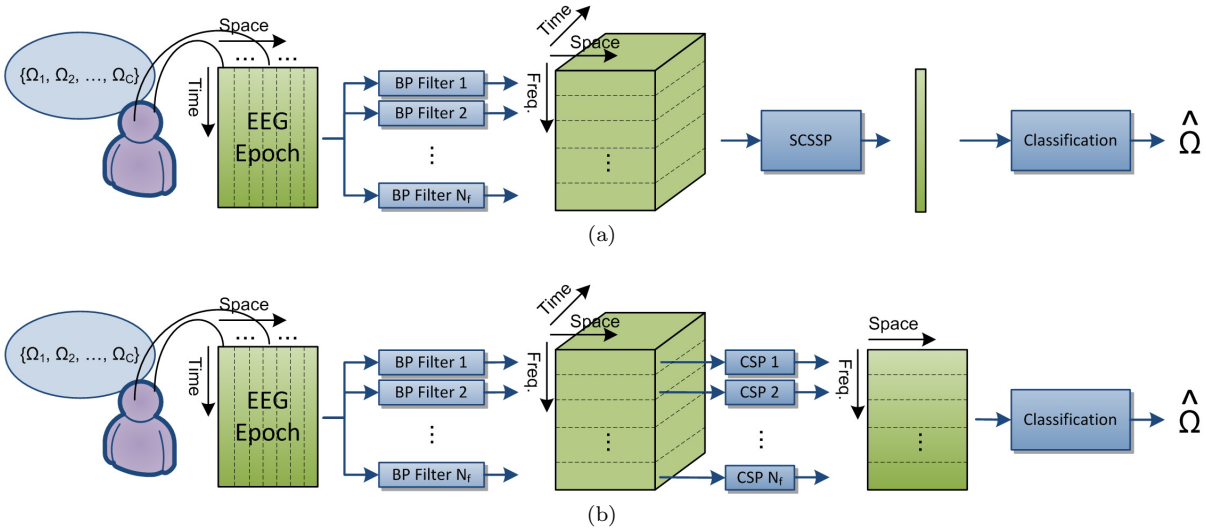


Figure 5.1: System model for spatio-spectral feature extraction schemes in (a) Separable common spatio-spectral pattern (SCSSP) method, and (b) Filter-bank common spatial pattern (FBCSP) method.

5.2 Separable Common Spatio-Spectral Patterns (SCSSP) Method

Consider a binary classification problem, i.e., $\Omega_i \in \{\Omega_1, \Omega_2\}$, and let $\mathbf{x} = \text{vec}(\mathbf{X})$ denote the feature vector which is formed by the column-wise concatenation of the elements in $\mathbf{X}$. The matrix-variate Gaussianity assumption in (5.1) implies that the feature vector $\mathbf{x}$ has a heteroscedastic vector-variate distribution as follows:

$$\mathbf{x}|\Omega_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i), \qquad i \in \{1, 2\} \tag{5.2}$$

where $\mathbf{\Sigma}_i = \mathbf{\Psi}_i \otimes \mathbf{\Phi}_i$. Moreover, recall from the discussion in Section 3.4 that any bilinear operation of the form $\mathbf{W}_L \mathbf{X} \mathbf{W}_R$ is equivalent to a linear operation of the form $\mathbf{W}^T\mathbf{x} = (\mathbf{W}_R \otimes \mathbf{W}_L)^T\mathbf{x}$.

Based on these properties, and following the general goal of the CSP approach, we look for a bilinear operation on $\mathbf{X}$, which simultaneously diagonalizes both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. In other words, we look for transformation matrices $\mathbf{W}_L$ and $\mathbf{W}_R$ which are the solutions to the following generalized eigenvalue problem:

$$\mathbf{\Sigma}_1 \mathbf{W} = (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\, \mathbf{W} \mathbf{\Lambda}, \tag{5.3}$$

$$\text{where} \quad \mathbf{W} = \mathbf{W}_R \otimes \mathbf{W}_L, \text{ and } \quad \mathbf{\Sigma}_i = \mathbf{\Psi}_i \otimes \mathbf{\Phi}_i$$

The next theorem provides the solution for (5.3).

Theorem 1: Let $\mathbf{x} = \text{vec}(\mathbf{X})$, where $\mathbf{X} \in \mathbb{R}^{N_f \times N_{ch}}$ has a matrix-variate Gaussian distribution as given

by (5.1). Then, the solution to (5.3) is given as follows:

$$\mathbf{\Lambda} = (\mathbf{\Lambda}_R \otimes \mathbf{\Lambda}_L)\left(\mathbf{\Lambda}_R \otimes \mathbf{\Lambda}_L + (\mathbf{I}_{N_{ch}} - \mathbf{\Lambda}_R) \otimes \left(\mathbf{I}_{N_f} - \mathbf{\Lambda}_L\right)\right)^{-1}$$
$$\mathbf{W} = \mathbf{W}_R \otimes \mathbf{W}_L$$

where $\mathbf{I}_K$ is the identity matrix of size K and the matrices $\mathbf{\Lambda}_R$, $\mathbf{W}_R$, $\mathbf{\Lambda}_L$ and $\mathbf{W}_L$ are the solutions to generalized eigenvalue problems for spatial and spectral covariances, respectively:

$$\mathbf{\Psi}_1 \mathbf{W}_R = (\mathbf{\Psi}_1 + \mathbf{\Psi}_2)\, \mathbf{W}_R \mathbf{\Lambda}_R, \tag{5.4}$$

$$\mathbf{\Phi}_1 \mathbf{W}_L = (\mathbf{\Phi}_1 + \mathbf{\Phi}_2)\, \mathbf{W}_L \mathbf{\Lambda}_L, \tag{5.5}$$

Proof: The proof is provided in Appendix.

Using this theorem, we can break the generalized eigenvalue problem of Equation (5.3) into the two lower-dimensional problems presented in Equations (5.4) and (5.5). Note that $\mathbf{W}_L$ provides the spectral transformation matrix, whereas $\mathbf{W}_R$ provides the spatial transformation matrix. These two transformations will be simultaneously applied to the matrix-variate data $\mathbf{X}$.

To provide a better insight into the result of Theorem 1, let λ_k, $1 \leq k \leq N_f N_{ch}$, denote the diagonal entries of $\mathbf{\Lambda}$ sorted in descending order. Theorem 1 implies that

$$\lambda_k = \frac{\lambda_{L,l[k]}\, \lambda_{R,j[k]}}{\lambda_{L,l[k]}\, \lambda_{R,j[k]} \,+\, (1 - \lambda_{L,l[k]})(1 - \lambda_{R,j[k]})} \tag{5.6}$$

where $\lambda_{L,l[k]}$ and $\lambda_{R,j[k]}$ are the corresponding eigenvalues in $\mathbf{\Lambda}_L$ and $\mathbf{\Lambda}_R$, with $1 \leq l[k] \leq N_f$ and $1 \leq j[k] \leq N_{ch}$. Also, the eigenvectors corresponding to λ_k are expressed as

$$\mathbf{w}_k = \mathbf{w}_{R,j[k]} \otimes \mathbf{w}_{L,l[k]},$$

where $\mathbf{w}_{R,j[k]}$ and $\mathbf{w}_{L,l[k]}$ are the eigenvectors in $\mathbf{W}_R$ and $\mathbf{W}_L$ corresponding to $\lambda_{R,j[k]}$ and $\lambda_{L,l[k]}$.

Note that for $k = 1$ and $k = N_f N_{ch}$ the pair of features $[y_1, y_{N_f N_{ch}}]^T$ provide the most discriminant power. Similarly, the features corresponding to $k = 2$ and $k = (N_f N_{ch} - 1)$ are the second most discriminant features, and so on. Based on these results, the following algorithm will be used for extraction of the "d" most discriminant spatio-spectral features:

1. Assuming that N_i training samples $\mathbf{X}_{i,n}$, $1 \leq n \leq N_i$, are available for each class Ω_i, estimate the

spatial covariance and spectral covariance of the data, using the following equations:

$$\hat{\mathbf{\Psi}}_i = \frac{1}{N_f N_i} \sum_{n=1}^{N_i} \mathbf{X}_{i,n}^T \mathbf{X}_{i,n}, \tag{5.7}$$

$$\hat{\mathbf{\Phi}}_i = \frac{1}{N_{ch} N_i} \sum_{n=1}^{N_i} \mathbf{X}_{i,n} \mathbf{X}_{i,n}^T. \tag{5.8}$$

2. Solve the generalized eigenvalue problems in (5.4) and (5.5) for the estimated spatial and spectral covariance matrices.

3. Using (5.6), calculate the eigenvalues λ_k and sort them in descending order to determine the corresponding indices $l[k]$ and $j[k]$.

4. Extract the d most discriminant features by calculating

$$y_k = \mathbf{w}_{L,l[k]}^T \mathbf{X} \mathbf{w}_{R,j[k]} \quad \text{for} \quad k \in \mathfrak{K} \tag{5.9}$$

where

$$\mathfrak{K} = \left\{ 1,\, N_f N_{ch},\, 2, (N_f N_{ch} - 1),\, \cdots,\, \frac{d}{2},\, (N_f N_{ch} - \frac{d}{2} + 1) \right\}.$$

Note that d is an even number here, similar to the CSP.

5. Calculate the normalized power of features over the length of epoch, in logarithmic scale, as follow:

$$z_k = \log \left(\frac{\text{var}\,(y_k)}{\Sigma_{k \in \mathfrak{K}} \text{var}\,(y_k)} \right) \tag{5.10}$$

where $\text{var}\,(y_k)$ function calculates the variance or power of y_k over N_t samples.

6. Construct the feature vector $\mathbf{z} = [z_1, z_{N_f N_{ch}}, \cdots]^T \in \mathbb{R}^{d \times 1}$ as the output of SCSSP algorithm.

It is worth mentioning that λ_k ranges between zero and one, and its value provides a measure for discriminant power of feature y_k. Similar to the conventional CSP method, values close to zero or one correspond to high discriminant features, whereas values close to $\frac{1}{2}$ correspond to low discriminant features. Thus, the pairs of extracted spatio-spectral features in $\mathbf{z}$ are sorted according to their discriminant power in descending order. These features are then passed to a classifier to determine the $\hat{\Omega}$. In our experimental studies, we consider two possible choices for classifier: (a) Naive Bayes classifier, (b) linear classifier.

5.3 A Comparative Discussion on The Theoretical Assumptions of FBCSP and SCSSP

This section briefly compares the SCSSP and FBCSP methods to provide the reader with a better understanding of the similarities and the differences between these two methods. Here, we use the following notation. Consider the matrix-variate data $\mathbf{X}$ at the output of the bandpass filterbank, and denote the f^{th} row-vector of $\mathbf{X}$ by $\mathbf{x}_f$, where $1 \leq f \leq N_f$. Also, let $\mathbf{x}' = [\mathbf{x}_1, \cdots, \mathbf{x}_{N_f}]$ denote the row-vector that is generated from the row-wise concatenation of the elements in $\mathbf{X}$, i.e., $\mathbf{x}' = (\text{vec}(\mathbf{X}^T))^T$. The class conditional covariance matrix of each row-vector $\mathbf{x}_f$ will be represented by $\mathbf{\Psi}_i^f$, and the class conditional covariance matrix of $\mathbf{x}'$ is represented by $\mathbf{\Sigma}_i^{'}$.

Recall that in the FBCSP approach, each row-vector $\mathbf{x}_f$ is processed independently from the other rows, using the projection matrix $\mathbf{W}_R^f$ which contains the generalized eigenvectors of $\mathbf{\Psi}_1^f$ and $\mathbf{\Psi}_1^f + \mathbf{\Psi}_2^f$. The projected feature pairs are then sorted in descending order of significance. Finally, the log-power of the resulting features are calculated during the epoch length and form the f^{th} row of the output feature matrix. In comparison of this approach with the SCSSP's approach, the following differences can be pointed out.

The assumption of matrix-variate Gaussianity which is used in SCSSP method implies that the covariance matrix of each row-vector $\mathbf{x}_f$ is equal, up to a scale, to the covariance matrix of other row-vectors. As a result, the SCSSP method only looks for one spatial filtering matrix $\mathbf{W}_R$ which will be commonly applied to all the row-vectors in $\mathbf{X}$. In contrast, the FBCSP method assumes that each row-vector $\mathbf{x}_f$ has a unique covariance matrix, and hence looks for a unique spatial filtering matrix $\mathbf{W}_R^f$ for each row.

The other important difference between FBCSP and SCSSP methods is in the spectral processing of the data. The FBCSP method assumes that different EEG rhythms in different frequency bands are independent from each other, and hence independently processes each rhythm. However, the SCSSP method calculates the class conditional spectral covariance matrix $\mathbf{\Phi}_i$ and uses this information along with the information from the spatial covariance matrix $\mathbf{\Psi}_i$ for extraction of the most discriminant spatio-spectral features. It is worth mentioning that owing to the matrix-variate Gaussianity assumption, the SCSSP method assumes that all EEG channels have the same spectral covariance matrices, up to a scale, and hence calculates a common spectral covariance matrix for all the channels.

In order to further clarify these points, consider the row vector $\mathbf{x}' = [\mathbf{x}_1, \cdots, \mathbf{x}_{N_f}]$, which contains all the elements of $\mathbf{X}$. The FBCSP method assumes a block-diagonal structure for the class conditional

covariance of $\mathbf{x}'$ as follows:

$$\mathbf{\Sigma}_i^{'} = \begin{bmatrix} \mathbf{\Psi}_i^1 & & & \\ & \mathbf{\Psi}_i^2 & & \\ & & \ddots & \\ & & & \mathbf{\Psi}_i^{N_f} \end{bmatrix} \tag{5.11}$$

whereas the SCSSP assumes the following block-wise structure for $\mathbf{\Sigma}_i^{'}$

$$\mathbf{\Sigma}_i^{'} = \begin{bmatrix} \phi_{11}\mathbf{\Psi}_i & \phi_{12}\mathbf{\Psi}_i & \cdots & \phi_{1N_f}\mathbf{\Psi}_i \\ \phi_{21}\mathbf{\Psi}_i & \phi_{22}\mathbf{\Psi}_i & \cdots & \phi_{2N_f}\mathbf{\Psi}_i \\ \vdots & & \ddots & \\ \phi_{N_f1}\mathbf{\Psi}_i & \phi_{N_f2}\mathbf{\Psi}_i & \cdots & \phi_{N_fN_f}\mathbf{\Psi}_i \end{bmatrix} \tag{5.12}$$

where ϕ_{mn} represents the $(m,n)^{th}$ element of the spectral covariance matrix $\mathbf{\Phi}$.

It is noteworthy that both assumption in (5.11) and (5.12) are restrictive models for the spatio-spectral covariance of the data. The FBCSP completely ignores the off-diagonal blocks of the $\mathbf{\Sigma}_i^{'}$, while trying to provide an accurate estimate for the diagonal blocks. In contrast, the SCSSP method takes into account the off-diagonal blocks of $\mathbf{\Sigma}_i^{'}$ by making the simplifying assumption that all the blocks in $\mathbf{\Sigma}_i^{'}$ are up to a scale equal to each other, where the scaling factor is determined by the elements of spectral covariance matrix.

5.4 Multiclass Extension of the SCSSP Method

The one-versus-rest strategy for multiband extension of CSP algorithm, which was explained in Section 4.3.3, can also be applied to the SCSSP method as follows. Consider the training phase of the SCSSP method, and let Ω_i' be the set of all motor-imagery tasks excluding the i^{th} task Ω_i. Starting from $i = 1$, the SCSSP method finds the bilinear transformation matrices $\mathbf{W}_L^{(i)}$ and $\mathbf{W}_R^{(i)}$ to extract d_{scssp} features that provide high discriminant power for classification of Ω_i versus Ω_i'. This procedure will be repeated for $i \in \{1, \cdots, C\}$, which results in a set of C spectral transformation matrices, and C spatial transformation matrices.

Now, consider the testing phase, and let $\mathbf{X} \in \mathbb{R}^{N_f \times N_{ch}}$ represent a test sample. The matrix $\mathbf{X}$ will be passed through C pairs of joint spatio-spectral transformation matrices, i.e., $\mathbf{T}_i = \{\mathbf{W}_L^{(i)}, \mathbf{W}_R^{(i)}\}$, $i \in \{1, \cdots, C\}$ to generate a set of $d_{scssp} * C$ features. The most discriminant features in this set consist of

the first pair of discriminant features obtained from each $\mathbf{T}_i$, which form a set of $2*C$ features. Similarly, the second pair of features from each $\mathbf{T}_i$ form the next $2*C$ discriminant features, and so on. Therefore, in the resulting feature vector, the first $2*C$ features will correspond to the most discriminant group of feature, and similarly the n^{th} groups of $2*C$ features represent the n^{th} most discriminant features.

5.5 Experimental Analysis

In this section we will study the performance of the proposed separable common spatio-spectral patterns (SCSSP) method and compere it with the conventional filterbank common spatial patterns (FBCSP) method, using Data set V from BCI competition III [147] and Data set 2a from BCI competition IV [148]. Similar to the experimental studies of Chapter 4, we will also study the effect of surface Laplacian (SL) filtering and channel selection on the performance of the SCSSP method.

Since the main focus in this chapter is on design of a domain-specific FE step, we will not consider any domain-agnostic FE after the SCSSP or FBCSP, and will directly pass the output of the domain-agnostic FE to the classifier. Recall that one of the main motivations behind the design of SCSSP method was to develop a DS-FE method that can effectively sort the extracted spatio-spectral features based on their discriminant power.

Therefore, the following processing steps will be considered for implementation of the SCSSP and FBCSP methods. First, the multichannel EEG signal will be passed through an optional stage of surface Laplacian (SL) filtering or channel selection. The resulting signal will then be passed through a bank of bandpass filters to generate the multiband EEG rhythms. At the next step either the SCSSP method or the FBCSP method will be applied to this multiband EEG data to extract a set of discriminant spatio-spectral features. These extracted features will then be directly passed to a classifier. The classifiers studied in this chapter are the simple linear (Lin) classifier and the naive Bayes Parzen window (NBPW) classifier, as defined in Section 3.3. This procedure results in a total of $16 = 2 \times 2 \times 2 \times 2$ different combinations for domain-specific FE and classification, namely SL(Yes/No), CS(Yes/No), FBCSP/SCSSP, and NBPW/Lin.

For comparative purposes, we will consider the performance of the FBCSP-MVLDA method from Chapter 4 as a benchmark, since it was shown to provide the best performance in different scenarios for both databases. However, in any comparison of the results of this chapter with the results of FBCSP-MVLDA method, it should be noted that after the optional SL/CS feature extraction, the FBCSP-MVLDA approach benefits from a two stage feature-extraction scheme, whereas the methods implemented in this chapter only extract the features in one step.

5.5.1 Experiment Setup

The motor-imagery experiments that are studied in this chapter are the same as the ones used in Chapter 4, which are explained in detail in Section 4.3.1. Thus, in this section we only provide a succinct recap of the main specifications of these experiments.

- BCI competition III, Data set V (Exp. 1): The goal of this experiment is to classify the following mental imagery tasks: left-hand movement (Ω_1), right-hand movement (Ω_2), and generation of words beginning with a random letter (Ω_3). The performance measure for this experiment is the correct classification rate (CCR)[1]. The dataset provided in this experiment contains four sessions. The last session can only be used as an unseen data for competition, and the first three sessions can be used for training and validation purposes. The training data in this experiment contains trials of length 15 seconds.

- BCI competition IV, Data set 2a (Exp. 2): The goal of this experiment is to classify the following motor-imagery tasks: left hand (Ω_1), right hand (Ω_2), both feet (Ω_3), and tongue (Ω_4) movement. The performance measure for this experiment is the kappa (κ) coefficient, defined as follows: $\kappa = (CCR - P_{rand})/(1 - P_{rand})$, where $P_{rand} = 0.25$ denotes the probability of random classification[2]. This data set contains only two sessions. The first session is used for training and validation purposes, and the second session will be used as unseen data for competition. The training session in this experiment contains trials of length 3 seconds.

The fist experiment is considered as a typical BCI application, where enough training data is recorded in multiple sessions, and includes long enough trials. In contrast, the second experiment is considered as an extreme scenario were the training trials are very short, and the algorithm only has access to one EEG recording session for training purposes. Although in most motor-imagery applications, the latter scenario will not occur, we have included this experiment to study the behaviour of the proposed SCSSP method in extreme cases. Finally, it should be mentioned that the surface Laplacian transformation in Exp. 2 is performed based on the approximate location of the EEG sensors that is provided in the data set, which affects the accuracy of the filtering output (ref. Section 4.3.1).

The bandpass filters used for both FBCSP and SCSSP algorithms in this chapter are the same as the ones designed in Section 4.3.2. For both experiments, Chebyshev Type-II filters with passband of width 4 Hz are utilized. A total of 6 filters are used in Exp. 1 to cover the frequency range of $8 - 32$

[1]The chance of random classification in Exp. 1 is $P_{rand} = 1/C = 0.33$, and the winning algorithm for this competition in the literature achieves a performance of %62.72 at the classifier output [149].

[2]The winning algorithm for this competition in the literature is the FBCSP-NBPW method.

Hz. Similarly, a total of 9 filters are used in Exp. 2 to cover the passband of $4-40$ Hz. These frequency ranges are selected based on the suggestions of the dataset providers and also the frequency ranges that the winning algorithm in each competition has considered, in order to provide a fair comparison with the previous works.

5.5.2 Cross-validation Results

The cross-validation schemes used in this chapter are the same as the ones explained in Chapter 4. In Exp. 1, we perform a three fold cross-validation to take advantage of the three distinct training sessions that are provided in this experiment. In Exp. 2, since only one training session is available, a 5×5 randomized cross-validation is performed. For each method, the optimal dimensionality of the feature space is determined based on the average performance of each subject over all the validation runs.

The validation results for Exp. 1 and Exp. 2 are presented in Table 5.1 and Table 5.2, respectively. In these tables, the results are presented in groups of size 5, in the following order: FBCSP-NBPW, FBCSP-Lin, FBCSP-MVLDA, SCSSP-NBPW, SCSSP-Lin. In each table, the first group of results corresponds to the case where no surface Laplacian (SL) or channel selection (CS) is applied to the data. Similarly, the next groups correspond to other possible combinations of surface Laplacian and channel selection. It should be mentioned that the results of FBCSP-NBPW, FBCSP-Lin, and FBCSP-MVLDA methods are the same as the ones reported in Chapter 4, and are presented here for comparison purposes.

The results of Table 5.1 and Table 5.2 are summarized in Figures 5.2(a) and 5.3(a), where the average validation performance over all the subjects, together with its corresponding standard error, are presented for every combination of domain-specific FE and classification. For more clarity, the results are categorized into four groups, based on whether or not surface Laplacian (SL) or channel selection (CS) are applied to the data. These figures show different trends for the performance of SCSSP method in Exp. 1 and Exp. 2.

For all combinations of surface Laplacian and channel selection in the first experiment, the SCSSP-Lin method always outperforms the FBCSP-Lin and FBCSP-NBPW methods, despite the fact that SCSSP has less computational cost compared to FBCSP method. Moreover, the SCSSP-Lin method even competes very closely with the FBCSP-MVLDA method which is a much more sophisticated algorithm and benefits from two consecutive stages of feature extraction. In the second experiment, however, the SCSSP method cannot compete with FBCSP-based methods. In order to describe this difference between the results of Exp. 1 and Exp. 2, it should be noted that the most important difference between these two experiments is the availability of the training data.

Recall from our discussions in Section 4.3.1 that Exp. 1 represents a typical BCI scenario where the training data is collected over multiple sessions and each training trial lasts for 15 seconds. In contrast, Exp. 2 represents the extreme case where only one training session is available and the training trials last for only 3 seconds. Considering these differences, the results of Figures 5.2 and 5.3 reveal that the performance improvement and computational cost efficiency of the SCSSP method are achieved at the cost of requiring more training samples, compared to the FBCSP algorithm.

Considering our discussions in Section 5.3 on the theoretical differences between the FBCSP and SCSSP methods, the higher sensitivity of SCSSP to the number of training samples can be explained as follows. The FBCSP only focuses on the diagonal block matrices of $\mathbf{\Sigma}_i{}'$ matrices, as defined in (5.11), whereas the SCSSP aims to provide an estimate of both diagonal and off-diagonal block matrices of $\mathbf{\Sigma}_i{}'$ matrices, as defined in (5.12). Therefore, when the number of training samples is extremely small, the SCSSP cannot reliably estimate the $\mathbf{\Sigma}_i{}'$, and consequently does not succeed in extracting discriminant features from the EEG data. The high performance of SCSSP method in Exp. 1 shows that the matrix-variate Gaussian model deployed by SCSSP algorithm, can very well describe the statistical characteristics of the EEG signals; however, reliable estimation of its parameters requires access to a large training set.

These results suggest that in order to benefit from the low computational cost and high performance of the SCSSP method, we need to provide this algorithm with enough training samples. As mentioned before, this condition is not restrictive in most motor-imagery BCI applications, since these BCIs are generally designed for longterm utilization, which guarantees access to large enough training sets. In such cases, the SCSSP method can reliably estimate the signal parameters, which allows for reducing the computational cost while improving the performance of the BCI system.

5.5.3 Test (Competition) Results

The performance results of different methods for the unseen competition data is presented in Table 5.3 and Table 5.4. These methods are categorized in groups of size five, depending on whether or not the surface Laplacian (SL) and channel selection are performed in the domain-specific FE stage, similar to Table 5.1 and Table 5.2. The feature space dimensionality for each method is set based on the value of d_{opt} in the validation phase. These results are summarized in Figures 5.2(b) and 5.3(b), where the average test performance over all the subjects are illustrated.

As mentioned in Section 4.3.5, the winning method in the literature for Exp. 1 uses a combination of surface Laplacian + channel selection + short-time Fourier transformation + LDA method and achieves

Table 5.1: Cross-validatoin performance results for different algorithms in Experiment-1. For each subject, the average correct classification rate (CCR) for the optimal dimension (d_{opt}) and its corresponding standard error are reported. The rightmost column represents the average performance over all the subjects. Note that the performance of a random classifier in this experiment is $\%CCR = \%33.3$.

DS-FE				**DA-FE**	**Clas.**	**Performance in the Cross-validation Phase**						
Spatial		**Spectral**	**Spatio-**			**Subject 1**		**Subject 2**		**Subject 3**		**Average**
SL	**CS**		**Spectral**			**%CCR**	d_{opt}	**%CCR**	d_{opt}	**%CCR**	d_{opt}	**%CCR**
No	No	BPF	FBCSP	-	NBPW	58.22±4.19	6*18*3	48.10±3.38	6*8*3	45.62±0.46	6*10*3	50.65±2.12
				-	Lin	63.25±1.99	6*12*3	50.42±3.37	6*12*3	48.73±1.00	6*10*3	54.13±2.03
				MVLDA	**Lin**	**69.50±2.81**	**2**	**53.94±0.99**	**2**	**46.75±1.20**	**27**	**56.73±1.56**
		BPF	SCSSP	-	NBPW	56.88±7.53	100*3	48.66±1.86	168*3	47.60±1.51	48*3	51.05±2.62
				-	**Lin**	**69.15±6.64**	**64*3**	**60.48±5.15**	**46*3**	**50.92±0.83**	**116*3**	**60.18±3.37**
	Yes	BPF	FBCSP	-	NBPW	60.64±2.56	6*6*3	46.62±3.27	6*8*3	45.83±2.31	6*6*3	50.97±1.76
				-	Lin	65.99±2.28	6*6*3	49.79±1.06	6*8*3	43.50±2.16	6*8*3	53.09±1.63
				MVLDA	**Lin**	**71.67±2.66**	**17**	**51.55±2.41**	**26**	**48.16±0.46**	**6**	**57.13±1.35**
		BPF	SCSSP	-	NBPW	61.21±5.45	48*3	48.59±6.65	48*3	45.34±1.16	46*3	51.71±3.68
				-	**Lin**	**67.82±4.20**	**38*3**	**49.37±5.57**	**48*3**	**50.00±3.24**	**32*3**	**55.72±4.15**
Yes	No	BPF	FBCSP	-	NBPW	58.71±4.01	6*18*3	48.10±2.86	6*14*3	46.40±1.71	6*16*3	51.07±2.11
				-	Lin	59.26±3.03	6*16*3	52.60±4.97	6*6*3	48.45±1.10	6*12*3	53.43±2.41
				MVLDA	**Lin**	**69.77±3.44**	**2**	**57.45±1.86**	**2**	**48.94±1.69**	**54**	**58.72±1.78**
		BPF	SCSSP	-	NBPW	59.06±7.44	174*3	52.25±2.14	44*3	49.65±2.63	106*3	53.65±2.24
				-	**Lin**	**68.24±7.23**	**192*3**	**61.32±4.58**	**44*3**	**51.55±0.42**	**100*3**	**60.37±3.45**
	Yes	BPF	FBCSP	-	NBPW	61.99±1.90	6*8*3	44.66±5.13	6*8*3	47.53± 2.20	6*8*3	51.39±0.83
				-	Lin	65.14±1.53	6*8*3	55.20±2.37	6*8*3	48.52±1.16	6*6*3	56.28±0.82
				MVLDA	**Lin**	**74.06±1.59**	**3**	**62.45±2.72**	**2**	**53.60±0.44**	**16**	**63.37±1.49**
		BPF	SCSSP	-	NBPW	59.32±3.32	44*3	45.64±4.93	48*3	49.93±2.08	38*3	51.63±2.20
				-	**Lin**	**69.71±3.41**	**46*3**	**52.46±3.77**	**48*3**	**54.31±2.22**	**38*3**	**58.82±2.63**

Table 5.2: Cross-validatoin performance results for different algorithms in Experiment-2. For each subject, the Kappa coefficient (κ) for the optimal dimension (d_{opt}) and its corresponding standard error are reported. The rightmost column represents the average performance over all the subjects. Note that the performance of a random classifier in this experiment is $\kappa = 0$.

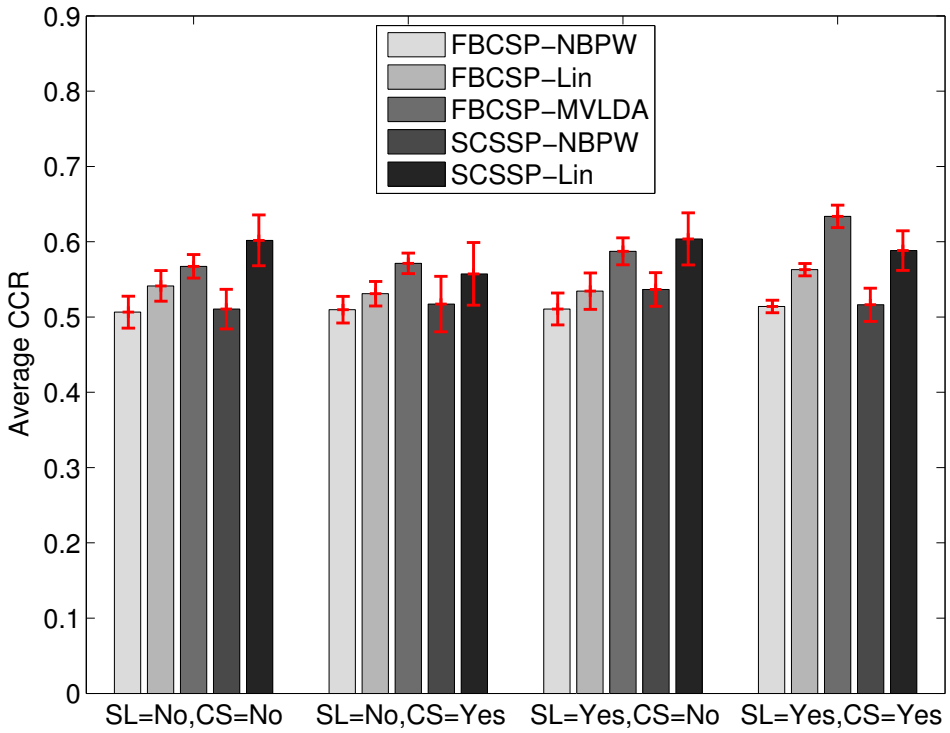
DS-FE						Performance in the Validation Stage (κ and d_{opt})									
Spatial		Spectral	Spatio-Spectral	DA-FE	Clas.										
SL	CS					Subj. 1	Subj. 2	Subj. 3	Subj. 4	Subj. 5	Subj. 6	Subj. 7	Subj. 8	Subj. 9	Average
No	No	BPF	FBCSP	-	NBPW	67.88±2.87 9*10*4	42.18±3.35 9*2*4	77.87±2.78 9*4*4	51.77±3.73 9*6*4	50.17±4.20 9*2*4	45.97±3.13 9*18*4	87.50±2.30 9*2*4	85.79±2.03 9*12*4	76.31±2.86 9*4*4	65.05±1.12
				-	Lin	68.76±2.56 9*8*4	29.21±3.02 9*12*4	72.07±2.67 9*2*4	49.24±2.45 9*20*4	52.54±3.82 9*2*4	48.91±3.26 9*2*4	84.63±2.40 9*6*4	85.96±2.41 9*8*4	76.46±2.89 9*2*4	63.09±0.90
				MVLDA	Lin	80.71±2.25 32	41.65±3.53 44	78.03±3.18 101	47.24±2.63 61	54.89±3.40 148	45.81±3.79 242	90.27±1.95 143	88.85±2.17 158	76.92±3.14 42	67.15±0.86
		BPF	SCSSP	-	NBPW	59.89±3.18 9*54*4	28.25±3.05 9*154*4	73.02±3.37 9*22*4	32.37±2.74 9*100*4	35.27±3.85 9*164*4	38.11±4.63 9*168*4	74.43±2.84 9*10*4	80.20±2.53 9*54*4	68.54±2.60 9*2*4	54.45±1.10
				-	Lin	64.98±3.40 184*4	25.65±3.97 172*4	72.44±2.89 10*4	32.16±2.90 164*4	34.43±3.66 48*4	36.82±3.06 184*4	73.68±3.51 16*4	85.60±2.62 8*4	72.28±2.63 2*4	55.34±0.92
	Yes	BPF	FBCSP	-	NBPW	65.69±3.06 9*4*4	42.32±2.36 9*2*4	80.37±2.41 9*2*4	40.12±3.71 9*2*4	39.04±3.59 9*6*4	50.58±3.57 9*10*4	78.57±2.84 9*8*4	77.67±3.16 9*4*4	76.60±3.12 9*4*4	61.22±1.15
				-	Lin	66.96±2.99 9*4*4	33.32±3.14 9*6*4	80.89±2.69 9*2*4	49.14±2.47 9*6*4	42.29±3.30 9*4*4	46.30±3.79 9*6*4	73.01±2.90 9*8*4	83.04±2.48 9*12*4	76.17±3.09 9*2*4	61.24±0.98
				MVLDA	Lin	75.42±2.47 16	38.39±3.37 20	80.01±2.65 157	44.29±2.50 80	35.03±4.20 90	53.43±4.47 96	79.85±2.78 19	85.64±2.62 11	75.89±3.11 12	63.11±1.14
		BPF	SCSSP	-	NBPW	53.40±2.54 9*42*4	28.74±4.46 9*116*4	80.32±2.40 9*12*4	36.77±3.00 9*18*4	19.59±4.04 9*106*4	35.03±4.33 9*62*4	60.95±3.33 9*46*4	75.81±2.75 9*82*4	69.79±3.25 9*40*4	51.15±1.36
				-	Lin	48.16±3.97 116*4	25.57±3.79 94*4	83.98±2.30 14*4	30.36±2.74 8*4	24.22±3.72 88*4	37.47±3.24 30*4	63.12±2.99 66*4	83.98±2.86 10*4	70.45±2.61 4*4	51.92±0.97
Yes	No	BPF	FBCSP	-	NBPW	71.45±1.65 9*4*4	40.10±1.78 9*4*4	71.00±1.72 9*4*4	39.30±2.05 9*4*4	57.56±1.86 9*2*4	30.37±2.26 9*6*4	81.43±1.23 9*2*4	77.98±0.94 9*4*4	73.93±1.71 9*6*4	60.35±0.62
				-	Lin	70.54±4.41 9*2*4	35.11±1.71 9*4*4	69.43±1.57 9*6*4	41.93±2.41 9*4*4	40.79±7.40 9*2*4	31.71±2.82 9*10*4	79.30±1.35 9*8*4	79.93±1.00 9*10*4	69.79±1.94 9*6*4	57.61±1.15
				MVLDA	Lin	54.39±8.97 5	37.88±1.67 112	71.32±4.51 41	31.82±4.99 136	24.03±9.03 126	28.17±3.82 154	74.01±6.64 195	68.90±7.86 243	64.46±7.36 5	50.55±2.42
		BPF	SCSSP	-	NBPW	63.48±1.50 9*48*4	35.27±1.59 9*122*4	65.12±1.68 9*26*4	35.21±1.89 9*188*4	31.78±2.21 9*32*4	25.84±1.94 9*140*4	68.78±1.59 9*32*4	72.28±1.13 9*40*4	67.68±1.66 9*58*4	51.72±0.65
				-	Lin	65.50±1.59 192*4	31.53±1.63 198*4	65.45±1.23 22*4	35.50±2.11 194*4	38.70±2.14 36*4	27.96±1.76 196*4	69.86±1.50 46*4	75.87±1.33 104*4	65.19±1.80 20*4	52.84±0.68
	Yes	BPF	FBCSP	-	NBPW	74.26±1.56 9*6*4	40.09±1.30 9*2*4	74.26±1.38 9*2*4	43.01±2.22 9*6*4	61.68±1.64 9*2*4	35.86±1.94 9*6*4	85.74±0.99 9*2*4	81.30±1.14 9*4*4	75.35±1.52 9*2*4	63.51±0.57
				-	Lin	76.37±1.52 9*4*4	39.70±1.63 9*6*4	78.13±1.26 9*4*4	46.67±2.22 9*6*4	59.49±1.72 9*4*4	35.39±2.29 9*6*4	84.74±1.50 9*6*4	82.38±1.14 9*4*4	71.56±1.77 9*2*4	63.82±0.49
				MVLDA	Lin	80.23±1.10 76	38.90±1.70 26	78.45±1.43 82	45.91±1.78 128	62.45±1.45 90	38.30±1.83 63	87.08±1.05 91	83.13±1.05 29	78.53±1.70 4	65.89±0.51
		BPF	SCSSP	-	NBPW	66.11±1.68 9*30*4	36.94±1.68 9*116*4	69.22±1.44 9*2*4	35.98±1.83 9*92*4	33.49±1.75 9*12*4	30.77±2.33 9*30*4	74.45±1.77 9*24*4	74.20±1.08 9*22*4	70.29±2.04 9*56*4	54.60±0.60
				-	Lin	67.27±1.44 114*4	35.93±1.67 112*4	68.28±1.40 22*4	39.13±2.05 106*4	41.76±1.73 112*4	30.81±2.04 96*4	74.39±1.38 116*4	77.16±1.04 10*4	67.21±1.57 20*4	55.77±0.51

Table 5.3: Test performance results for different algorithms in Experiment-1. The correct classification rate (CCR) for each subject and the total average over all the subjects are reported. Note that the performance of a random classifier in this experiment is $\%CCR = \%33.3$.

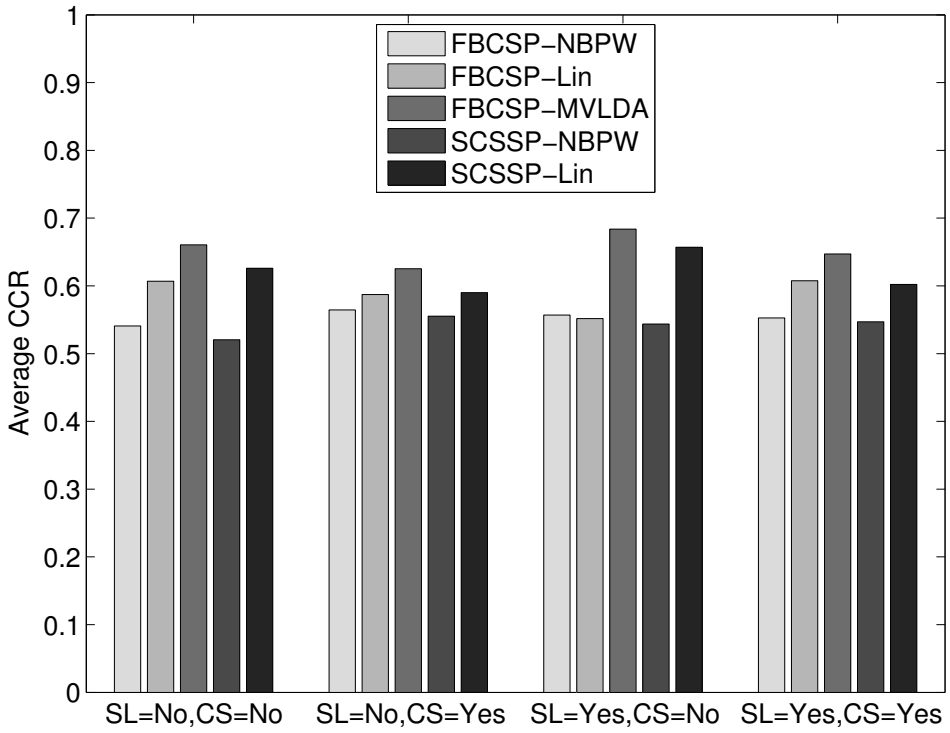
DS-FE									
Spatial		**Spectral**	**Spatio-**	**DA-FE**	**Classifier**	**Performance in the Test Stage (%CCR)**			
SL	**CS**		**Spectral**			**Subj. 1**	**Subj. 2**	**Subj. 3**	**Average**
No	No	BPF	FBCSP	-	NBPW	68.09	54.85	39.29	54.07
				-	Lin	71.06	62.66	48.32	60.68
				MVLDA	**Lin**	**80.64**	**68.99**	**48.53**	**66.05**
		BPF	SCSSP	-	NBPW	68.51	54.64	32.98	52.05
				-	**Lin**	**79.36**	**62.87**	**45.59**	**62.61**
	Yes	BPF	FBCSP	-	NBPW	64.04	55.91	49.37	56.44
				-	Lin	74.04	53.38	48.74	58.72
				MVLDA	**Lin**	**79.57**	**55.91**	**52.10**	**62.53**
		BPF	SCSSP	-	NBPW	65.32	53.80	47.48	55.53
				-	**Lin**	**75.53**	**51.05**	**50.42**	**59.00**
Yes	No	BPF	FBCSP	-	NBPW	66.17	60.55	40.34	55.68
				-	Lin	66.60	50.00	48.95	55.18
				MVLDA	**Lin**	**81.70**	**70.04**	**53.36**	**68.37**
		BPF	SCSSP	-	NBPW	69.36	54.64	39.08	54.36
				-	**Lin**	**80.43**	**67.72**	**48.95**	**65.70**
	Yes	BPF	FBCSP	-	NBPW	69.79	50.42	45.59	55.27
				-	Lin	72.77	60.97	48.53	60.76
				MVLDA	**Lin**	**79.15**	**60.97**	**53.99**	**64.70**
		BPF	SCSSP	-	NBPW	65.53	52.53	46.01	54.69
				-	**Lin**	**77.66**	**56.54**	**46.43**	**60.21**

Table 5.4: Test performance results for different algorithms in Experiment-1. The correct Kappa coefficient (κ)) for each subject and the total average over all the subjects are reported. Note that the performance of a random classifier in this experiment is $\kappa = 0$.

DS-FE															
Spatial		**Spectral**	**Spatio-Spectral**	**DA-FE**	**Clas.**	**Performance in the Test Stage (κ)**									
SL	**CS**					**Subj. 1**	**Subj. 2**	**Subj. 3**	**Subj. 4**	**Subj. 5**	**Subj. 6**	**Subj. 7**	**Subj. 8**	**Subj. 9**	**Average**
No	No	BPF	FBCSP	-	NBPW	63.67	31.20	60.10	36.73	23.33	19.43	64.20	57.07	61.39	46.35
				-	Lin	68.07	30.25	70.68	39.34	28.42	25.01	56.89	59.13	54.06	47.98
				MVLDA	**Lin**	65.94	23.84	68.41	39.54	41.77	21.00	68.18	63.43	57.76	49.98
		BPF	SCSSP	-	NBPW	50.12	21.03	62.16	31.40	0.00	12.13	33.35	53.31	63.30	36.31
				-	Lin	62.25	26.54	65.76	26.91	6.64	25.56	40.65	59.32	65.86	42.17
	Yes	BPF	FBCSP	-	NBPW	60.72	25.06	56.65	34.97	9.61	18.10	55.56	53.92	53.33	40.88
				-	Lin	64.46	26.79	60.81	39.07	11.19	24.63	66.53	53.97	54.84	44.70
				MVLDA	**Lin**	70.28	23.27	61.20	33.23	7.68	22.55	61.57	60.97	51.45	43.58
		BPF	SCSSP	-	NBPW	59.42	17.34	57.17	14.83	-1.95	8.77	28.63	52.13	51.78	32.01
				-	Lin	59.16	24.64	56.83	15.13	5.30	21.14	34.16	58.50	55.90	36.75
Yes	No	BPF	FBCSP	-	NBPW	64.11	32.17	65.95	43.33	27.37	22.15	64.58	31.93	56.23	45.31
				-	Lin	69.31	36.63	69.52	48.49	14.20	27.82	61.17	-24.06	57.17	40.03
				MVLDA	**Lin**	74.03	37.94	70.10	45.82	25.03	23.61	71.23	-25.65	54.76	41.88
		BPF	SCSSP	-	NBPW	51.19	19.85	52.04	26.60	14.58	18.89	38.41	42.56	48.62	34.75
				-	Lin	58.12	27.69	62.55	32.57	4.79	21.91	26.52	54.74	51.87	37.86
	Yes	BPF	FBCSP	-	NBPW	64.87	41.45	60.56	40.46	21.82	19.85	65.74	54.56	50.38	46.63
				-	Lin	67.55	40.17	66.60	43.54	23.69	25.84	62.14	53.12	49.05	47.97
				MVLDA	**Lin**	71.45	35.81	69.78	41.32	34.54	26.03	60.08	60.51	50.20	49.97
		BPF	SCSSP	-	NBPW	57.19	25.07	59.32	26.10	8.52	22.53	13.28	50.87	51.81	34.97
				-	Lin	59.09	32.34	58.72	25.41	5.27	23.65	23.72	59.48	54.99	38.07

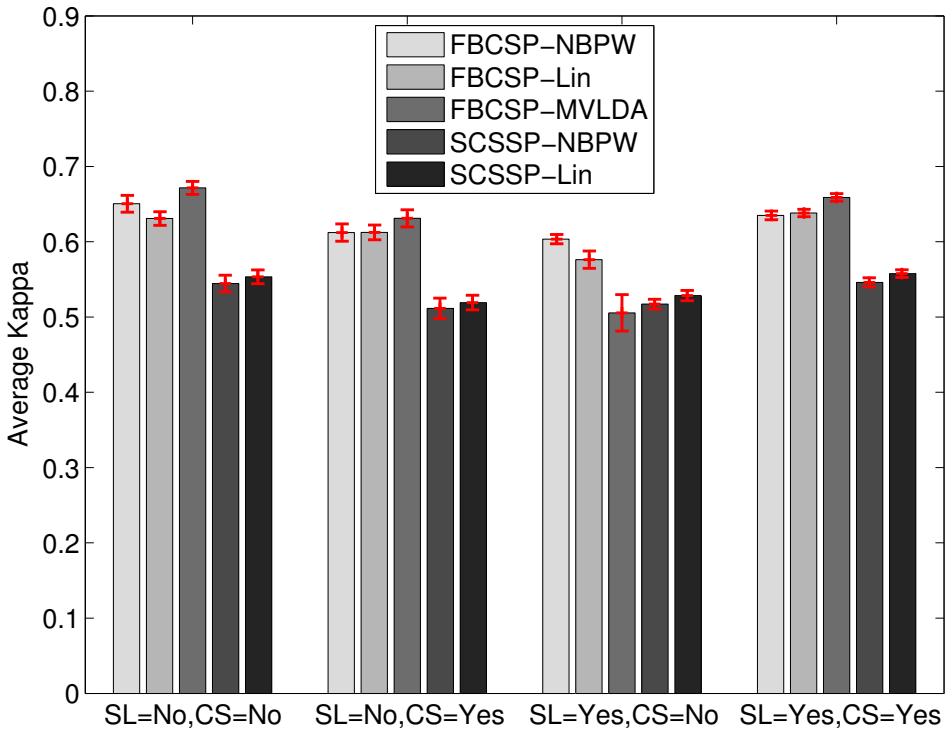


(a) Average performance over all the subjects in validation phase of Experiment-1.

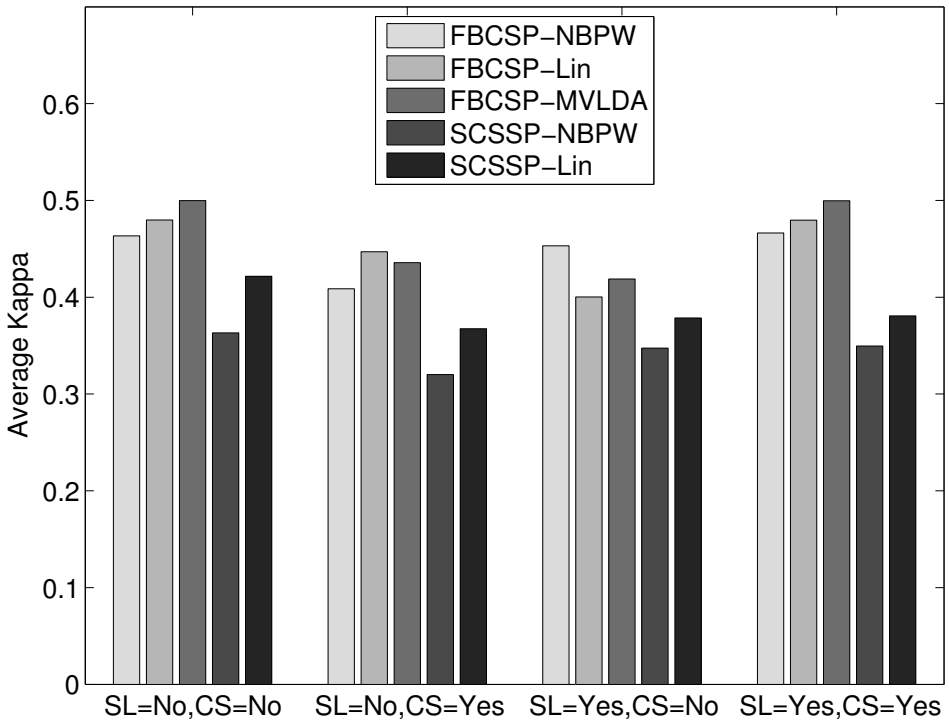


(b) Average performance over all the subjects in test phase of Experiment-1.

Figure 5.2: Comparison of the performance results for SCSSP-based and FBCSP-based solutions in (a) Validation phase of Experiment-1 and (b) Testing phase of Experiment-1. For validation results, the average performance of each method over all the subjects and all validation runs, together with its corresponding standard error, is plotted. For more clarity, the results are illustrated in four groups, depending on whether or not the surface Laplacian (SL) and channel selection (CS) are applied in the domain-specific feature extraction step. Note that the performance measure in Experiment-1 is the Correct Classification Rate (CCR), and a random classifier results in $\%CCR = \%33.3$.



(a) Average performance over all the subjects in validation phase of Experiment-2.



(b) Average performance over all the subjects in test phase of Experiment-2.

Figure 5.3: Comparison of the performance results for SCSSP-based and FBCSP-based solutions in (a) Validation phase of Experiment-2 and (b) Testing phase of Experiment-2. For validation results, the average performance of each method over all the subjects and all validation runs, together with its corresponding standard error, is plotted. For more clarity, the results are illustrated in four groups, depending on whether or not the surface Laplacian (SL) and channel selection (CS) are applied in the domain-specific feature extraction step. Note that the performance measure in Experiment-2 is the Kappa coefficient (κ), and a random classifier results in $\kappa = 0$.

an average performance of 62.72%, and the winning method in the literature for Exp. 2 is the FBCSP-NBPW approach without surface laplacian or channel selection.

The performance results on the test data shows a trend very similar to the performance results during the cross-validation phase. It can be seen that in the first experiment, the SCSSP-Lin method outperforms both FBCSP-Lin and FBCSP-NBPW methods, and exhibits a performance very close to the FBCSP-MVLDA method which has a two stage feature extraction scheme. In the second experiment, the SCSSP method cannot compete with other methods due to the lack of access to sufficient training information for reliable estimation of the model parameters.

5.5.4 The Effect of Surface Laplacian Filtering and Channel Selection

Comparison of Figures 5.2 and 5.3 shows that combination of surface Laplacian filtering with the SCSSP method slightly improves the classification performance in Exp. 1, but has adverse effect on the performance in Exp. 2. This difference in the trends is due to the approximate calculation of the surface Laplacian in the second experiment, which in turn is caused by the fact the accurate sensor locations are not available in Exp. 2.

Recall from our discussions in Chapter 4 that channel selection is mostly efficient when combined with the surface Laplacian filtering, even in the case of approximate surface Laplacian in Exp. 2. Therefore, let us compare the combined effect of channel selection and surface Laplacian on SCSSP-Lin and FBCSP-Lin methods in Figures 5.2 and 5.3. It can be seen that the FBCSP-Lin method achieves its highest performance when both surface Laplacian and channel selection are deployed, whereas the SCSSP-Lin achieves its highest performance when it is applied to the raw data, with only one exception which is the test phase of Exp. 1. In case of the FBCSP-Lin method, the combination of surface Laplacian and channel selection helps to manually reduce the dimensionality of the space in which the *spatial covariances* $\mathbf{\Psi}_i^f$ are calculated, without loosing the information relevant to the motor cortex area. This dimensionality reduction improves the accuracy of the spatial covariance estimation for each band, which in turn improves the performance of the system.

In case of the SCSSP method, however, it not necessarily desired to reduce the dimensionality of the data in the *spatial domain* while having the same dimensionality in the *spectral domain*. The main reason for this effect is as follows. The SCSSP method only calculates one common spatial covariance matrix for all the bands. As a result, SCSSP treats different rows of the matrix $\mathbf{X} \in \mathbb{R}^{N_f \times N_{ch}}$ as extra training samples for calculation of the covariance matrix. In other words, SCSSP method has access to $N_f * N_i$ training samples[3] for estimation of $\mathbf{\Psi}_i$, whereas FBCSP has only access to N_i samples for

[3]Here, N_i is the number of training matrices for class Ω_i

estimation of each $\mathbf{\Psi}_i^f$. On the other hand, SCSSP requires to estimate the common *spectral covariance* matrix $\mathbf{\Phi}_i$ by treating different columns of $\mathbf{X}$ as extra samples training samples, which leads to a total of $N_{ch} * N_i$ samples. As a consequence, any reduction in the number of EEG channels results in a significant reduction in the number of training samples for $\mathbf{\Phi}_i$.

In other words, in SCSSP method, the channel selection results in higher accuracy for spatial covariance estimation at the cost of reducing the accuracy for spectral covariance estimation. The results in Figures 5.2 and 5.3 suggest that these two opposite effects almost cancel out each other and there is marginal change in the performance of the SCSSP-based methods when channel selection and surface Laplacian are utilized together with SCSSP, as opposed to when SCSSP is directly applied to the raw data. Note that as long as channel selection does not deteriorate the overall performance of the system, it might still be beneficial since it reduces the computational cost of the feature extraction.

5.5.5 The Effect of Feature Space Dimensionality

In this section, we study the effect of number of extracted features on the performance of SCSSP-NBPW and SCSSP-Lin algorithms, and compare them with the FBCSP-NBPW and FBCSP-Lin solutions. The results for the first subjects in Exp. 1 and Exp. 2 are shown in Figure 5.4. The results for the rest of subjects in these two experiments are presented in Figures 5.5 and 5.6, respectively. The results in these three figures, correspond to the case where no surface Laplacian (SL) filtering or channel selection (CS) is applied to the EEG data. The effect of SL and CS will be studied later in this section.

Note that in all these four methods, no domain-agnostic FE has been used, and a total number of d most significant features are directly passed to the classifier. In contrast, the FBCSP-MVLDA method deploys a domain-agnostic FE scheme, which takes all the extracted spatio-spectral features, and further reduces the dimensionality of the feature space prior to classification. Therefore, if we want to present the results of FBCSP-MVLDA method in the same graph as the other four methods, it would correspond to only one value of $d = N_f * N_{ch} * C$, which is the maximum dimensionality of feature space for FBCSP-based solutions. The performance of FBCSP-MVLDA at this point will then depend on the number of features that are extracted by the MVLDA algorithm. Therefore, it is not meaningful to represent the performance of FBCSP-MVLDA versus d, which represents the the number of features that are extracted at the domain-specific FE. Thus, in order to provide the reader with a measure to compare the performance of FBCSP-MVLDA with the other four methods, we have marked the vertical access with point "A" and a red dashed line, which represents the optimal performance of FBCSP-MVLDA. However, before any comparison between this method and the other four methods, it should be noted

that FBCSP-MVLDA benefits from a two stage feature extraction scheme.

The results of Exp. 1 in Figure 5.4(a) and Figure 5.5 show that the SCSSP-Lin method outperforms all other methods (including the FBCSP-MVLDA) for most values of d. In contrast, the SCSSP-NBPW method has much lower performance and closely competes with the FBCSP-NBPW method. Note that the performance of SCSSP-Lin method peaks at a relatively low dimension, which shows that SCSSP has been able to capture the discriminant information of the data in a small number of features. In Exp. 2, where the training set is extremely limited, the SCSSP-based methods cannot compete with the FBCSP-based methods for most of the subjects. This significant difference between the two experiments is mostly due to the lack of training data in Exp. 2, as discussed in Section 5.5.2.

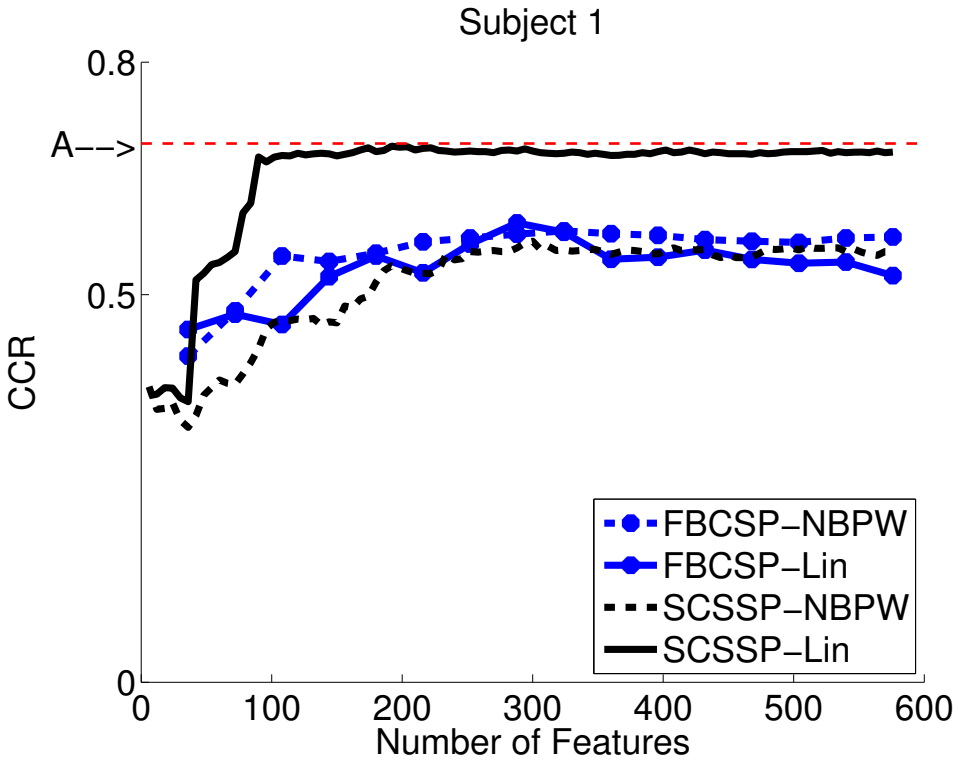
Finally, Figures 5.7 and 5.8 illustrate the effect of feature space dimensionality on the performance of SCSSP-NBPW and SCSSP-Lin methods, when they are utilized with different combinations of surface Laplacian (SL) and channel selection (CS). In these figures, the results for one high performing subject and one low performing subject are presented for each experiment. The results for other subjects show similar trends.

In case of the SCSSP-NBPW method in Figure 5.7, the surface laplacian filter is beneficial only in Exp. 1, in which case it improves the classification rate for a wide range of d values. The effect of channel selection on the raw EEG data is not consistent, however, when surface laplacian is applied to the data the channel selection always improves the overall performance of the system (compare the dashed red lines with the solid green lines). This trend confirms our previous discussions regarding the fact that surface Laplacian filtering pre-emphasizes the localized data, and hence is highly suggested in combination with the channel selection.

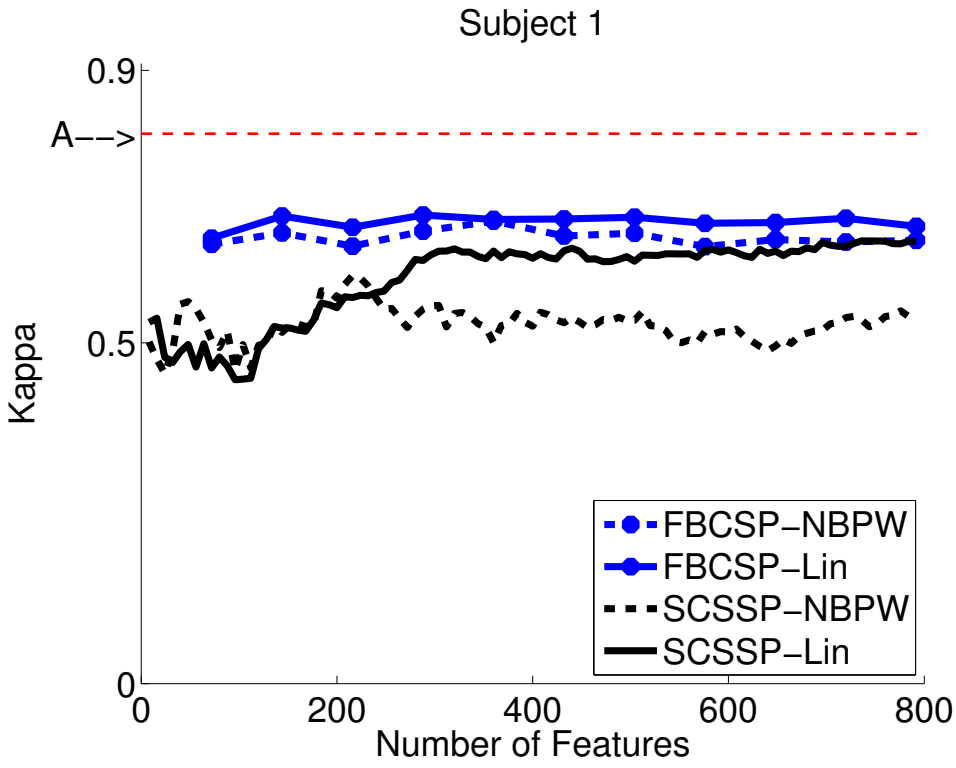
In case of the SCSSP-Lin method in Figure 5.8, the results from first experiment show that over most values of d, the surface laplacian has marginal effect on the overall performance unless it is combined with channel selection. In the second experiment, where small number of training samples are available, the surface laplacian and its combination with channel selection are very beneficial for improving the overall performance for low performing subjects, whereas they are not helpful in cases where the SCSSP is already achieving a high performance on the raw data. This trend is very similar to the trend for FBCSP-MVLDA discussed in Section 4.3.9.

5.6 Summary and Concluding Remarks

In this chapter, a new domain-specific FE method was proposed based on a heteroscedastic matrix-variate Gaussian model for the multiband EEG rhythms. In the proposed approach, the EEG signal is



(a) Correct Classification Rate (CCR) results for Subject 1 in Experiment-1.



(b) Kappa coefficient (κ) results for Subject 1 in Experiment-2.

Figure 5.4: Validation performance for SCSSP-based and FBCSP-based methods versus the number of features extracted by the domain-specific feature extraction method. in the validation phase for the first subject in (a) Experiment-1 and (b) Experiment-2. For comparison purposes, the performance of the FBCSP-MVLDA method is also marked on the vertical access by "A".

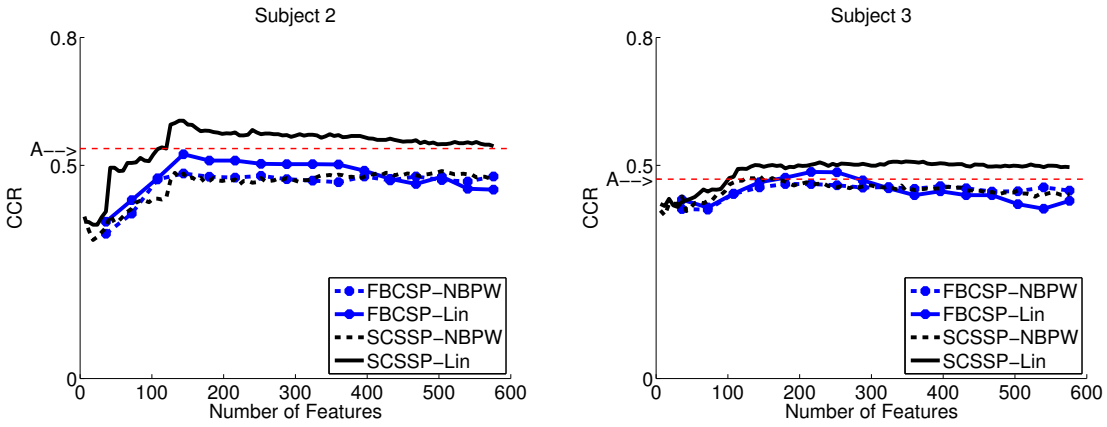


Figure 5.5: Correct Classification Rate (CCR) for different methods versus the number of features for all the subjects in the validation phase of Experiment-1. The illustrated results are for the case where both surface laplacian filtering and channel selection have been performed on the data. For comparison purposes, the performance of the FBCSP-MVLDA method is also marked on the vertical access by "A".

first passed through a bank of bandpass filters to extract different bands of EEG rhythms. The resulting signal is then passed through a joint spatio-spectral FE method, called separable common spatio-spectral patterns (SCSSP), which directly operates on the matrix-variate data.

The main advantage of the SCSSP method compared to the FBCSP algorithm is the fact that SCSSP jointly processes the data in both spectral and spatial domains, and hence can sort the extracted features across both domains; whereas FBCSP cannot sort the features that are extracted from different frequency bands. As a result, the SCSSP method does not need to be followed by a domain-agnostic FE stage, and its output can directly be passed to the classifier. The second advantage of SCSSP is its relatively low computational cost. The SCSSP involves only two generalized eigen decompositions (i.e., one for spectral covariances and one for spatial covariances); whereas the FBCSP requires a total of N_f generalized eigen decompositions (i.e., one for each frequency band).

The above advantages come at the cost that the SCSSP method requires a relatively larger training set, compared to the FBCSP method. The performance comparison of these two methods shows that in Exp. 1, the SCSSP-Lin method outperforms not only the FBCSP-Lin method, but also the FBCSP-MVLDA method that benefits from a two stage FE strategy. However, in Exp. 2 that the amount of training information is extremely limited, SCSSP cannot compete with FBCSP method. It is worth mentioning that the conditions in the second experiment does not typically happen in most motor-imagery BCI systems, and it is mostly considered here to study the behaviour of the SCSSP method in extreme scenarios. Since the motor-imagery BCIs are generally designed for longterm utilization by the user, it is a fair assumption that the BCI algorithm will have access to long enough training dataset.

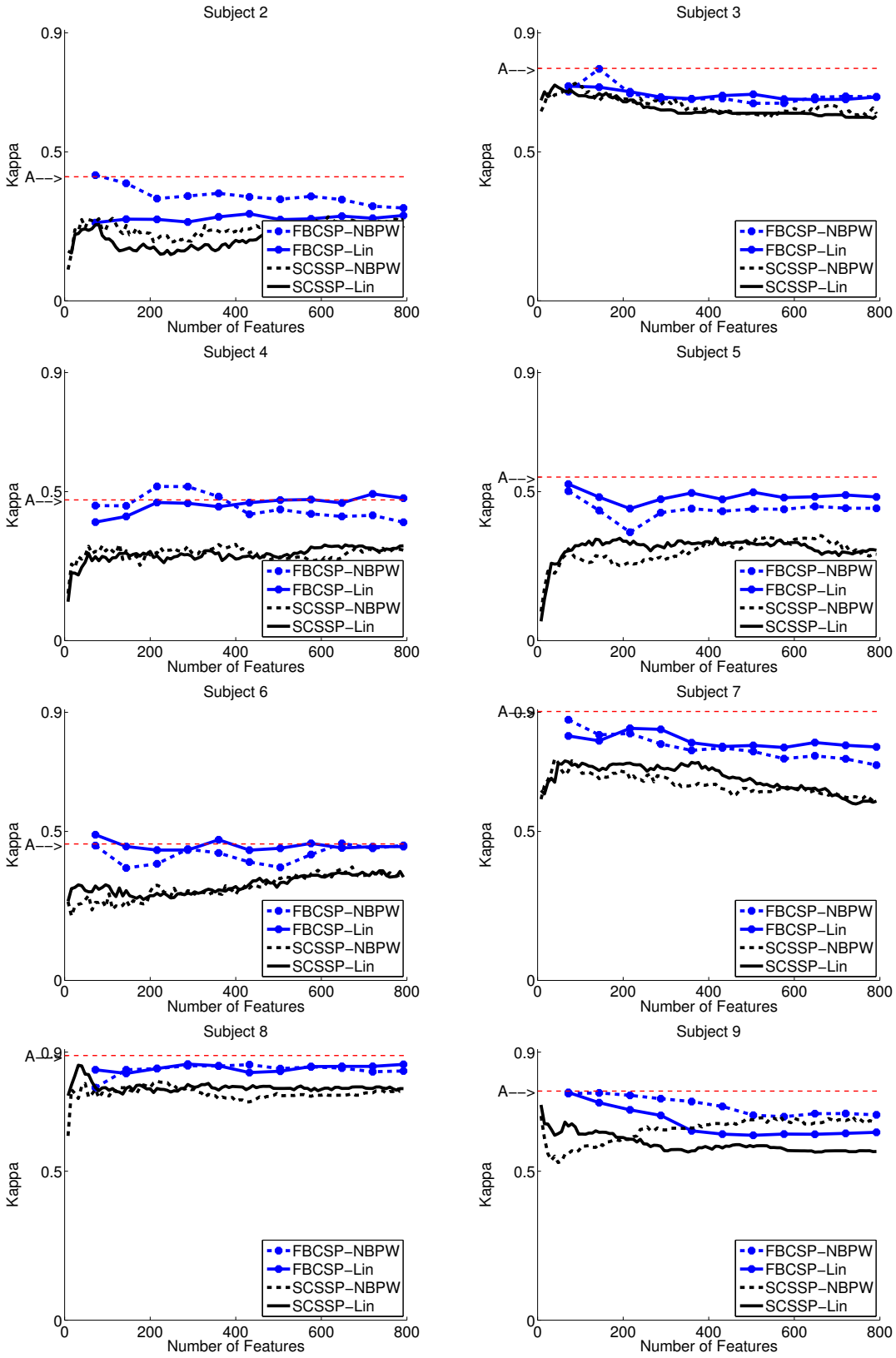
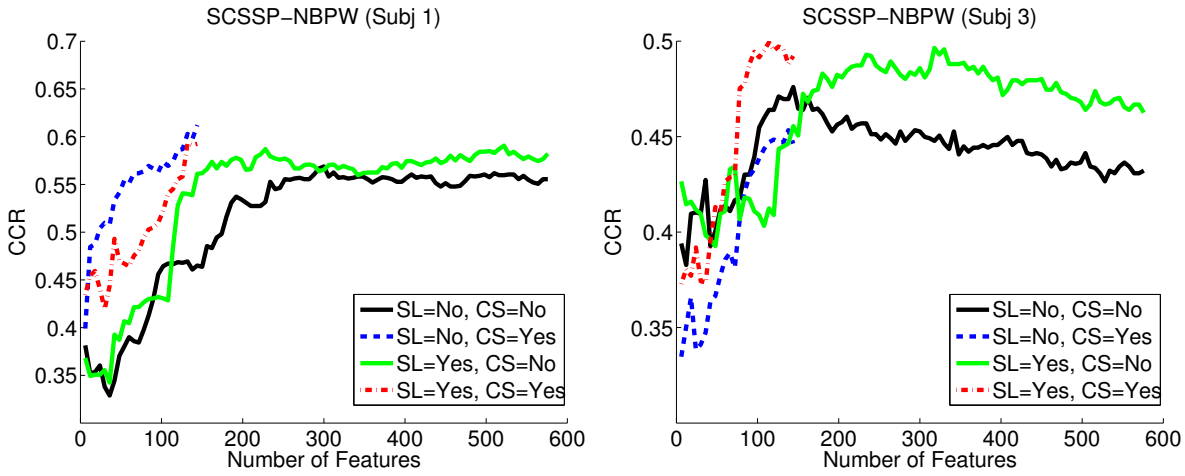
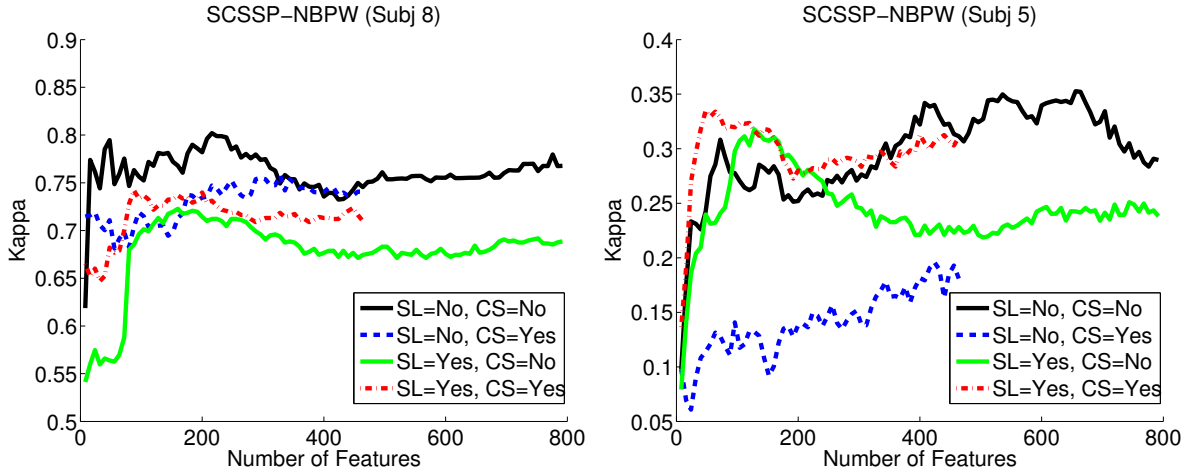


Figure 5.6: Kappa coefficient (κ) for different methods versus the number of features for all the subjects in the validation phase of Experiment-2. All these methods are applied to the raw EEG data, i.e., No surface laplacian or channel selection has been performed. Note that For comparison purposes, the performance of the FBCSP-MVLDA method is also marked on the vertical access by "A".

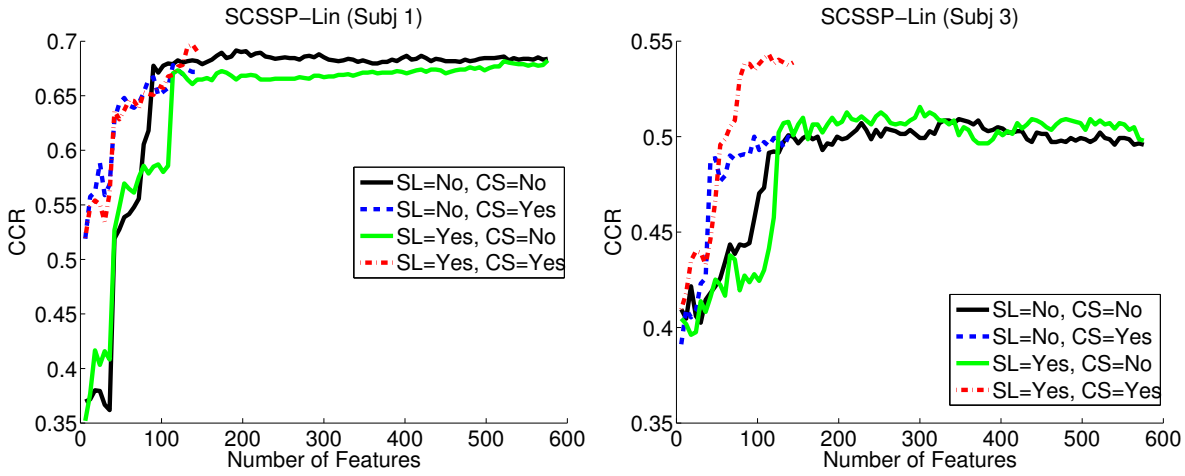


(a) Correct Classification Rate (CCR) Results, Experiment-1

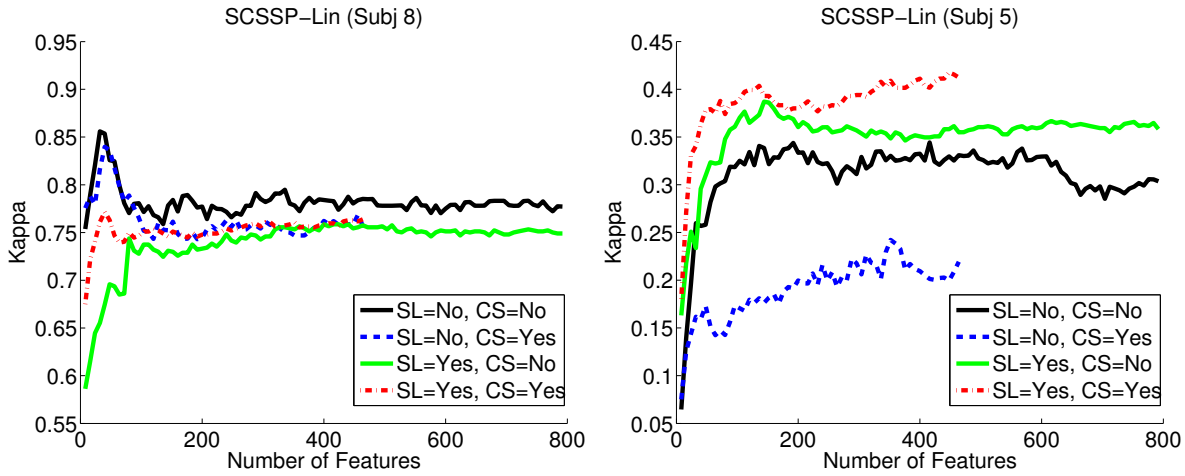


(b) Kappa coefficient (κ) Results, Experiment-2

Figure 5.7: The effect of surface Laplacian (SL) filtering and channel selection (CS) on the performance of SCSSP-NBPW method versus the number of extracted features in (a) Experiment-1 and (b) Experiment-2. For brevity, only the results of two subjects from each experiment are presented to illustrate the general trends in one high-performing subject and one low-performing subject.



(a) Correct Classification Rate (CCR) Results, Experiment-1



(b) Kappa coefficient (κ) Results, Experiment-2

Figure 5.8: The effect of surface Laplacian (SL) filtering and channel selection (CS) on the performance of SCSSP-Lin method versus the number of extracted features in (a) Experiment-1 and (b) Experiment-2. For brevity, only the results of two subjects from each experiment are presented to illustrate the general trends in one high-performing subject and one low-performing subject.

Chapter 6

Matrix-Variate Complex Gaussian Model for Spatio-Spectral Features Obtained Through Fourier Transformation

In the previous chapters, we studied the possibility of using the real-valued matrix-variate Gaussian model for designing various spatio-spectral feature extractors based on passing the EEG signal through a bank of bandpass filters and one or multiple CSP modules. As discussed in Chapter 3, there are several alternative approaches for extraction of spatio-spectral features. One of the most successful methods is the Fourier transformation of the data (ref. Section 3.1.2). A major difference between the features obtained through Fourier transformation with the features obtained through bandpass filtering is the complex-valued nature of the resulting features in the former case. As a result, the real-valued matrix-variate model presented in Section 3.4 is not directly applicable to such features.

One possible solution is to only consider the magnitude (or power) of the Fourier components and ignore their phase. This solution might be suitable if the phase of the EEG signal does not convey any information. However, several recent studies in neuroscience have revealed that there exist relevant information carried in the phase of electrical activities of the brain, both in microscopic level (the phase of neural firings) and in macroscopic level (the phase of EEG signals) [68–71]. Furthermore, recent studies on EEG source separation algorithms using independent component analysis (ICA) method have

shown that utilization of the complex-valued EEG spectrum, instead of power spectrum, significantly improves the performance of ICA algorithm [155].

It should be noted that most of the studies on the properties of the EEG signals in the Fourier domain are based on analysis of the power spectral density of EEG, which does not convey any information regarding the phase of the signal. Therefore, in this chapter we will specifically focus on the analysis of the complex valued spatio-spectral features obtained from Fourier transformation of the data. Motivated by the results of the previous two chapters, we are interested in studying the implications of separability and Gaussianity for these complex-valued features. To the best of our knowledge, there exists no theoretical work on this topic in the literature.

6.1 Complex-Valued Spectral Representation of EEG Data

Consider a multichannel EEG signal recorded during a trial while the subject is performing task Ω_i, as shown in Figure 6.1. Let $s(t, c|\Omega_i)$ denote the EEG signal recorded at time t from channel c. The $s(t, c|\Omega_i)$ notation is used in this chapter to emphasize the fact that the recorded EEG signal is a two dimensional stochastic signal whose statistical characteristics depend on the mental task Ω_i.

In order to obtain the frequency domain representation of this EEG signal at each time instant, an STFT is applied on each channel of the data. There exist two commonly used definitions for the STFT in the literature:

$$z(t, f, c|\Omega_i) = \int_{-\infty}^{\infty} s(\tau, c|\Omega_i) w(\tau - t) e^{-j2\pi f \tau} d\tau \tag{6.1}$$

$$\begin{aligned} z'(t, f, c|\Omega_i) &= \int_{-\infty}^{\infty} s(\tau, c|\Omega_i) w(\tau - t) e^{-j2\pi f(\tau - t)} d\tau \\ &= \int_{-\infty}^{\infty} s(\tau + t, c|\Omega_i) w(\tau) e^{-j2\pi f \tau} d\tau, \end{aligned} \tag{6.2}$$

where $w(t)$ is a window of length T_w. The definition given in Eq.(6.2) is a shift-invariant version of STFT which is more convenient for implementation. Eq.(6.2) can be implemented through applying a fixed windowed Fourier transformation to time-shifted versions of the EEG signal. However, it introduces a linear phase component to the original spectral representation given by Eq.(6.1), i.e.,

$$z'(t, f, c|\Omega_i) = z(t, f, c|\Omega_i) * e^{j2\pi f t}. \tag{6.3}$$

This phase shift conveys information about the amount of time elapsed since the starting of recording. Depending on the application, this information may or may not be useful:

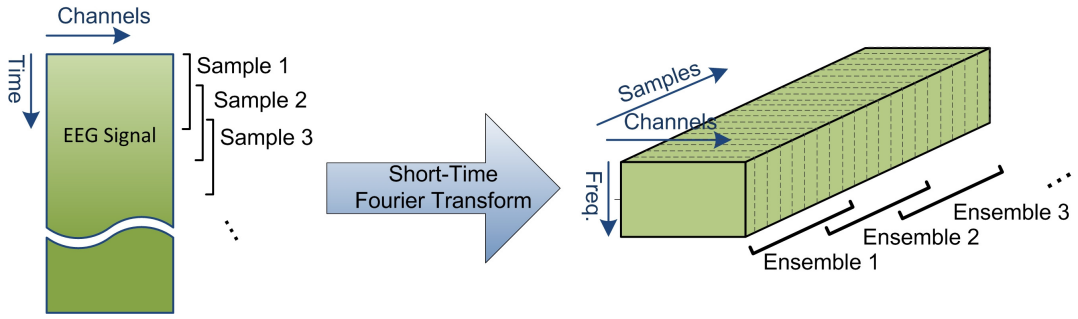


Figure 6.1: Complex-valued EEG spectral components obtained using short-time Fourier transformation of multichannel EEG data.

- If EEG signal is recorded to study the brain's response to an external stimulus (e.g., event related potentials), the linear phase shift term conveys information about the amount of time being passed from the *stimulus*. This information can be exploited when the STFT components $z'(t, f, c|\Omega_i)$ are analyzed.

- If the EEG is recorded while the brain is performing a repetitive task (e.g., motor imagery), the linear phase shift term conveys no relevant information other than the time passed since the *start of recording*, which can be shifted to any arbitrary time during the trial. In such a case, one should use $z(t, f, c|\Omega_i)$ spectral components in which the effect of the linear phase shift is removed from the spectrum.

Therefore, in this chapter we utilize the definition given in Eq.(6.1), which can also be obtained by adjusting the phase of $z'(t, f, c|\Omega_i)$ as follows: $z(t, f, c|\Omega_i) = z'(t, f, c|\Omega_i) * e^{-j2\pi ft}$. The resulting complex-valued spectral components can be decomposed as follows:

$$z(t, f, c|\Omega_i) = x(t, f, c|\Omega_i) + jy(t, f, c|\Omega_i), \tag{6.4}$$

$$z(t, f, c|\Omega_i) = r(t, f, c|\Omega_i) * \exp\{j\alpha(t, f, c|\Omega_i)\} \tag{6.5}$$

where $x(t, f, c|\Omega_i)$, $y(t, f, c|\Omega_i)$, $r(t, f, c|\Omega_i)$, $\alpha(t, f, c|\Omega_i)$ are respectively the real part, imaginary part, magnitude, and phase of $z(t, f, c|\Omega_i)$.

Using the above definitions, the complex-valued spatio-spectral feature matrix $\mathbf{Z}(t) \in \mathbb{C}^{N_f \times N_{ch}}$ can be formed as illustrated in Figure 6.1. Similar to the matrix-variate feature matrix defined in Section 3.4, the $(f, c)^{th}$ element of $\mathbf{Z}(t)$ contains $z(t, f, c|\Omega_i)$.

Assume that a subject is performing a specific mental imagery task during the time interval $t \in [t_1, t_2]$. The EEG spectral components $z(t, f, c|\Omega_i)$ are called stationary, if the probability density function (pdf) of $z(t, f, c|\Omega_i)$ only depends on the variables f and c and is constant over time $t \in [t_1, t_2]$. Similarly,

variables $z(t, f, c|\Omega_i)$ will be called quasi-stationary if their pdf changes very slowly with time, such that the pdf can be considered to be constant as long as $z(t, f, c|\Omega_i)$ is observed over a short period of time (i.e., $t_2 - t_1$ is small enough). In such a case, we consider the $z(t, f, c|\Omega_i)$ components that are observed during this short period of time to form a set of samples with the same pdf. We call this set of samples an *ensemble*. Based on the assumption of quasi-stationarity, we will omit the temporal index of $\mathbf{Z}(t)$ in the next section, and will assume that distribution of $\mathbf{Z}(t)$ does not change during an ensemble. Various implications of this assumption will be discussed later in this chapter.

6.2 Matrix-Variate Complex Gaussian Model for Z

In this section, we propose a matrix-variate Gaussian model for the complex-valued EEG spectrum. The main advantage of a Gaussian model is that complete characterization of this model only requires estimation of the first and second order statistics of the data. Furthermore, the studies in the previous chapters illustrated the potential benefits of utilizing matrix-variate Gaussian model in various stages of feature extraction in BCI systems. Finally, a matrix-variate Gaussian model provides a mathematically tractable framework for development of more efficient signal processing and feature extraction algorithms for analysis of the EEG spectrum.

Let $f(\mathbf{Z}|\Omega_i)$ denote the conditional probability of matrix $\mathbf{Z} \in \mathbb{C}^{N_f \times N_{ch}}$ under class Ω_i. A matrix-variate Gaussian model for the complex-valued feature matrix $\mathbf{Z}$ is denoted by:

$$\mathbf{Z}|\Omega_i \sim \mathcal{CN}(\mathbf{M}_i, \mathbf{\Phi}_i, \breve{\mathbf{\Phi}}_i, \mathbf{\Psi}_i, \breve{\mathbf{\Psi}}_i), \qquad 1 \leq i \leq C \tag{6.6}$$

Here, the matrices $\mathbf{M}_i, \mathbf{\Phi}_i, \breve{\mathbf{\Phi}}_i, \mathbf{\Psi}_i, \breve{\mathbf{\Psi}}_i$ denote the mean, spectral covariance, spectral pseudo-covariance, spatial covariance, and spatial pseudo-covariance of the class Ω_i. These matrices are defined as follows:

$$\mathbf{M}_i = \mathrm{E}_{\mathbf{Z}|\Omega_i}(\mathbf{Z}), \tag{6.7}$$

$$\mathbf{\Phi}_i = \mathrm{tr}^{-1}(\mathbf{\Psi}_i) * \mathrm{E}_{\mathbf{Z}|\Omega_i}\left((\mathbf{Z} - \mathbf{M}_i)(\mathbf{Z} - \mathbf{M}_i)^H\right), \tag{6.8}$$

$$\breve{\mathbf{\Phi}}_i = \mathrm{tr}^{-1}(\mathbf{\Psi}_i) * \mathrm{E}_{\mathbf{Z}|\Omega_i}\left((\mathbf{Z} - \mathbf{M}_i)(\mathbf{Z} - \mathbf{M}_i)^T\right), \tag{6.9}$$

$$\mathbf{\Psi}_i = \mathrm{tr}^{-1}(\mathbf{\Phi}_i) * \mathrm{E}_{\mathbf{Z}|\Omega_i}\left((\mathbf{Z} - \mathbf{M}_i)^H(\mathbf{Z} - \mathbf{M}_i)\right), \tag{6.10}$$

$$\breve{\mathbf{\Psi}}_i = \mathrm{tr}^{-1}(\breve{\mathbf{\Phi}}_i) * \mathrm{E}_{\mathbf{Z}|\Omega_i}\left((\mathbf{Z} - \mathbf{M}_i)^T(\mathbf{Z} - \mathbf{M}_i)\right). \tag{6.11}$$

Note that second order characterization of $\mathbf{Z}$ requires the knowledge of not only the spectral and spatial covariance matrices, but also the spectral and spatial pseudo-covariance matrices.

In order to explain the importance of the pseudo-covariances in second order characterization of $\mathbf{Z}$, consider the column-wise vectorized representation of $\mathbf{Z}$, denoted by $\mathbf{z} = \text{vec}(\mathbf{Z}) = [\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_{N_{ch}}^T]^T$, where $\mathbf{z}_n$ represents the n^{th} column vector in $\mathbf{Z}$. The conditional covariance and conditional pseudo-covariance of $\mathbf{z}$ are then defined as

$$\boldsymbol{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^H} = \text{E}\left\{(\mathbf{z}-\overline{\mathbf{z}})(\mathbf{z}-\overline{\mathbf{z}})^H|\Omega_i\right\} = \boldsymbol{\Psi}_i \otimes \boldsymbol{\Phi}_i, \tag{6.12}$$

$$\boldsymbol{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^T} = \text{E}\left\{(\mathbf{z}-\overline{\mathbf{z}})(\mathbf{z}-\overline{\mathbf{z}})^T|\Omega_i\right\} = \breve{\boldsymbol{\Psi}}_i \otimes \breve{\boldsymbol{\Phi}}_i, \tag{6.13}$$

It is well known in the literature that the knowledge of $\boldsymbol{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^H}$ and $\boldsymbol{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^T}$ is required for complete second order characterization of $\mathbf{z}$ (ref. [156–158]). Indeed, these two matrices convey information regarding the covariance of the real and imaginary parts of $\mathbf{z}$ as well as the cross covariance between the real and imaginary parts, as follows:

$$\boldsymbol{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^H} = \boldsymbol{\Sigma}^{(i)}_{\mathbf{x}\mathbf{x}^T} + \boldsymbol{\Sigma}^{(i)}_{\mathbf{y}\mathbf{y}^T} + j(\boldsymbol{\Sigma}^{(i)}_{\mathbf{y}\mathbf{x}^T} - \boldsymbol{\Sigma}^{(i)}_{\mathbf{x}\mathbf{y}^T}), \tag{6.14}$$

$$\boldsymbol{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^T} = (\boldsymbol{\Sigma}^{(i)}_{\mathbf{x}\mathbf{x}^T} - \boldsymbol{\Sigma}^{(i)}_{\mathbf{y}\mathbf{y}^T}) + j(\boldsymbol{\Sigma}^{(i)}_{\mathbf{x}\mathbf{y}^T} + \boldsymbol{\Sigma}^{(i)}_{\mathbf{y}\mathbf{x}^T}), \tag{6.15}$$

were $\mathbf{x}$ and $\mathbf{y}$ are the real and imaginary parts of $\mathbf{z} = \mathbf{x} + j\mathbf{y}$. Indeed, if we consider $\tilde{\mathbf{z}} = [\mathbf{x}^T, \mathbf{y}^T]^T$, then $\boldsymbol{\Sigma}_{\mathbf{z}\mathbf{z}^T}$ and $\boldsymbol{\Sigma}_{\mathbf{z}\mathbf{z}^H}$ can be uniquely determined by $\boldsymbol{\Sigma}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T}$, and vice versa [156, 157].

As a result, the probability density function (pdf) of $\mathbf{Z}$ can be determined in terms of $\boldsymbol{\Sigma}_{\mathbf{z}\mathbf{z}^H}$ and $\boldsymbol{\Sigma}_{\mathbf{z}\mathbf{z}^T}$ or alternatively in terms of $\boldsymbol{\Sigma}^{(i)}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T}$. Throughout this chapter, we use the following formulation for the pdf of vector $\mathbf{Z}$ in terms of the vector $\tilde{\mathbf{z}}$:

$$f(\mathbf{Z}|\Omega_i) = \left|2\pi\boldsymbol{\Sigma}^{(i)}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\tilde{\mathbf{z}}-\boldsymbol{\mu}_i)^T\left(\boldsymbol{\Sigma}^{(i)}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T}\right)^{-1}(\tilde{\mathbf{z}}-\boldsymbol{\mu}_i)\right\} \tag{6.16}$$

where $\boldsymbol{\mu}_i = \text{E}_{\tilde{\mathbf{z}}|\Omega_i}(\tilde{\mathbf{z}})$ is the conditional mean of $\tilde{\mathbf{z}}$.

6.2.1 Propriety of Z

By definition, random matrix $\mathbf{Z}$ will be called *proper* [158] or *circularly symmetric* [159] if $\boldsymbol{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^T} = \mathbf{0}$, i.e., $\breve{\boldsymbol{\Phi}}_i = \mathbf{0}$ and $\breve{\boldsymbol{\Psi}}_i = \mathbf{0}$; otherwise, it is called *improper* or *non circularly-symmetric*. From (6.15), it can be seen that a proper matrix $\mathbf{Z}$ has the following properties:

$$\boldsymbol{\Sigma}^{(i)}_{\mathbf{x}\mathbf{x}^T} = \boldsymbol{\Sigma}^{(i)}_{\mathbf{y}\mathbf{y}^T} \quad \text{and} \quad \boldsymbol{\Sigma}^{(i)}_{\mathbf{x}\mathbf{y}^T} = -\left(\boldsymbol{\Sigma}^{(i)}_{\mathbf{x}\mathbf{y}^T}\right)^T. \tag{6.17}$$

In the univariate case, therefore, a complex-valued random variable z_{mn} will be proper or circularly-symmetric if its real and imaginary parts are independent and have equal power. If z_{mn} is proper, its phase is uniformly distributed and conveys no information. Otherwise, the phase of z_{mn} would convey information relevant to brain activities, which should be taken into account in BCI systems. Thus, we will use the propriety of the EEG spectral components to measure whether or not its phase conveys any relevant information.

6.2.2 Sufficient Conditions for Separability of $\mathbf{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^H}$ and $\mathbf{\Sigma}^{(i)}_{\mathbf{z}\mathbf{z}^T}$

As mentioned in Section 3.4, the main difference between matrix-variate Gaussian distribution and the conventional multivariate Gaussian distribution is the Kronecker structure assumed for the covariance of the data, which implies the separability of the spectral and spatial covariances. The same property holds true for complex Gaussian distributions, in which case both covariance and pseudo-covariance of the features are required to have Kronecker structure, as defined in (6.12) and (6.13).

In Appendix A.4 a sufficient condition for separability of $\mathbf{Z}$ is provided. Based on this result, if the *spatio-temporal* covariance of the EEG data is separable, it is guaranteed that the spatio-spectral features obtained through Fourier transformation of the data are also separable. It is noteworthy that the same condition also guarantees that the spatio-spectral features obtained through bandpass filtering of the data are also separable (ref. Appendix A.4). Therefore, it is reasonable to assume that in both cases (i.e., Fourier transformation and bandpass filtering) the separability of the spatio-spectral features is caused by the same phenomenon. With minor modifications, similar conclusion can be made for most spectral analysis methods that are based on linear processing/filtering of the EEG data.

6.3 The Effect of Epoch Length on The Gaussianity of The Features

In the previous section, we suggested that if the duration of the observation window or epoch length (L_e) is short enough, the spatio-spectral matrix $\mathbf{Z}$ can be modelled by an improper matrix-variate complex Gaussian distribution. In this section, the validity of the suggested Gaussian model and possible ranges of L_e for which this model fits the data will be verified in three steps:

1. Validating the normality of individual components of $\mathbf{Z}$, denoted by z_{mn}, for different values of L_e and finding the maximum value of L_e over which all z_{mn} fit the complex-normal model;

2. Validating the joint-normality of each column vector of $\mathbf{Z}$, denoted by $\mathbf{z}_n$, over the observation length L_e determined in Step 1.

3. Validating the impropriety of each $\mathbf{z}_n$, over the observation length L_e determined in Step 1.

Ideally, it is also desired to perform a joint normality test on all the elements of $\mathbf{Z}$. However, due to the high dimensionality of the data and the limitations on the number of samples within each epoch, such a statistical test is not feasible. Indeed, as it will be discussed later in this section, even testing the joint normality of the elements in $\mathbf{z}_n$ is challenging.

6.3.1 Experiment Setup

Our analysis in this section are based on data set V of the BCI competition III [147]. As described in previous chapters, this data set consists of EEG signals of three normal subjects (persons) recorded during four non-feedback sessions. During each session, the subject sequentially imagines three different tasks: repetitive self-paced *left hand* movements (Task Ω_1), repetitive self-paced *right hand* movements (Task Ω_2), and *generation of words* beginning with the same random letter (Task Ω_3). The main benefit of data set V of BCI competition III in comparison to data set 2a of BCI competition IV is the long period of each motor imagery trial, which is 15 seconds in the former data set compared to 3 seconds in the latter one. This long trial length allows us to better study the changes in the EEG phase information over time.

Assuming that the EEG data is quasi-stationary over an observation window of L_e seconds, we divide the EEG recoding during each mental imagery task into several overlapping observation periods (i.e., $E_1, E_2, ...$) of length L_e. During each E_i, the signal is transformed from time-domain to the frequency-domain, using short-time Fourier transformation[1]. In this chapter, the $w(t)$ is chosen to be Tukey window of length 1 second with overlapping factor of 15/16 (i.e, window shift of 1/16 second) and $\alpha = 1/8$.

After STFT transformation, the spectral components in the range of $8 - 30$ Hz with a frequency resolution of 2 Hz are retained. This frequency band corresponds to the α rhythm ($8 - 12$ Hz) and β rhythm ($12 - 30$ Hz) of the brain which are known to be associated with mental imagery tasks. The resulting samples (i.e., $S_1, S_2, ...$) form an ensemble E_i. Each multichannel sample (S_i) in this ensemble can be represented by a complex-valued matrix $\mathbf{Z} \in \mathbb{C}^{13\times 8}$, where each column of $\mathbf{Z}$ represents the vector of 13 frequency components of an EEG channel (ref. Figure 6.1).

[1]For simplicity, we will consider the discrete Fourier spectral components derived from STFT.

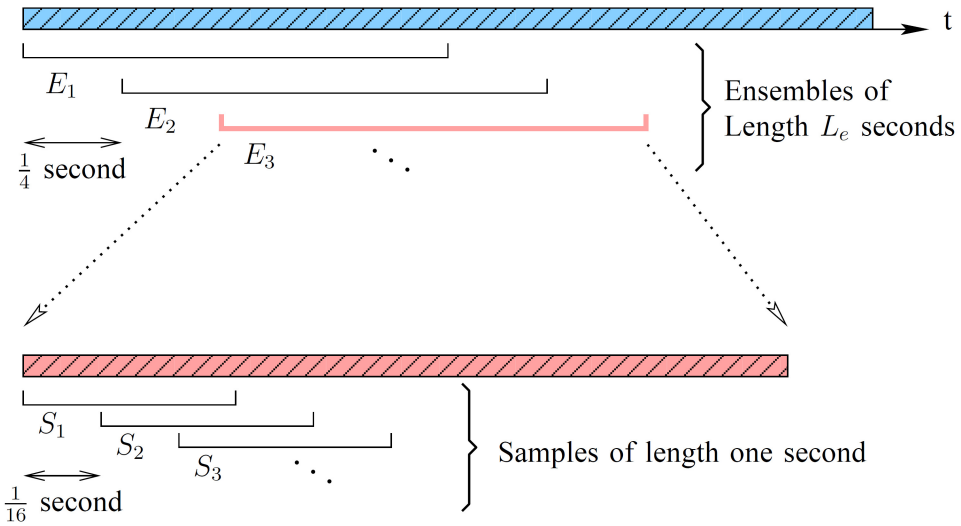


Figure 6.2: Observation windows of length L_e and their corresponding samples.

6.3.2 Testing the Normality of z_{mn}

This section studies the normality of each complex-valued frequency component of the multichannel EEG spectrum $\mathbf{Z}$. We test the following null hypothesis

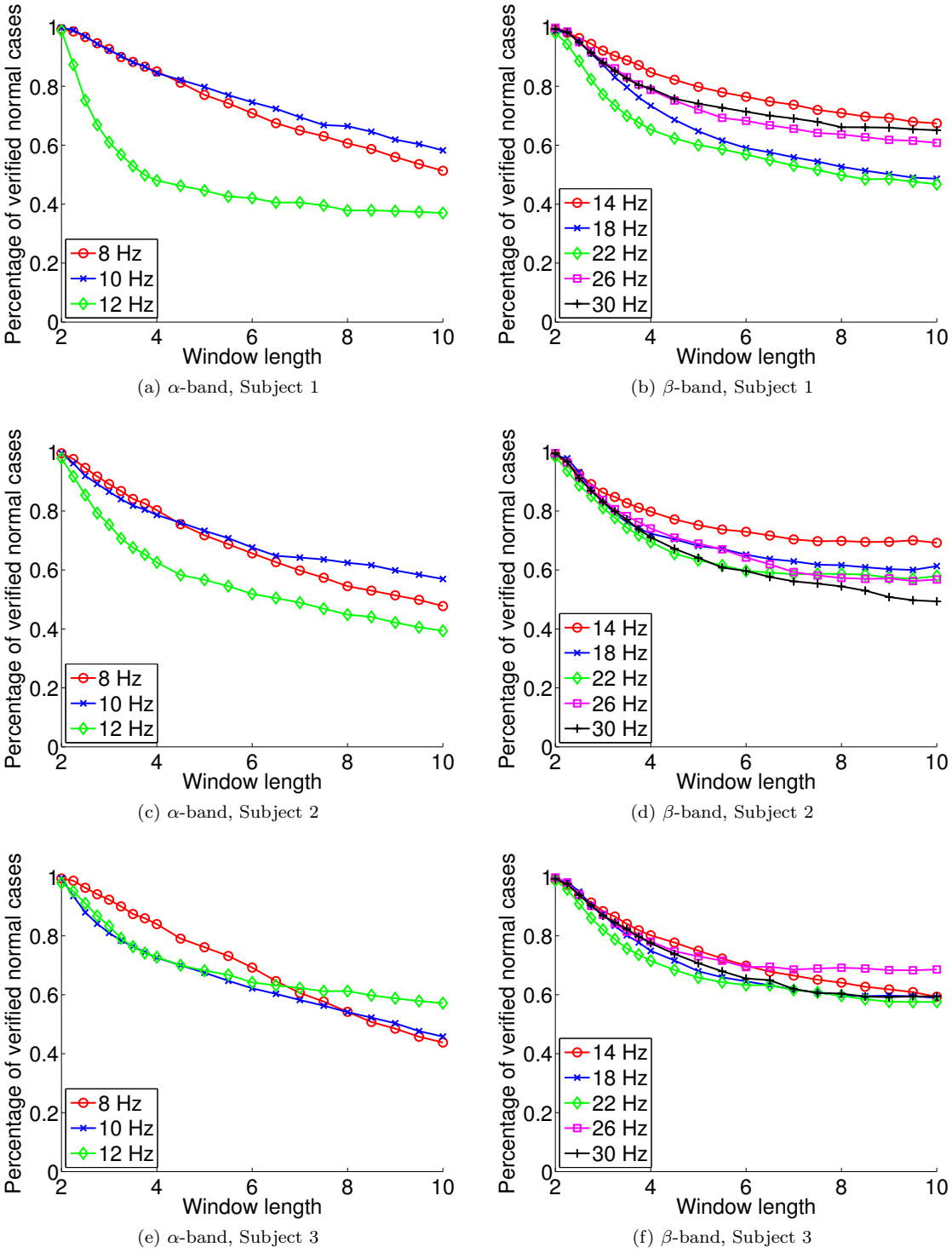
$$H_0 : z_{mn} = x_{mn} + jy_{mn} \sim \mathcal{CN}(\mu_z, \mathbf{\Sigma}_{zz^*}, \mathbf{\Sigma}_{zz}),$$

i.e., z_{mn} has a univariate complex-valued Gaussian distribution with unknown mean, variance, and pseudo-variance. As described in Section 6.2, H_0 is equivalent to the hypothesis

$$H_0' : \tilde{\mathbf{z}}_{mn} = \begin{bmatrix} x_{mn} \\ y_{mn} \end{bmatrix} \sim \mathcal{N}_2(\mu_{\tilde{\mathbf{z}}}, \mathbf{\Sigma}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T}),$$

i.e., $\tilde{\mathbf{z}}_{mn}$ has a bivariate real-valued Gaussian distribution with unknown mean and covariance.

We examine hypothesis H_0' using the well-known Mardia's multivariate normality test [160] with a significance level of 0.05. In order to find the maximum length L_e over which the EEG signal can be assumed to be quasi-stationary, we have repeated Mardia's test for various values of ensemble length L_e from 2 to 10 seconds. The test results for Task 1 of all three subjects are shown in Figure 6.3. We have reported the percentage of ensembles whose samples are verified to have Gaussian distribution. Parts (a-c) of this figure illustrate the results for α-band frequency components, and Parts (d-f) illustrate the results for β-band. It should be mentioned that since all the channels exhibited similar trends in the tests, the results reported in Figure 6.3 have been averaged over all 8 channels. This figure reveals that despite of the inter-subject and inter-frequency variability of the results, in all the situations, the complex-valued Gaussian model describes the experimental data more accurately as the length of L_e



(a) α-band, Subject 1

(b) β-band, Subject 1

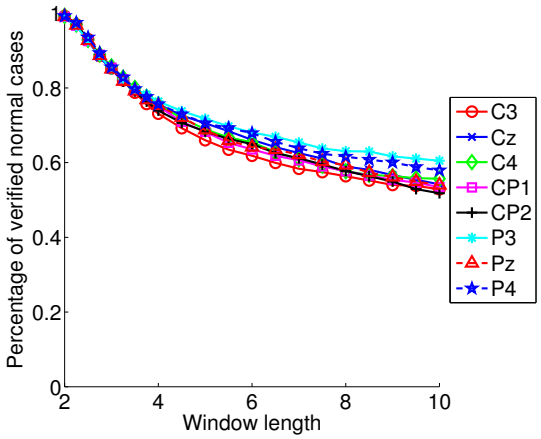
(c) α-band, Subject 2

(d) β-band, Subject 2

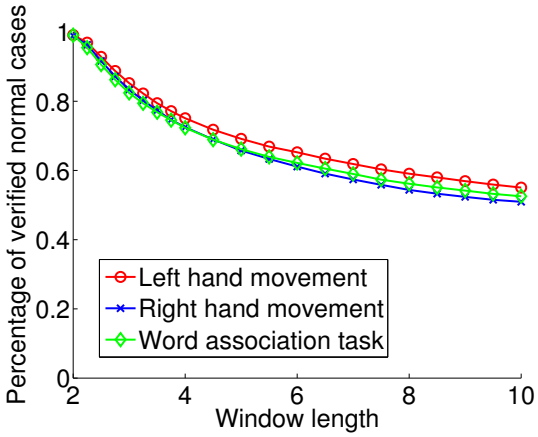
(e) α-band, Subject 3

(f) β-band, Subject 3

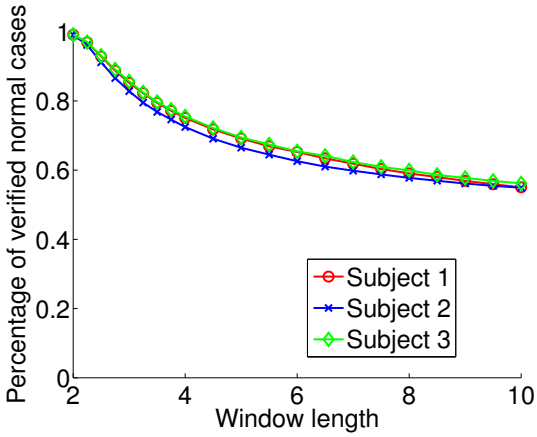
Figure 6.3: Percentage of verified normal EEG components for left hand movement task performed by different subjects is plotted for different frequencies averaged over channels. (For more clarity, only five frequency components of β-band are illustrated.)



(a) Left hand movement, Subject 1



(b) First subject, All tasks



(c) Left hand movement, All subjects

Figure 6.4: Percentage of verified normal EEG components for (a) left hand movement of Subject 1 in different channels, (b) different tasks of the first subject averaged over channels, and (c) left hand movement of different subjects averaged over channels. The values in all figures are averaged over frequencies.

decreases. Specifically, for $L_e = 3$ seconds, on average only %15 of the of the ensembles are rejected to have samples with normal distribution. The test results show similar trend for the other two tasks.

It is worthy to mention that when the resulting percentages are averaged over all frequencies, there is no significant variation between different tasks, different subjects, or different channels. As an example, Figure 6.4.a provides the average percentage of verified normal cases for first subject's left hand movement task, plotted for all the 8 channels. Figure 6.4.b compares the results of all three tasks for the first subject, which are again very close to each other. The same trend can be seen in Figure 6.4.c for one task over all three subjects.

As a result, we can conclude that H_0 is valid if the length of the observation window is small enough. Thus, we set $L_e = 3$ seconds in the rest of this chapter. It should be noted that even for large L_e, only

Table 6.1: Percentage of verified multi-variate complex normal EEG channels for different tasks in different subjects.

Task	Subj. 1	Subj. 2	Subj. 3
(1)	0.9978	0.9676	0.9899
(2)	0.9488	0.9939	0.9943
(3)	0.9303	0.9873	0.9938

Table 6.2: Average p-value of multi-variate normality test on EEG channels for different tasks in different subjects.

Task	P_1			P_2		
	Subj. 1	Subj. 2	Subj. 3	Subj. 1	Subj. 2	Subj. 3
(1)	0.9842	0.9740	0.9798	0.3118	0.3019	0.3139
(2)	0.9622	0.9776	0.9811	0.3042	0.3180	0.3135
(3)	0.9500	0.9803	0.9822	0.3010	0.3148	0.3118

in half of the cases H_0 is rejected.

6.3.3 Testing the Normality of $\mathbf{z}_n$

The results of the previous test showed each individual z_{mn} element can be modelled as a complex-valued Gaussian random variable when L_e is small enough. This is a necessary but not sufficient condition for *joint* Gaussianity of all elements of vector $\mathbf{z}_n$. This section examines the following hypothesis $H_0 : \mathbf{z}_n = \mathbf{x}_n + j\mathbf{y}_n \sim \mathcal{CN}_M(\mu_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}\mathbf{z}^H}, \mathbf{\Sigma}_{\mathbf{z}\mathbf{z}^T})$, where $M = 13$ is the number of frequency components in the EEG spectrum, and $\mathcal{CN}_M$ denotes the M-variate complex-valued Gaussian distribution. H_0 is equivalent to the following hypothesis:

$$H_0' : \tilde{\mathbf{z}}_n = \begin{bmatrix} \mathbf{x}_n \\ \mathbf{y}_n \end{bmatrix} \sim \mathcal{N}_{2M}(\mu_{\tilde{\mathbf{z}}}, \mathbf{\Sigma}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T}).$$

Assuming $L_e = 3$ seconds from previous section, we have 32 samples in each ensemble to examine the multivariate vector $\tilde{\mathbf{z}}$ of the relatively large dimension $2M = 26$. Consequently, Mardia's multivariate normality test cannot be utilized in this section. We use the multivariate normality test proposed by [161], which is designed to overcome this small-sample-size problem. The results of this test for a significance value of 0.05 are presented in Table 6.1. The average p-values of this test are also reported in Table 6.2.

The fact that multivariate-normality of $\mathbf{z}_n$ is only rejected in less than %10 of the cases, together with the fact that individual elements of $\mathbf{z}_n$ are shown to be normal, confirms with high confidence that our proposed multivariate normal model for the vector $\mathbf{z}_n$ fits the experimental data.

Table 6.3: Average impropriety score of EEG channels for different tasks in different subjects.

Task	Subj. 1	Subj. 2	Subj. 3
(1)	174.109	200.562	173.189
(2)	153.195	230.865	125.034
(3)	139.312	191.861	96.923

6.3.4 Testing the Propriety of $\mathbf{z}_n$

This section examines the propriety, or circular-symmetry, of the complex-valued EEG spectrum for each channel. As mentioned in Section 6.2, if $\mathbf{z}_n$ is proved to be improper, we can conclude that the phase of complex-valued spectrum obtained from STFT has relevant information which will be lost in the power spectral density representation. Therefore, we examine the hypothesis $H_0 : \mathbf{\Sigma}_{\mathbf{x}\mathbf{x}^T} = \mathbf{\Sigma}_{\mathbf{y}\mathbf{y}^T}$, which is a necessary condition for $\mathbf{z}_n$ to be proper. In other words, rejection of H_0 is a sufficient condition for $\mathbf{z}_n$ to be improper.

We use the test in [162], which examines the equality of two covariance matrices with small number of samples. The results for a significance value of 0.05 show that for *all* the cases, hypothesis H_0 is rejected. Indeed, all the resulting test statistics have large values, with an average of 165, whereas the critical value for hypothesis rejection is 1.645. These test statistics, which can be considered as the impropriety scores, are reported in Table 6.3.

6.4 Time-Varying Characteristics of the EEG Spectrum

In the previous section, it was shown that during a mental imagery task, the spectral components $z(t, f, c|\Omega_i)$ can be modelled with a quasistationary *noncircularly-symmetric complex-valued* Gaussian distribution, whose first and second order statistics can be defined with the following parameters:

$$\mu_z(t, f, c|\Omega_i) = E\{z(t, f, c|\Omega_i)\} = \mu_x(t, f, c|\Omega_i) + j\mu_y(t, f, c|\Omega_i) \tag{6.18}$$

$$\begin{aligned} \sigma_z^2(t, f, c|\Omega_i) &= E\left\{|z(t, f, c|\Omega_i) - \mu_z(t, f, c|\Omega_i)|^2\right\} \\ &= \sigma_x^2(t, f, c|\Omega_i) + \sigma_y^2(t, f, c|\Omega_i) \end{aligned} \tag{6.19}$$

$$\begin{aligned} \gamma_z^2(t, f, c|\Omega_i) &= E\left\{(z(t, f, c|\Omega_i) - \mu_z(t, f, c|\Omega_i))^2\right\} \\ &= \left(\sigma_x^2(t, f, c|\Omega_i) - \sigma_y^2(t, f, c|\Omega_i)\right) + j2\sigma_{xy}(t, f, c|\Omega_i) \end{aligned} \tag{6.20}$$

where $\mu_z(t,f,c|\Omega_i)$, $\sigma_z^2(t,f,c|\Omega_i)$, and $\gamma_z^2(t,f,c|\Omega_i)$ represent the time-varying mean, variance, and pseudo-variance of $z(t,f,c|\Omega_i)$.

The results of our statistical tests indicate that $\mu_x(t,f,c|\Omega_i)$, $\mu_y(t,f,c|\Omega_i)$, $\sigma_x^2(t,f,c|\Omega_i)$, and $\sigma_y^2(t,f,c|\Omega_i)$ parameters change very slowly during a trial and can be considered to be constant over observation intervals of length *three* seconds or less[2]. In other words, for $t \in [t_1, t_1+3]$ we have

$$x(t,f,c|\Omega_i) \sim \mathcal{N}\big(\mu_x(f,c,t_1|\Omega_i), \sigma_x^2(f,c,t_1|\Omega_i)\big) \quad \text{for} \quad t \in [t_1, t_1+3] \tag{6.21}$$

$$y(t,f,c|\Omega_i) \sim \mathcal{N}\big(\mu_y(f,c,t_1|\Omega_i), \sigma_y^2(f,c,t_1|\Omega_i)\big) \quad \text{for} \quad t \in [t_1, t_1+3] \tag{6.22}$$

In this section, we further study the time-varying nature of the complex-valued EEG spectral components. In particular, the time-varying properties of the mean and variance of the real and imaginary parts of the spectrum will be examined.

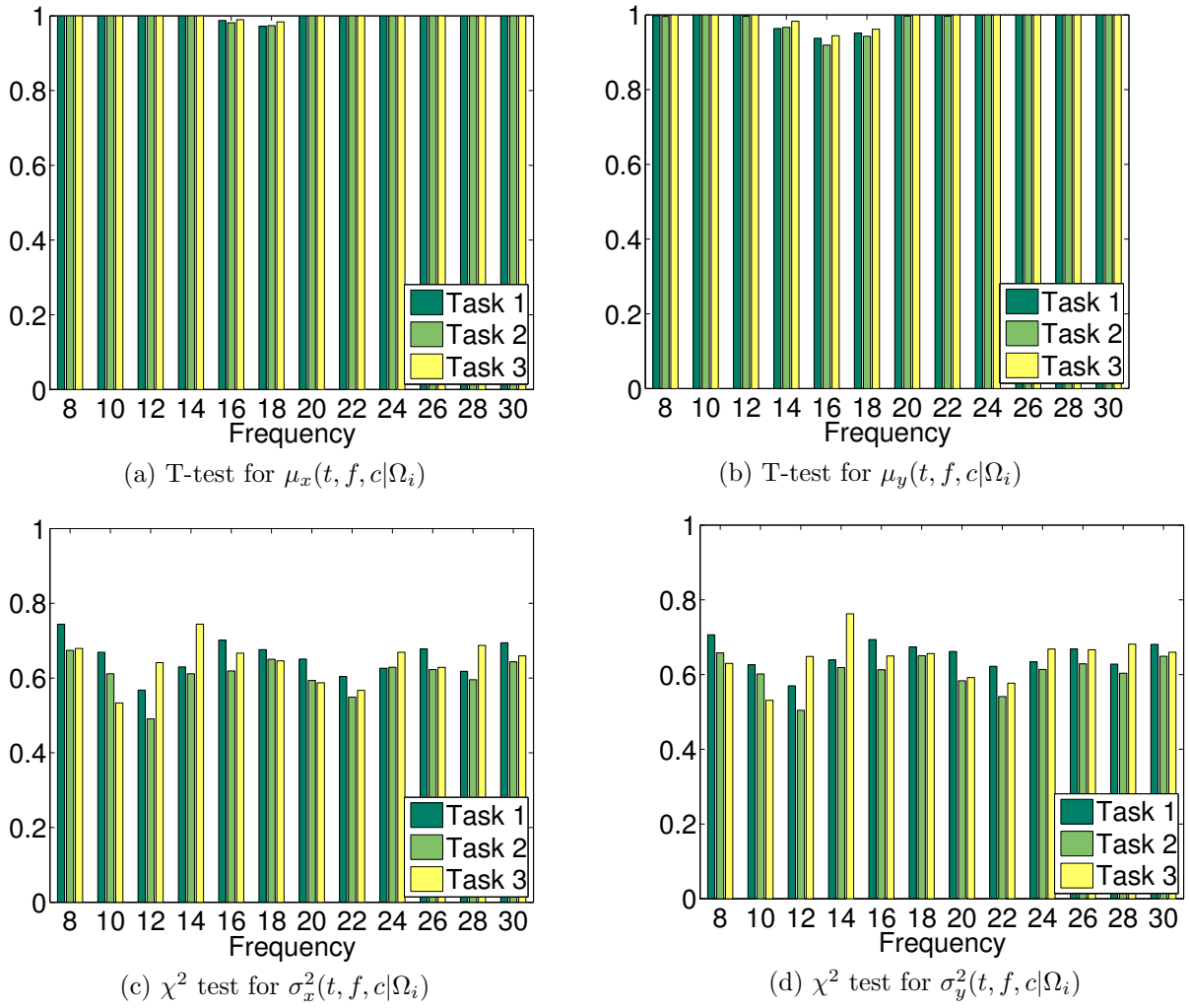
Let the STFT samples obtained during a trial be divided into overlapping ensembles of length three seconds. We perform the well known T-test and Chi-square variance test to determine if the mean/variance of the samples within each ensemble is equal to the overall trial mean/variance, denoted by $\mu(f,c|\Omega_i)$ and $\sigma^2(f,c|\Omega_i)$, which is empirically calculated from all the samples in the trial. Each of these tests is separately performed on the real part and imaginary part of spectral components.

In order to study the ensemble means, we use the T-test which examines the null hypothesis that the $x(t,f,c|\Omega_i)$ (or $y(t,f,c|\Omega_i)$) samples within an ensemble have a Gaussian distribution with mean $\mu_x(f,c|\Omega_i)$ (or $\mu_y(f,c|\Omega_i)$) and unknown variance. This test is repeated over all the trials in the database, for each specific frequency and each channel. A significance level of 0.05 is used for all the statistical tests performed in this section.

Figure 6.5.a shows the results of this test for the real part of the spectrum ($x(t,f,c|\Omega_i)$). In this figure, we have reported the percentage of ensembles for which the null hypothesis of T-test is not rejected. In other words, the percentage of ensembles which are verified to have the same mean as the overall trial mean are presented in this figure. Figure 6.5.b shows similar results for the imaginary part of the spectrum ($y(t,f,c|\Omega_i)$). Since all the subjects exhibited similar trends, only the results of Subject 1 are reported. These results are averaged over all channels.

The results of Figure 6.5 reveal that the mean of spectral components are highly stationary over each mental imagery trial. Therefore, we can assume that the $\mu_x(t,f,c|\Omega_i)$ and $\mu_y(t,f,c|\Omega_i)$ parameters are constant over each trial and do not change with time index t.

[2]For simplicity, in this chapter we assume $\sigma_{xy}(t,f,c|\Omega_i) = 0$ and only focus on the effect of $\sigma_x^2(t,f,c|\Omega_i)$ and $\sigma_y^2(t,f,c|\Omega_i)$.



(a) T-test for $\mu_x(t, f, c|\Omega_i)$

(b) T-test for $\mu_y(t, f, c|\Omega_i)$

(c) χ^2 test for $\sigma_x^2(t, f, c|\Omega_i)$

(d) χ^2 test for $\sigma_y^2(t, f, c|\Omega_i)$

Figure 6.5: Percentage of ensembles verified to have a mean (a-b) or a variance (c-d) equal to the overall empirical mean or variance calculated using all the samples in a trial. The test results are very similar for all subject, hence only the results of Subject 1 are presented here.)

In order to study the ensemble variances, we use the Chi-square variance test which examines the null hypothesis that the $x(t, f, c|\Omega_i)$ (or $y(t, f, c|\Omega_i)$) samples within an ensemble have a Gaussian distribution with $\sigma_x^2(f, c|\Omega_i)$ (or $\sigma_y^2(f, c|\Omega_i)$). Figures 6.5.c and 6.5.d show the results of this test for the real and imaginary parts of the spectrum ($x(t, f, c|\Omega_i)$ and $y(t, f, c|\Omega_i)$). Similar to the Figures a-b, the percentage of ensembles which are verified to have the same variance as the overall trial variance are presented in these figures. It can be seen that for more than %30 of the ensembles the null hypothesis is rejected, which shows that unlike the means, the variances are time-varying and cannot be assumed to be constant over the entire trial.

The above results together with the results of the previous section suggest that during a mental imagery trial, the complex-valued spectral components can be modelled with a time-varying noncircularly-

symmetric Gaussian model with a constant mean and a time-varying variance and pseudo-variance. In this model, the variations of the variance and the pseudo-variance are slow enough such that a Gaussian distribution with fixed parameters accurately models the spectral components observed during a short interval (of length 3 seconds or less). This motivates us to examine the possibility of using an *autoregressive conditional heteroscedastic (ARCH)* model for the time-varying variance of the spectral components in the next section.

6.5 ARCH Model for Spectral Components

The main challenge in dealing with the time-varying Gaussian model proposed in the previous section is to model the variations of $\sigma_x^2(t, f, c|\Omega_i)$ and $\sigma_y^2(t, f, c|\Omega_i)$ parameters over time. This section examines if an ARCH model [163] can be used for this purpose. The ARCH model assumes that: (a) The variance of the signal is not constant and changes over time; hence the term heteroscedastic. (b) The variance at each time instance is a linear function of the previous samples; hence the term conditional autoregressive. Let $\varepsilon_x(f, c, t) = x(t, f, c|\Omega_i) - \mu_x(f, c)$, then the ARCH model of order q implies that

$$\sigma_x^2(t, f, c|\Omega_i) = \alpha_0(f, c|\Omega_i) + \Sigma_{i=1}^{q} \alpha_i(f, c|\Omega_i)\ \varepsilon_x^2(f, c, t - i\Delta t|\Omega_i) \tag{6.23}$$

where $i\Delta t$ is the time lag between the current sample and the ith previous sample (As it was explained previously, $\Delta t = 1/16$ second). A similar model can also be used for $\sigma_y^2(t, f, c|\Omega_i)$ in terms of $\varepsilon_y(f, c, t|\Omega_i) = y(t, f, c|\Omega_i) - \mu_y(f, c|\Omega_i)$. In order to test the validity of this model we have performed the following steps for $x(t, f, c|\Omega_i)$ and $y(t, f, c|\Omega_i)$ over a trial.

1. Start with $q = 1$;
2. Assuming that an ARCH(q) model describes the variations of the variance of the spectral components over time, estimate the parameters $\{\alpha_0, ..., \alpha_q\}$ and find the log-likelihood objective function value (LLF_q) associated with the parameter estimates;
3. Assuming that an ARCH(q+1) model describes the variations of the variance of the spectral components over time, estimate the parameters $\{\alpha_0, ..., \alpha_{q+1}\}$ and find the value of LLF_{q+1};
4. Perform the likelihood ratio test to determine if there is enough statistical evidence to increase the ARCH order from q to $q + 1$.
5. If the likelihood ratio test confirms the order increase, then increase the ARCH order from q to $q + 1$ and go to step 2. Otherwise, the ARCH(q) model suffices for modelling the variations of the signals' variance.

Table 6.4: ARCH model order for different spectral components of each channel.

	Channels							
Freq	C3	Cz	C4	CP1	CP2	P3	Pz	P4
8 Hz	2*	1	2*	1	1	2*	2*	2*
10 Hz	1	1	1	1	1	1	2*	1
12 Hz	2	2	2*	2	2	2*	2*	2
14 Hz	1	1	1	1	1	1	1	1
16 Hz	1	1	2*	1	1	1	1	1
18 Hz	1	1	1	1	1	1	1	1
20 Hz	2	2	2	2	2	2	2	2
22 Hz	1	1	1	1	1	1	1	1
24 Hz	1	1	1	1	1	1	1	1
26 Hz	1	1	1	1	1	1	1	1
28 Hz	2	2	2	2	2	2	2	2
30 Hz	1	1	1	1	1	1	1	1

Let $q_x(f,c)$ and $q_y(f,c)$ denote the resulting order for $x(t,f,c|\Omega_i)$ and $y(t,f,c|\Omega_i)$. The validity of each of these model orders has been further confirmed using the Engle's test for residual heteroscedasticity [163] with a significance level of 0.05. Our analysis shows that (a) the value of $q_x(f,c)$ is equal to $q_y(f,c)$ in most of the cases (with only a few exceptions which will be mentioned later). Therefore, we will simply show the model order by $q(f,c)$. (b) The value of $q(f,c)$ does not change over different trials and tasks (with a few exceptions). Table 6.4 presents the results of $q(f,c)$ for subject 1 (The other two subjects show similar trends). The values marked by an asterisk are the ones for which the value of $q(f,c)$ was changing between 1 and 2 for the real/imaginary parts and/or for different tasks or trials.

The results of this table show that for all the frequency components the variations of $\sigma_x^2(t,f,c|\Omega_i)$ and $\sigma_y^2(t,f,c|\Omega_i)$, and hence the variations of $\sigma_z^2(t,f,c|\Omega_i)$ and $\gamma_z^2(t,f,c|\Omega_i)$,can be easily modelled using an ARCH model of order one or two. Therefore, we can conclude that the complex-valued spectral components $z(t,f,c|\Omega_i)$ of the EEG signal recorded during a mental imagery trial can be modelled by a noncircularly-symmetric Gaussian distribution with constant mean and time-varying variance and pseudo-variance that follow an ARCH(1) or ARCH(2) model (depending on the values of f and c).

6.6 Summary and Concluding Remarks

In this chapter, we focused on the complex-valued spatio-spectral feature matrices that are extracted through Fourier domain analysis of the EEG data, and proposed a matrix-variate complex Gaussian model for them. Due to the complex-valued nature of these features, the resulting matrix-variate Gaussian model requires the knowledge of spectral and spatial pseudo-covaraince of the data in addition to the conventional spectral and spatial covariances. A brief discussion on the implications of using

pseudo-covariances was also provided.

The matrix-variate complex Gaussian model in this chapter was proposed based on the assumption that the mean and covariance of the spatio-spectral features do not change during a short epoch, denoted by L_e. The effect of epoch length on the validity of this assumption was examined through a number of statistical tests on experimental data. It was shown that the longest acceptable epoch length, for which the aforementioned assumption holds true in this data set, is approximately 3 seconds. Then the temporal variations of the data over longer periods were modelled by an autoregressive conditional heteroscedastic (ARCH) model.

In this chapter, we also examined the non-circularity or impropriety of the complex-valued EEG features to study whether or not the phase of the EEG conveys relevant information during motor imagery tasks. The result of the statistical tests revealed a high degree of non-circularity, which in turn implies that there exists relevant information in the phase of the EEG data.

It should be noted that the statistical tests conducted in Sections 6.3 and 6.4 aimed to specifically examine the Gaussianity of EEG epochs, and they cannot be considered as general tests of stationarity. Since the Gaussian model used in this chapter is characterized based on the first and second order statistics of the data, the main purpose of these statistical tests was to examine the changes of the mean and covariance of the features over time. Therefore, the results of these test are only valid for analysis of the mean and covariance of the data, and shall not be generalized to the general stationarity properties of the EEG data. General analysis of the stationarity of the EEG features requires further statistical tests to directly examine the time-varying characteristics of the data, which is outside the scope of this thesis.

Chapter 7

Conclusions and Future Work

7.1 Research Summary

This research focused on the problem of spatio-spectral feature extraction for the motor-imagery brain-computer interface (MI-BCI) systems, and provided a general framework for FE in MI-BCI systems. In this framework, the feature extraction methods are categorized into domain-specific (DS-FE) and domain-agnostic (DA-FE) techniques. The former group takes advantage of the neurophysiological properties of the EEG signals to extract a set of discriminant features; whereas the latter group consists of generic feature extraction algorithms that are borrowed from the pattern recognition literature and are not necessarily based on any neurophysiological knowledge about the EEG data. The main purpose of the domain-agnostic FE is to reduce the redundancy of the spatio-spectral features and provide a smaller set of discriminant features.

Based on this framework, it was shown that the output of domain-specific FE stage has an inherent matrix-variate structure, which is currently ignored in the BCI literature. It was then proposed that this spatio-spectral feature set can be modelled by a matrix-variate Gaussian distribution. The matrix-variate Gaussian distribution can be considered as a special subset of the multivariate Gaussian model that requires a special Kronecker product structure for the covariance of the data. This extra requirement allows for the matrix-variate data to be analyzed in its inherent matrix structure through bilinear operations.

Using a matrix-variate Gaussian model for the spatio-spectral features, a new domain-agnostic FE approach for MI-BCIs was proposed in Chapter 4. In this approach, we proposed to use bilinear extensions of LDA method, namely 2DLDA and MVLDA methods, in conjunction with the filterbank

common spatial patterns (FBCSP) method. The experimental results showed that the combination of MVLDA with FBCSP highly improves the overall performance of the BCI system.

In Chapter 5, the matrix-variate Gaussian model was deployed at the domain-specific FE step. Based on this model, the separable common spatio-spectral patterns (SCSSP) was developed. This method has less computational cost compared to the FBCSP method and significantly outperforms FBCSP, provided that enough training data is available. Unlike the FBCSP method, SCSSP is able to sort the discriminant power of the extracted features along both spectral and spatial domains. This property of the SCSSP eliminates the need for a separate domain-agnostic FE step, and its output features can be directly passed to the classifier.

In Chapter 6, we focused on the complex-valued spatio-spectral features that are extracted through Fourier domain analysis of the EEG signals. Motivated by the results of the previous two chapters, we studied the possibility of modelling these features with a complex-valued matrix-variate Gaussian distribution. Towards this end, we studied the information conveyed in the phase of these complex-valued features, and statistically showed that during motor-imagery tasks the phase of the EEG conveys relevant information. This result agrees with the recent neurophysiological findings in the literature. Based on these results, a time-varying Gaussian model was proposed for these complex-valued spatio-spectral features, and the validity of this model was examined through statistical test.

The experimental results in Chapters 4 and 5 together with the statistical analysis in Chapter 6 and the preliminary studies in [164–166] highly suggest that

1. the Gaussian distribution is a fairly reasonable model for the spatio-spectral EEG features during motor imagery tasks, and

2. the covariance between any two rhythmic activities in two different EEG channels can be decomposed into two multiplicative components: (a) A spectral covariance term that only depends on the frequency of these two rhythms, and (b) A spatial component that only depends on the spatial location of these two EEG channels.

In order to investigate the validity of this conjecture, further statistical and neurophysiological studies on different EEG databases are required.

The statistical models that were introduced in this thesis, and the corresponding feature extraction approaches, can contribute in laying a mathematical foundation for more systematic analysis of the feature extraction methods for motor-imagery BCI systems. One of the main challenges in the BCI literature is that most of the proposed methods only focus on a certain aspect of the EEG features without providing a connection with other existing approaches or possible implications of the assumptions

made by their proposed method (ref. [167] and the discussions therein). Among other reasons, this problem can be partly attributed to the fact that there has been little focus over the past century on the mathematical modelling of the EEG signals and connecting them to the neurophysiological properties of the EEG signals (ref. [26] and the discussions therein). Nevertheless, recently there has been a growing interest in making connection between different areas of neuroscience and designing feature extraction methods based on a more systematic approach. The presented research work is part of this collective effort.

7.2 Future Work

The following is a short list of possible areas for future research related to this thesis:

- The matrix-variate Gaussian model that was adopted in this thesis can potentially be used in conjunction with various spatio-spectral feature extraction schemes, such as the spectral coherence or the directed transfer functions method to reduce the computational cost and improve the performance of these methods. However, in this thesis we mostly focused on the spatio-spectral features that are obtained through bandpass filtering of the EEG signals. One possible extension of this work is to study the validity of matrix-variate Gaussian distribution for modelling the features obtained from coherence analysis or the directed transfer function method. As mentioned in Chapter 3, the main challenge in utilization of these two methods is the high dimensionality of the feature space generated by them, which highly increases the computational cost of the BCI system and affects the performance of the subsequent classifier. From this perspective, the matrix-variate Gaussian distribution can be extremely helpful in reducing the computational cost of both domain-agnostic FE and domain-specific FE in such systems.

- In Chapters 4 and 5, we examined several factors that can affect the performance of the proposed feature extraction algorithms. However, due to the large number of parameters that are involved in such a design, we had to choose some parameters based on the conventional values used in the literature and our preliminary results on the experimental data. These parameters include: the bandwidth of bandpass filters, the choice of multiclass extension method for FBCSP and SCSSP, the epoch length, and the choice of the classifier. In the future, the aforementioned factors can be studied in more detail, and their effect on the performance both domain-specific FE and domain-agnostic FE stages can be examined.

- The SCSSP method presented in Chapter 5 was shown to be sensitive to the size of training

samples when extremely limited amount of training data is available. Although such situation is not common in motor-imagery BCI systems, it is worth to study the possible modifications in SCSSP method that can result in more robustness when training data is limited. One possible solution is to use multi-subject regularization methods, and aggregate data from multiple subjects to increase the amount of training data. The other possible solution is to combine the covariance assumptions of SCSSP and FBCSP methods in a regularized scheme to reduce the sensitivity of the SCSSP while benefiting from its low computational cost.

- One of the most important assumptions in Chapters 4 and 5 is the separability of the spatio-spectral features. The competitive performance of the MVLDA and SCSSP methods implicitly confirms this assumption. However, this assumption needs to be further validated in the future through statistical tests for separability of the covariance matrices. One of the most important challenges in such a statistical test is the time-varying nature of the EEG signals, which in turn poses a strict limitation on the number of samples that can be used for statistical test, similar to the case in Chapter 6. To the best of our knowledge, there are no statistical tests for such small sample size scenarios in the literature. Therefore, a possible future work can be to develop new statistical verification methods to verify the separability of spatio-spectral features in EEG signals.

- In order to increase the accuracy of the proposed motor-imagery BCI systems, fusion techniques can be deployed at different levels, as follows:

  - Signal acquisition level: As mentioned in Chapter 1, data from multiple brain-imaging techniques such as EEG and fNIR can be used simultaneously to mitigate some limitations of the EEG signals. Moreover, signals from other biological sources, such as heart-rate or eye-movement, can be used to improve the performance of the BCI system.

  - Feature extraction level: In Chapter 3, several alternative methods for feature extraction from EEG signals were discussed. In this thesis, we only focused on a specific subset of these methods. However, it is possible to utilize a number of different methods in parallel to obtain different feature sets, each of which focuses on certain characteristics of the data. These different feature sets can then be combined to provide a better representation of the underlying EEG characteristics.

  - Classification level: One of the commonly used methods for improving the performance of BCI systems is to utilize multiple classifiers in parallel, and combine the resulting decision outputs using different decision fusion techniques such as voting.

- One of the widely used techniques for improving the performance of the motor-imagery BCI systems is to provide the user with real-time feedback and adaptively train the feature extractors and the classifier. The feature extraction methods proposed in this thesis are specially suitable for such real-time implementation since they have very low computational cost for training. Therefore, in the future the adaptive extensions of the proposed solutions and their performance in real-time scenarios need to be studied.

- In Chapter 6, the matrix-variate complex Gaussian model for the Fourier domain representation of the spatio-spectral features was presented. However, since there had been no prior study in the literature on this topic, most of the analysis in this chapter was devoted to providing a concrete definition and verifying it through statistical tests. In the future, practical applications of this model need to be explored with more detail. In particular, the proposed model can be used to develop a complex-valued extension of the matrix-to-vector LDA (MVLDA) method which can be used in conjunction with the Fourier transformation.

- In the literature, there has been some heuristic solutions for extension of different domain-specific FE and domain-agnostic FE methods, such as ICA or CSP, for complex-valued features. The complex-valued model presented in Chapter 6 can be used as a mathematical framework for theoretical derivation of the complex-valued counterparts of these methods that can guarantee a certain set of desired properties.

- Finally, characterization of the time-varying spatio-spectral features in Chapter 6 requires further investigation through statistical tests that directly examine the stationarity of the data. Both parametric and non-parametric methods can be used for this purpose. Non-parametric statistical tests have the advantage that they do not depend on any prior assumption about the data, but they require relatively larger sample size, whereas parametric methods are based on certain presumptions to provide more accurate result even in the case where the number of samples are limited. Considering the fact that motor imagery datasets usually include short epochs of motor imagination data, parametric tests are more suitable for future analysis of the stationarity of the data.

Appendix A

Appendices

A.1 Generalized Eigen-decomposition and its Properties

Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ be real-valued symmetric matrices of size $n \times n$. A generalized eigen-decomposition problem, is the problem of finding vectors $\mathbf{v}_i$ that satisfy the following equation:

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{B}\mathbf{v}_i \quad 1 \leq i \leq n \tag{A.1}$$

The following is a short review of the properties of the generalized eigenvalues and eigenvectors.

- A nonzero vector $\mathbf{v}_i$ satisfying (A.1) is called a *right generalized eigenvector* for the eigenvalue λ_i. Since $\mathbf{v}_i$ satisfies $\mathbf{v}_i T\mathbf{A} = \lambda_i \mathbf{v}_i^T \mathbf{B}$ as well, it is also called a *left generalized eigenvector* for λ_i.
- All generalized eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ are real-valued. As a result, they can be sorted as follows: $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.
- For distinct eigenvalues, the corresponding eigenvectors are unique, up to a scale.
- The eigenvectors $\mathbf{v}_i$ may not in general be unique, but they can be chosen to be "$\mathbf{B}$-*orthogonal*" to each other, i.e.,

$$\mathbf{v}_i^T \mathbf{B}\mathbf{v}_j = \mathbf{0}, \qquad \text{for} \quad i \neq j \tag{A.2}$$

- If both $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite matrices, then all the generalized eigenvalues are strictly positive, i.e, $\lambda_i > 0$. If only $\mathbf{B}$ is positive definite, and $\mathbf{A}$ is positive semidefinite[1] then $\lambda_i \geq 0$.

[1] Here, we assume that either $\mathbf{A}$ or $\mathbf{B}$ are positive semidefinite. For a more general discussion, refer to [168].

- Let $\Lambda = \text{diag}[\lambda_1, \lambda_2, \cdots, \lambda_n]$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$. Then, the n equalities in (A.1) can be expressed as

$$\mathbf{AV} = \mathbf{BV}\Lambda \tag{A.3}$$

Similarly, we have

$$\mathbf{V}^T\mathbf{BV} = \Lambda_B \tag{A.4}$$

where $\Lambda_B = \text{diag}[\lambda_{B,1}, \lambda_{B,2}, \cdots, \lambda_{B,n}]$ is a real-valued diagonal matrix. Therefore, it can be easily shown that

$$\mathbf{V}^T\mathbf{AV} = \Lambda_A \tag{A.5}$$

where $\Lambda_A = \Lambda_B\Lambda$ is also a real-valued diagonal matrix of the form $\Lambda_A = \text{diag}[\lambda_{A,1}, \lambda_{A,2}, \cdots, \lambda_{A,n}]$.

A.2 Kronecker Product Definition and its Properties

Consider $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{p \times q}$, where a_{ij} and b_{ij} are the (i,j)th elements of $\mathbf{A}$ and $\mathbf{B}$, respectively. The *Kronecker product* of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \otimes \mathbf{B}$, is the $mp \times nq$ matrix defined by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} = \tag{A.6}$$

$$\left[\begin{array}{cccc|cccc|c|cccc} a_{11}b_{11} & a_{11}b_{12} & \cdots & a_{11}b_{1q} & a_{12}b_{11} & a_{12}b_{12} & \cdots & a_{12}b_{1q} & \cdots & a_{1n}b_{11} & a_{1n}b_{12} & \cdots & a_{1n}b_{1q} \\ a_{11}b_{21} & a_{11}b_{22} & \cdots & a_{11}b_{2q} & a_{12}b_{21} & a_{12}b_{22} & \cdots & a_{12}b_{2q} & \cdots & a_{1n}b_{21} & a_{1n}b_{22} & \cdots & a_{1n}b_{2q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ a_{11}b_{p1} & a_{11}b_{p2} & \cdots & a_{11}b_{pq} & a_{12}b_{p1} & a_{12}b_{p2} & \cdots & a_{12}b_{pq} & \cdots & a_{1n}b_{p1} & a_{1n}b_{p2} & \cdots & a_{1n}b_{pq} \\ \hline a_{21}b_{11} & a_{21}b_{12} & \cdots & a_{21}b_{1q} & a_{22}b_{11} & a_{22}b_{12} & \cdots & a_{22}b_{1q} & \cdots & a_{2n}b_{11} & a_{2n}b_{12} & \cdots & a_{2n}b_{1q} \\ a_{21}b_{21} & a_{21}b_{22} & \cdots & a_{21}b_{2q} & a_{22}b_{21} & a_{22}b_{22} & \cdots & a_{22}b_{2q} & \cdots & a_{2n}b_{21} & a_{2n}b_{22} & \cdots & a_{2n}b_{2q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ a_{21}b_{p1} & a_{21}b_{p2} & \cdots & a_{21}b_{pq} & a_{22}b_{p1} & a_{22}b_{p2} & \cdots & a_{22}b_{pq} & \cdots & a_{2n}b_{p1} & a_{2n}b_{p2} & \cdots & a_{2n}b_{pq} \\ \hline \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \ddots & \vdots & \vdots & & \vdots \\ \hline a_{m1}b_{11} & a_{m1}b_{12} & \cdots & a_{m1}b_{1q} & a_{m2}b_{11} & a_{m2}b_{12} & \cdots & a_{m2}b_{1q} & \cdots & a_{mn}b_{11} & a_{mn}b_{12} & \cdots & a_{mn}b_{1q} \\ a_{m1}b_{21} & a_{m1}b_{22} & \cdots & a_{m1}b_{2q} & a_{m2}b_{21} & a_{m2}b_{22} & \cdots & a_{m2}b_{2q} & \cdots & a_{mn}b_{21} & a_{mn}b_{22} & \cdots & a_{mn}b_{2q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ a_{m1}b_{p1} & a_{m1}b_{p2} & \cdots & a_{m1}b_{pq} & a_{m2}b_{p1} & a_{m2}b_{p2} & \cdots & a_{m2}b_{pq} & \cdots & a_{mn}b_{p1} & a_{mn}b_{p2} & \cdots & a_{mn}b_{pq} \end{array}\right] \tag{A.7}$$

The Kronecker product has the following properties:

- For any two scalars α and β,

$$(\alpha\mathbf{A}) \otimes (\beta\mathbf{B}) = \alpha\beta(\mathbf{A} \otimes \mathbf{B}) \tag{A.8}$$

- Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times n}$. For any matrix $\mathbf{C}$, we have

$$(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = (\mathbf{A} \otimes \mathbf{C}) + (\mathbf{B} \otimes \mathbf{C}), \tag{A.9}$$

$$\mathbf{C} \otimes (\mathbf{A} + \mathbf{B}) = (\mathbf{C} \otimes \mathbf{A}) + (\mathbf{C} \otimes \mathbf{B}) \tag{A.10}$$

- For any three matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, we have

$$(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) \tag{A.11}$$

- For any two matrices $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T \tag{A.12}$$

- For $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{p \times q}, \mathbf{C} \in \mathbb{R}^{nr}$, and $\mathbf{D} \in \mathbb{R}^{q \times s}$, we have

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}) \tag{A.13}$$

- For nonsingular matrices $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} \tag{A.14}$$

- For $\mathbf{A} \in \mathbb{R}^{m \times m}$ and $\mathbf{B} \in \mathbb{R}^{n \times n}$,

$$\det(\mathbf{A} \otimes \mathbf{B}) = \det(\mathbf{A})^n \, \det(\mathbf{B})^m \tag{A.15}$$

- For $\mathbf{A} \in \mathbb{R}^{m \times m}$ and $\mathbf{B} \in \mathbb{R}^{m \times m}$,

$$\mathrm{tr}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{tr}(\mathbf{A}) \, \mathrm{tr}(\mathbf{B}) \tag{A.16}$$

A.3 Proof of Theorem 1

Let $\mathbf{\Sigma}_i = \mathbf{\Psi}_i \otimes \mathbf{\Phi}_i$ for $i \in \{1, 2\}$ and $\mathbf{W} = \mathbf{W}_R \otimes \mathbf{W}_L$, where $\mathbf{W}_R$ and $\mathbf{W}_L$ satisfy the following equations:

$$\mathbf{\Psi}_1 \mathbf{W}_R = (\mathbf{\Psi}_1 + \mathbf{\Psi}_2)\, \mathbf{W}_R \mathbf{\Lambda}_R, \tag{A.17}$$

$$\mathbf{\Phi}_1 \mathbf{W}_L = (\mathbf{\Phi}_1 + \mathbf{\Phi}_2)\, \mathbf{W}_L \mathbf{\Lambda}_L, \tag{A.18}$$

In this section, we will prove that

$$\mathbf{\Sigma}_1 \mathbf{W} = (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\, \mathbf{W} \mathbf{\Lambda}, \tag{A.19}$$

where $\mathbf{\Lambda} = (\mathbf{\Lambda}_R \otimes \mathbf{\Lambda}_L) \left(\mathbf{\Lambda}_R \otimes \mathbf{\Lambda}_L + (\mathbf{I}_{N_{ch}} - \mathbf{\Lambda}_R) \otimes \left(\mathbf{I}_{N_f} - \mathbf{\Lambda}_L\right)\right)^{-1}$.

From the discussions in A.1, it can be easily shown that $\mathbf{W}_R$ jointly diagonalizes both $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$. Similarly, $\mathbf{W}_L$ jointly diagonalizes both $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$. Therefore, we have

$$\mathbf{\Psi}_1 \mathbf{W}_R = \mathbf{W}_R \mathbf{\Lambda}_R^{(1)}, \tag{A.20}$$

$$\mathbf{\Psi}_2 \mathbf{W}_R = \mathbf{W}_R \mathbf{\Lambda}_R^{(2)}, \tag{A.21}$$

$$\mathbf{\Phi}_1 \mathbf{W}_L = \mathbf{W}_L \mathbf{\Lambda}_L^{(1)}, \tag{A.22}$$

$$\mathbf{\Phi}_1 \mathbf{W}_L = \mathbf{W}_L \mathbf{\Lambda}_L^{(2)}, \tag{A.23}$$

where $\mathbf{\Lambda}_R^{(i)}$ and $\mathbf{\Lambda}_L^{(i)}$ are the corresponding diagonal eigenvalue matrices. By substituting Equations (A.20)-(A.23) into (A.17) and (A.18), it can be shown that

$$\mathbf{\Lambda}_R^{(1)} = \left(\mathbf{\Lambda}_R^{(1)} + \mathbf{\Lambda}_R^{(2)}\right) \mathbf{\Lambda}_R, \tag{A.24}$$

$$\mathbf{\Lambda}_L^{(1)} = \left(\mathbf{\Lambda}_L^{(1)} + \mathbf{\Lambda}_L^{(2)}\right) \mathbf{\Lambda}_L, \tag{A.25}$$

or equivalently,

$$\mathbf{\Lambda}_R^{(1)} \left(\mathbf{\Lambda}_R^{(1)} + \mathbf{\Lambda}_R^{(2)}\right)^{-1} = \mathbf{\Lambda}_R, \tag{A.26}$$

$$\mathbf{\Lambda}_R^{(2)} \left(\mathbf{\Lambda}_R^{(1)} + \mathbf{\Lambda}_R^{(2)}\right)^{-1} = \mathbf{I}_{N_{ch}} - \mathbf{\Lambda}_R, \tag{A.27}$$

$$\mathbf{\Lambda}_L^{(1)} \left(\mathbf{\Lambda}_L^{(1)} + \mathbf{\Lambda}_L^{(2)}\right)^{-1} = \mathbf{\Lambda}_L, \tag{A.28}$$

$$\mathbf{\Lambda}_L^{(2)} \left(\mathbf{\Lambda}_L^{(1)} + \mathbf{\Lambda}_L^{(2)}\right)^{-1} = \mathbf{I}_{N_f} - \mathbf{\Lambda}_L, \tag{A.29}$$

Using these definitions, the left-hand side of Equation (A.19) can be expanded as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_1\mathbf{W} &= (\boldsymbol{\Psi}_1 \otimes \boldsymbol{\Phi}_1)(\mathbf{W}_R \otimes \mathbf{W}_L) && \text{(A.30)}\\
&= (\boldsymbol{\Psi}_1\mathbf{W}_R) \otimes (\boldsymbol{\Phi}_1\mathbf{W}_L) && \text{(A.31)}\\
&= \left(\mathbf{W}_R\boldsymbol{\Lambda}_R^{(1)}\right) \otimes \left(\mathbf{W}_L\boldsymbol{\Lambda}_L^{(1)}\right) && \text{(A.32)}\\
&= (\mathbf{W}_R \otimes \mathbf{W}_L)\left(\boldsymbol{\Lambda}_R^{(1)} \otimes \boldsymbol{\Lambda}_L^{(1)}\right) && \text{(A.33)}
\end{aligned}$$

Similarly, the right-hand side of Equation (A.19) can be expanded as follows:

$$\begin{aligned}
(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{W}\boldsymbol{\Lambda} &= (\mathbf{W}_R \otimes \mathbf{W}_L)\left(\boldsymbol{\Lambda}_R^{(1)} \otimes \boldsymbol{\Lambda}_L^{(1)}\right)\boldsymbol{\Lambda} + (\mathbf{W}_R \otimes \mathbf{W}_L)\left(\boldsymbol{\Lambda}_R^{(2)} \otimes \boldsymbol{\Lambda}_L^{(2)}\right)\boldsymbol{\Lambda} && \text{(A.34)}\\
&= (\mathbf{W}_R \otimes \mathbf{W}_L)\left(\boldsymbol{\Lambda}_R^{(1)} \otimes \boldsymbol{\Lambda}_L^{(1)} + \boldsymbol{\Lambda}_R^{(2)} \otimes \boldsymbol{\Lambda}_L^{(2)}\right)\boldsymbol{\Lambda} && \text{(A.35)}
\end{aligned}$$

Therefore, (A.33) is equal to (A.35) when

$$\begin{aligned}
\boldsymbol{\Lambda} &= \left(\boldsymbol{\Lambda}_R^{(1)} \otimes \boldsymbol{\Lambda}_L^{(1)}\right)\left(\boldsymbol{\Lambda}_R^{(1)} \otimes \boldsymbol{\Lambda}_L^{(1)} + \boldsymbol{\Lambda}_R^{(2)} \otimes \boldsymbol{\Lambda}_L^{(2)}\right)^{-1} && \text{(A.36)}\\
&= (\boldsymbol{\Lambda}_R \otimes \boldsymbol{\Lambda}_L)\left(\boldsymbol{\Lambda}_R \otimes \boldsymbol{\Lambda}_L + (\mathbf{I}_{N_{ch}} - \boldsymbol{\Lambda}_R) \otimes \left(\mathbf{I}_{N_f} - \boldsymbol{\Lambda}_L\right)\right)^{-1} && \text{(A.37)}
\end{aligned}$$

where (A.37) is obtained by multiplying all $\boldsymbol{\Lambda}_R^{(i)}$ and $\boldsymbol{\Lambda}_L^{(i)}$ matrices in (A.36) by $\left(\boldsymbol{\Lambda}_R^{(1)} + \boldsymbol{\Lambda}_R^{(2)}\right)^{-1}$.

A.4 A Sufficient Condition for Separability of Spatio-Spectral Features

Theorem 2: Let $s(t, c|\Omega_i)$ denote the EEG signal recorded at time t from channel c. Without loss of generality, we assume that $s(t, c|\Omega_i)$ is zero mean, i.e., $E\{s(t, c|\Omega_i)\} = 0$. Also, let $z(t, f, c|\Omega_i)$ and $\check{z}(t, f, c|\Omega_i)$ respectively represent the spatio-spectral features obtained through short-time Fourier transformation and through bandpass filtering of $s(t, c|\Omega_i)$, defined as follows

$$z(t, f, c|\Omega_i) = \int_{-\infty}^{\infty} s(\tau, c|\Omega_i) w(\tau - t) e^{-j2\pi f \tau} d\tau \quad (A.38)$$

$$\check{z}(t, f, c|\Omega_i) = s(t, c|\Omega_i) * h_f(t) \quad (A.39)$$

where $w(t)$ denotes the STFT window, and $h_f(t)$ denotes the bandpass filter's impulse response.

The following condition is a sufficient condition for separability of both $z(t, f, c|\Omega_i)$ and $\check{z}(t, f, c|\Omega_i)$:

$$E\{s(t_1, c_1|\Omega_i)\, s(t_2, c_2|\Omega_i)\} = \rho(t_1, t_2|\Omega_i)\, \psi(c_1, c_2|\Omega_i) \quad (A.40)$$

where $\psi(c_1, c_2|\Omega_i)$ can be considered as spatial covariance between channels c_1 and c_2 during task Ω_i. Similarly, $\rho(t_1, t_2|\Omega_i)$ can be considered as temporal covariance, or correlation, between time instances t_1 and t_2.

Moreover, If the condition in (A.40) is satisfied, then we will have

$$E\{z(t, f_1, c_1|\Omega_i)\, z(t, f_2, c_2|\Omega_i)\} = \phi(t, f_1, f_2|\Omega_i)\, \psi(c_1, c_2|\Omega_i) \quad (A.41)$$

$$E\{z(t, f_1, c_1|\Omega_i)\, z^*(t, f_2, c_2|\Omega_i)\} = \phi(t, f_1, -f_2|\Omega_i)\, \psi(c_1, c_2|\Omega_i) \quad (A.42)$$

$$E\{\check{z}(t, f_1, c_1|\Omega_i)\, \check{z}(t, f_2, c_2|\Omega_i)\} = \check{\phi}(t, f_1, f_2|\Omega_i)\, \psi(c_1, c_2|\Omega_i) \quad (A.43)$$

where $\phi(t, f_1, f_2|\Omega_i)$ can be considered as the spectral covariance between frequency components (or bands) f_1 and f_2 at time t. Similarly, $\phi(t, f_1, f_2|\Omega_i)$ denotes the spectral pseudo-covariance between f_1 and f_2 at time t.

Proof For STFT Features: For spatio-spectral features obtained through short-time Fourier trans-

formation, we have

$$
\begin{aligned}
&E\big\{z(t,f_1,c_1|\Omega_i)\; z(t,f_2,c_2|\Omega_i)\big\} \\
&\quad = \quad E\Bigg\{ \left(\int_{-\infty}^{\infty} s(\tau_1,c_1|\Omega_i)\; w(\tau_1-t)\; e^{-j2\pi f_1\tau_1}\; d\tau_1\right)\cdot\left(\int_{-\infty}^{\infty} s(\tau_2,c_2|\Omega_i)\; w(\tau_2-t)\; e^{-j2\pi f_2\tau_2}\; d\tau_2\right)\Bigg\} \\
&\quad = \quad E\Bigg\{ \iint_{-\infty}^{\infty} s(\tau_1,c_1|\Omega_i)\; s(\tau_2,c_2|\Omega_i)\; w(\tau_1-t)\; w(\tau_2-t)\; e^{-j2\pi(f_1\tau_1+f_2\tau_2)}\; d\tau_1\; d\tau_2\Bigg\} \\
&\quad = \quad \iint_{-\infty}^{\infty} E\big\{s(\tau_1,c_1|\Omega_i)\; s(\tau_2,c_2|\Omega_i)\big\}\; w(\tau_1-t)\; w(\tau_2-t)\; e^{-j2\pi(f_1\tau_1+f_2\tau_2)}\; d\tau_1\; d\tau_2 \\
&\quad = \quad \iint_{-\infty}^{\infty} \rho(\tau_1,\tau_2|\Omega_i)\; \psi(c_1,c_2|\Omega_i)\; w(\tau_1-t)\; w(\tau_2-t)\; e^{-j2\pi(f_1\tau_1+f_2\tau_2)}\; d\tau_1\; d\tau_2 \\
&\quad = \quad \psi(c_1,c_2|\Omega_i)\iint_{-\infty}^{\infty} \rho(\tau_1,\tau_2|\Omega_i)\; w(\tau_1-t)\; w(\tau_2-t)\; e^{-j2\pi(f_1\tau_1+f_2\tau_2)}\; d\tau_1\; d\tau_2 \\
&\quad = \quad \psi(c_1,c_2|\Omega_i)\; \phi(t,f_1,f_2|\Omega_i) \qquad\qquad (A.44)
\end{aligned}
$$

where $\phi(t,f_1,f_2|\Omega_i)$ represents the two-dimensional (or bivariate) short-time Fourier transformation of $\rho(t_1,t_2|\Omega_i)$ evaluated at time $t_1=t_2=t$, i.e.,

$$
\phi(t,f_1,f_2|\Omega_i) \quad = \quad \iint_{-\infty}^{\infty} \rho(\tau_1,\tau_2|\Omega_i)\; w(\tau_1-t)\; w(\tau_2-t)\; e^{-j2\pi(f_1\tau_1+f_2\tau_2)}\; d\tau_1\; d\tau_2 \qquad (A.45)
$$

Similarly, it can be shown that

$$
\begin{aligned}
&E\big\{z(t,f_1,c_1|\Omega_i)\; z^*(t,f_2,c_2|\Omega_i)\big\} \\
&\quad = \quad \psi(c_1,c_2|\Omega_i)\iint_{-\infty}^{\infty} \rho(\tau_1,\tau_2|\Omega_i)\; w(\tau_1-t)\; w(\tau_2-t)\; e^{-j2\pi(f_1\tau_1-f_2\tau_2)}\; d\tau_1\; d\tau_2 \\
&\quad = \quad \psi(c_1,c_2|\Omega_i)\; \phi(t,f_1,-f_2|\Omega_i) \qquad\qquad (A.46)
\end{aligned}
$$

Note that Equation (A.44) represents the pseudo-covariance between the complex-valued signals $z(t,f_1,c_1|\Omega_i)$ and $z(t,f_2,c_2|\Omega_i)$, whereas Equation (A.46) represents the covariance between these two signals.

Proof For Bandpass Filter Features: For spatio-spectral features obtained through bandpass filter-

ing (or any other similar filtering) the EEG signal, we have

$$
\begin{aligned}
& E\big\{\check{z}(t, f_1, c_1|\Omega_i)\ \check{z}(t, f_2, c_2|\Omega_i)\big\} \\
&\quad = E\Bigg\{ \bigg(\int_{-\infty}^{\infty} s(\tau_1, c_1|\Omega_i)\ h_{f_1}(t-\tau_1)\ d\tau_1\bigg) \cdot \bigg(\int_{-\infty}^{\infty} s(\tau_2, c_2|\Omega_i)\ h_{f_2}(t-\tau_2)\ d\tau_2\bigg) \Bigg\} \\
&\quad = E\Bigg\{ \iint_{-\infty}^{\infty} s(\tau_1, c_1|\Omega_i)\ s(\tau_2, c_2|\Omega_i)\ h_{f_1}(t-\tau_1)\ h_{f_2}(t-\tau_2)\ d\tau_1\ d\tau_2 \Bigg\} \\
&\quad = \iint_{-\infty}^{\infty} E\big\{ s(\tau_1, c_1|\Omega_i)\ s(\tau_2, c_2|\Omega_i)\big\}\ h_{f_1}(t-\tau_1)\ h_{f_2}(t-\tau_2)\ d\tau_1\ d\tau_2 \\
&\quad = \iint_{-\infty}^{\infty} \rho(\tau_1, \tau_2|\Omega_i)\ \psi(c_1, c_2|\Omega_i)\ h_{f_1}(t-\tau_1)\ h_{f_2}(t-\tau_2)\ d\tau_1\ d\tau_2 \\
&\quad = \psi(c_1, c_2|\Omega_i) \iint_{-\infty}^{\infty} \rho(\tau_1, \tau_2|\Omega_i)\ h_{f_1}(t-\tau_1)\ h_{f_2}(t-\tau_2)\ d\tau_1\ d\tau_2 \\
&\quad = \psi(c_1, c_2|\Omega_i)\ \check{\phi}(t, f_1, f_2|\Omega_i) \qquad \text{(A.47)}
\end{aligned}
$$

where $\check{\phi}(t, f_1, f_2|\Omega_i)$ denotes the two-dimensional (or bivariate) convolution of $\rho(t_1, t_2|\Omega_i)$ with the function $H_{f_1,f_2}(t_1, t_2) = h_{f_1}(t-\tau_1)\ h_{f_2}(t-\tau_2)$, evaluated at time $t_1 = t_2 = t$, i.e.,

$$
\check{\phi}(t, f_1, f_2|\Omega_i) \quad = \quad \iint_{-\infty}^{\infty} \rho(\tau_1, \tau_2|\Omega_i)\ h_{f_1}(t-\tau_1)\ h_{f_2}(t-\tau_2)\ d\tau_1\ d\tau_2 \qquad \text{(A.48)}
$$

Bibliography

[1] "The brain from top to bottom," available online at `http://thebrain.mcgill.ca/flash/index_d.html`.

[2] J. Malmivuo and R. Plonsey, *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields*. New York: Oxford University Press, 1995, available online: `http://www.bem.fi/book/in/in.htm`.

[3] C. J. Stam, B. Jelles, H. A. M. Achtereekte, S. A. R. B. Rombouts, J. P. J. Slaets, and R. W. M. Keunen, "Investigation of EEG non-linearity in dementia and parkinson's disease," *Electroencephalography and clinical neurophysiology*, vol. 95, no. 5, pp. 309–317, 1995.

[4] L. Bonanni, A. Thomas, P. Tiraboschi, B. Perfetti, S. Varanese, and M. Onofrj, "EEG comparisons in early alzheimer's disease, dementia with lewy bodies and parkinson's disease with dementia patients with a 2-year follow-up," *Brain*, vol. 131, no. 3, pp. 690–705, 2008.

[5] J. Gotman, "Automatic recognition of epileptic seizures in the EEG," *Electroencephalography and clinical neurophysiology*, vol. 54, no. 5, pp. 530–540, 1982.

[6] M. L. V. Quyen, J. Martinerie, V. Navarro, P. Boon, M. D'Hav, C. Adam, B. Renault, F. Varela, and M. Baulac, "Anticipation of epileptic seizures from standard EEG recordings," *Lancet*, vol. 357, no. 9251, pp. 183–188, 2001.

[7] H. Ocak, "Optimal classification of epileptic seizures in EEG using wavelet analysis and genetic algorithm," *Signal Processing*, vol. 88, no. 7, pp. 1858–1867, 2008.

[8] C. Grnling, M. Ligges, R. Huonker, M. Klingert, H. J. Mentzel, R. Rzanny, W. A. Kaiser, H. Witte, and B. Blanz, "Dyslexia: The possible benefit of multimodal integration of fMRI- and EEG-data," *Journal of neural transmission*, vol. 111, no. 7, pp. 951–969, 2004.

[9] L. M. Oberman, E. M. Hubbard, J. P. McCleery, E. L. Altschuler, V. S. Ramachandran, and J. A. Pineda, "EEG evidence for mirror neuron dysfunction in autism spectrum disorders," *Cognitive Brain Research*, vol. 24, no. 2, pp. 190–198, 2005.

[10] S.-K. Yeom, H.-I. Suk, and S.-W. Lee, "Person authentication from neural activity of face-specific visual self-representation," *Pattern Recognition*, vol. 46, no. 4, pp. 1159–1169, 2013.

[11] S. Marcel and J. Millan, "Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 743–748, 2007.

[12] R. Palaniappan and D. P. Mandic, "Biometrics from brain electrical activity: A machine learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 738–742, 2007.

[13] R. Palaniappan, *Electroencephalogram signals from imagined activities: A novel biometric identifier for a small population.* Lecture Notes in Computer Science, 2006, vol. 4224 LNCS.

[14] R. B. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles, "The electroencephalogram as a biometric," in *Canadian Conference on Electrical and Computer Engineering*, vol. 2, 2001, pp. 1363–1366.

[15] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[16] Dornhege, Milla, Hinterberger, McFarland, and Muller, *Toward Brain-Computer Interfacing.* The MIT Press, 2007.

[17] D. Huang, K. Qian, D.-Y. Fei, W. Jia, X. Chen, and O. Bai, "Electroencephalography (EEG)-based brain-computer interface (BCI): A 2-D virtual wheelchair control based on event-related desynchronization/ synchronization and state control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 3, pp. 379–388, 2012.

[18] G. Muller-Putz, R. Scherer, G. Pfurtscheller, and R. Rupp, "EEG-based neuroprosthesis control: A step towards clinical practice," *Neuroscience Letters*, vol. 382, no. 1-2, pp. 169–174, 2005.

[19] R. Scherer, G. Muller, C. Neuper, B. Graimann, and G. Pfurtscheller, "An asynchronously controlled EEG-based virtual keyboard: Improvement of the spelling rate," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 979–984, 2004.

[20] A. D. Gerson, L. C. Parra, and P. Sajda, "Cortically coupled computer vision for rapid image search," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 174–179, 2006.

[21] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda, "Response error correction - a demonstration of improved human-machine performance using real-time EEG monitoring," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 173–177, 2003.

[22] L. C. Parra, C. Christoforou, A. D. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. G. Philiastides, and P. Sajda, "Spatiotemporal linear decoding of brain state," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 107–115, 2008.

[23] J. Wang, E. Pohlmeyer, B. Hanna, Y. Jiang, P. Sajda, and S. Chang, "Brain state decoding for rapid image retrieval," in *ACM Int. Conf. on Multimedia*, Beijing, China, Oct. 2009, pp. 945–954.

[24] G. Pfurtscheller, R. Leeb, C. Keinrath, D. Friedman, C. Neuper, C. Guger, and M. Slater, "Walking from thought," *Brain research*, vol. 1071, no. 1, pp. 145–152, 2006.

[25] Q. Zhao, L. Zhang, and A. Cichocki, "EEG-based asynchronous BCI control of a car in 3D virtual reality environments," *Chinese Science Bulletin*, vol. 54, no. 1, pp. 78–87, 2009.

[26] P. L. Nunez and R. Srinivasan, *Electric fields of the brain : the neurophysics of EEG*, 2nd ed. New York: Oxford University Press, 2006.

[27] G. Pfurtscheller, B. Z. Allison, C. Brunner, G. Bauernfeind, T. Solis-Escalante, R. Scherer, T. O. Zander, G. Mueller-Putz, C. Neuper, and N. Birbaumer, "The hybrid BCI," *Frontiers in neuroscience*, vol. 4, 2010.

[28] S. Fazli, J. Mehnert, J. Steinbrink, G. Curio, A. Villringer, K. R. Muller, and B. Blankertz, "Enhanced performance by a hybrid NIRS-EEG brain computer interface," *NeuroImage*, vol. 59, no. 1, pp. 519–529, 2012.

[29] Y. Liu, H. Ayaz, A. Curtin, B. Onaral, and P. A. Shewokis, *Towards a hybrid P300-based BCI using simultaneous fNIR and EEG*, ser. Lecture Notes in Computer Science, 2013, vol. 8027 LNAI.

[30] M. Jeannerod and V. Frak, "Mental imaging of motor activity in humans," *Current opinion in neurobiology*, vol. 9, no. 6, pp. 735–739, 1999.

[31] L. Kauhanen, P. Jylanki, J. Lehtonen, P. Rantanen, H. Alaranta, and M. Sams, "EEG-based brain-computer interface for tetraplegics," *Computational Intelligence and Neuroscience*, vol. 2007, 2007.

[32] R. Leeb, S. Perdikis, L. Tonin, A. Biasiucci, M. Tavella, M. Creatura, A. Molina, A. Al-Khodairy, T. Carlson, and J. D. Millan, "Transferring brain-computer interfaces beyond the laboratory: Successful application control for motor-disabled users," *Artificial Intelligence in Medicine*, vol. 59, no. 2, pp. 121–132, 2013.

[33] A. Kubler and N. Birbaumer, "Brain-computer interfaces and communication in paralysis: Extinction of goal directed thinking in completely paralysed patients?" *Clinical Neurophysiology*, vol. 119, no. 11, pp. 2658–2666, 2008.

[34] M. Conson, S. Sacco, M. Sar, F. Pistoia, D. Grossi, and L. Trojano, "Selective motor imagery defect in patients with locked-in syndrome," *Neuropsychologia*, vol. 46, no. 11, pp. 2622–2628, 2008.

[35] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[36] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, 2000.

[37] G. Pfurtscheller and C. Neuper, "Motor imagery direct communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[38] K. J. Miller, E. C. Leuthardt, G. Schalk, R. P. N. Rao, N. R. Anderson, D. W. Moran, J. W. Miller, and J. G. Ojemann, "Spectral changes in cortical surface potentials during motor movement," *Journal of Neuroscience*, vol. 27, no. 9, pp. 2424–2432, 2007.

[39] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, CA, USA: Academic Press Professional Inc., 1990.

[40] M. Arvaneh, C. Guan, K.K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in EEG-based BCI," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1865–1873, 2011.

[41] C. Brunner, M. Naeem, R. Leeb, B. Graimann, and G. Pfurtscheller, "Spatial filtering and selection of optimized components in four class motor imagery EEG data using independent components analysis," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 957–964, 2007.

[42] F. Cincotti, D. Mattia, C. Babiloni, F. Carducci, S. Salinari, L. Bianchi, M. Marciani, and F. Babiloni, "The use of EEG modifications due to motor imagery for brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 131–133, 2003.

[43] G. Rodriguez-Bermudez and P. Garcia-Laencina, "Automatic and adaptive classification of electroencephalographic signals for brain computer interfaces," *Journal of Medical Systems*, pp. 1–13, 2012, article in Press.

[44] S. Hu, Q. Tian, Y. Cao, J. Zhang, and W. Kong, "Motor imagery classification based on joint regression model and spectral power," *Neural Computing and Applications*, pp. 1–6, 2012, article in Press.

[45] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 4, pp. 317–326, 2008.

[46] G. Fabiani, D. McFarland, J. Wolpaw, and G. Pfurtscheller, "Conversion of EEG activity into cursor movement by a brain-computer interface (BCI)," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 12, no. 3, pp. 331–338, 2004.

[47] D. Garrett, D. Peterson, C. Anderson, and M. Thaut, "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 141–144, 2003.

[48] F. Babiloni, F. Cincotti, L. Bianchi, G. Pirri, J. Del R. Millan, J. Mourino, S. Salinari, and M. Grazia Marciani, "Recognition of imagined hand movements with low resolution surface Laplacian and linear classifiers," *Medical Engineering and Physics*, vol. 23, no. 5, pp. 323–328, 2001.

[49] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers in Neuroscience*, no. MAR, 2012.

[50] L. Astolfi, F. Cincotti, D. Mattia, F. D. V. Fallani, A. Tocci, A. Colosimo, S. Salinari, M. G. Marciani, W. Hesse, H. Witte, M. Ursino, M. Zavaglia, and F. Babiloni, "Tracking the time-varying cortical connectivity patterns by adaptive multivariate estimators," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 902–913, 2008.

[51] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541–1548, 2005.

[52] C. Gerloff, J. Richard, J. Hadley, A. Schulman, M. Honda, and M. Hallett, "Functional coupling and regional activation of human cortical motor areas during simple, internally paced and externally paced finger movements," *Brain*, vol. 121, no. 8, pp. 1513–1531, 1998.

[53] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals," *Journal of Neural Engineering*, vol. 4, no. 2, 2007.

[54] "Taxonomy of Feature Extraction and Translation Methods for BCI," available online at `http://www.cs.colostate.edu/ eeg/taxonomy.html`.

[55] H. Xiong, M. Swamy, and M. Ahmad, "Two-dimensional FLD for face recognition," *Pattern Recognition*, vol. 38, no. 7, pp. 1121 – 1124, 2005.

[56] M. Li and B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527 – 532, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865504002272

[57] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature reduction via generalized uncorrelated linear discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1312 –1322, oct. 2006.

[58] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press, 2005, pp. 1569–1576.

[59] J. Yang, D. Zhang, X. Yong, and J. yu Yang, "Two-dimensional discriminant transform for face recognition," *Pattern Recognition*, vol. 38, no. 7, pp. 1125 – 1129, 2005.

[60] K. Inoue and K. Urahama, "Non-iterative two-dimensional linear discriminant analysis," in *18th Int. Conf. Pattern Recognition (ICPR)*, vol. 2, 2006, pp. 540 –543.

[61] S. Chowdhury, J. Sing, D. Basu, and M. Nasipuri, "Face recognition by generalized two-dimensional FLD method and multi-class support vector machines," *Applied Soft Computing Journal*, vol. 11, no. 7, pp. 4282–4292, 2011.

[62] S. Noushath, G. H. Kumar, and P. Shivakumara, "(2D)2LDA: An efficient approach for face recognition," *Pattern Recognition*, vol. 39, no. 7, pp. 1396 – 1400, 2006.

[63] J. Zhao, P. Yu, L. Shi, and S. Li, "Separable linear discriminant analysis," *Computational Statistics and Data Analysis*, vol. 56, no. 12, pp. 4290–4300, 2012.

[64] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Muller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 11, pp. 2274–2281, 2006.

[65] W. Wu, X. Gao, B. Hong, and S. Gao, "Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL)," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 6, pp. 1733–1743, 2008.

[66] H. Zhang, Z. Y. Chin, K. K. Ang, C. Guan, and C. Wang, "Optimum spatio-spectral filtering network for brain-computer interface," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 52–63, 2011.

[67] H.-I. Suk and S.-W. Lee, "A probabilistic approach to spatio-spectral filters optimization in brain-computer interface," in *2011 IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, October 2011, pp. 19–24.

[68] J. Huxter, N. Burgess, and J. O'Keefe, "Independent rate and temporal coding in hippocampal pyramidal cells," *Nature*, vol. 425, no. 6960, pp. 828–832, 2003.

[69] P. Fries, D. Nikolic, and W. Singer, "The gamma cycle," *Trends in neurosciences*, vol. 30, no. 7, pp. 309–316, 2007.

[70] N. A. Busch, J. Dubois, and R. VanRullen, "The phase of ongoing EEG oscillations predicts visual perception," *Journal of Neuroscience*, vol. 29, no. 24, pp. 7869–7876, 2009.

[71] T. Masquelier, E. Hugues, G. Deco, and S. J. Thorpe, "Oscillations, phase-of-firing coding, and spike timing-dependent plasticity: An efficient learning scheme," *Journal of Neuroscience*, vol. 29, no. 43, pp. 13 484–13 493, 2009.

[72] A. S. Aghaei, M. S. Mahanta, and K. N. Plataniotis, "A general framework for spatio-spectral feature extraction in motor imagery BCI systems," *IEEE Trans. Biomed. Eng.*, submitted to.

[73] M. S. Mahanta, A. S. Aghaei, and K. N. Plataniotis, "A bayes optimal matrix-to-vector approach to 2DLDA," *Pattern Recognition*, submitted to.

[74] A. S. Aghaei, M. S. Mahanta, and K. N. Plataniotis, "Separable common spatio-spectral pattern algorithm for classification of EEG signals," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing.*, 2013, pp. 988–992.

[75] ——, "Separable common spatio-spectral patterns for brain computer interface systems," *IEEE Trans. Biomed. Eng.*, submitted to.

[76] A. S. Aghaei, M. S. Mahanta, K. N. Plataniotis, and S. Pasupathy, "Statistical characterization of complex-valued EEG spectrum during mental imagery tasks," in *33rd Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*, 2011.

[77] A. S. Aghaei, K. N. Plataniotis, and S. Pasupathy, "A time-varying gaussian model for the complex-valued EEG spectrum during mental imagery tasks," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing.*, 2012, pp. 669–672.

[78] S. Shoham, E. Halgren, E. M. Maynard, and R. A. Normann, "Motor-cortical activity in tetraplegics," *Nature*, vol. 413, no. 6858, p. 793, 2001.

[79] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, no. 7099, pp. 164–171, 2006.

[80] D. Mattia, F. Cincotti, L. Astolfi, F. de Vico Fallani, G. Scivoletto, M. G. Marciani, and F. Babiloni, "Motor cortical responsiveness to attempted movements in tetraplegia: Evidence from neuroelectrical imaging," *Clinical Neurophysiology*, vol. 120, no. 1, pp. 181–189, 2009.

[81] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *NeuroImage*, vol. 34, no. 4, pp. 1600–1611, 2007.

[82] "BCI competitions," available online at `http://www.bbci.de/competition/`.

[83] T. P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects," *Clinical Neurophysiology*, vol. 111, no. 10, pp. 1745–1758, 2000.

[84] R. Vigrio, J. Srel, V. Jousmaki, M. Hmlinen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 5, pp. 589–593, 2000.

[85] G. Buzsaki, *Rhythms of the brain.* Oxford University Press, 2006.

[86] C. Neuper and G. Pfurtscheller, "Event-related dynamics of cortical rhythms: Frequency-specific features and functional correlates," *International Journal of Psychophysiology*, vol. 43, no. 1, pp. 41–58, 2001.

[87] G. Pfurtscheller, C. Brunner, A. Schlogl, and F. H. Lopes da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, 2006.

[88] S. L. Gonzalez, R. G. de Peralta, G. Thut, J. R. Milln, P. Morier, and T. Landis, "Very high frequency oscillations (VHFO) as a predictor of movement intentions." *NeuroImage*, vol. 32, no. 1, pp. 170–179, 2006.

[89] F. Perrin, J. Pernier, O. Bertrand, and J. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalography and Clinical Neurophysiology*, vol. 72, no. 2, pp. 184–187, 1989.

[90] G. Wahba, "Spline interpolation and smoothing on the sphere," *SIAM Journal on Scientific and Statistical Computing*, vol. 2, no. 1, pp. 5–16, 1981.

[91] P. Nunez, R. Silberstein, P. Cadusch, R. Wijesinghe, A. Westdorp, and R. Srinivasan, "A theoretical and experimental study of high resolution EEG based on surface laplacians and cortical imaging," *Electroencephalography and Clinical Neurophysiology*, vol. 90, no. 1, pp. 40–57, 1994.

[92] J. Kayser and C. Tenke, "Principal components analysis of laplacian waveforms as a generic method for identifying ERP generator patterns: I. evaluation with auditory oddball tasks," *Clinical Neurophysiology*, vol. 117, no. 2, pp. 348–368, 2006.

[93] ——, "Principal components analysis of laplacian waveforms as a generic method for identifying ERP generator patterns: II. adequacy of low-density estimates," *Clinical Neurophysiology*, vol. 117, no. 2, pp. 369–380, 2006.

[94] J. Kayser, "Current source density (CSD) interpolation using spherical splines - CSD Toolbox," available online: `http://psychophysiology.cpmc.columbia.edu/Software/CSDtoolbox`, 2009.

[95] V. Garoosi and B. H. Jansen, "Development and evaluation of the piecewise prony method for evoked potential analysis," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 12, pp. 1549–1554, 2000, tradenames: piecewise prony method; Cited By (since 1996): 15.

[96] O. David and K. J. Friston, "A neural mass model for meg/eeg: Coupling and neuronal dynamics," *NeuroImage*, vol. 20, no. 3, pp. 1743–1755, 2003.

[97] C. J. Stam, "Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field," *Clinical Neurophysiology*, vol. 116, no. 10, pp. 2266–2301, 2005.

[98] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[99] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*, 2nd ed. Springer, 1991.

[100] K. P. Burnham and D. R. Anderson, *Model selection and multi-model inference: a practical information-theoretic approach*. Springer, 2002.

[101] A. Subasi, "Selection of optimal AR spectral estimation method for EEG signals using cramer-rao bound," *Computers in biology and medicine*, vol. 37, no. 2, pp. 183–194, 2007.

[102] B. H. Jansen, J. R. Bourne, and J. W. Ward, "Autoregressive estimation of short segment spectra for computerized EEG analysis," *IEEE Trans. Biomed. Eng.*, vol. 28, no. 9, pp. 630–638, 1981.

[103] J. S. Barlow, "Methods of analysis of nonstationary EEGs, with emphasis on segmentation techniques: A comparative review," *Journal of Clinical Neurophysiology*, vol. 2, no. 3, pp. 267–304, 1985.

[104] P. J. Franaszczuk, K. J. Blinowska, and M. Kowalczyk, "The application of parametric multichannel spectral estimates in the study of electrical brain activity," *Biological cybernetics*, vol. 51, no. 4, pp. 239–247, 1985.

[105] C. W. Anderson, E. A. Stolz, and S. Shamsunder, "Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 277–286, 1998.

[106] J. R. Wolpaw, D. J. McFarland, T. M. Vaughan, and G. Schalk, "The wadsworth center brain - computer interface (BCI) research and development program," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 204–207, 2003.

[107] J. D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, no. 7, pp. 523–534, 2006.

[108] B. Blankertz, G. Dornhege, C. Schafer, R. Krepki, J. Kohlmorgen, K. R. Muller, 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 Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 127–131, 2003.

[109] B. Blankertz, G. Dornhege, M. Krauledat, K. R. Muller, V. Kunzmann, F. Losch, and G. Curio, "The berlin brain-computer interface: EEG-based communication without subject training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 147–152, 2006.

[110] J. Wolpaw and E. W. Wolpaw, *Brain-computer interfaces: principles and practice*. Oxford University Press, 2012.

[111] E. Gysels and P. Celka, "Phase synchronization for the recognition of mental tasks in a brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 12, no. 4, pp. 406–415, 2004.

[112] X. Bao and J. Hu, "Phase synchronization for classification of motor imagery EEG," *Journal of Information and Computational Science*, vol. 5, no. 2, pp. 949–955, 2008.

[113] Q. Wei, Y. Wang, X. Gao, and S. Gao, "Amplitude and phase coupling measures for feature extraction in an EEG-based brain-computer interface," *Journal of Neural Engineering*, vol. 4, no. 2, pp. 120–129, 2007.

[114] J. Ito, A. R. Nikolaev, and C. V. Leeuwen, "Spatial and temporal structure of phase synchronization of spontaneous alpha EEG activity," *Biological cybernetics*, vol. 92, no. 1, pp. 54–60, 2005.

[115] S. Pockett, G. E. J. Bold, and W. J. Freeman, "EEG synchrony during a perceptual-cognitive task: Widespread phase synchrony at all frequencies," *Clinical Neurophysiology*, vol. 120, no. 4, pp. 695–708, 2009.

[116] K. Sekihara and B. Scholz, "Generalized wiener estimation of three-dimensional current distribution from biomagnetic measurements," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 3, pp. 281–291, 1996.

[117] B. D. Van Veen, W. Van Drongelen, M. Yuchtman, and A. Suzuki, "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 9, pp. 867–880, 1997.

[118] J. Gross and A. A. Ioannides, "Linear transformations of data space in MEG," *Physics in Medicine and Biology*, vol. 44, no. 8, pp. 2081–2097, 1999.

[119] M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss, "Beamforming in noninvasive brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1209–1219, 2009.

[120] M. Ahn, J. H. Hong, and S. C. Jun, "Feasibility of approaches combining sensor and source features in brain-computer interface," *Journal of neuroscience methods*, vol. 204, no. 1, pp. 168–178, 2012.

[121] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Muller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, 2008.

[122] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Muller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993–1002, 2004.

[123] Y. Wang, S. Gao, and X. Gao, "Common spatial pattern method for channel selection in motor imagery based brain-computer interface," in *2005 27th Annual International Conference of the Engineering in Medicine and Biology Society, IEEE-EMBS 2005*, vol. 7 VOLS, Sep. 2005, pp. 5392–5395.

[124] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial patterns with generic learning for EEG signal classification," in *2009 31th Annual International Conference of the Engineering in Medicine and Biology Society, IEEE-EMBS 2009*, Sep. 2009, pp. 6599–6602.

[125] J. Kalcher, D. Flotzinger, C. Neuper, S. Golly, and G. Pfurtscheller, "Graz brain-computer interface II: Towards communication between humans and computers based on online classification of three different EEG patterns," *Medical and Biological Engineering and Computing*, vol. 34, no. 5, pp. 382–388, 1996.

[126] N. Hazarika, J. Z. Chen, A. C. Tsoi, and A. Sergejew, "Classification of EEG signals using the wavelet transform," *Signal Processing*, vol. 59, no. 1, pp. 61–72, 1997.

[127] V. J. Samar, A. Bopardikar, R. Rao, and K. Swartz, "Wavelet analysis of neuroelectric waveforms: A conceptual tutorial," *Brain and language*, vol. 66, no. 1, pp. 7–60, 1999.

[128] V. Bostanov, "BCI competition 2003 - data sets Ib and IIb: Feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1057–1061, 2004.

[129] E. Sitnikova, A. E. Hramov, A. A. Koronovsky, and G. van Luijtelaar, "Sleep spindles and spike-wave discharges in EEG: Their generic features, similarities and distinctions disclosed with fourier transform and continuous wavelet analysis," *Journal of neuroscience methods*, vol. 180, no. 2, pp. 304–316, 2009.

[130] M. J. Kaminski and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," *Biological cybernetics*, vol. 65, no. 3, pp. 203–210, 1991.

[131] M. Kaminski, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[132] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8659 – 8666, 2010.

[133] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs," *Pattern Recognition*, vol. 45, no. 6, pp. 2137–2144, 2012.

[134] G. Coelho, C. Barbante, L. Boccato, R. Attux, J. Oliveira, and F. Von Zuben, "Automatic feature selection for BCI: An analysis using the davies-bouldin index and extreme learning machines," in *Int. Joint Conf. Neural Networks*, 2012.

[135] R. Corralejo, R. Hornero, and D. Alvarez, "Feature selection using a genetic algorithm in a motor imagery-based brain computer interface," in *2011 Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2011, pp. 7703–7706.

[136] P. E. H. Richard O. Duda and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley, 2001.

[137] O. Hamsici and A. Martinez, "Bayes optimality in linear discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 647 –657, april 2008.

[138] F. Lotte, M. Congedo, A. Lcuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, pp. R1–R13, 2007.

[139] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[140] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, 1997.

[141] A. Gupta and D. Nagar, *Matrix variate distributions*. Chapman & Hall/CRC, 1999.

[142] W.-S. Zheng, J.-H. Lai, and S. Z. Li, "1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based?" *Pattern Recognition*, pp. 2156–2172, 2008.

[143] Z. Liang, Y. Li, and P. Shi, "A note on two-dimensional linear discriminant analysis," *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2122 – 2128, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865508002377

[144] B. Yang and Y. Bu, "A comparative study on vector-based and matrix-based linear discriminant analysis," *Journal of Computers*, vol. 6, no. 4, pp. 818–824, 2011.

[145] P. Dutilleul, "The MLE algorithm for the matrix normal distribution," *Journal of Statistical Computation and Simulation*, vol. 64, no. 2, pp. 105–123, 1999.

[146] M. Srivastava, T. von Rosen, and D. von Rosen, "Models with a kronecker product covariance structure: Estimation and testing," *Mathematical Methods of Statistics*, vol. 17, no. 4, pp. 357–370, 2008.

[147] J. Millan, "On the need for on-line learning in brain-computer interfaces," in *Proc. Int. Joint Conf. on Neural Networks*, vol. 4, July 2004, pp. 2877–2882 vol.4.

[148] M. Tangermann, K. -R Muller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Muller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlogl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI competition IV," *Frontiers in Neuroscience*, no. JULY, 2012.

[149] F. Galán, F. Oliva, and J. Guàrdia, "Using mental tasks transitions detection to improve spontaneous mental activity classification," *Med. Bio. Eng. Comput.*, vol. 45, no. 6, pp. 603–609, 2007.

[150] C.-J. Lin and M.-H. Hsieh, "Classification of mental task from EEG data using neural networks based on particle swarm optimization," *Neurocomputing*, vol. 72, no. 4-6, pp. 1121–1130, 2009.

[151] J.-M. Cano-Izquierdo, J. Ibarrola, and M. Almonacid, "Improving motor imagery classification with a new bci design using neuro-fuzzy s-dfasart," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 1, pp. 2–7, 2012.

[152] A. Schlogl, J. Kronegg, J. Huggins, and S. Mason, *Toward Brain-Computer Interfacing.* Cambridge: MIT Presss, 2007.

[153] S. Knecht, B. Drager, M. Deppe, L. Bobe, H. Lohmann, A. Floel, E. . Ringelstein, and H. Henningsen, "Handedness and hemispheric language dominance in healthy humans," *Brain*, vol. 123, no. 12, pp. 2512–2518, 2000.

[154] I. Taylor, *Psycholinguistics : learning and using language.* Englewood Cliffs, N.J.: Prentice Hall, 1990.

[155] A. Hyvrinen, P. Ramkumar, L. Parkkonen, and R. Hari, "Independent component analysis of short-time fourier transforms for spontaneous EEG/MEG analysis," *NeuroImage*, vol. 49, no. 1, pp. 257–271, 2010.

[156] B. Picinbono, "Second-order complex random vectors and normal distributions," *IEEE Trans. Signal Process.*, vol. 44, no. 10, pp. 2637–2640, Oct. 1996.

[157] B. Picinbono and P. Bondon, "Second-order statistics of complex signals," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 411–420, Feb. 1997.

[158] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1293–1302, Jul. 1993.

[159] B. Picinbono, "On circularity," *IEEE Trans. Signal Process.*, vol. 42, no. 12, pp. 3473–3482, Dec. 1994.

[160] K. V. Mardia, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, no. 3, pp. 519–530, 1970.

[161] J. Liang, R. Li, H. Fang, and K.-T. Fang, "Testing multinormality based on low-dimensional projection," *Journal of Statistical Planning and Inference*, vol. 86, no. 1, pp. 129 – 141, 2000.

[162] J. R. Schott, "A test for the equality of covariance matrices when the dimension is large relative to the sample sizes," *Comput. Stat. Data Anal.*, vol. 51, no. 12, pp. 6535–6542, 2007.

[163] R. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, vol. 50, pp. 987–1008, 1982.

[164] M. Mahanta, A. Aghaei, K. Plataniotis, and S. Pasupathy, "Spatio-spectral sufficient statistic for mental imagery EEG signals," in *Int. Joint Conf. on Neural Networks*, 2010, pp. 1–7.

[165] M. S. Mahanta, A. S. Aghaei, and K. N. Plataniotis, "A bayes optimal matrix-variate LDA for extraction of spatio-spectral features from EEG signals," in *2012 Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2012, pp. 3955–3958.

[166] M. Mahanta, A. S. Aghaei, and K. N. Plataniotis, "Regularized LDA based on separable scatter matrices for classification of spatio-spectral EEG patterns," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing.*, 2013, pp. 1237–1241.

[167] "Five methodological challenges in cognitive electrophysiology," *NeuroImage*, 2013.

[168] Z. Bai, *Templates for the solution of algebraic eigenvalue problems: a practical guide*. Siam, 2000, vol. 11.