NUS

National University of Singapore

Founded 1905

MODELLING AND CLASSIFICATION OF MOTOR IMAGERY EEG FOR BCI

LI XINYANG

(B. Eng)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Li Xinyang 28/11/2014

Li Xinyang

28 November 2014

Acknowledgments

I would like to express my deep and sincere gratitude to my supervisor, Associate Professor Ong Sim Heng. He trusted me and provided me a great opportunity to be under his supervision when I was faced with difficulties. This was invaluable and meant a lot to me. Prof. Ong was very responsible, patient and considerate. He even revised my manuscripts on the weekends, when I did not finish them early enough before the deadline. Moreover, he helped me a lot when I failed to take everything into consideration. The most important thing I learnt from him was how to be responsible and professional in research, which will definitely benefit me in my future work.

I would like to express my deepest gratitude to Dr. Guan Cuntai. Without Dr. Guan's help, guidance and understanding, I would never have finished my Ph.D. work and achieved what I have achieved. Although he is very busy, he spent a lot of time with students like me to give us guidance and help on our research. He taught me to think wide and to have a higher and clearer goal for research, while in practical works he guided me to make progress step by step. It is really fortunate for me to work in his team. It is a great and invaluable experience for me to meet and learn from top scientists and researchers in BCI, brain science and neuroscience.

My sincere gratitude goes to the NUS Graduate School for Integrative Sciences and Engineering (NGS) for providing me with a great opportunity and financial support to pursue my Ph.D. degree. I specially would like to thank Associate Professor Tang Bor Luen, Professor Ding Jeak Ling and Professor Philip Moore, who gave me great help and support when I was encountered

with difficulties. Their encouragement and trust are really meaningful to me.

I would like to express my gratitude to Professor Li Xiaoping, who is my thesis advisory committee chair. He has provided me invaluable advices and assistance in my research study.

My sincere gratitude and respect go to Dr. Ang Kai Keng and Dr. Zhang Haihong, who gave me a lot of guidance for my research, and helped me improve my scientific writing skill. I would like to express my gratitude to Dr. Pan Yaozhang for her help and guidance when I just started my Ph.D. knowing nothing about BCI.

I want to say that before I started my Ph.D., I was really curious about the attributes of a scientist. All these people taught me not only what a good and professional scientist should be but more importantly how to be a good and professional scientist.

I also want to thank Ms. Irene Christina Chuan and Ms. Ivy Wee for their help and patience on handling tedious paper work for me.

My sincere gratitude and respect go to all members in the Brain Computer Interface Lab for making this lab such a wonderful place to do research. And my thanks goes to my colleagues, Ms Atieh Bamdadian, Dr. Sidath Ravindra Liyanage, Dr. Mahanaz Arvaneh, Mr. Siavash Sakhavi and Ms Foong Ruyi. I really enjoyed discussing and talking with all of them, although I might not appear to be that way.

I would like to express my gratitude to Singapore, and all the adorable animals (owls, squirrels, pangolins and monkeys, etc.), trees and flowers here, which make me feel that the world is really wonderful.

At last but not least, I give my dearest gratitude to my family, especially my mom, who always believes I am better than what I think of myself, and

probably better than whom I actually am.

Contents

Summary

This thesis describes the construction of discriminative models for motor imagery EEG classification in brain computer interfaces (BCIs). Two types of methods are introduced to address the issues from the perspectives of model generalization and model adaptation.

The computational model for motor imagery EEG feature extraction needs to be a discriminative function conforming to the underlying dynamics of motor imagery, and robust against nonstationarity inherent in EEG. There exist successful methods that extract the event-related (de)synchronization (ERD/ERS) effects by designing spatial filters that maximize differences between EEG signals from different classes. However, in the presence of causal relationships and neuronal propagation, spatial filters in the instantaneous mixing model are not capable of describing such dynamics. To this end, a novel computational model for discriminative learning of propagation and spatial pattern is proposed. By introducing a convolutive model, the causal relationship could be covered in extracting ERD/ERS related features. Experimental studies on a two-class motor imagery data validate the effectiveness of the model, and indicate that the proposed model is better for background-noise attenuation. An ensemble learning method is proposed to improve the feature extraction model by addressing the biased estimates of covariance matrix. The mismatch between the data and the feature extraction model are used to re-sample the training trials, and different models are generated for different sub-sets of trials. The spatial filters are obtained by ensembling multiple models, and discrepancies between samples can be

addressed. The experimental results demonstrate that the ensemble learning model can improve the classification accuracy.

The large variation in EEG signals recorded on different days makes learning such nonstationarity within training data ineffective. It is necessary for the computational model constructed from the training data to adapt to the test data. The key challenge involved in computational model adaptation is how to construct a metric that measures this mismatch between test data and training model without test labels. To solve this problem, we construct a data-model mismatch metric to evaluate the feature extraction model, which is used to guide the adaptation toward reducing data-model mismatch in the proposed model adaption method. Experimental results show that the quantified mismatch is closely related to the classification accuracy, and comparison with other state-of-the-art spatial filter design methods validates the proposed model adaption method. To further understand the nonstationarity inherent in EEG and its implication on feature distribution change, a theoretical analysis is performed from the perspective of discriminative subspace of the EEG covariance matrix. By establishing the relationship between the shift of the discriminative subspace and that in feature space, a model adaptation method is proposed with the discriminative subspace updated for the test data. To take the risk from semi-supervised learning into consideration, a cross-validation-based loss function is proposed to evaluate the adaption direction. Experimental results show that compared to the adaptation method based on normalization, the proposed adaptation method can further enhance the classification results.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Brain Computer Interface

The discovery that electrical signals produced by the human brain could be recorded from the scalp implies the possibility of communicating with external devices via brain independent of muscle, and subsequently, makes brain computer interface (BCI) research a burgeoning field [1, 2]. By measuring central nervous system (CNS) activity, a BCI system enables people to access and understand the ongoing brain activities, and also provides alternative brain output pathways that are independent of normal brain outputs such as peripheral nerves. Applications of BCI range from modulating normal CNS output to facilitating new interactions between CNS and the environment [3].

There exist many kinds of brain signals, which can be categorized by the type of measurement technique being used or the nature of the brain activity being measured. For instance, activation, communication and information transfer in the CNS are fulfilled by neuronal action potentials (or spikes), which also give rise to neuronal electrical activities in the cerebral cortical surface [4]. Such electric fields are accessible to magnetic recording, such as

magnetoencephalography (MEG), and various types of electric recordings at different spatial scales, including electroencaphalography (EEG), electrocorticography (ECoG), and multielectrode arrays implanted in the brain tissue [5, 6, 7, 8]. Besides electric signals, chemical processes involved in brain activities can also be measured, e.g., using positron emission tomography (PET) [9, 10]. In addition, the metabolic process involved in the energy consumption during different brain activities can be revealed by the change in hemoglobin, which is regarded as the blood-oxygen-level-dependent (BOLD) response [11, 12]. Based on the BOLD response, there are metabolic signal measurements including functional near-infrared spectroscopy (fNIRS) and functional magnetic resonance imaging (fMRI) [13, 14, 15, 16, 17, 18].

Among all the aforementioned different measurement techniques, EEG is the most popular and widely-used measurement in BCI systems [19]. Compared to EEG, both fMRI and MEG are more expensive and call for much more complicated implementation. Also, PET, fNIRs, and fMRI suffer from poor temporal resolution and delayed responses, which make these measurements less feasible for most of the BCI applications in reality. In contrast, electrical signals usually have relatively high temporal resolution and fast response. However, electric signal measurements except EEG, i.e., ECoG and implanted electrodes, are also less practical and convenient, because as invasive methods these measurements need surgical operations. In conclusion, EEG-based BCI is the most widely studied and applied BCI paradigm, which can be attributed to the following advantages of EEG:

i) EEG provides real-time measurements for on-going brain activities;

ii) EEG can be implemented under relatively lower cost; and

iii) EEG recording is non-invasive.

EEG-based BCI systems vary depending on the EEG signals used to drive the system, which can be categorized by the type of the signal generation. One kind of the EEG signals is generated by external stimulus, and is regarded as evoked potentials (EPs). For example, P300 is a kind of endogenous event-related brain potentials (ERPs) in EEG, and it occurs over the central-parietal scalp around 300 milliseconds after a rare stimulus appears in the typical "odd-ball" experiment paradigm [20, 21, 22]. The speller based on P300 with the "odd-ball" paradigm is one of its most important applications, and it functions in a similar way to a standard computer keyboard. In the experiment, a subject is presented with a matrix of characters, and required to attend to one of the elements in it. By successively and randomly intensifying either a row or a column of the matrix, the "odd-ball" event is created when the intensification event is relevant to the element with the subject's attention. Thus, P300 can be triggered and observed from EEG when such events occur [23, 24]. By eliciting and detecting P300, a "virtual keyboard" BCI system is created as a helpful alternative communication or control approach for the disabled people who cannot use normal control devices [25, 26, 27, 28].

In contrast to the signals that are generated as the direct results of external stimulus, another kind of commonly used EEG signals are spontaneous changes in rhythmic activity recorded over the sensorimotor cortex known as sensorimotor rhythms (SMRs) [29, 30]. Changes in the SMRs are typically associated with motor cortex activation [31]. In particular, decrease in SMRs, known as ERD, has been discovered during motor behaviors, followed

by the discovery that increase in SMRs, known as the ERS, is also related to sensorimotor events [32, 33, 34].

Not only real motor movements, imagination of certain movements (regarded as motor imagery) can also be revealed by ERD/ERS in EEG signals, which has attracts even more attention in EEG-based BCI research [30, 35, 36]. As a dynamic state facilitated by the motor system, motor imagery relates to intending and preparing movements. It is also generally assumed that internally motor imagery can cause the same motor representations as the corresponding motor execution [37, 38]. Many findings suggest that there exist parallels between the motor imagery and the executed movement, i.e., close temporal coupling between motor imagery and executed movement [39, 40, 41]. Moreover, motor imagery can even lead to performance improvements for athletes, and previous studies also suggest the effectiveness of motor imagery training for functional recovery of stroke patients [42].

Motor imagery related SMR has been extensively studied and exploited in BCI for supportive and therapeutic purpose, and is a highly attractive research area. For example, the motor impairment caused by stroke is one of the major causes of permanent disabilities, and active movement training (AMT) is usually used to restore the patients' motor function [42]. However, this kind of traditional therapy is quite labor intensive and expensive. To this end, motor-imagery-based BCI provides promising solutions. By detecting and quantifying ERD and ERS associated with motor imagery, BCI can translate motor imagery of certain actions into commands for possible orthosis to perform predefined tasks, which is illustrated by Figure 1.1. On the one hand, the motor-imagery-based BCI system can be used as a sub-

stitute of neuromuscular functions for environment control or interaction. On the other hand, patients could restore their motor functions gradually through AMT provided by BCI rehabilitation systems with less assistance from therapists [43, 44, 45]. For example, it has been reported that patients with spinal cord injuries could regain hand grasp function with a motor imagery BCI-based rehabilitation system, and many studies show that stroke patients' motor functions could be improved with BCI-based rehabilitation [46, 47, 48]. In short, motor-imagery-based BCI does not require any voluntary muscle control, and can be used to develop alternative supportive and therapeutic systems that call for less manpower [49, 50, 51, 52].

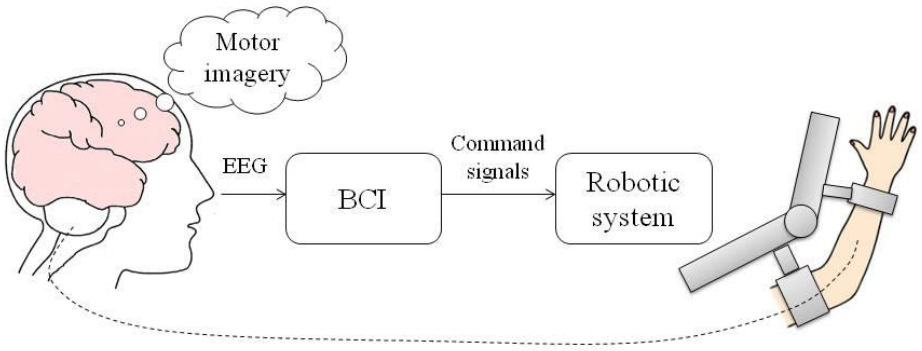


Figure 1.1: An example of motor-imagery-based BCI rehabilitation system

1.1.2 Processing Procedures in a BCI system

To learn the mental condition from the EEG signals is the core for the usability, information transfer, and robustness of BCI systems [51, 53, 54]. Especially for the aforementioned BCI-based rehabilitation system, the effectiveness of the rehabilitation is largely depending on classifying EEG signals corresponding to the correct motor imagery task. Generally speaking,

for motor-imagery-EEG-based BCI, it takes three steps in processing raw EEG signals to obtain the classification results regarding the motor imagery condition, which are the preprocessing step, feature extraction step and classification step.

Preprocessing aims at increasing the signal-to-noise ratio of the input signals, and it usually includes temporal filters and spatial filters. A temporal filter is used to obtain EEG signal components from frequency bands with the strongest SMR effect obtained by bandpass filtering. For example, it is generally believed that ERD/ERS is more distinctive at $4-30$Hz, and subsequently, a bandpass filter of around $4-30$Hz is usually applied to raw EEG signals. Given the drawbacks of EEG with regards to poor spatial resolution, spatial-filtering is used to localize EEG signals recorded from multiple channels. Common average reference (CAR) filtering subtracts a common sample average of all the remaining channels from one specific channel. Similarly, in Laplacian filtering, the average of the four neighbouring channels is used as the reference for a specific channel. Both CAR and Laplacian filtering are commonly-used spatial filtering techniques to alleviate spatial mixture and enhance localizing information as preprocessing [55].

Feature extraction is the process of obtaining signal characteristics that can represent certain mental conditions for discriminative purposes. The characteristics containing useful information in a compact and efficient form is referred to as a feature or a feature vector. There are frequency-analysis methods that extract the frequency parameters pertaining to ERD/ERS as features, such as the Fourier transform, the wavelet transform, the Hilbert-Huang transform, and the autoregressive model [56, 57, 58, 59, 60]. Moreover, entropy measurements such as Kolmogorov entropy have also been used as

features to quantify ERD/ERS [61]. Recently, the analysis of neuronal connectivity is gaining more attention in neuroscience because it describes the general functioning of the brain and communication between its different regions [62, 63, 64]. For example, causal connectivity is found in motor related core regions such as the primary motor cortex (M1) and supplementary motor area (SMA) during motor imagery [39]. Therefore, scalp connectivity or intra-channel synchronization measurements have been used as features for motor imagery analysis [65, 66]. Synchronization features derived from the phase locking value (PLV) and from the spectral coherence have been examined for classifying mental tasks in [66]. Similarly, in [65], nonlinear regressive coefficients and PLV are used as features of amplitude and phase coupling between different brain regions, and prior neurophysiological knowledge is used to determine the pairs of electrodes of interest. Furthermore, common spatial pattern analysis (CSP) is a type of feature extraction method based on spatial filter design. By maximizing the power differences between different motor imagery conditions, spatial filters of CSP can capture the ERD/ERS associated with motor imagery with the power of the signal in the pseudo-scalp space as features [67, 53]. Because of its discriminative function, CSP differs from the other spatial filtering methods, and is regarded as a feature extraction method, while filtering methods such as CAR and Laplacian filters are usually regarded as preprocessing procedures.

The final stage of a BCI is usually classification, where the mental state corresponding to the type of the motor imagery being performed is predicted by classifying the features of the brain signals. Classification methods in machine learning have been widely explored and applied in BCI. The most commonly-used and successful classifiers in BCI include linear discriminant

analysis (LDA), support vector machine (SVM) and naive Bayesian classifier. In [68, 69], the authors compare the performances of several classifiers.

Figure 1.2 shows a block diagram of the three procedures in a typical BCI system: preprocessing, feature extraction and classification. Usually, to detect motor imagery condition accurately, machine learning is applied for both feature extraction and classification to build effective discrimination models. To formulate the relationship between EEG data and the label, i.e., the ground truth of the corresponding motor imagery task being performed, the computational models used in these two steps need to be trained by training data with labels, which is usually regarded as the calibration or training stage. After the training stage, the resultant feature extraction model and classification model are applied to the test data to predict the labels, which is regarded as the evaluation or test stage. It should be noted that preprocessing, feature extraction and classification are not always separate steps. As introduced before, frequency analysis and spatial filter design can be used for both preprocessing and feature extraction. In some studies, part of the preprocessing and feature extraction are formulated in one model, while in some other studies the feature extraction model is optimized by the same optimization function in the classifier [70, 71].

1.2 Objectives

Because EEG data is typically represented by a large matrix of multi-channel time series, which cannot be fed to the classifier directly, the feature extraction step is of particular importance for the accuracy and reliability of BCI. As introduced earlier, measured from electrodes on the scalp that are

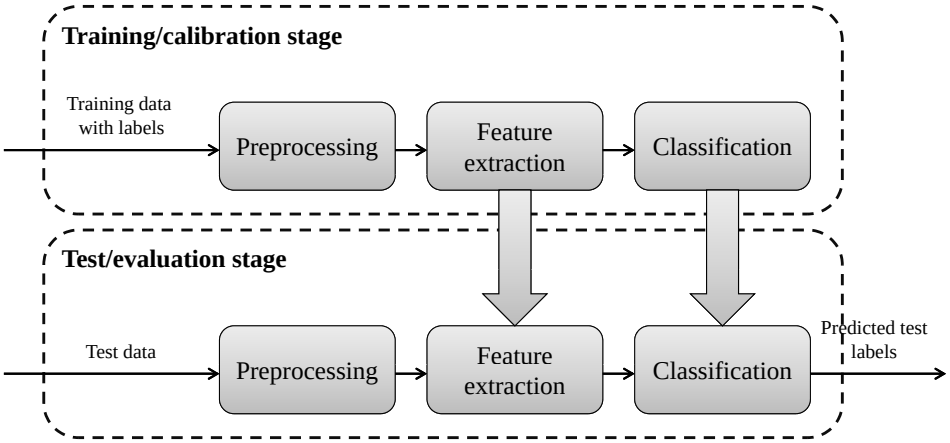


Figure 1.2: EEG processing procedures involved in BCI

far from the neurons, EEG fails to provide precise positions of the activated brain areas, and is very prone to artefacts, such as electrooculography (EOG) and electromyography (EMG) interference. Apart from the mixing nature of brain signals, the characteristics of the brain signal of a specific mental task may vary largely from trial to trial, and from session to session. Such nonstationarity inherent in EEG makes capturing information related to motor imagery even more difficult [72, 50]. Therefore, it is very challenging to obtain discriminative and effective features from EEG by distinguishing between the change in SMRs due to motor imagery and the irrelevant changes from background noise. Thus, the main motivation of this thesis is to enhance the performance of BCI with the focus on feature extraction for motor imagery EEG.

The computational model for feature extraction needs to be a discriminative function that is in accordance with underlying dynamics and phenomena of brain activities during motor imagery while robust against the nonstationary nature of EEG. The challenge for such a computational model aiming at

motor imagery EEG classification in BCI arises mainly from two aspects:

i) complex dynamics and phenomena of brain activities during motor imagery revealed by accumulating neuroscience findings need to be taken into consideration; and

ii) nonstationary nature of EEG and low signal-to-noise ratio cause ineffective feature extraction with resultant inaccurate prediction.

It is worthwhile noting that these two aspects are not independent of each other. A model depicting the dynamics more accurately is more robust to a certain extent because it is better in capturing activities relevant to motor imagery from varying noises. Regarding these two aspects, limitations of existing research studies can be summarized below:

i) despite computational models combining frequency, temporal and spatial analysis, causal connectivity between different brain areas caused by possible neuronal propagation effect during motor imagery is not fully investigated; and

ii) most existing works that address nonstationarity focus on measuring data variations, while the data-model mismatch has not been addressed directly and sufficiently.

The main aim of this study is to propose computational models for feature extraction regarding research issues arising from the two aforementioned aspects so as to enhance the performance of the BCI in classifying motor imagery EEG. More specifically, the objectives of this thesis are:

i) to introduce a convolutive computational model to depict the more complex underlying causal relationships involved in ERD/ERS effects;

ii) to build an ensemble learning model with a re-sampling approach that takes the mismatch between model and data into consideration;

iii) to propose a model adaptation method using a novel quantification of data-model mismatch between the training model and the test data; and

iv) to present a discriminative subspace tracking method for model adaptation with theoretical investigation of the data-model mismatch from the perspective of subspace.

The outcomes of this study may improve the capabilities of BCI in detecting and classifying motor imagery EEG:

i) with more complex underlying dynamics of motor imagery being covered, the computational model is more accurate and is better in background noise attenuation;

ii) ensemble learning of multi-model improves the performance of the model with the mismatch between data and model being considered; and

iii) the nonstationarity inherent in EEG data could be addressed by adapting the model for the test data.

1.3 Structure of the Thesis

In the context of feature extraction for motor imagery EEG classification in BCI, this thesis addresses the following problems: model generalization and model adaptation. In Chapter 2, we give a literature review of feature

extraction methods for motor imagery EEG. In Chapter 3, we propose a computational model to account for neuronal propagation effect in spatial pattern analysis, and estimate the propagation and volume conduction jointly and iteratively in the proposed unified model. In Chapter 4, an ensemble learning of spatial filter design is proposed to address the nonstationarity issue, which takes the mismatch between samples and the model into consideration. In Chapter 5, we propose a model adaptation method by introducing a quantification of the mismatch between training model and test data based on a tensor formulation. In Chapter 6, a discriminative subspace tracking algorithm is proposed for model adaptation. In summary, the whole structure of the thesis is shown in Figure 1.3.

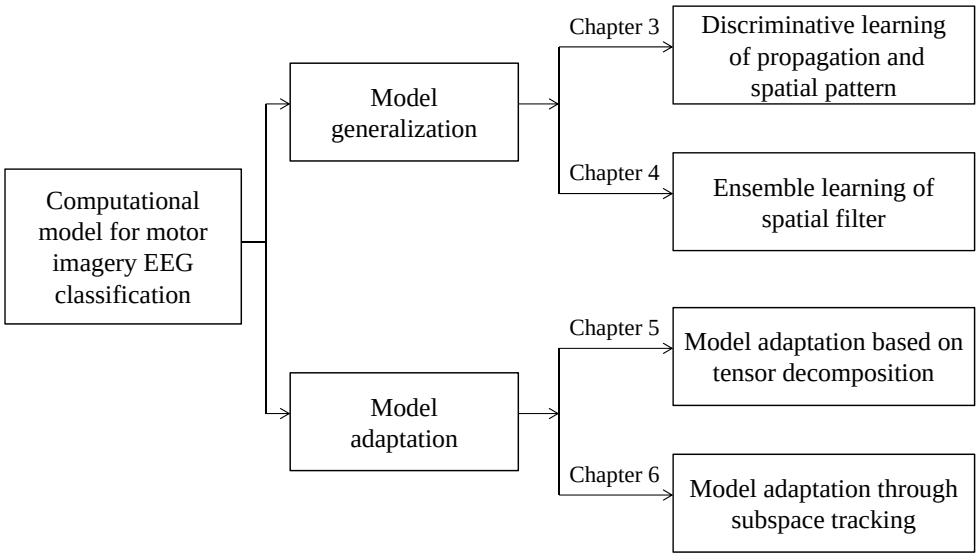


Figure 1.3: Thesis structure

Chapter 2

Literature Review

A vast number of studies addressing feature extraction have been proposed to determine the most distinctive characteristics in EEG that represent the mental task of interest. For motor imagery EEG classification, spatial filtering has been widely used to localize EEG signals with the strongest ERD/ERS involved in motor imagery, and probably the most recognized feature extraction technique is the spatial filter design based on CSP [53, 67].

2.1 Common Spatial Pattern Analysis

In CSP, the desired spatial filters are constructed as projection matrices. The prominent ERD/ERS can be extracted by maximizing the variance of the projected signal under one condition while minimizing it under another so that the EEG signal could be classified by its power in the projected space [73, 74].

Let $R^{+/-} \in \mathbb{R}^{n_c \times n_c}$ be the pooled estimate of the covariance matrix of the band-pass filtered EEG signal measured from n_c channels under condition $+$ or $-$ (e.g., the left hand imagination or right hand imagination), i.e.,

$$R^c = \frac{1}{|\mathcal{Q}^c|} \sum_{i \in \mathcal{Q}^c} \frac{X^i (X^i)^T}{\mathrm{tr}[X^i (X^i)^T]}, \ c \in \{-, +\} \tag{2.1}$$

where X^i is the data matrix of a short segment of the band-pass filtered EEG signal for trial i, and tr[$\cdot$] is the trace of a matrix. Considering the binary classification problem, the two classes are indexed by $c \in \{+, -\}$. $\mathcal{Q}^c$ denotes the set of trials that belongs to class c such that $\mathcal{Q}^+ \bigcap \mathcal{Q}^- = \varnothing$, and $|\mathcal{Q}^c|$ denotes the total number of samples belonging to set $\mathcal{Q}^c$. Let

$$R = R^+ + R^- \tag{2.2}$$

with the eigen-decompositions as

$$R = U_R \Gamma U_R^T \tag{2.3}$$

where U_R is the matrix of eigenvectors and Γ is the diagonal matrix of eigenvalues. Thus, the whitening transformation P can be obtained as

$$P = \Gamma^{-\frac{1}{2}} U_R^T \tag{2.4}$$

Let

$$\Sigma^+ = P R^+ P^T \tag{2.5}$$

$$\Sigma^- = P R^- P^T \tag{2.6}$$

Σ^+ and Σ^- in (2.5) and (2.6) have two key properties for the discrimination of motor imagery EEG. Firstly, they share common eigenvectors:

$$\Sigma^+ = U \Lambda^+ U^T \tag{2.7}$$

$$\Sigma^- = U \Lambda^- U^T \tag{2.8}$$

Secondly, the sum of the corresponding eigenvalues is 1:

$$\Lambda^- + \Lambda^+ = I \tag{2.9}$$

where $I \in \mathbb{R}^{n_c \times n_c}$ is the identity matrix. Then, the spatial filter W in CSP can be obtained as

$$W = (P^T U)^T \tag{2.10}$$

such that

$$WR^+W^T = \Lambda^+ \tag{2.11}$$

$$WR^-W^T = \Lambda^- \tag{2.12}$$

The significance of the transformation in (2.11) and (2.12) lies in the fact that the diagonal elements λ_j^c, $j = 1, 2, ...n_c$, in Λ^c are the variances of the signal after the projection by W. Given that $\lambda_j^+ + \lambda_j^- = 1$, if λ_j^+ is close to one, λ_j^- is close to zero. In other words, the corresponding spatial filter $\mathbf{w}_j$ yields signals of class + with high variance and signals of class - with low variance in the surrogate space, and vice versa. Therefore, if we sort λ_j^c, $j = 1, 2, ...n_c$ in a descending order (or an ascending order) and pick $\mathbf{w}_j$ corresponding to the largest and the smallest λ_j^c, we can extract features of variances with the strongest discriminative information. Usually, λ_j^c is sorted in a descending (or an ascending) order, and correspondingly the top and bottom rows of W

are used, which yields the feature $\mathbf{f}$ as

$$\mathbf{f}_j^i = \log\frac{\mathbf{w}_j X^i (X^i)^T \mathbf{w}_j^T}{\sum_j \mathbf{w}_j X^i (X^i)^T \mathbf{w}_j^T},\ j = 1, \ldots, r, n_c - r + 1, \ldots, n_c \quad (2.13)$$

where $\mathbf{f}_j^i$ is the j-th element of $\mathbf{f}^i$, and r is the number of pairs of spatial filters being used.

The calculation of W can also be expressed as solving the optimization problem of maximizing the Rayleigh coefficient between R^+ and R^-, i.e.,

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \frac{\mathbf{w} R^+ \mathbf{w}^T}{\mathbf{w} R^- \mathbf{w}^T} \quad (2.14)$$

(2.2)-(2.10) and the optimization function (2.14) can be solved as a generalized eigenvalue decomposition problem, i.e.,

$$R^+ \mathbf{w}^T = \lambda R^- \mathbf{w}^T \quad (2.15)$$

In short, optimizing $\mathbf{w}_j$ to minimize or maximize λ_j is equivalent to finding the spatially-filtered signal with the strongest ERD/ERS effects, which is the reason why CSP is successful in extracting discriminative features from motor imagery EEG. Thus, efforts have been made to improve CSP, which will be discussed in the following sections.

2.2 Theoretical Analysis of CSP

There exist many works that analyze CSP from different angles for further understanding and possible improvement in addition to its discriminative function in maximizing power differences between classes. In [75], the au-

thors build a two-stage hierarchical Bayesian structure to model the underlying brain activities during motor imagery, and show that CSP is a maximum likelihood (ML) estimate of the model under certain assumptions. In [76], the authors establish the theory linking spatial filtering directly to Bayes classification error in the CSP feature space. In particular, it is proved that Bayes error can be reduced by minimizing the Rayleigh quotient in (2.14). Moreover, it is proved in [77] that spatial filters in CSP project the EEG data into subspaces where the Kullback-Leibler divergence (KL-divergence) between the data distributions from two classes is maximized. Therefore, the objective function of CSP can also be formed in a divergence-based framework. The significance of this work lies in the fact that it is a unifying framework for CSP with different kinds of regularization.

2.3 Joint Optimization of Spatial Temporal and Spectral Parameters

Despite the importance of spatial filtering, temporal and spectral analysis are also critical for motor imagery EEG classification [32, 78, 30]. For example, it is generally believed that broad band around 8 - 40Hz is the band with the most distinctive ERD/ERS effects. However, the specific discriminative bands for different subjects may be different. Besides frequency components, time segmentation is also critical for effective feature extraction. Even in the BCI experiments with a specific cue to instruct subjects to start performing motor imagery, it is difficult to know the exact time when the motor imagery begins. Therefore, many works improve the feature extraction model by

combining temporal and spectral analysis with the spatial pattern analysis, i.e., CSP.

In [70], common spatio-spectral pattern (CSSP) is proposed by optimizing spatial filters by adding a one-time-delayed sample, which is equivalent to increasing the number of the channels. In [79], common sparse spectral spatial pattern (CSSSP) extends CSSP by implementing the optimization of a complete global spatial-temporal filter into the objective function of CSP. In [71], iterative spatio-spectral patterns learning (ISSPL) is designed to automatically learn spatio-spectral filters and the classifier sequentially from labeled multichannel EEG data in an iterative fashion. In each iteration, the spatial filters are calculated via CSP based on the spectral filters optimized in the preceding iteration. Coefficients of the temporal filters can be regarded as the feature coefficients, and are optimized using the optimization function of SVM.

In addition to the optimization of a single frequency band, in [80, 68, 81, 82], EEG signals are decomposed into several frequency bands, and spatial filters based on multiple bands are explored. In particular, filter bank CSP (FBCSP) employs multiple bandpass filters, which is denoted as a filter bank, to bandpass filter the EEG raw data into different frequency bands [80, 68]. CSP is implemented on each of these bands. Thus, each pair of bandpass and spatial filter yields CSP features that are specific to the frequency range of the bandpass filter. After calculation of features of each band, feature selection based on mutual information is applied. As a result, only those effective spatial filters corresponding to the selected features are used for test data, which reduces the computational complexity. In [82], sub-band CSP (SBCSP) employs a Gabor Fourier-based filterbank and calculates a

sub-band score for each spatial filter based on classifiers. In other words, different from FBCSP selecting and combining features from different bands, SBCSP fuses the bands according to the scores by a recursive band elimination based on classification algorithms. Based on FBCSP, an optimum spatio-spectral filtering network (OSSFN) is proposed by jointly learning the bandpass filters and spatial filters to maximize the mutual information between feature vector variables and the class label [83]. Instead of designing the candidate filters in advance, in [84], the authors propose discriminative filter bank CSP (DFBCSP), which designs finite impulse response filters and the associated spatial weights simultaneously. Different from the mutual information, the objective function used in [84] is Rayleigh quotient, and both spatial filters and temporal filters are solved in a sequential manner with the parameters optimized one by one.

As stated earlier, the effectiveness of CSP in feature extraction for motor imagery EEG lies in discriminating power difference, which is in accordance with the ERD/ERS phenomenon involved in motor imagery. Recently, increasing neuroscience findings based on fMRI or EEG suggest that brain activities of neuronal connectivities exist during motor imagery in brain areas such as M1 or SMA [85, 39]. In particular, the analysis of neuronal connectivity is gaining more attention because it describes the general functioning of the brain and communication between its different regions [62, 64]. In the presence of neuronal propagation and causal relationship during motor imagery, connectivity measurements have been explored as features [65, 66]. Given the importance of the synchronization or coupling features, in [86], PLV is combined with FBCSP in the filter bank feature combination (FBFC) model.

2.4 Extensions of CSP for Nonstationarity

Brain signals are typically mixtures of significant noises, extraneous information, and the components of interest, and all these components could vary due to different experimental setups, subjects' conditions and some other factors, which contribute to the nonstationarity inherent in EEG. Recent neuroimaging studies have shown that nonstationarity may be partially caused by low frequency spontaneous fluctuations in brain signals that are coherent within resting state networks (RSNs) [87, 88]. As reported in [87], intrinsic brain activity of RSNs persists during task performance and contributes to variability in evoked brain responses, as well as in human behaviours. Besides, electrode impedance and positioning, and subjects' different response behaviours would also result in the drastic signal variation. However, because the calibration procedure is tedious and time-consuming, it is not feasible to account for such data variation by calibrating the computational model in every session for patients who are undergoing continuous rehabilitation. For practical BCI-based rehabilitation, only the computational model obtained from the calibration session is available for all the following rehabilitation sessions [49]. Hence, it would be useful to propose methodologies that address the nonstationarity issue in motor imagery EEG classification.

Among a number of algorithms that have been proposed to address the nonstationary issue, one category considers improving the robustness of the model using calibration data only, such that this may translate to better generalization in processing unseen test data [67, 73, 74].

As EEG could be regarded as a mixture of underlying stationary and nonstationary sources, it could be helpful to distinguish between stationary

and nonstationary contributions for constructing a more robust model. To this end, in stationary subspace analysis (SSA), the observed signal is modeled as a linear superposition of stationary and nonstationary sources, and the aim is to separate the two groups by estimating the mixing matrix [89]. In [90], SSA has been applied to motor imagery EEG classification as a kind of preprocessing procedure. Given the stationary components identified by SSA, CSP is applied to the stationary sources for feature extraction, and it is found that the classifier performance is significantly improved by the preprocessing of SSA. However, as pointed out in [77], this kind of two-step approach suffers from the loss of discrimination information in the first step, i.e., the SSA step. To avoid the loss in the two-step combination of SSA and CSP, regularization-based methods have been studied extensively to incorporate stationarity constraint into the discriminative objective function in CSP.

The regularization method refers to adding certain terms to the denominator of the CSP objective function in a Rayleigh coefficient form (2.14). In this way, this term, denoted as the regularization term, can be penalized in the objective function [91]. A regularization-based robust model was first proposed in [72], which is denoted as invariant CSP (iCSP). In iCSP, the invariant property of CSP is achieved by adding disturbance covariance matrices as the regularization term. Therefore, iCSP is robust against disturbances whose covariance could be anticipated from prior physiological knowledge or extra measurements like EOG or EMG. However, the extra recordings or prior knowledge about noise are usually not available or reliable. In [92], stationary CSP is proposed to address nonstationary noise in a more general case. Instead of using additional recordings to estimate the

nonstationary artefacts, nonstationarity is estimated as the sum of absolute differences between the mean variance and variance of a certain trial in the projected space. By penalizing the cross-trial differences, spatial filters can keep the variance features as stable as possible across trials while differentiating variances between two conditions. In [93], the authors introduce a different penalizing term that measures the KL-divergence of distributions of EEG data across trials, and subsequently, the learning algorithm can minimize within-class dissimilarities while maximizing inter-class separation. In [94], the nonstationary projection directions are estimated based on the principal component analysis (PCA) using cross-subject data, and then penalized in the objective function to build subject-specific spatial filters. Similarly to [94], cross-subject data are also used in [95] to enhance the robustness of the spatial filters. In particular, instead of estimating the directions of the nonstationary components, average covariance matrices of multiple subjects are directly incorporated in the denominator of the Rayleigh coefficient as a kind of ground truth of the covariance matrix estimate. In this way, an inaccurate model could be avoided when only very few EEG data from a single subject are available. In [77], those methods regularizing nonstationarity measurements are unified in a divergence-based framework, and different divergence measurements, such as KL-divergence, symmetric KL-divergence and beta-divergence, are compared and discussed.

Different from the aforementioned methods regularizing covariance matrix estimates or nonstationary components, another category of regularization method imposes constraints on the solution to mitigate the influence of artifacts. In [91], CSP with Tikhonov regularization (TRCSP) is proposed by penalizing the l_2 norm of the solutions so that the channels with large

weights can be penalized. Given that the importance of different channels is different for motor imagery classification, a weighted version of TRCSP is also introduced in the same work. In particular, different penalty levels are assigned to different channels, and the penalty level is defined according to the activation levels of different brain regions for a given mental task in the literature. Moreover, the sparsity of the spatial filters is used as the constraint in [96] based on the l_1/l_2 norm. Based on the sparse spatial filters obtained, only a few channels are selected to perform feature extraction.

Another category of methods investigates the actual variations across sessions and then adapts detection models accordingly. While motor imagery EEG detection algorithms usually consist of a feature extraction step and a classification step, some methods focus on the classification step and study the shift of CSP features with fixed spatial filters [97, 98, 99, 100, 101]. Studies in [97] show that the two-class motor imagery EEG classification accuracy could increase significantly among more than 90% of the subjects by using simple adaptive procedures such as bias adaptation. The shortfall of this methodology is that adapting a classifier is not effective when the test features are inseparable.

To address the issue of feature separability, another category of adaptation methods investigates the adaptation of the feature extraction model, i.e., spatial filters. Variations of EEG data across sessions can be taken into consideration by incorporating data from test sessions to adapt the projection matrix in CSP [102]. In particular, since the solution of the spatial filters in CSP is based on the joint diagonalization of the average covariance matrices, such adaptation can be achieved by updating the covariance matrices by using test data. In [100], both the feature extraction model and classifier are

adapted with more test data available using an expectation-maximization method.

Another approach assumes that there is a domain-invariant subspace (a domain refers to a training space or a test space [103]), where the classifier trained by training data could be equally effective to test data [104]. In [104], this domain-invariant subspace is assumed to be the whitened subspace, where the whitened training data and test data have the same (or similar) marginal distributions, and the posterior distributions of the labels are the same across domains. Therefore, the whitening part in the spatial filter is updated based on test data, which is equivalent to projecting both training data and test data to the invariant whitened space. Similarly, in domain space adaptation (DSA) in [103], a linear mapping matrix is estimated to project the test data into the training data space based on minimizing the KL-divergence between data in the two spaces, and the unsupervised case of DSA is shown to be equivalent to [104]. As pointed out in [104], this domain-invariant assumption on the whitened space holds only when the linear transformation between the two domains is symmetric.

2.5 Conclusion

In this chapter, we present a brief review of feature extraction methods for motor imagery EEG. By maximizing the differences of signal powers between different conditions, CSP can be regarded as the most successful method in capturing ERD/ERS effects in motor imagery EEG, and it has been introduced in detail. Moreover, there are a large number of BCI research studies aiming at improving CSP, and these methods are introduced from three per-

spectives: theoretical analysis of CSP, joint optimization of different parameters in CSP, and enhancement of CSP regarding nonstationary issues.

As stated earlier, the main signal processing issues for classifying motor imagery EEG are the complex dynamics and phenomena involved in the motor imagery and the nonstationary nature of EEG, and there are still research gaps arising from these two aspects in the feature extraction model development. Despite spatial filter design combined with frequency and temporal analysis, investigation of causal relationship between EEG signals during motor imagery for feature extraction is not given adequate attention. Moreover, regarding EEG nonstationarity, data variation measurements are often adopted with very few works addressing the mismatch between the data and the model directly. Thus, in the following chapters, novel computational models regarding these two issues are proposed for motor-imagery-based BCIs from the perspectives of model generalization and model adaptation.

Chapter 3

Discriminative Learning of Propagation and Spatial Pattern

As introduced earlier, multiple brain regions cooperate during motor imagery [85, 39]. To investigate such connectivity or causal relationship, the directed transfer function (DTF) has been used to evaluate the causal flow between any given pair of channels in a multi-channel EEG in the frequency domain [105, 106, 107]. The estimation of the DTF is based on a multivariate autoregressive model (MVAR) and, more importantly, it has been applied to EEG data of voluntary finger movement and motor imagery for event-related causal flow investigation [108, 109]. Based on DTF, it has been found that there is a rapid increase in information outflow from electrodes FC3 and C3 caused by ERS, and propagation of β-synchronization from FC3 and FC1 to C3, C1, Cz, CP3 and CP1, which provides the evidence of communication between different sensorimotor areas [110]. The causal flow or time-lagged correlation is assumed to be caused by possible neuronal propagation [111]. However, looking at only the time profiles of ERD/ERS, it is difficult to judge which is the primary source of activity.

In the presence of neuronal propagation and causal relationship during motor imagery, conventional spatial filter design is not adequate to capture

the underlying brain activities [112, 113]. It is worthwhile noting that, although some of the connectivity measurements mentioned earlier have been explored in existing works [65, 66], only scalp connectivity or intra-channel synchronization measurements are directly used as features, whereas volume conduction effects are not rigorously addressed. One consequence is that bandpower variations are misinterpreted as changes in connectivity [114]. Therefore, rather than ignoring the connectivity or propagation between sources in spatial filter design or using scalp connectivity directly as features, we would like to promote a computational model that can more accurately describe the underlying processes by considering both neuronal propagation and volume conduction effects.

In this chapter, we present a novel feature extraction model for motor imagery EEG based on a multi-variate convolutive process with an analysis of the spurious effects in classifying ERD/ERS based on an instant linear mixture model. The effectiveness of introducing a time-lagged demixing matrix to produce time-decorrelated data is analyzed theoretically from the perspective of background noise elimination. Furthermore, the demixing matrices accounting for propagation and volume conduction are estimated jointly and iteratively in the proposed model. Through the experimental study, we evaluate the efficiency of the proposed method in terms of classification accuracy in a two-class motor imagery EEG classification problem. We also analyze the effectiveness of the proposed method for background noise elimination using the KL-divergence measurement.

This chapter is organized as follows. In Section 3.1, limitations of conventional spatial filter design are discussed, with the necessity of taking the causal propagation into consideration. Then, the details of the proposed dis-

criminative learning of propagation and spatial pattern are given in Section 3.2. The investigation in background noise is performed in Section 3.3. In Section 3.4, the validity of the proposed method is verified by experimental studies on two-class motor imagery classification. Concluding remarks are given in Section 3.5.

3.1 Data Model and Problem Formulation

Let $X(t)$ be the time-series of a multi-channel EEG signal, with each component in $X(t)$ representing a particular EEG channel measured at time t. Considering the complex temporal dynamics, in particular the latent causal relations in $X(t)$, we describe the observed data $X(t)$ as an n_c-dimension linear convolutive mixture process of order l [112, 115], i.e.,

$$X(t) = \sum_{\tau=0}^{l} \Phi(\tau) S(t-\tau) \tag{3.1}$$

where $S(t)$ is the source signal of interest, $\Phi(\tau)$ is the projection matrix of the order τ, and l is the maximum time-lagged order. When $l = 0$, the observed data $X(t)$ is an instant mixing process. For simplicity of description, the additive EEG noise can be described by an component in $S(t)$. Conventionally, it is assumed in motor imagery EEG classification that $X(t)$ is an instant linear mixture of source signals. This leads to an instant de-mixing solution to the estimation of $S(t)$:

$$\hat{S}(t) = WX(t) \tag{3.2}$$

In (3.2), W does not necessarily be the CSP projection matrix but could be any projection or demixing matrix based on an instant linear model. Because the thesis focuses on feature extraction model, W in CSP is used in this section.

Interestingly, we note that the estimate $\hat{S}(t)$ given by (3.2) is also a mixture of the time-lagged components, i.e.,

$$\hat{S}(t) = \sum_{\tau} \Phi_{\mathbf{w}}(\tau) S(t-\tau) \tag{3.3}$$

where $\Phi_{\mathbf{w}}(\tau) = W\Phi(\tau)$ is a mixing matrix. In discriminative analysis, W is designed to extract the most discriminative signal $\hat{S}(t)$. However, as shown in (3.3), discriminative signals could still be mixed with non-discriminative ones in $\hat{S}(t)$.

A perfect solution would be that $\Phi_{\mathbf{w}}(\tau)$ takes an identity matrix form for $\tau = 0$ and a zero matrix form for any $\tau \neq 0$. This is generally impossible except for the case that $\Phi(\tau)$=0 for $\tau \neq 0$, or in other words, when the convolutive mixture model in (3.1) reduces to an instant mixing model. Therefore, it is necessary to take the causal flow into consideration together with spatial filter design in a unified model to have a better estimation of $S(t)$, which is the motivation of the work in this chapter.

Solving the reconstruction problem of $S(t)$ from (3.1) may lead to a solution in the form of an infinite impulse response (IIR) filter. As we will elaborate shortly and also for practical use, we simplify the problem into a finite impulse response (FIR) filter given by

$$\hat{S}(t) = W(X(t) - \sum_{\tau=1}^{p} A(\tau) X(t-\tau)) \tag{3.4}$$

where $A(\tau)$ is the demixing matrix of the order τ that accounts for the time-lagged propagation effect.

For the convenience of analysis, we divide the reconstruction problem of $S(t)$ into two parts. First, we define

$$\widetilde{X}(t) \;=\; X(t) - \sum_{\tau=1}^{p} A(\tau)X(t-\tau) \tag{3.5}$$

where $\widetilde{X}(t)$ is the signal processed by a finite multi-variate FIR filter of order p. For the simplicity of presentation, we refer to it as the time-decorrelated data in the following discussion. The source signal can be recovered from the time-decorrelated data $\widetilde{X}(t)$ by

$$\hat{S}(t) = W\widetilde{X}(t) \tag{3.6}$$

Although calculating $\hat{S}(t)$ based on (3.5) and (3.6) resembles the causal connectivity estimation based on MVAR analysis, the objective of this work is discriminative learning, different from the connectivity identification in [112, 111, 116]. For connectivity analysis, $S(t)/\hat{S}(t)$ is usually regarded as the innovation process which is a temporally and spatially uncorrelated time sequence. In contrast, $\hat{S}(t)$ based on (3.5) and (3.6) is assumed to be the discriminative signal with the ERD/ERS effects enhanced by the demixing matrix $A(\tau)$. Detailed discussions of the differences and relationship between the connectivity analysis and the proposed method can be found in Appendix A.2. Thus, based on the convolutive model, propagation effects can be addressed in the discriminative model. The joint estimation of $A(\tau)$ and W in

(3.5) and (3.6) for the objective of classification is introduced in the following section.

3.2 Joint Estimation of Propagation and Spatial Pattern

We adopt the principle of CSP in the joint estimation of propagation and spatial pattern. The normalized sample covariance matrix R^i of trial i is obtained as

$$R^i = \frac{X^i (X^i)^T}{\mathrm{tr}[X^i (X^i)^T]} \tag{3.7}$$

Suppose that the signal power is to be maximized for class $+$, the objective function in CSP can be given in the form of optimization by

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \ \mathbf{w} R^+ \mathbf{w}^T \ \text{ s.t. } \ \mathbf{w}(R^+ + R^-)\mathbf{w}^T = 1 \tag{3.8}$$

To extract ERD/ERS, we deal with the estimation of $S(t)$ in the proposed model by adopting the variance discriminative objective in CSP. To embed the estimation of $A(\tau)$ in (3.4) into the objective function (3.8), we rewrite (3.5) to make the relationship between raw EEG data X and the time-decorrelated data $\widetilde{X}$ more compact by defining

$$\hat{A}(\tau) = \begin{cases} I, & \tau = 0; \\ -A(\tau), & \tau > 0. \end{cases} \tag{3.9}$$

which we refer to as the time-lagged demixing matrix for simplicity. There-

fore, $\widetilde{X}(t)$ in (3.5) becomes

$$\widetilde{X}(t) \;=\; \sum_{\tau=0}^{p} \hat{A}(\tau) X(t-\tau) \tag{3.10}$$

Similarly, the covariance matrix of $\widetilde{X}(t)$ is

$$\widetilde{R}^i = \frac{\widetilde{X}^i (\widetilde{X}^i)^T}{\mathrm{tr}(\widetilde{X}^i (\widetilde{X}^i)^T)} \tag{3.11}$$

and the average covariance based on $\widetilde{X}(t)$ for each class is

$$\widetilde{R}^c = \frac{1}{|\mathcal{Q}^c|} \sum_{i \in \mathcal{Q}^c} \widetilde{R}^i, \quad c \in \{-, +\} \tag{3.12}$$

Replacing R^c in (3.8) with $\widetilde{R}^c$ and considering (3.10) and (3.11), the optimization problem becomes

$$\begin{aligned} &\max_{\mathbf{w}, \hat{A}(\tau)} \mathbf{w} \left(\sum_{\tau_1=0}^{p} \sum_{\tau_2=0}^{p} \hat{A}(\tau_1) R^+(\tau_\Delta) \hat{A}(\tau_2)^T \right) \mathbf{w}^T, \quad \text{s.t.} \\ &\mathbf{w} \left(\sum_{\tau_1=0}^{p} \sum_{\tau_2=0}^{p} \hat{A}(\tau_1) (R^+(\tau_\Delta) + R^-(\tau_\Delta)) \hat{A}(\tau_2)^T \right) \mathbf{w}^T = 1 \end{aligned} \tag{3.13}$$

where $R^c(\tau_\Delta) = \frac{1}{|\mathcal{Q}^c|} \sum_{i \in \mathcal{Q}^c} X^i(t-\tau_1)(X^i(t-\tau_2))^T$. In this way, the estimation of model (3.4) can be achieved by solving the optimization problem in (3.13). Moreover, as shown in (3.13), only one $\hat{A}(\tau)$, as a part of the feature extraction model, is obtained upon the completion of the optimization since the calculation is conducted with the averaged matrix $R^c(\tau_\Delta)$ over all the trials. This is very different from the regression model in connectivity analysis, where the estimated models are usually different for different trials.

Since the above objective function can be highly nonlinear, we adopt an iterative procedure to estimate the spatial filter $\mathbf{w}$ and the time-lagged demixing matrix $\hat{A}(\tau)$. We alternatively update one while fixing the other. The spatial filter $\mathbf{w}$ can be obtained (with fixed $\hat{A}(\tau)$) by solving (3.8) with R^c substituted with $\widetilde{R}^c$ as

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \ \mathbf{w}\widetilde{R}^{+}\mathbf{w}^T \ \text{ s.t. } \ \mathbf{w}(\widetilde{R}^{+} + \widetilde{R}^{-})\mathbf{w}^T = 1 \tag{3.14}$$

For $\hat{A}(\tau)$, we calculate the j-th column of $\hat{A}(\tau)$, $[\hat{a}_{1j}, \hat{a}_{2j}, \ldots, \hat{a}_{n_c j}]^T$, separately based on a fixed spatial filter $\mathbf{w}$ and $[\hat{a}_{1m}, \hat{a}_{2m}, \ldots, \hat{a}_{n_c m}]^T$ ($m = 1, \ldots, n_c$ and $m \neq j$) from the last iteration. In this way, the information flow from different channels is optimized individually, and the update of $\hat{A}(\tau)$ finishes upon the completion of estimating $[\hat{a}_{1j}, \hat{a}_{2j}, \ldots, \hat{a}_{n_c j}]^T$ for $j = 1, \ldots, n_c$. The implementation of the proposed discriminative learning algorithm of propagation and spatial patterns is summarized in Algorithm 1. The loop will not stop until the convergence criteria are met.

Note that during the optimization, only one spatial filter $\mathbf{w}$ is used. After completion of the optimization, $\widetilde{X}$ can be obtained from (3.10), and subsequently $\widetilde{R}^c$ can be obtained based on (3.11). Therefore, by replacing R^c with $\widetilde{R}^c$, we can calculate the projection matrix W as in (2.2)-(2.10), and select r pairs of spatial filters corresponding to the r largest/smallest eigenvalues in (2.11) as in the usual CSP procedure. Replacing X^i in (2.13) with $\widetilde{X^i}$, the feature $\tilde{\mathbf{f}}^j$ for trial i is obtained as

$$\tilde{\mathbf{f}}_j^i \ = \ \log\frac{\mathbf{w}_j \widetilde{X}^i (\widetilde{X}^i)^T \mathbf{w}_j^T}{\sum_j \mathbf{w}_j \widetilde{X}^i (\widetilde{X}^i)^T \mathbf{w}_j^T}, \ j = 1, \ldots, r, n_c - r + 1, \ldots, n_c \tag{3.15}$$

where $\widetilde{\mathbf{f}}_j^j$ is the j-th element of $\widetilde{\mathbf{f}}^j$.

Algorithm 1 Discriminative learning of propagation and spatial pattern

Input: Training EEG data with class labels;
Output: Spatial filter $\mathbf{w}$ and time-lagged demixing matrix $\hat{A}(\tau)$.
begin
  Set the initial parameters of the spatiotemporal filters $\hat{A}(\tau)$ as zero matrices;
  while $k < n_k$ **do**
    Compute $\widetilde{X}$ based on $\hat{A}(\tau)$ using (3.10);
    Compute $\mathbf{w}$ by solving the optimization problem in (3.14);
    % *Update the spatial filter* $\mathbf{w}$
    for $j = 1 : n_c$ **do**
      Compute $[\hat{a}_{1j}, \hat{a}_{2j}, \ldots, \hat{a}_{n_c j}]^T$ based on the updated spatial filter $\mathbf{w}$ by solving the optimization problem in (3.13);
      % *Update* $\hat{A}(\tau)$.
    Compute the change in the norm $\hat{A}(\tau)$ by $\delta = \frac{\|\hat{A}(\tau)^k\| - \|\hat{A}(\tau)^{k-1}\|}{\|\hat{A}(\tau)^{k-1}\|}$;
    if $\delta < \zeta$ *(ζ is a small preset constant)* **then**
      Stop.
    k=k+1;

3.3 Background Noise Separation

In this section, we investigate the effectiveness of introducing the time-lagged demixing matrix $\hat{A}(\tau)$ into the estimation of the ERD/ERS source. To further analyze and evaluate the proposed model, the difference between the time-decorrelated EEG signal $\widetilde{X}(t)$ in (3.5) and original EEG data $X(t)$ is investigated. Suppose $X(t)$ is described by the following MVAR model

$$X(t) = \sum_{\tau=1}^{q} B(\tau)X(t-\tau) + N(t) \tag{3.16}$$

where $N(t)$ is the prediction error. It is also regarded as the innovation process because it is spontaneous and cannot be totally predicted by past observations [111]. Note that $B(\tau)$ is the mixing matrix based on the regression model, which is different from $A(\tau)$ estimated in the proposed model for discriminative purpose and q is the order of the MVAR model. Similarly, (3.16) is rearranged in the following form to make the input-output relationship more compact

$$N(t) = \sum_{\tau=0}^{q} \hat{B}(\tau) X(t-\tau) \tag{3.17}$$

where

$$\hat{B}(\tau) = \begin{cases} I, & \tau = 0; \\ -B(\tau), & \tau > 0. \end{cases} \tag{3.18}$$

Transforming (3.17) into the frequency domain yields

$$N(f) = B(f)X(f) \tag{3.19}$$

$$B(f) = \sum_{\tau=0}^{q} \hat{B}(\tau) e^{-i2\pi f\tau} \tag{3.20}$$

where f is the frequency. Therefore, the transfer function of the system $H(f)$ can be described by

$$H(f) = B^{-1}(f) \tag{3.21}$$

such that $X(f) = H(f)N(f)$.

By substituting (3.16) into (3.5) and following the steps from (3.19) to

(3.21), we obtain

$$\begin{aligned} \widetilde{X}(f) &= (I - A(f))X(f) \\ &= (H(f) - \frac{A(f)}{B(f)})N(f) \end{aligned} \tag{3.22}$$

where

$$A(f) = \sum_{\tau=0}^{p} \hat{A}(\tau) e^{-i2\pi f\tau} \tag{3.23}$$

Let $\widetilde{H}(f) = H(f) - \frac{A(f)}{B(f)}$, which is the transfer function from $N(f)$ to $\widetilde{X}(f)$. Since the causal flow measurement DTF is defined based on the transfer function [107], we see that the proposed method changes the information flow by changing the transfer function from $H(f)$ to $\widetilde{H}(f)$. Moreover, comparison of the transfer functions of $\widetilde{X}$ and X in (3.22) shows its similarity to the classical signal-plus-noise (SPN) model. In particular, in [117] the observed EEG data containing ERP $X_E(f)$ is usually formulated as

$$X_E(f) = \Phi_E S_E(f) + Z(f) \tag{3.24}$$

where $S_E(f)$ is the ERP of interest, Φ_E is the projection matrix, and $Z(f)$ is the background noise or the ongoing activity.

As discussed in [117], the background noise is not a noise despite its noise-like appearance but represents ongoing brain activity rich in oscillatory content. In the light of the above discussion, we can interpret (3.22) from a similar perspective. As indicated in (3.22), the frequency component removed from X is an oscillatory signal with a transfer function $\frac{A(f)}{B(f)}$

and it can be regarded as an estimate of ongoing activity. In this way, the ERD/ERS components are enhanced in the proposed model with the oscillatory background noise attenuated.

The KL-divergence is a measure of probability divergence given two probability distributions, and it has been utilized to evaluate nonstationarity in motor imagery EEG classification problem [103, 93, 77]. Therefore, to verify that the component removed from X is the background noise, we adopt the KL-divergence as the criterion.

As the Gaussian model is usually used to model EEG data, we consider the KL-divergence between two Gaussian distributions. Assume two Gaussian distributions $\mathcal{N}^0(\mu_N^0, \Sigma_N^0)$ and $\mathcal{N}^1(\mu_N^1, \Sigma_N^1)$. Then, the KL-divergence between them is

$$\begin{aligned} &D_{\mathrm{KL}}(\mathcal{N}^0(\mu_N^0, \Sigma_N^0)||\mathcal{N}^1(\mu_N^0, \Sigma_N^0)) = \tfrac{1}{2}(\mathrm{tr}((\Sigma_N^1)^{-1}\Sigma_N^0) \\ &-(\mu_N^1 - \mu_N^0)^T(\Sigma_N^1)^{-1}(\mu_N^1 - \mu_N^0) - \ln(\tfrac{\det \Sigma_N^0}{\det \Sigma_N^1} - k_N)) \end{aligned} \tag{3.25}$$

where $\mu_N^1(\mu_N^0)$ and $\Sigma_N^1(\Sigma_N^0)$ are respectively the mean and covariance of the distribution $\mathcal{N}^1(\mu_N^1, \Sigma_N^1)(\mathcal{N}^0(\mu_N^0, \Sigma_N^0))$. It is reasonable to assume that the improved separation of background noise will result in more stationary data with less within-class dissimilarities. We therefore adopt KL-divergence to measure such within-class dissimilarities. The smaller the KL-divergences within trials from the same class, the less the variation of the data, which generally relates to better classification results. Since EEG data is usually processed to be centered and the dimension k_N of the distribution is the number of channel n_c, for every trial i from class c, we use $D_{\mathrm{KL}}(\mathcal{N}(0, R^i)||\mathcal{N}(0, R^c))$ to measure the dissimilarity of the distribution of this trial from the mean

distribution of the class c as

$$D_{\mathrm{KL}}(\mathcal{N}(0,R^i)||\mathcal{N}(0,R^c)) = \frac{1}{2}(\mathrm{tr}((R^i)^{-1}R^c) - \ln(\frac{\det R^i}{\det R^c}) - n_c) \quad (3.26)$$

and subsequently we obtain an average probability divergence D for EEG data X as

$$D = \sum_{c=+,-} \frac{1}{|\mathcal{Q}^c|} \sum_{i \in \mathcal{Q}^c} D_{\mathrm{KL}}(\mathcal{N}(0,R^i)||\mathcal{N}(0,R^c)) \quad (3.27)$$

Similarly, we obtain $\widetilde{D}$ based on $\widetilde{X}$ as

$$\widetilde{D} = \sum_{c=+,-} \frac{1}{|\mathcal{Q}^c|} \sum_{i \in \mathcal{Q}^c} D_{\mathrm{KL}}(\mathcal{N}(0,\widetilde{R}^i)||\mathcal{N}(0,\widetilde{R}^c)) \quad (3.28)$$

In this way, by comparing D and $\widetilde{D}$, we can evaluate the quality of X and $\widetilde{X}$ in terms of within-class dissimilarities.

Moreover, The proposed method addresses a more complicated dynamics of motor imagery EEG but does not depend on the very critical explanation of the generation of ERD/ERS. On the one hand, it is possible that propagation effects that contribute to the generation of ERD/ERS exist. On the other hand, discriminative sources could correlate with noise in a convolutive manner. Blind source separation or connectivity estimation methods, as discussed before, may not be effective for this classification problem because it is difficult to differentiate between these two kinds of propagation effects. The proposed model, which is formulated in a phenomenological form (3.22), takes both cases into consideration.

3.4 Experimental Study

3.4.1 Experiment Set-Up and Data Description

A total of 16 subjects participated in the study with informed consent. Ethics approval was obtained beforehand from the Institutional Review Board of National University of Singapore. EEG from 27 channels were obtained using Nuamps EEG acquisition hardware with unipolar Ag/AgCl electrodes channels. The sampling rate was 250 Hz with a resolution of 22 bits for the voltage range of $\pm$ 130 mV. A bandpass filter of 0.05 to 40 Hz was set in the acquisition hardware.

In the experiment, the training and test sessions were recorded on different days with the subjects performing motor imagery. During the EEG recording process, the subjects were asked to avoid physical movement and eye blinking. Additionally, they were instructed to perform kinesthetic motor imagery of the chosen hand in two runs. During the idle state, they did mental counting to make the resting EEG signal more consistent. Each run lasted for approximately 16 minutes and comprised 40 trials of motor imagery and 40 trials of idle state. Each training session consisted of 2 runs while the test session consisted of 2-3 runs. Details of the scalp map of the 27 channels and the segmentation of one trial can be found in Appendix A.1, and more details of experiment setup can be found in [118].

3.4.2 Data Processing

We select the time segments from 0.5s to 2.5s after the cue [96]. The raw data is pre-filtered by a 8-35Hz band pass filter that covers rhythms related

to motor imagery. The filtered training data is used to train the feature extraction model based on the proposed method as described in Section 3.2, where $\zeta = 0.02$ and $n_k = 30$. The number of spatial filters in W is chosen as 2 ($r = 2$ in (3.15)). As discussed in [69], for BCI tasks with explicit known cue, the linear support vector machine (SVM) shows advantages. Thus, in this work, we adopt the linear SVM with a soft margin, which is trained by the extracted training features first and applied to test features to obtain the predicted labels.

3.4.3 Investigation on the Order of the Time-Lagged Demixing Matrix

To determine the order p of $\hat{A}(\tau)$ in (3.10), we fit the MVAR model to EEG data as in (3.16). Although the orders p and q have different meanings, the analysis of the order q of the mixing matrix $B(\tau)$ in (14) gives the time-lagged level at which the propagation effects are stronger. From (20) and the analysis given in Section 3.3, it is reasonable to choose the order p of $\hat{A}(\tau)$ in accordance with q, the order of $B(\tau)$, as $\hat{A}(\tau)$ corresponds to certain components of $B(\tau)$ in frequency domain. Therefore, the analysis of the mixing matrix $B(\tau)$ can be used to initialize the order p of $\hat{A}(\tau)$ in the proposed model. The Swartz Bayesian criterion is used to automatically select the model order that best matches the data [119]. We found that for every subject, order 5 for q is selected for most of the trials and order 4 or 6 is selected for the remaining of the trials. Therefore, we restrict the investigation to orders 4, 5 and 6.

Figure 3.1 illustrates the result of one subject in the dataset introduced

in Section 3.4.1. The y-axis indicates the value of the norm of mixing matrix $B(\tau)$ in (3.16) of different orders and the x-axis indicates the order τ. The coefficient matrices are obtained under MVAR models with q equal to 4, 5 or 6, and averaged over the training set and test set respectively, resulting in the six lines in Figure 3.1. We see that, in all six cases, the norms of the coefficient matrices of orders 2 and 3 are the highest, which means that the data at time t is most influenced by the data at time $t-2$ and time $t-3$. Therefore, the order p of $\hat{A}(\tau)$ should include these two time lags, and subsequently the proposed discriminative learning model addresses the most influential propagation effects. Furthermore, we focus on investigating the feasibility of the proposed model with order 4 and below. It is sufficient that $\hat{A}(\tau)$ covers only the major components in $\hat{B}(\tau)$ to aim at the most influential propagation effect.

3.4.4 Classification Results

Tables 3.1 summarizes the performance of the proposed feature extraction method, compared with CSP as the baseline. In the table, we refer to the proposed method as discriminative propagation and spatial pattern analysis (DPSP), and results of DPSP with $p = 1, 2, 3, 4$ are included.

As shown by the classification results, the proposed feature extraction method improves the performance of the classifier, and the improvements are significant when the order of $\hat{A}(\tau)$ in DPSP is 2 or 3, which is in agreement with the previous analysis based on the MAVR model. Specifically, the average classification accuracy for order 2 is 68.30% and the accuracy for order 3 is 67.91%, both of which are higher than that of CSP (65.56%).

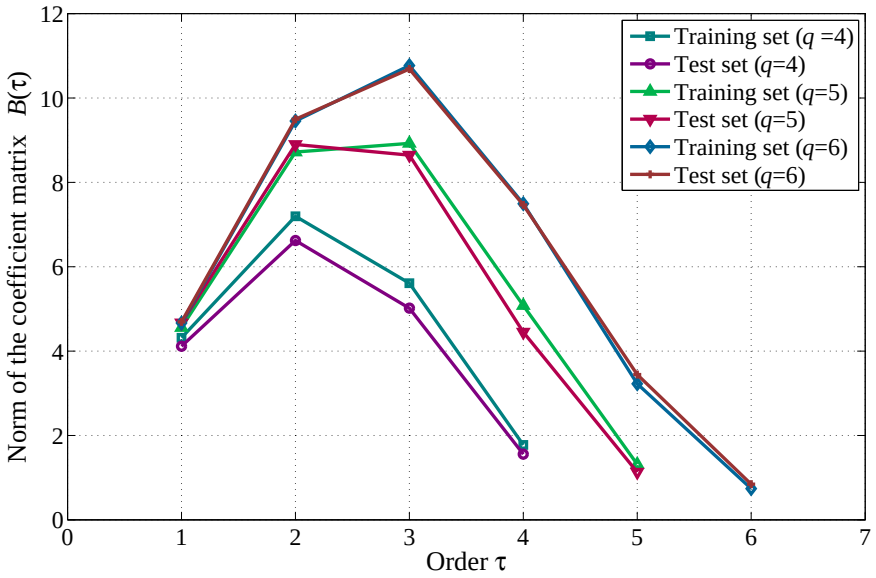


Figure 3.1: Norms of coefficient matrices under MVAR model. The x-axis represents the order τ and y-axis represents the norm of $B(\tau)$. Three MVAR models with orders q from 4 to 6 are used to fit EEG data of training and test sets separately, yielding six lines. And the peak points of the six lines correspond to either $\tau = 2$ or $\tau = 3$.

The paired t-test confirms the significance of the improvement at a 5% level with p-values equal to 0.008 and 0.040, corresponding to the cases of $p = 2$ and $p = 3$, respectively. The accuracy for order 4 is 66.01%, which are not significantly different. Moreover, the accuracy for order 1 is almost the same as that of CSP, which also confirms our previous analysis, i.e., it is necessary and sufficient for $\hat{A}(\tau)$ to cover the major components of $\hat{B}(\tau)$. The propagation effect is strongest at orders 2 and 3 so that the optimization based on $\hat{A}(\tau)$ for order 1 has little effect and results in almost the same result. The optimization based on $\hat{A}(\tau)$ of order 4 accounts for most of the propagation effect, but using more parameters will pose a risk of over-fitting. Theoretically, the higher the order of $\hat{A}(\tau)$, the better the results would be,

since more information is taken into consideration. However, the increased number of parameters would give rise to over-fitting, which would adversely affect the classification performance. A good balance between accounting for the propagation effects and over-fitting is obtained by covering as few major components of propagation as possible, which comes from orders 2 and 3 in this experiment.

Figure 3.2 is used to compare the results in a more intuitive manner. Each plot in Figure 3.2 shows the test accuracy under DPSP with order p against that under CSP. The x-axis represents the accuracy results under CSP and the y-axis represents that under DPSP. In each plot, a circle above the diagonal line marks a subject for which DPSP outperforms CSP.

3.4.5 Analysis of Background Noise Separation

To further verify the validity of DPSP, we have evaluated the classwise KL-divergence (Section 3.3), and results averaged among all subjects are shown in Table 3.2 and Figure 3.3. Note that for the computation of D_{KL} of both training set and test set, the average covariance matrix R^c ($\widetilde{R}^c$) is the mean of the training set since under the single-trial analysis setting we cannot obtain the mean of the test set. Therefore, the fact that the average divergence D of the test set is larger than that of the training set in all cases reflects the differences between the test set and the training set, as indicated by Table 3.2. This is mainly caused by the session-to-session transfer effects. According to the results, the proposed DPSP algorithm decreases the KL-divergence within the same class for both the training set and the test set, which means that, compared to the EEG data X, the data processed by

Table 3.1: Session-to-session transfer test results (%)

Subject	**CSP**	**DPSP**			
		p=1	p= 2	p=3	p=4
1	65.00	65.41	62.91	66.66	67.08
2	51.25	51.25	54.17	52.08	52.08
3	55.00	55.00	57.50	55.83	55.00
4	66.67	66.67	70.41	71.25	77.08
5	54.58	54.16	67.08	70.41	58.33
6	67.08	67.50	72.50	69.16	69.58
7	77.08	77.08	77.92	76.66	72.5
8	94.16	94.16	92.50	96.25	95.41
9	74.58	75.00	75.83	75.83	74.58
10	61.66	61.25	60.41	60.83	60.00
11	46.25	46.67	49.16	53.33	47.08
12	77.00	77.08	81.25	79.58	73.33
13	51.25	51.25	54.58	51.25	50.00
14	72.08	72.08	79.16	73.75	74.58
15	65.83	65.58	67.50	64.16	64.58
16	69.58	69.60	70.00	68.75	65.00
mean	65.56	65.59	68.30	67.91	66.01
std	12.26	12.28	11.57	11.79	12.35
p**-value**	-	0.64	**0.008**	**0.040**	0.63

p is the order of $\hat{A}(\tau)$ in DPSP. The larger p is, the better the results would be with more propagation effects taken into consideration. However, the increased number of parameters would give rise to over-fitting. A good balance is to cover as few major components of propagation as possible, which could be the reason why $p = 2$ and $p = 3$ yield better results than $p = 1$ and $p = 4$.

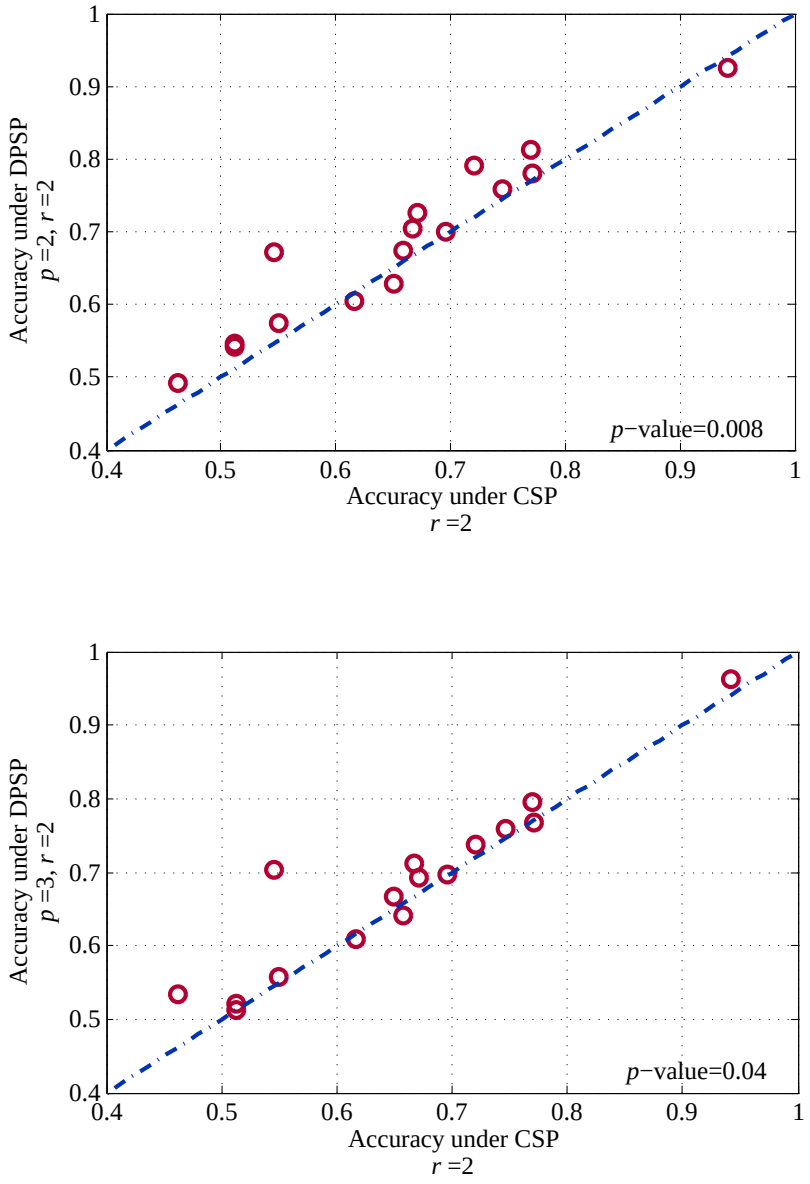


Figure 3.2: Test classification accuracy comparison. The x-axis represents the accuracy result under CSP and the y-axis represents that under DPSP with different orders p. The $y = x$ line is denoted in dotted-dashed line. In each plot, a circle above the $y = x$ line marks a subject for which DPSP outperforms CSP. It can be seen from the plots that improvements of DPSP for order 2 and 3 are significant.

Table 3.2: KL-divergence comparison(%)

		p=2		p=3		p=4	
	D	$\widetilde{D}$	$1-\frac{\widetilde{D}}{D}$	$\widetilde{D}$	$1-\frac{\widetilde{D}}{D}$	$\widetilde{D}$	$1-\frac{\widetilde{D}}{D}$
Training set	4.96	4.09	17.68%	4.25	14.39%	4.84	2.55%
Test set	64.3	25.2	60.84%	36.68	42.98%	57.09	11.24%

p is the order of $\hat{A}(\tau)$ in DPSP. The decreases in the KL-divergence in $\widetilde{X}$ of different orders compared to X are shown in percentage. Great decrease in the KL-divergence indicates that $\widetilde{X}$ is more stationary than X. The decrease is more significant for the test data.

DPSP $\widetilde{X}$ is more stationary. A more significant decrease is achieved for the test set, which means that the proposed method is more stationary against the session-to-session transfer effects. Moreover, the comparison between different orders indicates that a better performance is achieved with order 2, which is in accordance with the classification accuracy results.

Figure 3.4 illustrates the correlation between the decrease of KL-divergence and the increase of the classification accuracy at the subject level. The linear correlation coefficient r_c equals to 0.30 and 0.31 corresponding to $p = 2$ and $p = 3$, respectively. Due to the large variety across subjects, their KL-divergence may lie in different feature spaces. The decrease of KL-divergence and the increase of classification performance may not correlate linearly. It can be seen that almost all the points lie in the first quadrant, indicating that the decrease in the KL-divergence contributes to the increase in the classification accuracy to a certain extent. Nevertheless, there could be additional factors that contribute to the increase in classification accuracy, and this would be an interesting topic for future work.

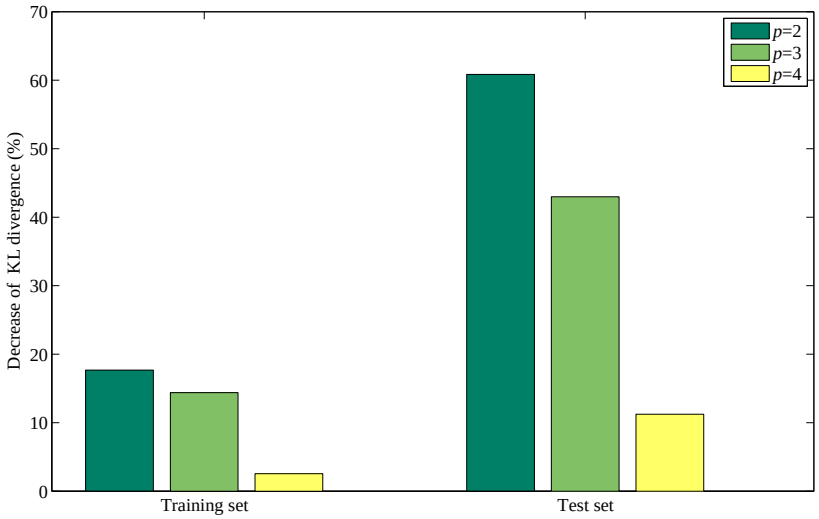
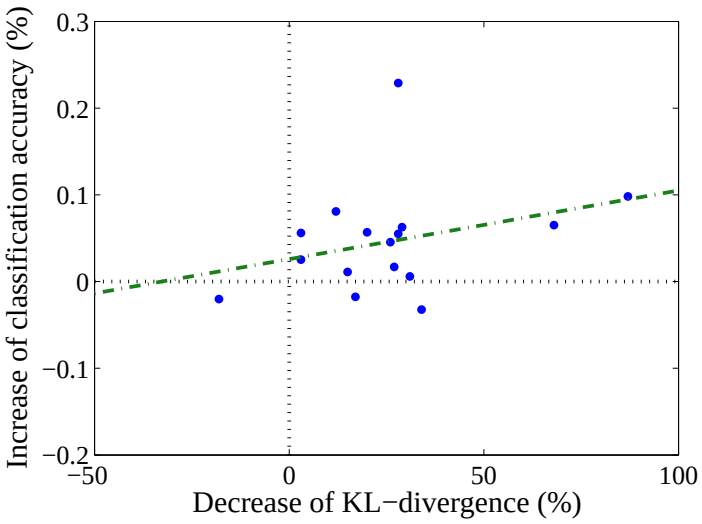
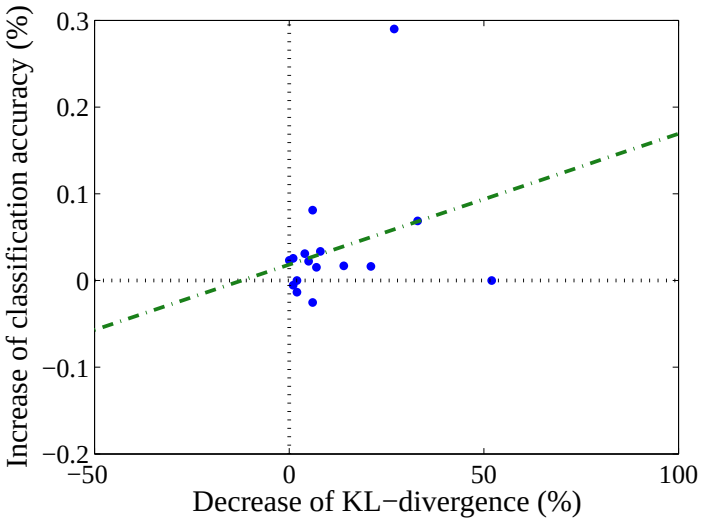


Figure 3.3: Decrease in the KL-divergence. The decreases in the KL-divergence in $\widetilde{X}$ of different orders compared to X are shown in percentage. Great decrease in the KL-divergence indicates that $\widetilde{X}$ is more stationary than X. Therefore, the proposed DPSP algorithm can reduce varying background noise and session-to-session transfer effects.



(a) $r_c = 0.30$ $(p = 2)$



(b) $r_c = 0.31$ $(p = 3)$

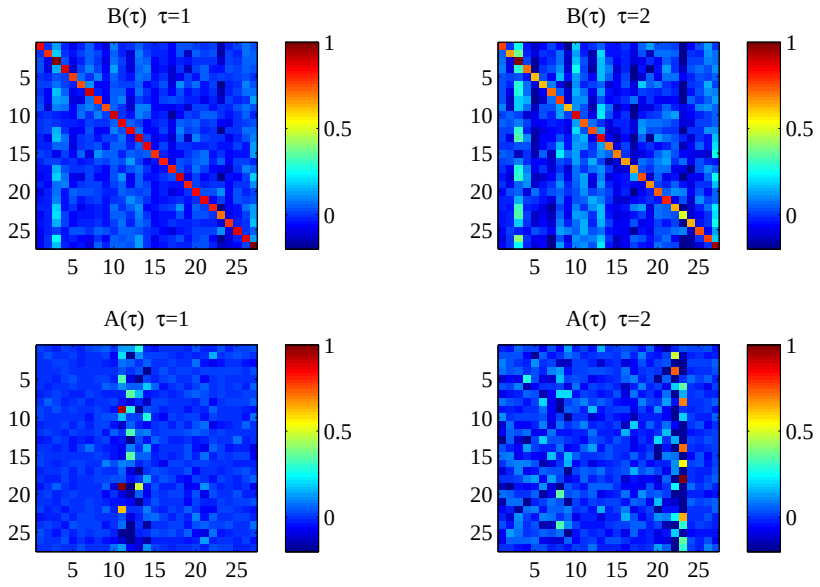
Figure 3.4: Correlation between the decrease of the KL-divergence and the increase of the classification accuracy. The x-axis represents the decrease of the KL-divergence and y-axis represents the increase of the classification accuracy. Subfigures (a) and (b) correspond to $p = 2$ and $p = 3$, respectively.

3.4.6 Discussion

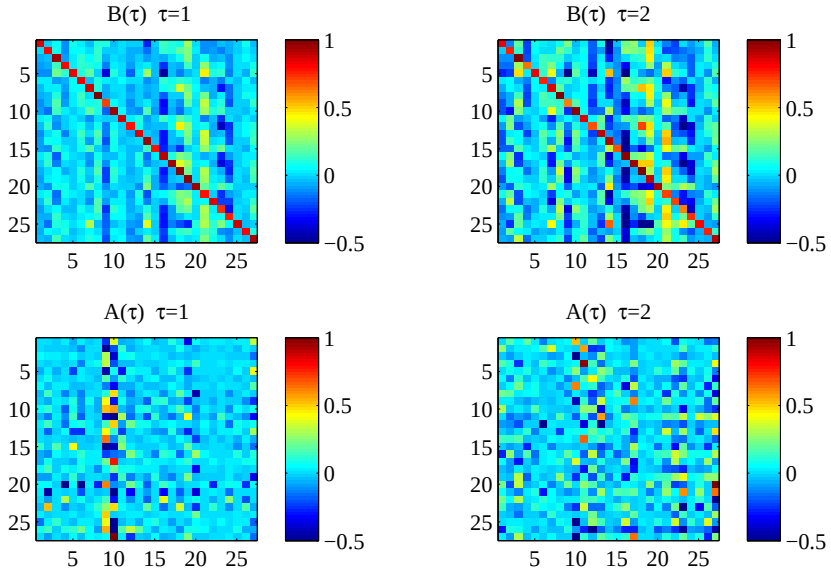
Figure 3.5 displays $A(\tau)$ for two subjects. For a better comparison of differences between the proposed method and the MVAR model, mixing matrices $B(\tau)$ based on the MVAR model of the two subjects are also provided. As seen in Figure 3.5, the diagonal elements of $B(\tau)$ are much higher than the off-diagonal elements, because the auto spectrum of the signal is usually stronger than the cross spectrum between the EEG signals from different channels. However, there are no large differences between diagonal elements and off-diagonal elements in $A(\tau)$. Since the diagonal elements of $A(\tau)$ are not significantly larger than the off-diagonal ones, the auto spectrum of the signal is not modulated radically by $A(\tau)$. Moreover, elements of higher values concentrate in certain columns in $A(\tau)$, which means that the propagation effects from a certain channel are modified more substantially than that from other channels.

Moreover, a comparison between Figures 3.5 (a) and (b) shows that the coefficient matrices $A(\tau)$ are quite different for different subjects due to the large inter-subject variability. With more parameters optimized for each subject, the proposed method may not be suitable for the inter-subject task, which is one limitation of the proposed method.

Regarding the number of spatial filters r, usually 2 or 3 pairs of spatial filters are used, i.e., $r = 2, 3$ [91, 68]. Experimental studies have also been conducted with $r = 3$. The proposed method shows significant improvements when $p = 2$ and $p = 3$, which is similar to the case with $r = 2$. Regarding the parameters used in the iteration, $n_k = 30$ is chosen based on extensive tests. For ζ, as long as the relative difference of the norm of $A(\tau)$ is small, e.g.,



(a) Comparison between $A(\tau)$ and $B(\tau)$ for subject 7



(b) Comparison between $A(\tau)$ and $B(\tau)$ for subject 14

Figure 3.5: Comparison of coefficient matrices obtained by the proposed method, $A(\tau)$, and the mixing matrices in MVAR, $B(\tau)$. For both subjects, the diagonal elements of $B(\tau)$ are much higher than the off-diagonal elements. For $A(\tau)$, elements of higher values are found in certain columns.

$\delta < 0.02$, the resultant differences of the feature extraction would be limited. Therefore, the sensitivity of the proposed method to these two parameters is acceptable.

3.5 Conclusion

Co-existence of brain connectivity and volume conduction may have complicated effects in EEG measurements, and poses technical challenge to detecting the correct motor imagery condition. Conventional linear spatial filters designed with the instantaneous mixing model are not sufficient in addressing such complicated dynamics. Due to the causal relationship, reconstructed ERD/ERS signals based on the instantaneous demixing may not be optimal in terms of discrimination.

Moreover, the propagation effects are closely related to the background noise and nonstationarity of EEG. It is possible that an electrode that actually contains no discriminative information could be given a high weight due to information flow from signals containing ERD/ERS and such dependence could be very unstable compared with the original ERD/ERS source. The above analysis is the motivation to propose the computational model for the discriminative learning of propagation and spatial patterns.

We have reported in this chapter a novel computational model that accounts for both time-lagged correlations between signals and the volume conduction effect. Experimental results have shown statistically significant improvement in classification accuracy under the proposed learning method. Moreover, the effectiveness of the background noise attenuation is also confirmed with a significant decrease of KL-divergence of EEG data of the same

class, especially for test data.

Chapter 4

Ensemble Learning of Spatial Filter Design

CSP designs spatial filters by jointly diagonalizing the average covariance matrices from two different classes, and subsequently, the effectiveness of feature extraction is somewhat sensitive to the estimates of the covariance matrices. In other words, biased estimates of covariance matrices may result in an inaccurate feature extraction model.

To this end, in this chapter we introduce an ensemble learning framework for spatial filter design by considering the mismatch between data and model. In particular, there exist some training trials for which the projection matrix in CSP fails to extract discriminative features. This may be caused by the fact that the covariance matrices used to construct the CSP model are biased estimates for those trials. These trials in the training set are utilized to re-estimate the covariance matrices and projection matrices. Instead of giving different weights to the training trials, the projection matrix for feature extraction is obtained by integrating multiple projection matrices that are estimated using different subsets of the training trials. Based on the integrated model, feature extraction is carried out and is followed by classification. The validity of the proposed method is verified through experimental

studies with two sets of two-class motor imagery data. The results show that the proposed method is able to generate more discriminative features.

This chapter is organized as follows. In Section 4.1, the problem of mismatch between model and data is discussed, followed by the details of the proposed spatial filter design method based on ensemble learning. In Section 4.2, the validity of the proposed method is verified by experimental studies on the two-class motor imagery classification problem. Concluding remarks are given in Section 4.3.

4.1 Spatial Filter Design Based on Ensemble Learning

4.1.1 Problem Formulation

The projection matrix W in CSP is computed from the average covariance matrices of the band-passed EEG signals from different classes. However, due to the discrepancies between the covariance matrices and the estimates, W may fail to extract discriminative features for certain trials.

Figure 4.1 shows an example of a 2D feature distribution obtained by CSP before taking the logarithm. The dataset used to obtain the results in Figure 4.1 are described in Section 4.2. Given the function of CSP projection matrix W in (2.2)-(2.10), ideally all the features of class + should be around the lower-right side, with $\mathbf{f}_1$ maximized and $\mathbf{f}_2$ minimized, and the other way around for class $-$. The line $x = y$ is a natural classifier for the features from two classes. However, as shown in Figure 4.1, there exist many features lying on the wrong side, indicated by red and blue crosses, for which the classifier

would very likely give wrong labels.

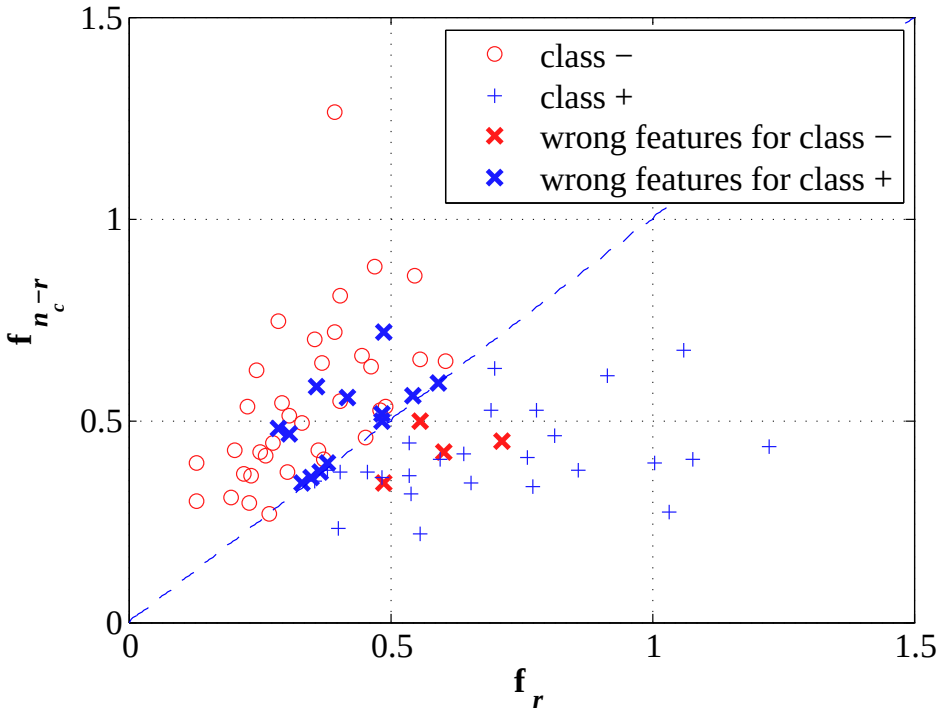


Figure 4.1: An example of a 2D feature distribution obtained by CSP. The line $x = y$ is denoted in dashed line, which can be regarded as a classifier. Red and blue crosses represent features lying on the wrong side.

The example shows that there exists a mismatch between some samples and the CSP model based on the average covariance matrices. Such a mismatch can be attributed to the covariances of those exceptional samples being very different from the average ones. Since neither the distribution of the raw EEG data nor the covariance matrices can be measured directly, it is difficult to evaluate whether the average covariance matrices are biased, and would give rise to the mismatch.

To this end, in the following section, we will introduce a method to reduce

the mismatch caused by covariance matrix discrepancies, and thus improve the separability of features.

4.1.2 Spatial Filter Design

To take the mismatch between certain trials and the feature extraction model into consideration, we propose a spatial filter design method by ensemble learning. Figure 4.2 shows the flow chart of the proposed method. In particular, subsets of training data consisting of exceptional trials are formed, different spatial filters are generated based on different subsets of trials, and finally the feature extraction model, W_e, is obtained by combining these models.

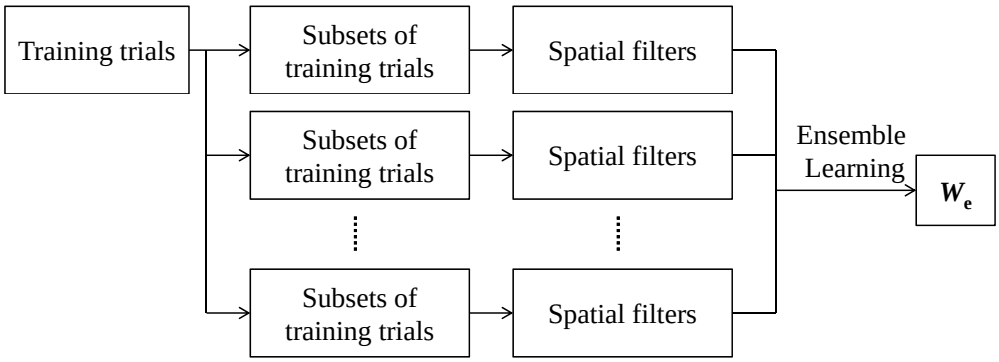


Figure 4.2: Flow chart of the proposed method. Subsets of training data consisting of exceptional trials are formed, different spatial filters are generated based on different subsets of trials, and finally the feature extraction model, W_e, is obtained by combining these models.

The following sections are dedicated to introduce the proposed spatial filter design method in two steps: selection of exceptional samples and ensemble learning of spatial filter design.

4.1.2.1 Selection of Exceptional Samples

In this section, we will discuss the criterion to select the exceptional samples. The feature reflects the mismatch between the model and the data caused by the covariance matrix discrepancies. As described in Section 2.1, $\mathbf{f}^i$ is the variance feature extracted for trial i which contains $2r$ elements. Suppose that the first r features are generated by the spatial filters maximizing the activity of class $+$. According to CSP, if trial i is from class $+$, the first r elements in $\mathbf{f}^i$ should be close to the largest elements in Λ^+, while the last r elements close to the smallest elements in Λ^+. The elements in Λ^+ constitute the mean feature of class $+$. In the general cases, if $r = 1$, $\mathbf{f}_1^i$ would be around one, and $\mathbf{f}_2^i$ around zero. Theoretically, features from two classes can be divided by the line $x = y$. Thus, trials lying on the wrong sides can be considered as exceptional trials, as seen in Figure 4.1. Those trials suffer from relatively larger covariance matrix discrepancies regarding the mean covariance matrices, as the feature extraction model fails to generate proper features for them. Besides, there are also some trials with both feature values close to zero, which are also considered as trials with the mismatch. Be selecting these two kinds of exceptional trials, we can obtain subsets of indexes referred to the exceptional samples for the two classes, $\mathcal{Q}_b^+$ and $\mathcal{Q}_b^-$. The total number of selected trials of class $+$ and class - are denoted as $|\mathcal{Q}_b^+|$ and $|\mathcal{Q}_b^-|$, respectively, similar to $\mathcal{Q}^+$ and $\mathcal{Q}^-$ in (2.1). The following equation describes the re-sampling criteria

$$i \in \begin{cases} \mathcal{Q}_b^+, \text{ when } i \in \mathcal{Q}^+ \text{ and } \mathbf{f}_j^i < \mathbf{f}_{n_c-j+1}^i \text{ or } \mathbf{f}_j^i < \xi, \\ \mathcal{Q}_b^-, \text{ when } i \in \mathcal{Q}^- \text{ and } \mathbf{f}_{n_c-j+1}^i < \mathbf{f}_j^i \text{ or } \mathbf{f}_{n_c-j+1}^i < \xi. \end{cases} \tag{4.1}$$

where $j = 1, 2, ..., r$ denotes the index of feature, and ξ is the parameter controlling which features close to zero to be chosen.

Instead of selecting trials that are misclassified by the classifier, we propose the criterion in (4.1). Generally speaking, the training classification results are not very proper for feature extraction model evaluation. Considering training classification accuracy, it is possible that the classifier could predict the labels correctly for the training trials with less discriminative features. In this case, the classifier is also prone to over-training. Therefore, the features are selected based on the principle of CSP, which is a more direct evaluation of the mismatch between the feature extraction model and the samples.

4.1.2.2 Ensemble Learning of Spatial Filters

After the trial re-sampling step, we calculate the new covariance matrices for trials belonging to $\mathcal{Q}_b^c$ with

$$R_b^c = \frac{1}{|\mathcal{Q}_b^c|} \sum_{i \in \mathcal{Q}_b^c} \frac{X^i (X^i)^T}{\text{tr}[X^i (X^i)^T]}, \ c \in \{+, -\} \tag{4.2}$$

With R_b^+ and R_b^-, projection matrices for trials from $\mathcal{Q}_b^+$ and $\mathcal{Q}_b^-$ can be obtained. W_b^+ is computed using R_b^+ and R^-, while W_b^- is calculated using R^+ and R_b^-, which can be represented in the form of generalized eigenvalue

decomposition similar to (2.15), as follows

$$R_b^+(\mathbf{w}_b^+)^T = \lambda R^-(\mathbf{w}_b^+)^T \tag{4.3}$$

$$R_b^-(\mathbf{w}_b^-)^T = \lambda R^+(\mathbf{w}_b^-)^T \tag{4.4}$$

where $\mathbf{w}_b^-(\mathbf{w}_b^+)$ is a spatial filter. Assuming that the first r spatial filters maximizing the power of EEG signals from class + and the last r spatial filters maximizing that from class -, the computation of the new projection matrix W_e is described by

$$\mathbf{w}_{e,j} = \frac{|\mathcal{Q}^+| - |\mathcal{Q}_b^+|}{|\mathcal{Q}^+|}\mathbf{w}_j + \frac{|\mathcal{Q}_b^+|}{|\mathcal{Q}^+|}\mathbf{w}_{b,j}^+ \tag{4.5}$$

$$\mathbf{w}_{e,n_c-j+1} = \frac{|\mathcal{Q}^-| - |\mathcal{Q}_b^-|}{|\mathcal{Q}^-|}\mathbf{w}_{n_c-j+1} + \frac{|\mathcal{Q}_b^-|}{|\mathcal{Q}^-|}\mathbf{w}_{b,n_c-j+1}^- \tag{4.6}$$

where $\mathbf{w}_{e,j}$ is the j-th row of W_e with $j \leq r$. Similarly, $\mathbf{w}_j$, $\mathbf{w}_{b,j}^+$ and $\mathbf{w}_{b,j}^-$ are respectively used to denote the j-th rows of W, W_b^+ and W_b^-. As shown in (4.5) and (4.6), W_e is the weighted combination of the three projection matrices. The reason why the summation is not conducted between the covariance matrices R_b^+, R_b^-, R^+ and R^- is that it would be equivalent to giving different weights to different trials. Possible cancelling effects of covariance matrix summation would undermine the effect of re-estimation. Since the change in W with respect to the change in covariance matrices is nonlinear and complicated, it is more direct and effective to integrate the model at the projection matrix level. With the projection matrix W_e based on ensemble learning, feature extraction can be conducted. The procedure of calculating W_e in the proposed method is summarized in Algorithm 2.

Algorithm 2 Ensemble learning of spatial filters.

Input: Training data and training labels;
Output: W_e.
begin
Apply CSP to the whole training data to obtain training features;
Select exceptional trials in training data according to the criteria indicated in (4.1);
Compute W_b^+ and W_b^- for the selected samples $X^i,\ i \in \mathcal{Q}_b^+$ or $i \in \mathcal{Q}_b^-$ as indicated in (4.2) to (4.4);
Obtain W_e from W, W_b^+ and W_b^- as indicated in (4.5).
end

4.2 Experimental Study

4.2.1 Experiment Set-Up and Data Description

In this study, two sets of data are used to evaluate the performance of the proposed spatial filter design. The first set is the open dataset BCI competition III Dataset IVa, which contains five subjects. For each subject, a total 280 single-trials of samples, including training and test sets, are recorded using 118 channels. For each trial, the subjects were instructed to perform one of two motor imagery tasks with right hand or foot. In this work, 280 samples of data are divided equally into training and test sets, different from the competition setting that is aimed at the problem of a small number of training samples. Moreover, for computation efficiency, 28 (F3, F1, Fz, F2, F4, FC5, FC3, FC1, FCz, FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, CP6, O1, O2) of 118 channels are used.

The other dataset contains 16 subjects with one motor imagery (MI) session and one passive movement (PM) session collected on the same day. The motor imagery session is as described in Section 3.4.1. During the passive movement session, EEG data were collected from the subjects while passive

movement of the chosen hand was performed using the haptic knob robot. Similarly, the other class is the idle state, during which the subjects also did mental counting as instructed to make the EEG signals more constant. Each passive movement session consists of 2 runs, each of which comprises 40 trials of the passive movement class, and 40 trials of the idle class.

4.2.2 Data Processing

For each trial of data, time segments of 0.5 to 2.5s after the cue were used following most of the works that are employed in this dataset, such as [96, 120]. The raw signal is filtered by band-pass filters of 8-35Hz for the same reason. The filtered signal is used to extract features as described in Section 4.1.

First, CSP is applied to the band passed EEG signals from the training set to obtain projection matrix W. Subsequently, the signals are spatially filtered by the first and last two spatial filters, i.e., $r = 2$, and the variance features are extracted as in (2.4). Then, a new projection matrix W_e is derived according to (4.1) and (4.5), where ξ is set to 0.1. Finally, the first and last two spatial filters in W_e are applied to both training and test data, and the re-calculated variance features are classified by the support vector machine (SVM) classifier.

4.2.3 Classification Results

Tables 4.1 and 4.2 summarize the classification results of the two datasets, where W indicates the baseline method, i.e., CSP, and W_e indicates the proposed ensemble learning method. In Table 4.2, "MI-MI" denotes using

the first run of the motor imagery data as the training data and the second one as the test data, and "PM-MI" denotes using the passive movement data as the training data and the motor imagery data as test data. Usually, MI calibration is more difficult and tedious to perform, and PM data is relatively easier to obtain [118]. Therefore, it would be more desirable for practical implementation that the computational model in BCI trained by PM data could perform well in classifying motor imagery data. This is the reason why using the PM model to test MI data is investigated. Moreover, the incorrectness level $e_r = \frac{\sum_{c=+,-} |\mathcal{Q}_b^c|}{\sum_{c=+,-} |\mathcal{Q}^c|}$ has also been included in Tables 4.1 and 4.2.

It is shown by the comparison that the proposed spatial filter design method improves the performance of the classifier in terms of average classification accuracies with lower standard deviation values in all three cases. Moreover, the proposed method shows more improvements in the "PM-MI" case, which indicates that it could perform better with the experiment paradigm change. We also applied the paired t-test to classification accuracy results to validate the effectiveness of the proposed method. As discussed in [121], some subjects have difficulties in performing BCI, which is termed "BCI illiteracy". It is very difficult to capture the modulation of SMRs for these subjects during motor imagery, and training a classifier with an acceptable accuracy will not be possible. Therefore, the reliability in classifying motor imagery of those BCI illiterate subjects is quite important for BCI. Usually, subjects with error rates higher than 30% (BL) are regarded as BCI illiterate [121, 103]. Therefore, in our experiments, the illiterate subjects with baseline classification accuracies lower than 70% are investigated separately with their t-test results listed in Table 4.3. The results show that the proposed

method is more effective for the subjects with relatively poorer BCI performance, with a p-value of 0.012 at the 5% confidence level. In particular, for the dataset IVa, the proposed method achieves a greater improvement for the subject av. It is observed that the lower the baseline classification accuracy, the higher the incorrectness level e_r, which indicates that the proposed method is more effective for the illiterate subjects. Relatively, fewer improvements are achieved in the dataset of 16 subjects. The two classes in dataset IVa are hand and foot movement motor imageries. It is possibly because that the re-estimation of covariances is more effective with both classes being motor imageries, the data of which are more consistent than that of the idle state. One class in the dataset of 16 subjects is the idle condition which is more noisy and unstable. The re-estimation of the covariance matrices could be less effective, as the covariance matrix discrepancies are still relatively large within the selected subset $\mathcal{Q}_b^c$ in (4.1). Nevertheless, there are still improvements in average classification accuracies for this dataset.

Figure 4.3 shows the comparison of the results summarized in Table 4.1 and 4.2, where results from different datasets are plotted in different shapes. As the x-axis represents the accuracies under CSP and the y-axis represents the accuracies under the proposed method, the more dots above the line $y = x$, the greater the improvements that are achieved with the proposed method. Generally, there are more dots above the line $y = x$ indicating that the proposed method could improve the performance of the classifier. Moreover, on the left side of Figure 4.3, there are more subjects with improvements by using the proposed method, and these are the subjects with BCI illiteracy.

An example of features corresponding to the first and the last spatial filters is shown in Figure 4.4 to illustrate class-wise feature distribution. As

Table 4.1: Competition III Dataset IVa test results (140-140) (%)

Subject	aa	al	av	aw	ay	**mean**	**std**
e_r	22.86	0.00	38.57	16.43	16.43	16.71	15.13
W	75.00	95.00	61.43	87.68	95.71	83.00	14.64
W_e	71.43	95.00	68.57	90.00	97.43	84.29	13.29

Table 4.2: Test results (16-subject dataset) (%)

	MI-MI			PM-MI		
Subject	e_r	W	W_e	e_r	W	W_e
1	21.25	70.00	72.50	22.50	61.88	62.50
2	36.67	61.67	68.33	30.00	59.17	63.33
3	41.67	73.33	65.00	48.33	60.00	60.00
4	26.67	85.00	83.33	9.17	68.33	68.33
5	10.00	71.67	71.67	18.33	62.50	64.17
6	11.67	96.67	96.67	15.83	82.50	85.00
7	26.67	76.67	80.00	18.33	70.83	68.33
8	6.67	93.33	93.33	5.00	98.33	98.33
9	15.00	88.75	88.75	15.00	86.88	84.38
10	28.33	46.67	45.00	30.00	55.00	61.67
11	25.00	46.67	55.00	45.00	49.17	50.00
12	3.33	78.33	78.33	7.50	83.33	84.17
13	33.75	52.50	52.50	38.75	59.38	56.25
14	10.00	93.75	93.75	3.75	81.88	82.50
15	43.75	68.75	73.75	43.75	52.50	61.88
16	15.00	80.00	78.75	16.25	81.88	70.76
mean	22.21	73.98	74.79	22.97	69.60	70.76
std	12.60	15.99	15.13	14.42	13.35	14.68

Table 4.3: T-test results for different groups of subjects.

	all subjects	illiterate subjects
***p*-value** (W vs. W_e)	0.050	**0.012**

The results show that the proposed method is more effective for the subjects with relatively poorer BCI performance, with a p-value of 0.012 at the 5% confidence level.

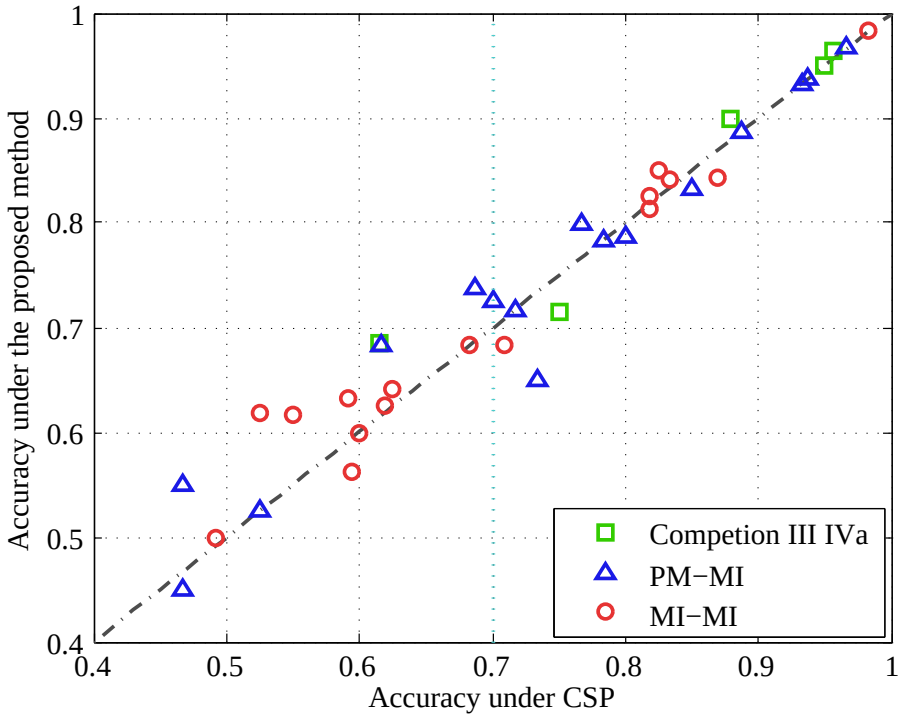


Figure 4.3: Test classification accuracy comparison. The x-axis represents the accuracies under CSP, and the y-axis represents the accuracies under the proposed method. Generally, there are more dots above the line $y = x$. Moreover, on the left side of the figure there are more subjects with improvements by using the proposed method, who are the subjects with BCI illiteracy.

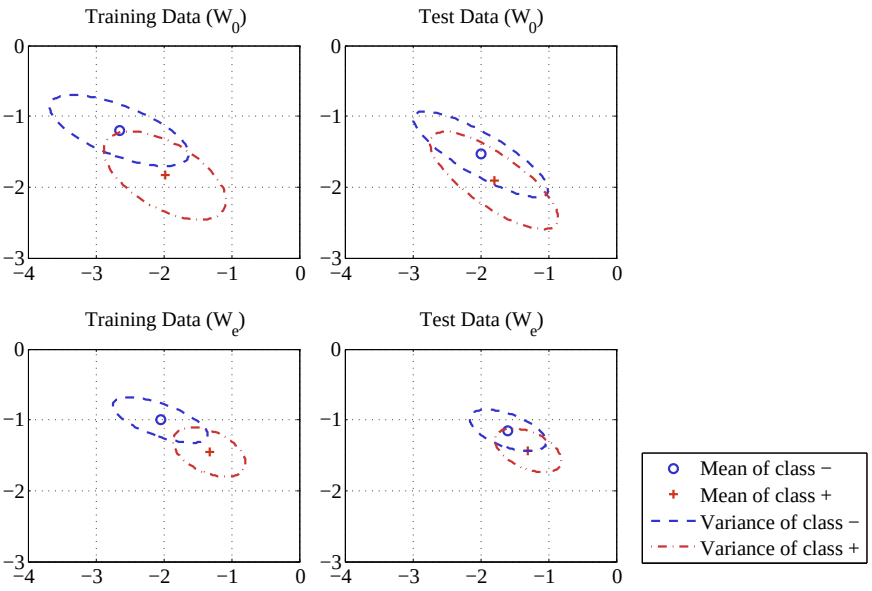
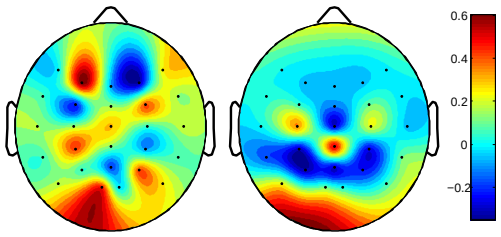


Figure 4.4: An example of feature distribution comparison (subject av). The 2D features correspond to the first and the last spatial fitters in W or W_e. The overlap of features from the two classes is reduced by using the proposed method for both training set and test set.

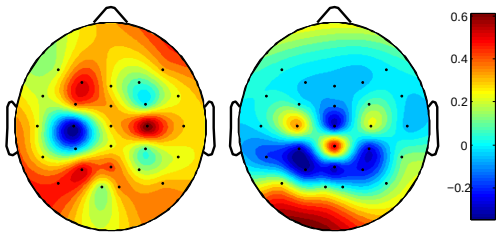
illustrated by Figure 4.4, the overlap of features from the two classes is reduced by using the proposed method for both training and test sets. Moreover, the within-class feature dissimilarities are also reduced as the variances of the features are smaller by using the proposed method, which indicates that the ensemble learning of spatial filter could alleviate the sample discrepancy problem to a certain extend.

4.2.4 Spatial Filter Comparison

Figure 4.5 shows an example of spatial filter weights in projection matrices W and W_e. The spatial filters from W are shown in subfigure (a) and that from W_e in subfigure (b). The left plots in subfigures (a) and (b) correspond



(a) W



(b) W_e

Figure 4.5: An example of spatial filter weights in projection matrices W and W_e (subject 2). In W_e, the weights of the spatial filter maximizing the right hand motor imagery are more concentrated on the left hemisphere compared with that in W.

to the spatial filters maximizing the power of EEG signals with motor imagery of right-hand movements, while the right plots correspond to that of the idle states. It is seen in these figures that in W_e the weights are more concentrated on the left hemisphere compared with that in W, which implies an improvement as the right hand motor imagery relates to the activities in the left hemisphere. There is not much difference between W and W_e for the spatial filters maximizing the power of the idle state.

If W/W_e is interpreted from the perspective of channel selection, those channels with larger weights should contain more discriminative powers. It is possible that some channels are given relatively higher weights based on the average covariance matrices, while for certain trials these channels are not discriminative or very noisy. Therefore, the resulting features of these

trials may contain wrong information. With the proposed ensemble learning method, weights of channels are combined based on different groups of trials so that certain exceptional patterns can be taken into consideration.

4.2.5 Discussion

The proposed method aims at reducing the error caused by covariance matrix discrepancies. The motivation is straightforward: if the projection matrix W is ineffective in generating a proper feature for a certain trial, there is a mismatch between the model and the data. This would be a rare case if samples from the same class are fairly consistent. However, as shown in the feature distribution in Figure 4.1, discrepancies exist within the same class, since the EEG data are quite nonstationary. However, how to discriminate two classes while considering within-class discrepancies is difficult for the spatial filter design. In the EEG data space or the covariance matrix space, it is difficult to evaluate the distribution of trials. To this end, we utilize features to evaluate the model, and select the exceptional trials with the mismatch problem.

W functions by maximizing the differences of the variances between two classes after the projection. Given certain trials selected as exceptional samples from one class, which trials from the other class should be used for covariance matrix estimation cannot be determined, as discrepancies also exist in the trials from the other class. In other words, there exist several combinations of the covariance estimates to calculate W considering different patterns in different trials. To overcome this problem, a new projection matrix W_b^+/W_b^- for the exceptional samples is computed based on the mean covari-

ance matrix of these samples and the mean covariance matrix of all samples from the other class. In this way, the new projection matrix W_b^+/W_b^- aims at discriminating the exceptional trials from all trials of the other class. The virtue of this strategy lies in avoiding totally different projection matrices for ensemble learning while maintaining one of the covariance matrices.

With ensemble learning, biased estimates could be taken into consideration. For example, in subject av, the reason why there is a significant accuracy improvement could be that the average covariance matrices are biased greatly by noise, and the exceptional samples are actually trials containing the discriminative information. By the re-estimating process, such a bias is reduced. It is also possible that there exist several kinds of patterns within one class, and the patterns that dominate the training set are very different from that dominate the test data. The exceptional samples are more similar to test data. Therefore, by balancing those different patterns with combined projection matrices, the model generalization could be improved.

Regarding the parameter ξ in (4.1), we did not tune it although other values of ξ could result in better results for certain subjects. Moreover, if methods such as cross-validation are used to tune ξ based on training data, it is also possible that the over-fitting problem could emerge.

In this work, despite the study on the "PM-MI" paradigm, the proposed method is evaluated by classifying data recorded on the same day. Preliminary results on classifying test data recorded on different days show limited improvements. As in the proposed method, the mismatch between data and the model is evaluated within training set, it is possible that more significant nonstationarity cannot be fully removed, which is the limitation of this work to be addressed in the future work.

4.3 Conclusion

This work has addressed the problem of mismatch between model and samples brought by within-class covariance matrix discrepancy. To take the discrepancies of covariance matrices and their estimates into consideration, an approach to design spatial filters based on ensemble learning has been studied. In particular, the mismatch between the model and data has been evaluated based on the features, and subsequently, the exceptional trials for which the current projection matrix cannot extract proper features have been selected. By ensembling models recalculated based on different subsets of training data, more patterns within one class can be taken into consideration in a very direct and convenient manner. The experimental results have shown that the proposed spatial filter design method yielded a better classification accuracy. The significance of this improvement has been validated using t-test especially for the BCI illiterate subjects.

The proposed method combines different projection matrices in a weighted way so that the mismatch between data and the model can be alleviated, while the problem of nonstationarity of data recorded on different day cannot be fully solved by the proposed method. Given that the data-model mismatch is critical for the performance of BCI systems, we will focus on the development of adaptive learning schemes considering the mismatch. In particular, we will investigate how to evaluate the mismatch between the training model and test data without test labels and use it to adapt the model. In this regard, instead of improving the model generalization, we can implement adaptation by addressing the mismatch problem of the training model and test data, which will be introduced in the following chapters.

Chapter 5

Model Adaptation Based on Tensor Decomposition

Given the significant data variation between sessions, learning the nonstationarity within the training data is not effective enough. The mismatch between the model obtained from training data and test data is more critical. Thus, it is important to construct a metric that measures this mismatch between test data and the model obtained from training data, and make use of the mismatch metric to guide the adaptation of feature extraction models.

This chapter presents a systematic attempt to quantify the data-model mismatch and use the mismatch metric as a basis for the model adaptation. We apply a tensor model to the covariance matrices of EEG data so that the ERD/ERS effects of multi-trial data as well as the projection matrix can be formulated in a unified model [122]. Interpreted from a regression perspective, the residual part in this tensor model reflects the fitness of the projection matrix in describing the ERD/ERS effects underlying the covariance matrices. Therefore, this residual error can be used to evaluate the feature extraction model.

As it is difficult to achieve the residual error minimization and the discrimination objective simultaneously, we propose a two-step approach where

the residual error is estimated in the first place and then combined with the discriminative objective function in a regularized manner. For model adaptation, the major challenge in the first step lies in learning the mismatch relevant to the discriminative task without the true labels of new sessions. To address this issue, we adopt a semi-supervised learning approach to take the class information into consideration instead of the conventional error minimization used in regression model estimation. In this way, the performance of feature extraction model can be enhanced by the adaptation toward reducing the data-model mismatch.

This chapter is organized as follows. In Section 5.1, spatial pattern analysis with a tensor model is presented, followed by the introduction of the adaptation method based on the quantification of the mismatch between model and data. In Section 5.2, we present the investigation into the correlation between the classification performance and data-model mismatch metric as well as the validation of the proposed method in a two-class motor imagery classification problem. Concluding remarks are given in Section 5.3.

5.1 Spatial Filter Adaptation Based on Tensor Decomposition

5.1.1 Spatial Filtering in Tensor Decomposition Form

For convenience, we will follow the conventional notations and definitions in the area of multi-linear algebra. Thus, in this study, tensors are denoted by calligraphic letters [123]. For the details of the definitions and notations, please refer to Appendix A.3.

Let V be an arbitrary projection matrix that maps EEG data from the scalp space to a surrogate channel space, where the resulting covariance matrix is

$$\Lambda_v^i = V^T R^i V \tag{5.1}$$

The covariance matrix R^i of trial i can be written as

$$R^i = V^{-T} \Lambda_v^i V^{-1} \tag{5.2}$$

where Λ_v^i is usually assumed to be diagonal for ERD/ERS feature extraction [75, 124].

To describe multiple trials in one model, we adopt the tensor model to describe the mapping relationship in (5.2). Let $\mathcal{R}$ be a tensor including the covariance matrices of totally n_i trials as $\mathcal{R} \in \mathbb{R}^{n_c \times n_c \times n_i}$. Then, the i-th frontal slice of $\mathcal{R}$ is the covariance matrix R^i for trial i, and (5.2) for all trials can be formulated as

$$\mathcal{R} = \mathcal{I} \times_1 V \times_2 V \times_3 \Lambda_d + \mathcal{E} \tag{5.3}$$

where $\mathcal{I} \in \mathbb{R}^{n_c \times n_c \times n_c}$ is the cubic tensor with ones along the super diagonal, and $\mathcal{E} \in \mathbb{R}^{n_c \times n_c \times n_i}$ is the tensor of residual error components. Each of the frontal slices of $\mathcal{E}$ is denoted by E^i. $\Lambda_d = [\lambda_d^1, \lambda_d^2, \ldots, \lambda_d^{n_i}]^T \in \mathbb{R}^{n_i \times n_c}$, where $\lambda_d^i,\ i \in 1, ..., n_i$ is the vector containing the diagonal elements of Λ_v^i in (5.2). In addition, Λ_d can be regarded as the matrix containing the variances of the signals of all trials after projection.

The objective of the discriminative spatial pattern learning is to estimate

spatial filter V in (5.3) so that the reconstructed signal can be classified. In CSP, the solution can be obtained as a generalized eigen-decomposition of covariance matrices of two classes (2.15). Define $\bar{\mathcal{R}} \in \mathbb{R}^{n_c \times n_c \times 2}$ as a tensor such that R^+ and R^- are frontal slices [122]. And CSP can be written in a tensor form as

$$\bar{\mathcal{R}} = \mathcal{I} \times_1 V \times_2 V \times_3 \bar{\Lambda}_d \tag{5.4}$$

with the solution $V = W^T$. $\bar{\Lambda}_d = [\lambda^+, \lambda^-]^T \in \mathbb{R}^{2 \times n_c}$, where λ^+ and λ^- are, respectively, vectors consisting of diagonal elements of Λ^+ and Λ^- in (2.11). In other words, they are the eigenvalues of R^+ and R^+ upon the joint diagonalization.

An interesting term in (5.3) but absent in (5.4) is $\mathcal{E}$. It is the residual part of modelling which is not taken into consideration in CSP. It is often neglected in conventional spatial filter design methods, where the multi-way structure of the data is simplified by averaging covariance matrices. In [122], this non-jointly-diagonalized term has been explored and it is assumed to be related to the quality of the EEG trials. Compared with parameters that measure the data variation, the residual part $\mathcal{E}$ provides a natural data-model mismatch metric in a more direct way. In other words, the residual part $\mathcal{E}$ can be used to evaluate the performance of the spatial filter because it reflects how accurate the model is in describing the ERD/ERS process. Based on this motivation, we utilize the tensor model of the covariance matrices for the data-model mismatch metric estimation, which is used to guide the spatial filter adaptation.

5.1.2 Tensor Decomposition Based Adaptation

As the residual part $\mathcal{E}$ can be regarded as a quantification of the mismatch between model and data, the mismatch between the calibration model and test data from different sessions is of particular interest, which is formulated as

$$\mathcal{E}_{te} = \mathcal{R}_{te} - \mathcal{I} \times_1 W_{tr} \times_2 W_{tr} \times_3 \Lambda_{d,te} \tag{5.5}$$

where $\mathcal{R}_{te}$ is the tensor of covariance matrices of all test trials and W_{tr} is the solution of CSP in (5.4) obtained from the calibration session. Then, $\Lambda_{d,te}$ contains the variances of the signals after projection, and $\mathcal{E}_{te}$ is the tensor of residual error components, i.e., the mismatch metric between the calibration model and the test data. The error part of test data, $\mathcal{E}_{te}$, is usually much larger than that of the training data, i.e.,

$$\mathcal{E}_{tr} = \mathcal{R}_{tr} - \mathcal{I} \times_1 W_{tr} \times_2 W_{tr} \times_3 \Lambda_{d,tr} \tag{5.6}$$

Examples will be shown in Section 5.2.

To address the session-to-session transfer problem, W_{tr} should be adapted toward minimizing the residual error with respect to the test data while keeping power differences between classes maximized. However, it is difficult to combine the objective function that minimizes the residual error with the one maximizing the Rayleigh coefficient in CSP, as both W and Λ are dependent on each other. To this end, we propose a two-step approach where the residual error is estimated at the first step and then combined with the objective function of CSP in a regularized manner.

5.1.2.1 Residual Error Estimation

In (5.5), $\Lambda_{d,te}$ corresponds to the variance features used for classification of the test data (details of variance feature extraction can be found in [73]). The estimation of $\mathcal{E}_{te}$ is not useful for the adaptation of the discrimination model, if $\Lambda_{d,te}$ is not separable. To solve this problem, we propose a semi-supervised learning approach to evaluate and adapt the discrimination model as shown in Algorithm 3, instead of using $\mathcal{E}_{te}$ in (5.5) directly. Details of the derivation of the updating equations (5.8) and (5.9) can be found in Appendix A.4 and [122]. Different from the iteration in [122], the class information is taken into consideration in the estimation to obtain the data-model mismatch metric relevant to the discriminative objective.

As shown in (5.4), $\bar{\Lambda}_d$ consists of λ^+ and λ^-, which are the vectors comprising, respectively, the eigenvalues of R^+ and R^- upon joint diagonalization. Generally speaking, λ^+ and λ^- are the centres of distributions of the training features. It is desirable that the test features are close to the corresponding centers in a class-wise way. Therefore, we adopt λ^+ and λ^- as the references of variance features of the two classes by using pseudo labels of the test data, denoted by $\hat{y}$ in (5.7). Upon the class-wise initialization, (5.8) and (5.9) are iterated in a data-driven manner so that this estimation process is not relying on the predicted labels totally. In other words, by combining the semi-supervised initialization and iteration procedure, we can balance the trade-off between the discrimination objective and the risk of semi-supervised learning. This approach also allows that intrinsic variations remain, and only the residual parts that cannot be jointly diagonalized will be penalized.

Algorithm 3 Estimation of residual error

Input: Training data, a batch of test EEG data w/o class label, and maximum number of iteration n_k;

Output: Data-model mismatch metric $\hat{\mathcal{E}}_{te}$.

begin

Train a feature extraction model based on the training data;

Obtain features of both training data and test data;

Train a classifier based on the training features;

Classify the test features to obtain the estimated label $\hat{y}$;

Initiate $\Lambda^0_{d,te}$ as

$$\lambda^{i,0}_{d,te} = \begin{cases} \lambda^+, & \hat{y}^i = +; \\ \lambda^-, & \hat{y}^i = -. \end{cases} \tag{5.7}$$

where $\lambda^{i,0}_{d,te}$ is the i-th column of $\Lambda^0_{d,te}$.

Initiate $V^0 = W^T_{tr}$;

while $k < n_k$ **do**

Update V^k as

$$V^k = R_{te,(2)}\{(\Lambda^{k-1}_{d,te} \odot V^{k-1})^T\}^\dagger \tag{5.8}$$

where † denotes the pseudo-inverse of a matrix.

Update $\Lambda^k_{d,te}$ as

$$\Lambda^k_{d,te} = R_{te,(3)}\{(V^k \odot V^k)^T\}^\dagger \tag{5.9}$$

$k = k + 1$;

Compute

$$\hat{\mathcal{E}}_{te} = \mathcal{R}_{te} - \mathcal{I} \times_1 V^k \times_2 V^k \times_3 \Lambda^k_{d,te} \tag{5.10}$$

5.1.2.2 Regularization of the Error Term

The residual error term $\hat{\mathcal{E}}_{te}$ estimated in Algorithm 3 cannot be regularized directly because it may not be positive-definite, and in this case the regularization actually increases the mismatch as discussed in [92]. In this section, we introduce two methods to guarantee that the penalty term to be positive, and the results comparison and discussion will be given in the next section.

Let $\hat{E}_{te}^{i}$ be the i-th frontal slice of $\hat{\mathcal{E}}_{te}$. To guarantee that the penalty term is positive, we consider the penalty term in the form

$$P_s(\mathbf{w}) = \mathbf{w}(\sum_{i=1}^{n_{te}}(\hat{E}_{te}^{i}\hat{E}_{te}^{i^T}))\mathbf{w}^T \tag{5.11}$$

where n_{te} is number of test trials available for adaptation. The penalty term in (5.11) may fail to penalize appropriate elements of W in certain cases, as pointed out in [125, 77]. To solve this problem, we propose a novel operator $\mathcal{F}^*$. Let $E \in \mathbb{R}^{n_c \times n_c}$ be an arbitrary error term with eigen-composition

$$E = U_e D_e U_e^T, \tag{5.12}$$

Then, we have

$$\mathcal{F}^*(E) = \sum_{m=1}^{n_c} |d_{e,m}| \begin{bmatrix} u_{e,1m}^2 & \dots & |u_{e,1m}u_{e,n_cm}| \\ \vdots & \ddots & \vdots \\ |u_{e,1m}u_{e,n_cm}| & \dots & u_{e,n_cm}^2 \end{bmatrix} \tag{5.13}$$

where $u_{e,nm},\ m,n = 1, ..., n_c$ is the element of the n-th row and m-th column of U_e. Detailed discussion of operation $\mathcal{F}^*$ and its relationship with the "flipping" method in [92] can be found in Appendix A.5. The penalty term

based on (5.13) is

$$P_f(\mathbf{w}) = \mathbf{w}(\sum_{i=1}^{n_{te}} \mathcal{F}^*(\hat{E}_{te}^i)^T)\mathbf{w}^T \tag{5.14}$$

With the regularization terms, the regularized objective functions based on CSP become

$$J^+(\mathbf{w}) = \frac{\mathbf{w}R^+\mathbf{w}^T}{\mathbf{w}(R^+ + R^-)\mathbf{w}^T + \mu P(\mathbf{w})} \tag{5.15}$$

$$J^-(\mathbf{w}) = \frac{\mathbf{w}R^-\mathbf{w}^T}{\mathbf{w}(R^+ + R^-)\mathbf{w}^T + \mu P(\mathbf{w})} \tag{5.16}$$

where $\mu \in [0, 1]$ is the tuning parameter. By maximizing (5.15) and (5.16), spatial filters that respectively maximize the power of class + and − can be obtained [91]. $P(\mathbf{w})$ in (5.15) and (5.16) represents a penalty term. By replacing $P(\mathbf{w})$ with $P_s(\mathbf{w})$ in (5.11) or $P_f(\mathbf{w})$ in (5.13), we can obtain objective functions based on different transformation methods. More forms of $P(\mathbf{w})$ will be introduced in Section (5.2.2) for comparing the proposed method with other regularization-based spatial filtering methods.

Note that while R^+ and R^- are computed using training data only, $P(\mathbf{w})$ is calculated based on a batch of unlabelled test data as presented in Algorithm 3, and (5.11)-(5.13). Therefore, (5.15) and (5.16) are applied to update the spatial filters and it can be considered as adaptation. By penalizing $P(\mathbf{w})$ in the objective function, the residual part $\mathcal{E}$ can be minimized in the updated CSP space. Subsequently, the updated model fits the new data better, and the performance of feature extraction can be improved.

5.2 Experimental Study

5.2.1 Experiment Set-Up and Data Description

Please refer to Section 3.4.1.

5.2.2 Data Processing and Feature Extraction

Since FBCSP is one of the most successful feature extraction methods for motor imagery EEG classification, we implement the proposed adaptation method based on FBCSP. First, we train FBCSP and the Naive Bayesian Parzen Window (NBPW) classifier with the training data as in [80, 68]. Then, data from the test session is divided equally into two batches, and as described in Section 5.1.2, $\hat{\mathcal{E}}_{te}$ is estimated based on the first batch of the test data and the projection matrix W_a is obtained using different penalization terms as in Section 5.1.2.2. Note that during the adaptation procedure the true labels of the test data are not available. This adaptation procedure is only applied to the bands selected in FBCSP for the sake of efficiency. Finally, the updated projection matrices were applied to the training data and the classifier was re-trained by the updated training features. Test data from the second batch is classified by the updated model. For the convenience of presentation, we refer to the batch of test data used to estimate the error term as the adaptation batch, and the rest of test data as the evaluation batch.

To compare the proposed method with other regularization based methods and adaptation methods, we implement Tikhonov (Tik) regularized CSP, spatially regularized (SP) CSP [91], unsupervised data space adaptation

(DSA) [103], naive regularization using average covariance of the test set (nvCSP), and stationary CSP (sCSP) [92]. For Tik and SP, we use cross-validation results of the training set to select the best regularization term, as in [91]. In DSA [103], the space adaptation matrix is calculated using the test data from the adaptation batch

$$W_{DSA} = W_{tr}\bar{R}_{tr}^{\frac{1}{2}}\bar{R}_{te}^{-\frac{1}{2}} \tag{5.17}$$

where $\bar{R}_{tr}$ and $\bar{R}_{te}$ are average covariance matrices of training set and adaptation batch, respectively. In nvCSP, $\bar{R}_{te}$ is used as the regularization term, as below

$$P_{nv}(\mathbf{w}) = \mathbf{w}\bar{R}_{te}\mathbf{w}^T \tag{5.18}$$

Note that for nvCSP we use the ratio between the number of the training trials and test trials to determine the regularization coefficient, i.e., $\mu = \frac{n_{te}}{n_{tr}}$, where n_{tr} denotes the number of training trials. For a better comparison, sCSP is implemented in an adaptive manner using data from the adaptation batch

$$P_{st}(\mathbf{w}) = \mathbf{w}(\sum_{i}^{n_{te}} \mathcal{F}(R_{te}^{i} - \bar{R}_{tr}))\mathbf{w}^T \tag{5.19}$$

where $\mathcal{F}$ denotes the "flipping" operator introduced in [92]. Moreover, to validate the necessity of Algorithm 3, we use $\mathcal{E}_{te}$ in (5.5) as the regularization term, by substituting E_{te}^{i} into (5.11) and (5.14) for $\hat{E}_{te}^{i}$.

Since sCSP and the proposed method are used for adaptation, the cross-validation based training set cannot be used to select μ. Thus, we choose to

cross-validate the classification performance in a leave-one-subject-out manner. In particular, μ is pre-set as $\mu \in \{0.1, 0.2, ..., 1\}$, and for a current subject the value of μ is chosen as the one with the best average performance for the rest of the subjects. All methods are implemented with FBCSP in the same way, i.e., they are all applied to the bands selected by FBCSP.

5.2.3 Analysis of Residual Error

In this section, we investigate the residual error $\mathcal{E}$ to validate the proposed method in measuring the mismatch between the feature extraction model and data. In particular, we perform the correlation test between $||\mathcal{E}_{tr/te}||$ and the classification accuracy. 5-by-5 cross-validation accuracies are used for training data and session-to-session transfer classification accuracies are used for test data. Figure 5.1 illustrates the correlation between the classification accuracy based on FBCSP and average $||\mathcal{E}_{tr/te}||$ of trials from the training/test set. Pearson's correlation coefficient r_c equals -0.60 for the training data with p-value 0.01. Therefore, we can see that the accuracy for the training data significantly correlates to $||\mathcal{E}||$ in a negative way. The p-value for the test data is not significant (0.19) but the correlation is also negative (-0.34). The correlation for the test data is not significant, which is possibly because the session-to-session transfer classification accuracy is subject to more complicated factors. We see that the regression lines for the test data and training data in Figure 5.1 are almost parallel. Generally, there is a trend that a higher $||\mathcal{E}_{tr/te}||$ may correspond to a lower classification accuracy and vice versa.

In addition, the change of $\hat{\mathcal{E}}_{te}$ with respect to the iteration number is also

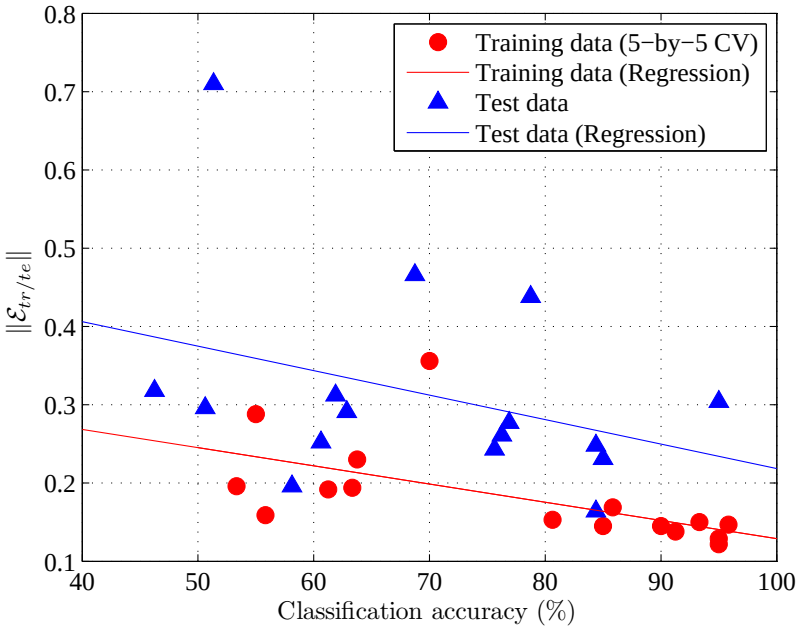


Figure 5.1: Relation between the residual error and classification accuracy. Each circle or triangle marks one subject. The x-axis represents the classification accuracy and the y-axis represents $||\mathcal{E}_{tr}||$ or $||\mathcal{E}_{te}||$. For both training data and test data, there is a trend that a larger $||\mathcal{E}_{tr}||$ or $||\mathcal{E}_{te}||$ may correspond to a lower classification accuracy. Pearson's correlation test shows a significant correlation for training data with coefficient r_c equal to -0.60 and p-value equal to 0.01.

investigated because there is an iteration procedure in Algorithm 3. Figure

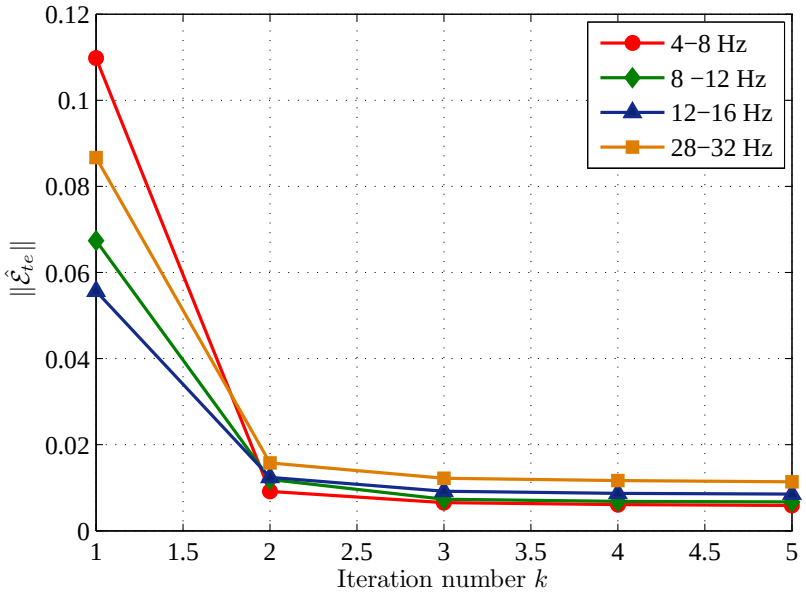


Figure 5.2: The change in $||\hat{\mathcal{E}}_{te}||$ with respect to the iteration number k. As shown in this figure, the change in $||\hat{\mathcal{E}}_{te}||$ becomes very small after 2 iterations. Thus, for the efficiency of computation, it is reasonable to run the iterations twice.

5.2 shows an example of the change in $||\hat{\mathcal{E}}_{te}||$ during the process of iteration, where the 4 frequency bands are selected by mutual information for this subject. As shown in Figure 5.2, the change in $||\hat{\mathcal{E}}_{te}||$ is very small after 2 iterations, and this trend exists for every subject. Thus, it is reasonable to run the iterations twice, and this setting is applied to all subjects to obtain the classification results in the following section.

5.2.4 Classification Results

In this section, we present the classification results using the proposed tensor decomposition adaptation (TDA) method. Table 5.1 summarizes the per-

formance of the methods mentioned in Section 5.2.2 compared with FBCSP without any adaptation or regularization as the baseline. Note that all classification accuracies are based on the evaluation batch. We use TDA_s or TDA_f to indicate that (5.11) or (5.13) is used in the proposed method to transform E into a positive definite matrix. And for simplicity, $P_s(E_{te})$ or $P_f(E_{te})$ indicates that the direct differences between test data and model E_{te} in (5.5) are used with different transforming methods. Generally, all adaptation methods improve the performance of FBCSP while spatial-smoothing methods ("Tikhonov" and "SP") fail to do so. Paired t-test results show that only TDA_s and TDA_f outperform the baseline in a significant way, and TDA_s achieves the highest accuracy of 74.41% which indicates the effectiveness of the proposed methods. One reason for the better results of TDA_s could be that $\hat{E}_{te}^{j}\hat{E}_{te}^{j^T}$ is simpler so it is closer to the original error while operation $\mathcal{F}^*(E)$ modifies the term substantially and becomes less accurate. Moreover, the iteration in Algorithm 3 actually decreases $||\hat{\mathcal{E}}_{te}\hat{\mathcal{E}}_{te}||_F^2$, equalling $\sum_j^{n_{te}} \text{tr}(\hat{E}_{te}^{j}\hat{E}_{te}^{j^T})$ (Appendix A.4), which could also be a reason that $\hat{E}_{te}^{j}\hat{E}_{te}^{j^T}$ in (5.11) matches TDA better.

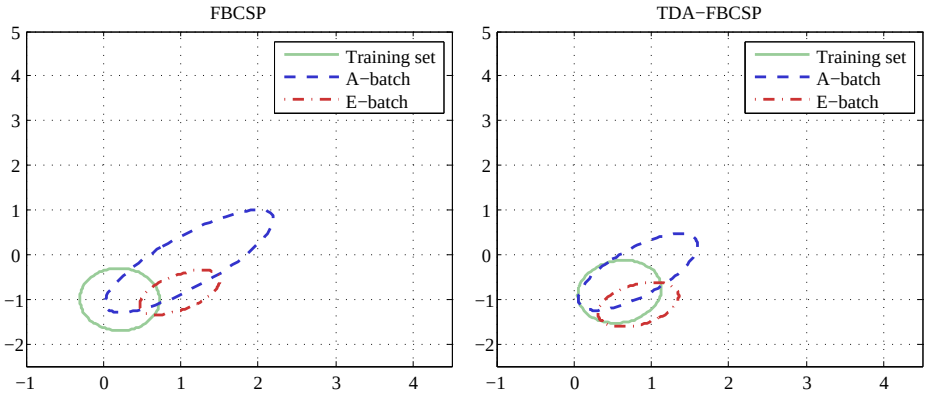
Table 5.1: Session-to-session transfer classification results on the evaluation batch (%).

Subject	BL	SP	Tik	nvCSP	DSA	sCSP	$P_s(E_{te})$	$P_f(E_{te})$	TDA_s	TDA_f
1	67.50	68.75	68.75	61.25	73.50	67.50	66.25	75.00	71.25	76.25
2	58.75	55.00	47.50	68.75	60.00	53.75	56.25	56.25	56.25	56.25
3	50.63	50.63	59.49	70.89	67.09	60.76	63.29	60.76	70.89	70.89
4	71.25	71.25	71.25	61.25	83.75	86.25	78.75	77.50	80.00	87.50
5	75.00	77.50	80.00	60.00	72.50	77.50	82.50	78.75	82.50	78.75
6	82.50	82.50	82.50	77.50	81.25	82.50	81.25	81.25	81.25	82.50
7	80.00	80.00	73.75	51.25	56.25	68.75	73.75	76.25	82.50	75.00
8	93.33	93.33	93.33	95.00	93.33	95.00	96.67	95.00	96.67	95.00
9	78.75	78.75	83.75	72.50	83.75	85.00	78.75	78.75	81.25	82.50
10	65.00	63.29	65.00	51.25	62.03	58.23	63.29	61.25	73.75	63.75
11	50.00	51.25	52.50	51.25	53.75	50.00	50.00	45.00	50.00	51.25
12	78.75	77.50	78.75	77.50	77.50	76.25	80.00	81.25	85.00	80.00
13	53.95	51.25	51.25	51.25	71.25	70.00	68.75	63.75	62.50	65.00
14	71.25	71.25	71.25	80.00	80.00	76.25	73.75	76.25	75.00	72.50
15	57.50	60.00	66.25	52.50	58.75	63.75	60.00	61.25	60.00	60.00
16	73.75	75.00	75.00	76.25	81.25	77.50	77.50	80.00	77.50	76.25
mean	69.24	69.20	69.63	66.15	72.73	71.81	71.92	71.77	74.14	73.34
***p*-value**	-	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05	**0.0023**	**0.029**

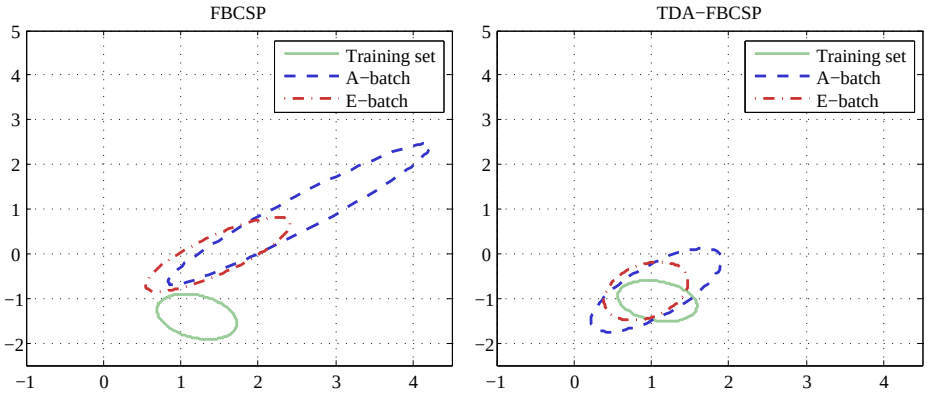
All classification accuracies are based on the evaluation batch. FBCSP without any adaptation or regularization is used as the baseline (BL). Tikhonov (Tik) regularized CSP, spatially (SP) regularized CSP, data space adaptation (DSA), naive regularization using average covariance of the test set (nvCSP), and stationary CSP (sCSP) are introduced in Section 5.2.2. $P_{s/f}(E_{te})$ indicates that E_{te} in (5.5) is used as the penalty term. Subscript s or f indicates that (5.11) or (5.13) is used to transform penalty term into a positive definite matrix in the proposed methods.

The changes in the feature distribution between sessions are shown in Figure 5.3. In particular, in each subfigure of Figure 5.3, the distributions of the 2D features in different sessions are plotted and the corresponding subjects are listed. Those features correspond to the most discriminative spatial filters selected using mutual information in the FBCSP procedure. The terms "a-batch" and "e-batch" are used to represent the adaption batch and the evaluation batch. We can see that without adaptation, the feature distributions shift greatly. It is clearly shown that such a shift has been reduced significantly by TDA, and subsequently, the feature distributions become more consistent across sessions. More importantly, we find that the variances of the features are also reduced by TDA, which means that the proposed method can also reduce the within-session nonstationarity.

Visualization of class-wise feature distribution is shown in Figure 5.4 to compare the separations of features from different classes with and without adaptation. The non-linear classification boundary in NBPW classifier is presented by the contrast between colors blue (class $-$) and red (class $+$). Comparing the left and right columns in Figure 5.4, it can be seen that by employing the proposed method, more features lie on the correct side of the classifier. In particular, for subject 1, it is observed that the reduced shifts between training and test features contribute to the improvements. For subject 4, separability of the test features is improved more significantly. Therefore, besides the shift of the average distance, the proposed method is able to capture the cross-session nonstationarity that makes the feature extraction model fail to extract discriminative features for certain subjects. This improvement is a more meaningful adaptive behaviour of the feature extraction model.



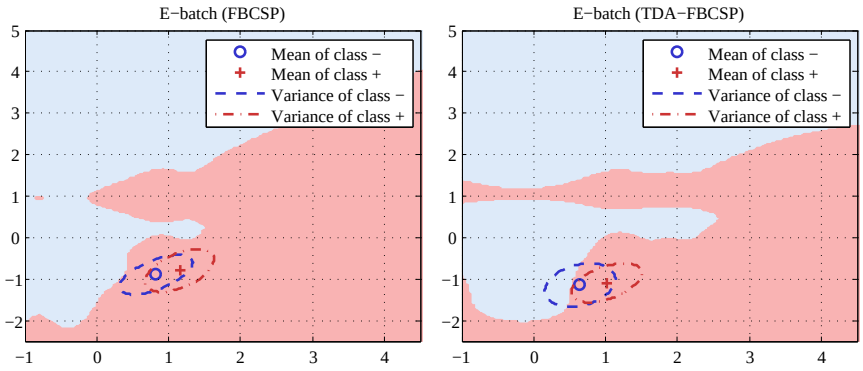
(a) Subject 1 (BL) (b) Subject 1 (TDA$_f$)



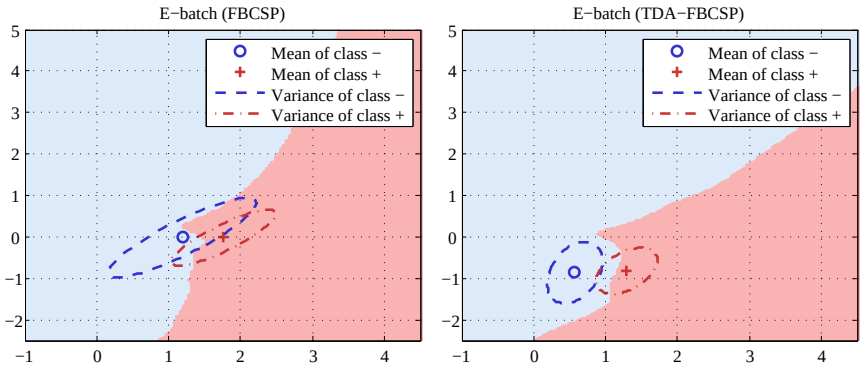
(c) Subject 4 (BL) (d) Subject 4 (TDA$_s$)

Figure 5.3: Tracking the nonstationary feature space across sessions. Comparing the feature distributions extracted from the training session and two test batches, we observe that the feature distributions become more consistent across sessions by employing TDA, with the distances between training features and test features significantly reduced.

Figures 5.5 (a) and (b) show the change of $||\mathcal{E}_{tr/te}||$ with different values of tuning parameter μ. The x-axis represents the value of μ and the y-axis $||\mathcal{E}_{tr/te}||$. Note that in this analysis $||\mathcal{E}_{tr/te}||$ is calculated by substituting W_a using different μ into (5.6) or (5.5). Therefore, when $\mu = 0$, $||\mathcal{E}_{tr/te}||$ equals to that in (5.6) or (5.5), respectively. The baseline values are given by dotted/dashed lines. For the two sets of test data, $||\mathcal{E}_{te}||$ decreases first and

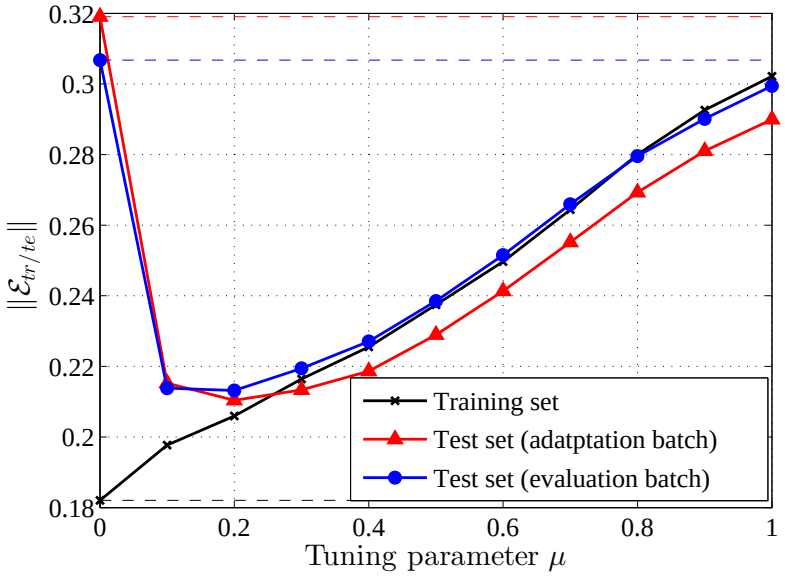


(a) Subject 1 (BL) (b) Subject 1 (TDA$_f$)

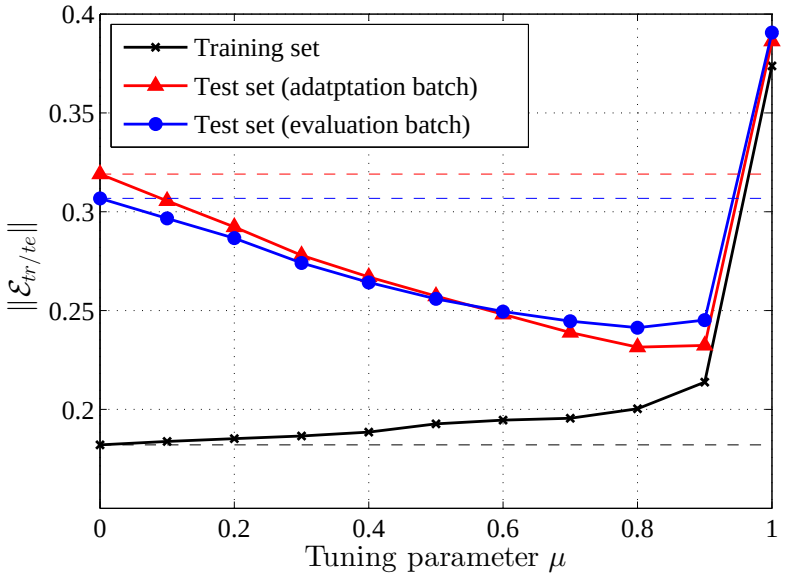


(c) Subject 4 (BL) (d) Subject 4 (TDA$_s$)

Figure 5.4: Visualization of class-wise feature distributions. The non-linear classification boundary in NBPW classifier is presented by the contrast of different color patterns. By employing TDA, more features fall in the corresponding side of the boundary.

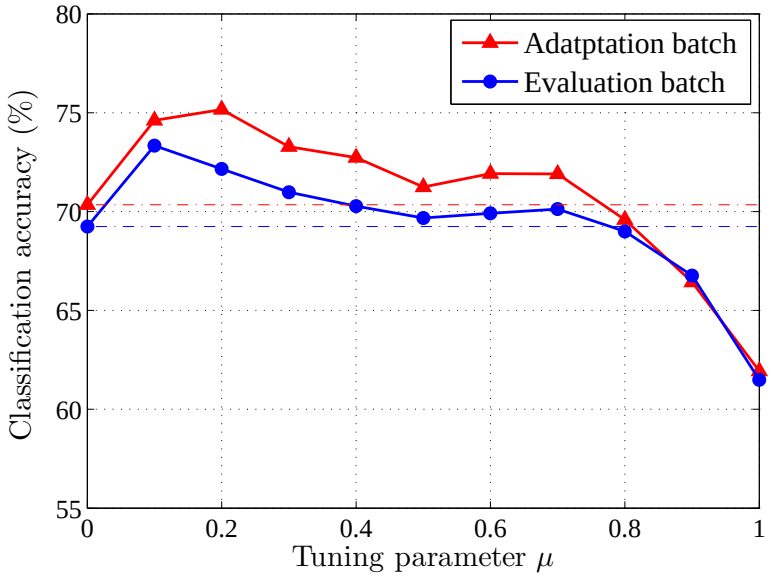


(a) TDA_f

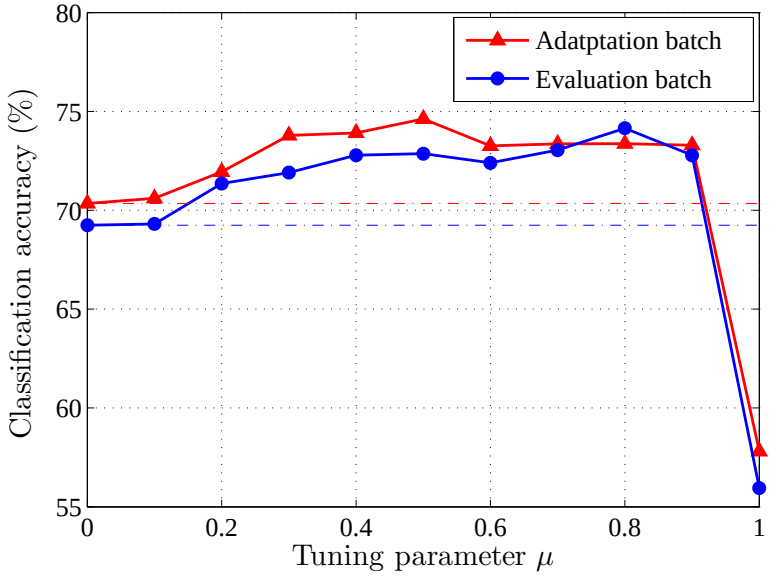


(b) TDA_s

Figure 5.5: Change of $||\mathcal{E}||$ with respect to μ. The x-axis represents the value of μ, and the y-axis represents $||\mathcal{E}_{tr/te}||$ averaged across subjects. $||\mathcal{E}_{tr/te}||$ based on FBCSP without any adaptation are denoted with dotted-dashed lines.

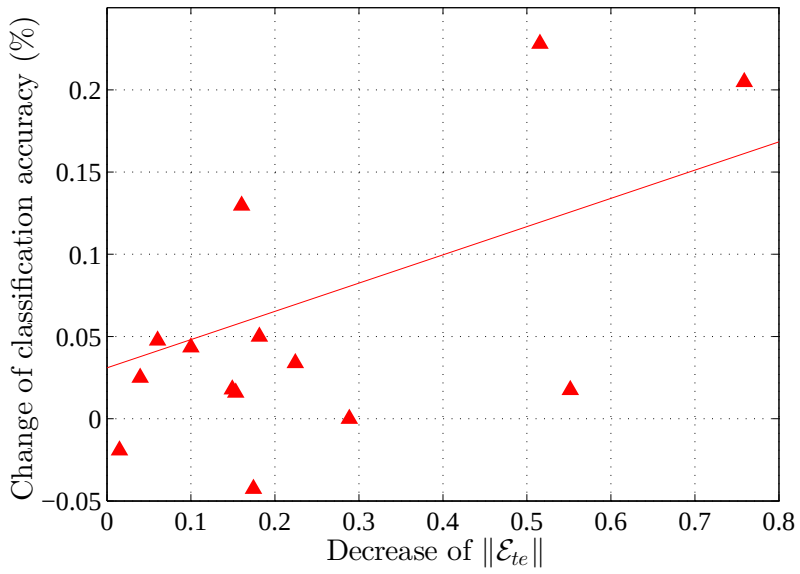


(a) TDA_f

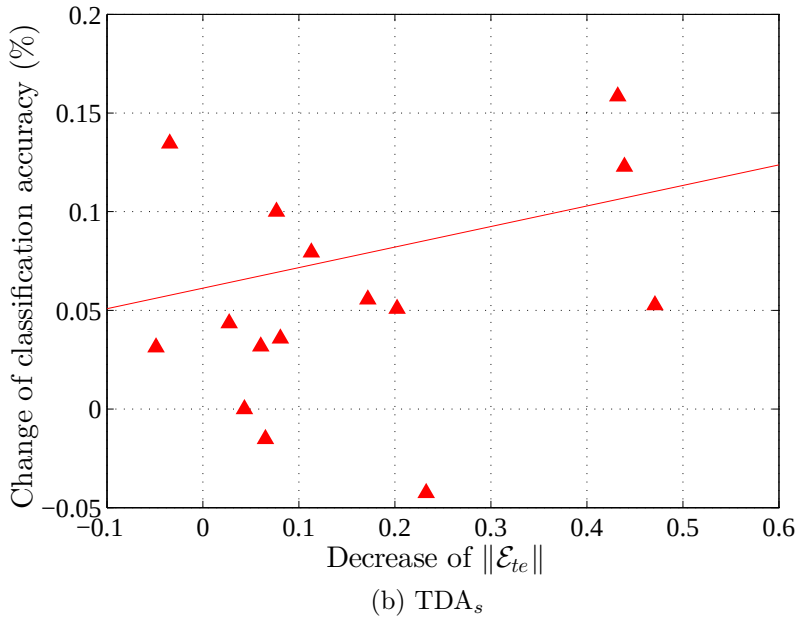


(b) TDA_s

Figure 5.6: Change of accuracy with respect to μ. The x-axis represents the value of μ, and the y-axis represents accuracy averaged across subjects.



(a) TDA$_f$



(b) TDA$_s$

Figure 5.7: Change of accuracy with respect to change of $||\mathcal{E}||$. The x-axis represents the decrease of $||\mathcal{E}||$, and the y-axis represents change of accuracy. Each triangle marks one subject.

then increases. The trends for TDA_s and TDA_f are different, the reason for which could be that after squaring the scale of the elements in the penalty terms changes greatly. Figures 5.6 (a) and (b) show the change of accuracy with respect to μ. Comparing Figures 5.5 and 5.6, we see that in general the lower the value of $||\mathcal{E}_{te}||$, the higher the accuracy. Since $\mathcal{E}$ reflects the mismatch between model and data, when a high weight is given to the penalty term, we sacrifice the fitness of that model for training data. The value of μ actually controls the balance between test data and training data. As shown in Figures 5.5 (a) and (b), $\mu = 0.1$ for TDA_f and $\mu = 0.8$ for TDA_s can be deemed as "equilibrium" points, where the decrease of $\mathcal{E}_{te}$ is significant while $\mathcal{E}_{tr}$ is not increased greatly. This is the reason why these two parameters yield the best accuracy improvements in Figures 5.6 (a) and (b). Figures 5.7 (a) and (b) show classification improvements with decrease in $||\mathcal{E}_{te}||$. In both cases, we find that improvements increase with decrease in $||\mathcal{E}_{te}||$, which is not significant in the Pearson's correlation test. As we have discussed earlier, since the improvements are subject to both $||\mathcal{E}_{tr}||$ and $||\mathcal{E}_{te}||$, it is reasonable that such unilateral correlations are not significant.

5.2.5 Discussion

As described in Section 5.1, the role of the regularization term of TDA can be viewed as minimizing the regression error of the model. A natural idea is to use the residual parts of the training data to regularize the model to improve model generalization. However, from the experimental study, it is found that the classification performance of such an implementation is not significantly higher than that of FBCSP without any regularization. The reason is that,

since the average covariance matrices are obtained from training data, the residual parts are trivial, as shown in Figure 5.1. Therefore, it is more effective to utilize the residual error from the test data to adapt the model. By improving the model from the perspective of fitness, the classification performance can be enhanced simultaneously.

Regarding the choice of parameters, for most of the regularization based methods, the most time-consuming part is related to optimizing parameters using cross-validation. However, since the proposed method is designed for adaptation, such cross-validation based on the training set is not appropriate. Therefore, we adopt leave-one-out to choose different regularization terms μ. Moreover, our analysis on the relationship between $||\mathcal{E}_{tr/te}||$ and accuracy improvements in Figures 5.6 and 5.5 also provide insights into the selection of μ by balancing $||\mathcal{E}_{te}||$ and $||\mathcal{E}_{tr}||$. For the number of iterations in estimating $\mathcal{E}_{te}$ in Algorithm 3, we show that only after 2 iterations, the change of $||\mathcal{E}_{te}||$ becomes quite small. In addition, more iterative steps could be redundant, because we wish to maintain the discriminative property of Λ_d. Therefore, we choose the number of iterations as 2, which satisfies the requirements and also reduces the computational burden. Based on the above discussion, for these two parameters there exist feasible values based on which general improvements can be achieved. It is not necessary to tune the parameters for every subject individually, although there may exist better classification results for certain subjects by setting them differently. Regarding the necessity of the tensor formulation and the iteration, we have performed the adaptation using $\mathcal{E}_{te}$ in (5.5) and there is no significant improvement, which validates our consideration that penalizing $\mathcal{E}_{te}$ could be ineffective since $\Lambda_{te,d}$ in (5.5) may not be discriminative. Moreover, we would like to address the effectiveness

of the proposed method as it can be combined with FBCSP easily with low computational complexity and achieves performance improvements. In particular, 4 frequency bands are typically selected for a subject in FBCSP. For example, it takes 0.0574s for a 4.00 GB CPU to process one trial to obtain the mismatch estimates for 4 bands in MATLAB. The rest time between trials is around 5s and usually much longer between runs. Thus, such a computational time is acceptable for the proposed method to be implemented online.

As described earlier, there exist other works addressing the nonstationarity problem by utilizing data from other subjects [95, 94]. However, based on FBCSP, usually different frequency bands are selected for different subjects, which makes such multi-subject strategies difficult to implement. Recently, a generic framework is proposed in [77], in which CSP and its regularization methods are unified based on divergence. The divergence-based regularization objective function needs to be solved by a geodesic searching approach or a deflation method. In addition, FBCSP addresses the stationarity problem by selecting bands using mutual information, and subsequently, improvements gained by regularization based on training data could be limited. Therefore, in this work, we focus on the regularization objective function that could be solved by eigen-decomposition in one step for the sake of the computational efficiency. For a similar reason, the signals after projection are assumed to have diagonal covariance matrices in (5.2) as in CSP. Given the neuroscience findings about source connectivities, a possible extension of the proposed method could be measuring the data-model mismatch for the computational model based on convolutive sources model in Chapter 3.

5.3 Conclusion

For practical BCI systems, a computational model obtained from the training/calibration session is required to be applied to test sessions conducted on different days, while data variation between sessions often leads to the inaccuracy of the computational model. Despite the effort made on adaptive BCI, the quantification of mismatch between test data and training model needs to be investigated. In this work, we present a systematic attempt to quantify the data-model mismatch, and use the mismatch metric to guide the model adaptation.

To capture the multidimensional structure of EEG, we adopt a tensor model to formulate the mapping between the variances of the source signals and covariance matrices of scalp EEG signals. The residual error of this model proves to be an effective quantification of the mismatch between model and data. Different from the conventional regression models, the mismatch metric needs to be relevant to the discrimination function. However, in adaptation, true class labels of test data are not available in this discriminative estimation of the mismatch metric. To solve this problem, the estimation is accomplished by a semi-supervised learning approach. Then, the feature extraction model can be updated accordingly toward reducing the data-model mismatch.

We implement the proposed adaptation method combined with FBCSP, which improves the session-to-session transfer classification accuracy significantly as confirmed by the statistical test. Moreover, our correlation analysis also validates the effectiveness of the proposed metric as a quantification of mismatch between model and data.

Chapter 6

Model Adaptation through Subspace Tracking

As shown in Section 2.1, the projection matrix in CSP consists of a whitening part and an orthogonal part, and there are methods that adapt the projection matrix by re-estimating the whitening part. This whitening approach is equivalent to projecting both training data and test data to an invariant subspace, which is the orthogonal part in the projection matrix [104, 103]. The cross-session invariance of the subspace holds under the assumption that the linear transformation between the two domains is symmetric.

However, due to the significant cross-session data variation, the discriminative subspaces also vary from the training data to the test data. The adaptation issue for a more general case, i.e., the asymmetric transformation case, should be taken into consideration for feature extraction. In fact, it is not feasible to seek an invariant subspace where both training data and test data are discriminative. The major challenge is adapting discriminative subspaces for the test data while keeping the feature spaces consistent from session to session. To solve this problem, in this work, we propose a novel adaptation approach based on the divergence framework [77]. The cross-session change can be taken into consideration by searching the discriminative subspaces

for the test data on a manifold of orthogonal matrices in a semi-supervised manner. The model adaptation is based on the divergence measurement of distributions of the training data and the test data in different subspaces. In particular, the adaptation objective is to maximize inter-class divergence between the test data distribution in the adapted subspaces and the training data distribution in the original subspaces. By adding a regularization term, within-class divergence could also be taken into consideration. In this way, although different projection matrices are applied to training data and test data, the feature space is more consistent and the performance of the classifier can be improved without the adaptation of the classifier.

This chapter is organized as follows. In Section 6.1, the problem of discriminative subspace shift in feature extraction is investigated and discussed. In Section 6.2, the adaptation method based on the divergence framework for the spatial filter design is presented. In Section 6.3, the shift of the discriminative subspace is further investigated by a numerical study, and the validity of the proposed method is verified by experimental studies on a two-class motor imagery classification problem. Concluding remarks are given in Section 6.4.

6.1 Problem Formulation

6.1.1 Spatial Filter Adaptation Based on Normalization

To address the nonstationarity of EEG data from different sessions, we use R_{tr} to denote the average covariance matrix of the training data, and R_{te} to denote the test data as computed in (2.2). Assuming that the prior proba-

bilities of the two classes are equal, $R_{tr/te}$ can be obtained as

$$R_{tr/te} = \frac{1}{|\mathcal{Q}_{tr/te}|} \sum_{i \in \mathcal{Q}_{tr/te}} R^i \tag{6.1}$$

where $\mathcal{Q}_{tr/te}$ denotes the training/test set. Given the composition of W as in (2.10), the projection matrix obtained based on the training set is

$$W_{tr} = U_{tr}^T P_{tr} \tag{6.2}$$

where P_{tr} and U_{tr} are the whitening part and the orthogonal part based on the training set, respectively. In [104], it has been established that the projection matrix can be adapted by replacing the whitening part $P_{tr} = R_{tr}^{-\frac{1}{2}}$ with $P_{te} = R_{te}^{-\frac{1}{2}}$ so that the updated projection matrix becomes

$$\begin{aligned} W_n &= W_{tr} P_{tr}^{-1} P_{te} \\ &= U_{tr}^T P_{te} \end{aligned} \tag{6.3}$$

where W_n denotes the adapted projection matrix based on the method in [104], which is usually referred to as the normalization-based adaptation. As shown in (6.3), by only updating the whitening part, the orthogonal part U_{tr} in W_{tr} is maintained in W_n. It is also pointed out in [104] that the orthogonal part U_{tr} is kept constant across sessions if and only if

$$X_{te} = C R_{tr}^{-\frac{1}{2}} X_{tr} \tag{6.4}$$

where C is an arbitrary symmetric positive definite matrix, and X_{tr} and X_{te} correspond to EEG data from the test session and the training session,

respectively. The virtue of adapting W by normalization lies in the fact that the estimation of R_{te} can be assessed without the labels of the data of the test session as

$$W_n = W_{tr} R_{tr}^{\frac{1}{2}} R_{te}^{-\frac{1}{2}} \tag{6.5}$$

6.1.2 From Discriminative Subspace to Feature Space

Each column u_j of the orthogonal part U in $W = U^T P$ can be regarded as a subspace. The vectors that correspond to the largest and smallest eigenvalues in (2.7) to (2.9) are the most discriminative subspaces. In this section, we investigate the relationship between the discriminative subspace and the features space. In other words, how the change of the discriminative subspaces influence that of the feature space. Given W in (2.10), the covariance matrix after the projection can be rewritten as

$$\begin{aligned} \Lambda^i &= W X^i (X^i)^T W^T \\ &= (P^T U)^T X^i (X^i)^T P^T U \\ &= U^T P X^i (X^i)^T P^T U \end{aligned} \tag{6.6}$$

Σ^i is used to denote the covariance matrix of trial i after whitening as

$$\Sigma^i = P X^i (X^i)^T P^T \tag{6.7}$$

Apply eigenvalue decomposition to Σ^i so that

$$\Sigma^i = U^i V^i U^{i^T} \tag{6.8}$$

where U^i is a matrix containing the eigenvectors of Σ^i as columns, and V^i is a diagonal matrix containing the eigenvalues of Σ^i as diagonal elements. Thus, the covariance matrix after projection can be rewritten as

$$\Lambda^i = U^T U^i V^i U^{iT} U \tag{6.9}$$

Then, the j-th feature of trial i corresponding to the j-th spatial filter in W (2.10) becomes

$$\mathbf{f}_j^i = \sum_{m=1}^{n_c} v_m^i u_j^T u_m^i u_m^{i\ T} u_j \tag{6.10}$$

where u_j is the j-th column of U, u_m^i is the m-th column of U^i, and v_m^i is the m-th diagonal element of V^i. Suppose trial i belongs to class $+$, and let $\bar{\mathbf{f}}_j^+$ be the mean of the j-th feature of class $+$. As shown in (2.2)-(2.10), $\bar{\mathbf{f}}_j^+ = \lambda_j^+$, while it can be also written in the following form

$$\begin{aligned} \bar{\mathbf{f}}_j^+ &= \sum_{m=1}^{n_c} \lambda_j u_j^T u_m^i u_m^{i\ T} u_j \\ &= \lambda_j u_j^T u_j u_j^T u_j \end{aligned} \tag{6.11}$$

The distance between $\mathbf{f}_j^i$ and $\bar{\mathbf{f}}_j^+$ is

$$\mathbf{f}_j^i - \bar{\mathbf{f}}_j^+ = u_j^T (\lambda_j u_j u_j^T - \sum_{m=1}^{n_c} v_m^i u_m^i u_m^{i\ T}) u_j \tag{6.12}$$

After the whitening, the range of eigenvalues λ_j and v_m, should be between 0 and 1, and subsequently the differences between λ_j and v_m^i would be very

small, which means

$$\begin{aligned} \mathbf{f}_j^i - \bar{\mathbf{f}}_j^+ &\approx u_j^T (\sum_{m=1}^{n_c} v_m (u_m^i + u_j)(u_m^i - u_j)^T) u_j \\ &\propto \sum_{m=1}^{n_c} v_m^i < u_j, u_m^i > \end{aligned} \tag{6.13}$$

From (6.13), we can see that the nonstationarity of the features is related to the nonstationarity of U. To be specific, the larger the angle between u_j and u_m^i, $m = 1, ..., n_c$, the larger the distance between the features. By only updating the whitening part, the orthogonal part U_{tr} in W_{tr} is maintained in the normalization approach. For the test data from a different session, the covariances matrices R_{te} could be very different from R^+ and R^- that are estimated using the training data, so that U^i could be very different from U. Thus, large $< u_j, u_m^i >$ would induce inseparable test features, and only adapting the whitening part P_{tr} in W_{tr} could not be effective enough for feature extraction. Moreover, to address the adaptation issue for a more general case, i.e., the asymmetric transformation case, it is necessary to adapt the discriminative subspaces for the test data. Based on this motivation, the objective of this work is to develop the adaption method that updates U in the projection matrix.

6.2 Spatial Filter Adaptation through Subspace Tracking

6.2.1 Preliminary of Divergence-Based CSP

To make this chapter self-contained, the divergence-based CSP is introduced in this section. It is proved in [77] that spatial filters W in CSP project the EEG data into subspaces where the KL-divergence between the data distributions from two classes is maximized. Thus, the objective function of the divergence-based CSP (divCSP) with regularization is in the form of

$$\mathcal{L}_0 = (\mu - 1)\tilde{D}_{kl}(WR^+W^T||WR^-W^T) + \mu\Delta \tag{6.14}$$

where $R^{+/-} \in \mathbb{R}^{n_c \times n_c}$ is the average covariance matrix in (2.1). $\tilde{D}_{kl}$ is the symmetric KL-divergence, and with the KL-divergence defined in (3.25) it is defined as

$$\tilde{D}_{kl}(\mathcal{N}^0||\mathcal{N}^1) = D_{kl}(\mathcal{N}^0||\mathcal{N}^1) + D_{kl}(\mathcal{N}^1||\mathcal{N}^0) \tag{6.15}$$

In (6.14), $\tilde{D}_{kl}(W^TR^+W||W^TR^-W)$ is the objective function of CSP in the form of symmetric KL-divergence, Δ is the regularization term, and μ is the regularization parameter. The solution of minimizing (6.14) with $\mu = 0$ is equivalent to that of (2.10). Δ is also based on the KL-divergence and it is defined according to the type of nonstationarity to be minimized, e.g.,

$$\Delta = \sum_{c=+,-} \frac{1}{|\mathcal{Q}^c|} \sum_{i \in \mathcal{Q}^c} D_{kl}(WR^iW^T||WR^cW^T) \tag{6.16}$$

(6.16) is an example of the regularization term representing within-class nonstationarity, which measures the average divergence between the trials and mean data distribution for each class separately. By replacing the divergence term in (6.16) with different measurements, we can choose to penalize different types of nonstationarity, such as cross-subject nonstationarity or cross-session nonstationarity.

The major challenge is adapting discriminative subspaces for the test data while keeping the feature spaces consistent from session to session. To solve this problem, in this work, we propose a novel adaptation approach based on the divergence framework (6.14), the advantage of which lies in the fact that it can measure the distribution divergence in different subspaces. Therefore, the cross-session change can be taken into consideration by searching a new discriminative subspace for test data, while the subspace for training data remains the same. The formulation of the objective function in the proposed adaptation approach will be introduced in the next section.

6.2.2 Subspace Tracking

To ensure that the test features are in the same space with the classifier, we propose the following objective function for adaptation

$$\mathcal{L} = (\mu - 1)\mathcal{L}_{csp} + \mu\Delta \tag{6.17}$$

where

$$\begin{aligned} \mathcal{L}_{csp} &= \frac{1}{2}\tilde{D}_{kl}(W_{te}R_{te}^{+}W_{te}^{T}||W_{tr}R_{tr}^{-}W_{tr}^{T}) + \\ &\quad \frac{1}{2}\tilde{D}_{kl}(W_{te}R_{te}^{-}W_{te}^{T}||W_{tr}R_{tr}^{+}W_{tr}^{T}) \end{aligned} \tag{6.18}$$

$$\begin{aligned}\Delta &= \frac{1}{2}D_{kl}(W_{te}R_{te}^{+}W_{te}^{T}||W_{tr}R_{tr}^{+}W_{tr}^{T}) + \\ &\quad \frac{1}{2}D_{kl}(W_{te}R_{te}^{-}W_{te}^{T}||W_{tr}R_{tr}^{-}W_{tr}^{T}) \end{aligned} \tag{6.19}$$

Instead of measuring the distribution divergence between two classes using the test data, the distribution divergence between the test data and training data is formulated in (6.18) and (6.18). In this way, inter-class and within-class divergence between the test data in the adapted subspaces and the training data in the original subspaces could be maximized and minimized, respectively. This is to guarantee that the classifier trained by training features could be effective for the test features. Given P_{te}, the covariance matrix of test data after whitening is

$$\Sigma_{te}^{+/-} = P_{te}R_{te}^{+/-}P_{te}^{T} \tag{6.20}$$

Note that for the adaptation without test labels, $R_{te}^{+/-}$ could be estimated using the predicted labels, while P_{te} is calculated without predicted or true test labels under the assumption of balanced dataset. Since r pairs of spatial filters will be used for feature extraction, based on (6.20), (6.17)-(6.18) could be rewritten as

$$\mathcal{L}(U) = (\mu - 1)\mathcal{L}_{csp}(U) + \mu\Delta(U) \tag{6.21}$$

$$\begin{aligned}\mathcal{L}_{csp}(U) &= \frac{1}{2}\tilde{D}_{kl}(I_d^T U^T \Sigma_{te}^{+} U I_d||WR_{tr}^{-}W^T) \\ &\quad + \frac{1}{2}\tilde{D}_{kl}(I_d^T U^T \Sigma_{te}^{-} U I_d||WR_{tr}^{+}W^T) \end{aligned} \tag{6.22}$$

$$\begin{aligned}\Delta(U) &= \frac{1}{2}D_{kl}(I_d^T U^T \Sigma_{te}^{+} U I_d||WR_{tr}^{+}W^T) + \\ &\quad \frac{1}{2}D_{kl}(I_d^T U^T \Sigma_{te}^{-} U I_d||WR_{tr}^{-}W^T) \end{aligned} \tag{6.23}$$

where $I_d \in \mathbb{R}^{d \times n_c}$ is the identity matrix truncated to the first d columns with $d = 2r$ as the number of spatial filters.

6.2.3 Semi-Supervised Gradient Descent Searching

To solve (6.21), we adopt a subspace searching approach based on a gradient descent on the manifold of orthogonal matrices [77, 126]. In the training stage, the subspace searching can be performed with labels and stopped upon the convergence of the loss function $\mathcal{L}_0$. For the adaptation without test labels, convergence of the loss function $\mathcal{L}$ could be problematic if the predicted labels are used. Adaptation until the convergence of $\mathcal{L}$ is prone to performance drops caused by incorrect predicted labels. To avoid these problems owing to semi-supervised learning, in the proposed adaptation design, the objective function used to update U is calculated based on a subset of available test trials. Let

$$R_{te,b}^{+/-} = \frac{1}{|\mathcal{Q}_{te,b}^{+/-}|} \sum_{i \in \mathcal{Q}_{te,b}^{+/-}} R^i \tag{6.24}$$

$$\Sigma_{te,b}^{+/-} = P_{te} R_{te,b}^{+/-} P_{te}^T \tag{6.25}$$

where $\mathcal{Q}_{te,b}^{+/-}$ denotes a subset of available test data for adaptation. By replacing $\Sigma_{te}^{+/-}$ with $\Sigma_{te,b}^{+/-}$ in (6.21) to (6.23), we can obtain the function for adaptation, which is denoted as $\mathcal{L}_b$. Therefore, the adapted orthogonal matrix is

$$U_{te} = \arg\min_{U} \quad \mathcal{L}_b(U) \tag{6.26}$$

During the gradient descent search, the loss function is also evaluated based on (6.21) to (6.23) using all available test trials at each iteration step, which is denoted by $\mathcal{L}(U)$. Stopping the iteration depends on the changes in both $\mathcal{L}_b(U)$ and $\mathcal{L}(U)$. In particular, the search is stopped if

$$\mathcal{L}(U^{k+1}) > \mathcal{L}(U^k) \tag{6.27}$$

where U^k is the orthogonal matrix U at the k-th step. By using (6.27), some of the trials used to evaluate the change of loss function are independent of the adaptation. Details of the semi-supervised adaptation is summarised in Algorithm 4. After U_{te} is obtained, the projection matrix W_{te} can be calculated as

$$W_{te} = U_{te}^T P_{te} \tag{6.28}$$

6.3 Experimental Study

6.3.1 Experiment Set-Up and Data Description

Please refer to Section 3.4.1.

6.3.2 Data Processing and Feature Extraction

First, we train a CSP model and the Naive Bayesian Parzen Window (NBPW) classifier with the training data as in [80, 68]. Then, as described in Section 6.2, with the predicted labels of a batch of the test data from the new session, the projection matrix W_{te} is calculated and applied to test data for feature

Algorithm 4 Subspace searching based on gradient descent

Input: training data and adaptation data;
Output: U_{te}.
begin
Compute P_{te};
Initialize $U^0 = U_{tr}$;
while $k < n_k$ **do**
Compute the gradient matrix M of $\mathcal{L}_b(U^k)$ with respect to U^k;
Compute

$$H = \begin{pmatrix} 0 & M \\ -M^T & 0 \end{pmatrix} \tag{6.29}$$

Let $t_u = [0.9^5, 0.9^6, ..., 0.9^{10}]$.
Determine the optimal step size

$$\hat{t}_u = \arg\min_{t_u} \mathcal{L}_b(\exp(t_u H)U^k) \tag{6.30}$$

Update the rotation matrix

$$U^{k+1} = \exp(\hat{t}_u H)U^k \tag{6.31}$$

Compute $\delta = \mathcal{L}(U^{k+1}) - \mathcal{L}(U^k)$;
Compute $\delta_b = \mathcal{L}_b(U^{k+1}) - \mathcal{L}_b(U^k)$;
if $\delta > 0$ *or* $|\delta_b| < \zeta$ *(ζ is a small preset value)* **then**
break.
end
k=k+1;
end
$U_{te} = U$.
end

extraction. The number of the pairs of spatial filters $r = 3$, which means that the dimension of the subspace used $d = 6$. Finally, test features are classified by the classifier trained by the training data. In this work, we use the first 1/5 of the test data as adaptation batch and the remaining 4/5 as the evaluation batch. The subset of the adaptation trials , $\mathcal{Q}_{te,b}^{+/-}$, used to calculate $\mathcal{L}_b(U)$ is chosen as 50% of the adaptation trials with higher posterior probabilities for each class given by the NBPW classifier. n_k is set to 100.

6.3.3 Numerical Study

In this section, we investigate how the change of the discriminative subspace influences the feature distribution. In order to visualize the discriminative subspace, we select only 3 channels, C3, Cz, and C4, which are known as the 3 most discriminative channels for motor imagery EEG classification [30]. We select subject 8 from the dataset introduced in Section 6.3, whose training classification based on only C3, Cz, and C4 is the best among all the subjects. Then, we calculate the whitening matrix $P \in \mathbb{R}^{3\times3}$, $U \in \mathbb{R}^{3\times3}$, and projection matrix $W \in \mathbb{R}^{3\times3}$ using (2.1)-(2.10). By listing the diagonal elements of Λ^+ in an ascending order (2.11), the first column, u_1, and the last column, u_3, of U correspond to the spatial filters maximizing the variance of the EEG signals of class - and class +, respectively. Therefore, the most discriminative feature pair comprising $\mathbf{f}_1$ and $\mathbf{f}_3$ is used. The 2D-feature distribution is shown in Figure 6.1, where the features from class + and class - are presented by triangles and circles, respectively. And the mean of each class is presented by a solid triangle/circle. The line representing $x = y$ is denoted in a dashed line for referencee.

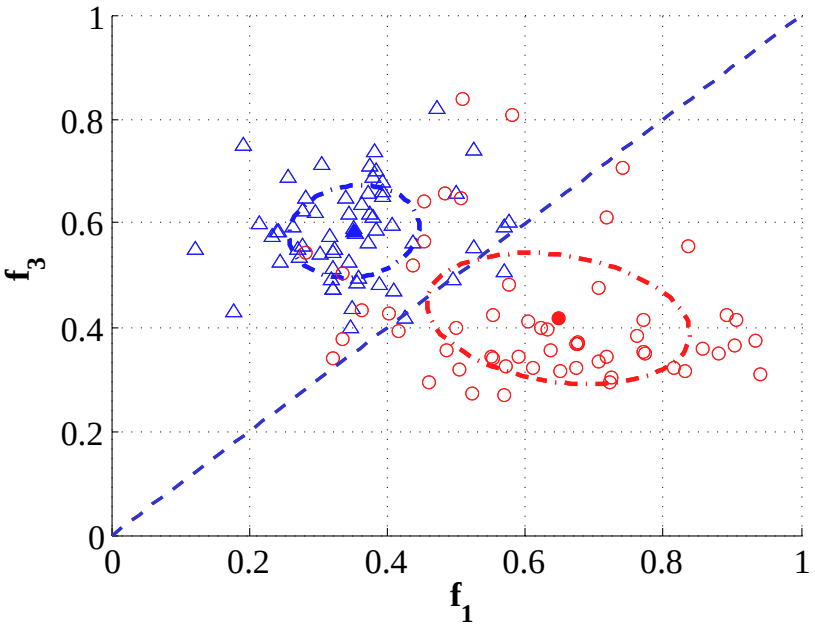


Figure 6.1: An example of 2D feature distribution using channels C3, C4 and Cz, where the features from class + and class - are presented by triangles and circles, respectively. And the mean of each class is presented by a solid triangle/circle.

The subspaces u_1, u_2, and u_3 are illustrated in Figure 6.2. Then, we rotate the subspace U^i around the axis with the direction of u_m, $m \in \{1, 2, 3\}$ by an angle θ, and the rotation matrix is denoted as $R_t(\theta, u_m)$. Details of calculating $R_t(\theta, u_m)$ can be found in Appendix A.6. An example of rotating U around u_2 with $\theta = 0, \frac{\pi}{30}, \frac{\pi}{15}, ..., \frac{\pi}{6}$ is given in Figure 6.2 by the intermediate colors from blue/red to yellow/pink.

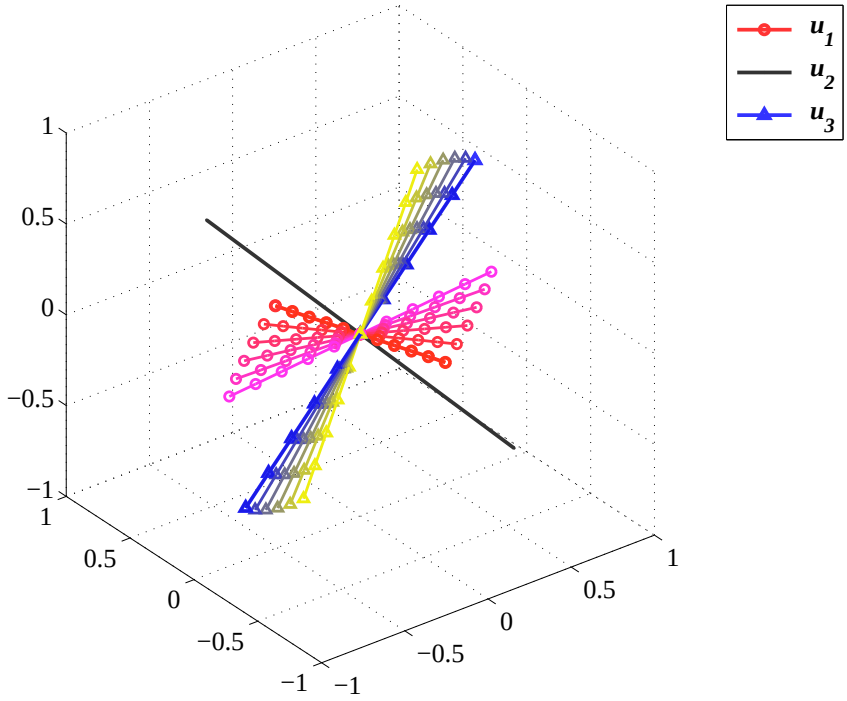


Figure 6.2: The subspaces u_1, u_2, and u_3 in U. An example of rotating U around u_2 with $\theta = 0, \frac{\pi}{30}, \frac{\pi}{15}, ..., \frac{\pi}{6}$ is given by the intermediate colors from blue/red to yellow/pink.

Replacing U^i in (6.9) with $R_t(\theta, u_m)U^i$, we can obtain the j-th feature of trial i with U^i rotated around u_m by θ as

$$\mathbf{f}_j^i(\theta, u_m) = u_j^T (R_t(\theta, u_m)U^i)V^i(R_t(\theta, u_m)U^i)^T u_j \tag{6.32}$$

The distributions of $\mathbf{f}_j^i(\theta, u_m)$ with $m = 1, 2, 3$ are shown respectively in Figures 6.3 to 6.5. For a clearer presentation, only the mean features (solid points) and the covariances of features (ellipses) are shown in Figures 6.3 to 6.5. And rotation angle $\theta = 0, \frac{\pi}{12}, \frac{\pi}{6}..., \frac{\pi}{2}$ is presented by intermediate colors from red/blue ($\theta = 0$) to pink/yellow ($\theta = \frac{\pi}{2}$). As shown in Figures 6.3 and 6.5, when the axis of the rotation is the same as the direction of a discriminative subspace, i.e., u_1 or u_3, the feature corresponding to this direction will not be affected by the rotation. In this case, if the classifier could be rotated appropriately, the features could still be classified. In other words, it can be seen from Figure 6.3 or 6.5 that, when $\theta = \frac{\pi}{2}$, the ideal classifier becomes a vertical or horizontal line, which means that only the feature dimension corresponding to the rotation axis is still discriminative. However, when the axis of the rotation is the same as the direction of u_2, it is impossible to achieve the same classification accuracy by modifying the classifier, as shown in Figure 6.4. In particular, when $\theta = \frac{\pi}{4}$, the feature distributions of the two classes are completely overlapped by each other.

In the case shown in Figures 6.1 to 6.5, the features without any rotation can be regarded as the training data, while the test features are the features after rotation due to the nonstationarity. The objective of this work can be regarded as finding an adapted projection matrix $W_r(\theta, u) = U^T R_t^{-1}(\theta, u_m) P$ so that the adapted feature

$$\begin{aligned} & \mathbf{f}_{a,j}^i(\theta, u_m) \\ = \; & u_j^T R_t^{-1}(\theta, u_m) R_t(\theta, u_m) U^i V^i (U^i)^T R_t^T(\theta, u_m) R_t^{-T}(\theta, u_m) u_j \\ = \; & \mathbf{f}_j^i(\theta, u_m)|_{\theta=0} \end{aligned} \tag{6.33}$$

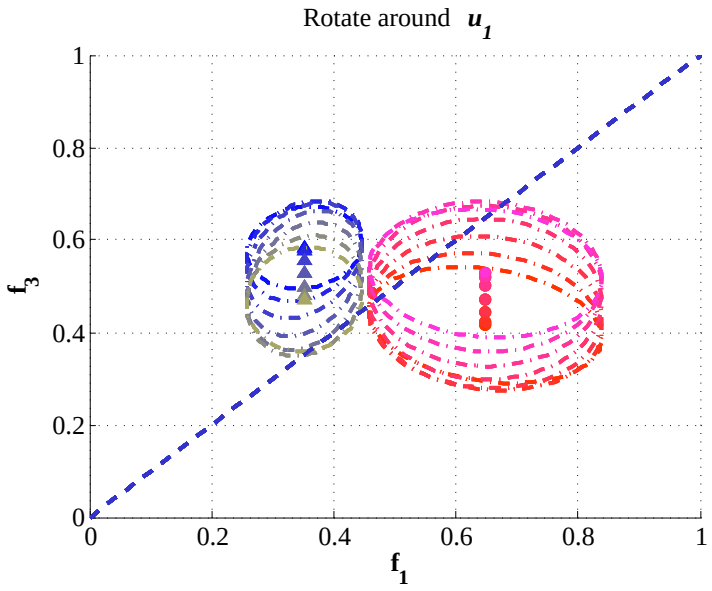


Figure 6.3: Change of the distributions of $\mathbf{f}_j(\theta, u_1)$ with θ. The discrimination of the feature dimension $\mathbf{f}_1$ is not affected by the rotation. The ideal classifier becomes a vertical line when $\theta = \frac{\pi}{2}$.

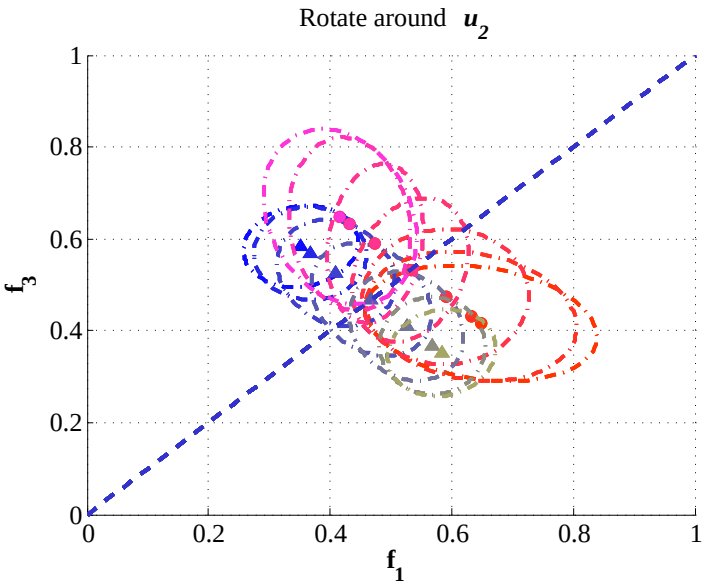


Figure 6.4: Change of the distributions of $\mathbf{f}_j(\theta, u_2)$ with θ. Both feature dimensions are affected by the rotation. It is impossible to achieve the same classification accuracy by changing the classifier only.

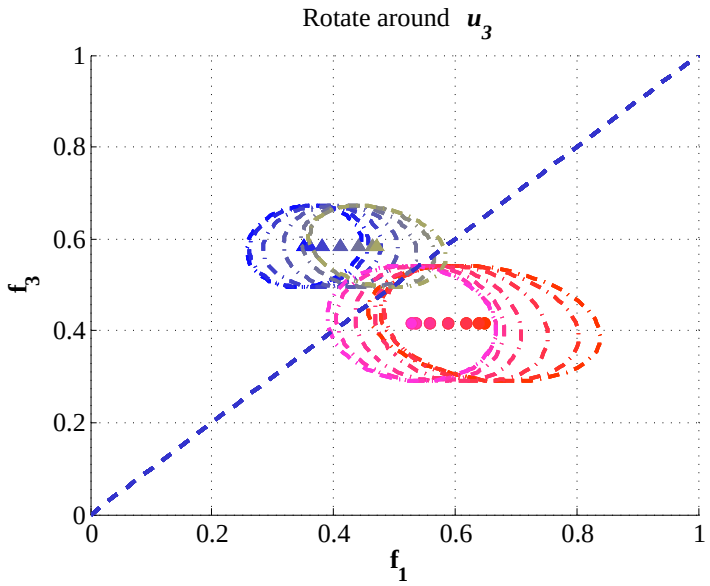


Figure 6.5: Change of the distributions of $\mathbf{f}_j(\theta, u_3)$ with θ. The discrimination of the feature dimension $\mathbf{f}_3$ is not affected by the rotation. The ideal classifier becomes a horizontal line when $\theta = \frac{\pi}{2}$.

When (6.33) holds, the classifier trained by the features without rotation could be equally effective for the features after rotation with projection matrix properly adapted, which is the goal we want to achieve for adaptation.

In this numerical study, we only show the feature distribution change with the rotation around the axis parallel to the direction of either one of the discriminative subspaces. In real cases, neither the rotation axis nor the angle is available, and rotation would be more complicated. For example, the direction of the rotation axis would be a combination of u_m, such as $\sum_{m=1,2,3} g_m u_m$, where g_m is the scalar coefficient. Therefore, it is difficult to find out W_r in the form of $W_r(\theta, u) = U^T R_t^{-1}(\theta, u_m)P$ explicitly, so we propose the method in Section 6.2 to search the discriminative subspace of the projection matrix.

6.3.4 Classification Results

Figure 6.6 summarizes the results of the proposed adaptation method, denoted by W_{te}, compared with the adaptation method based on normalization without updating the orthogonal part in the projection matrix, denoted by W_n as in (6.3). Note that all classification accuracies are based on the evaluation batch. As shown by Figure 6.6, for most of the subjects the proposed adaptation method yields improvements with very few drops compared to the normalization approach in [104]. Besides, the average accuracy of the proposed method using W_{te} is 67.42%, which is higher than that of using W_n, i.e., 66.41%.

The changes in $\mathcal{L}_b$ and $\mathcal{L}$ with respect to the iteration number k are shown in Figures 6.7 and 6.8, respectively. And the change in classification accuracy

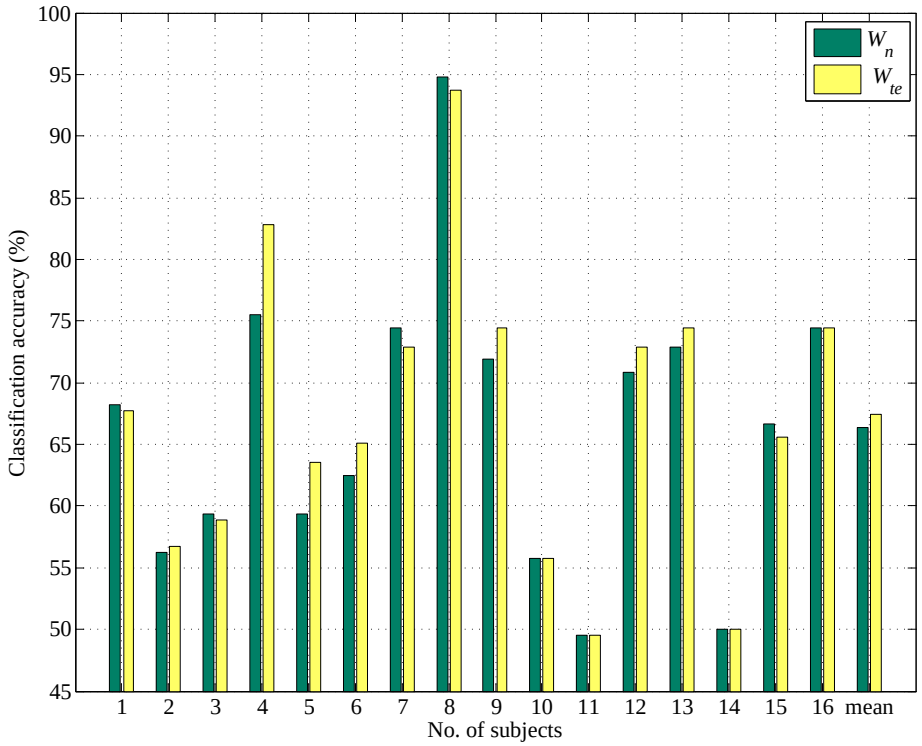


Figure 6.6: Accuracy comparison. The average accuracy of the proposed method using W_{te} is 67.42%, which is higher than that of using W_n, i.e., 66.41%.

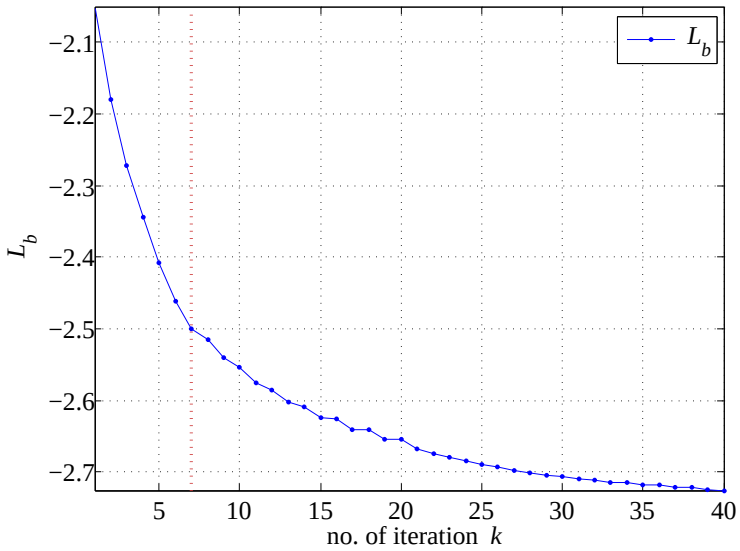


Figure 6.7: Change in $\mathcal{L}_b$ with respect to iteration number k.

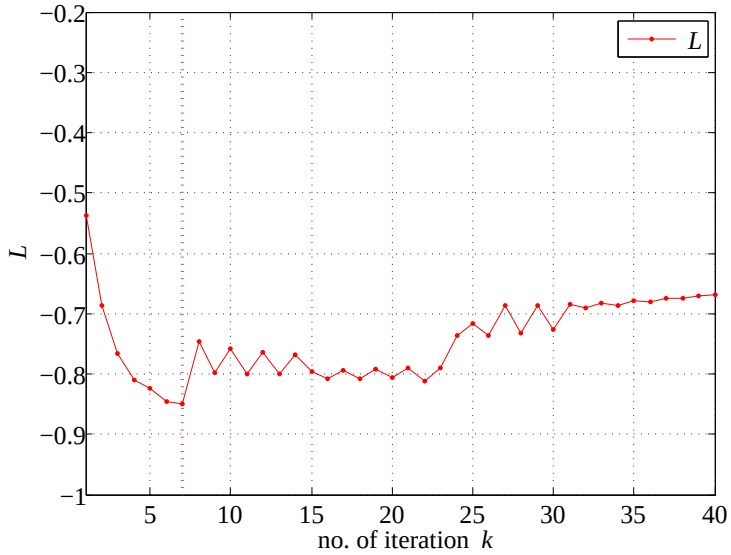


Figure 6.8: Change in $\mathcal{L}$ with respect to iteration number k.

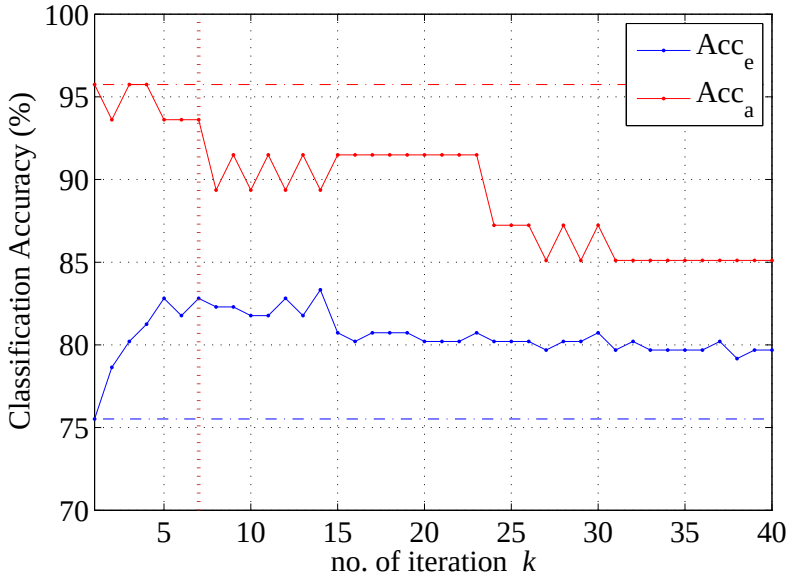


Figure 6.9: Change in classification accuracy with respect to the iteration number k. The x-axis represents the value of k, and the y-axis represents classification accuracy. Acc_a and Acc_e represent the classification accuracies of adaptation batch and evaluation batch, respectively, and the baselines of the normalization approach are denoted by dotted-dashed lines.

with respect to the iteration number k is shown in Figure 6.9. In Figure 6.7, $\mathcal{L}_b$ decreases with respect to k until the convergence at $k = 40$. And in Figure 6.8, $\mathcal{L}$ decreases with respect to k first and then increases after $k = 7$. In Figure 6.9, for the evaluation batch, the classification accuracies first increase and then decrease, and for the adaptation batch, the classification accuracies decrease more significantly after $k = 7$. If the adaptation is stopped upon the convergence of $\mathcal{L}_b$, the classification accuracy of neither batch is optimal. As illustrated in Figure 6.8, $\mathcal{L}$ decreases first and begins to increase from $k = 7$, which means that the adaptation could no longer benefit the unselected trial after $k = 7$. Thus, in the proposed method, the adaptation is stopped when $\mathcal{L}^k > \mathcal{L}^{k-1}$, i.e., $k = 7$ in this case. As shown in Figure 6.9, the classification accuracies when $k = 7$ of both batches are higher than that when $k = 40$, i.e., the convergence of $\mathcal{L}_b$. This shows the effectiveness of the stop criterion in the proposed adaptation design. As shown in Figure 6.6, the proposed method fails to improve the performance for some subjects. To investigate the underlying reason, we perform similar analysis of the change in the loss function and the classification accuracy for the subjects with little improvement in performance. We find that for subjects 10, 11 and 15 the adaptation is stopped at the very beginning of the iteration, which yields results very similar to the baseline. A possible reason for subjects 10 and 11 is that the classification accuracy of the adaptation batch is similar to that obtained purely by chance. Hence, with very few correctly predicted labels it is difficult to find a right adaptation direction and the iteration stops at the beginning. The good side of this result is that the adaptation toward a wrong direction is avoided. In our future work, we would focus on solving this problem with a better searching strategy.

6.4 Conclusion

To address the nonstationarity issue for a more general case, the shift of the discriminative subspaces should be investigated. In particular, the influence of such a shift on the feature space has been analyzed theoretically and investigated by a numerical study. To solve the problem of nonstationary discriminative subspaces, this study investigates the feasibility of updating the spatial filters by adapting the discriminative subspaces for test data. The adaptation is facilitated by the gradient searching on the manifold of orthogonal matrices based on the divergence-based framework in a semi-supervised manner. In this way, the orthogonal part could be adapted together with the update of the whitening part in the spatial filters, and the cross-session data variation with the asymmetric data transformation could be taken into consideration. To account for the risk in the semi-supervised learning, the adaptation trials are divided into two subsets. Only one subset is used to obtain the adaptation direction, while the search is stopped by the change of loss function of all adaptation trials. The advantage of this cross-validation-like design is to have independent validation of the adaptation and to avoid possible over-fitting. Experimental studies show that the proposed method further enhances the BCI performance compared to the normalization adaptation approach.

Chapter 7

Conclusion and Future Work

In this chapter, the results of the research work are summarised, and the major contributions of this work are highlighted, following which suggestions for future work are presented.

7.1 Conclusion

The thesis has investigated modelling and classification of motor imagery EEG for BCI. Model generalization has been studied with the discriminative learning of propagation and spatial pattern, and ensemble learning of spatial filters presented.

(i) Conventional spatial filter design based on instant mixing model is not capable of describing complex dynamics such as neuronal propagation, as accumulating neuroscience findings suggest that cooperation of multiple brain regions is involved in motor imagery. To take the causal relationship during motor imagery into consideration, in Chapter 3, we propose a novel discriminative algorithm for joint learning of propagation and spatial pattern with an iterative optimization approach. In particular, a convolutive model is used to describe the relationship between source signals and scalp EEG. Experimental studies validated

the effectiveness of the proposed discriminative learning of propagation and spatial pattern analysis. Moreover, the oscillatory background noise related to ongoing activity has been analyzed by comparing the proposed model and the MVAR model in the frequency space. Based on KL-divergence measurements, we find that nonstationarity of the EEG data can be reduced by the proposed method, which confirms our analysis of background noise reduction.

(ii) As shown in the background noise analysis, the nonstationarity inherent in EEG signals poses a big challenge for modelling EEG in BCI. Biased estimates of covariance matrices would lead to the ineffectiveness of the spatial filters. To overcome this problem, an ensemble learning of spatial filter design has been proposed to improve the feature extraction model in Chapter 4. The mismatch between data and model is evaluated using features, and samples that are more likely to be misclassified are selected. Multiple spatial filters are constructed based on different groups of samples, and the final projection matrix for feature extraction is designed as a weighted summation of different spatial filters. In this way, the biased estimates as well as the sample discrepancies can be taken into consideration. The experimental results showed the improved classification accuracy of the proposed method. Significant improvements for the subjects with relatively poorer BCI performance indicate the effectiveness of the ensemble learning of spatial filter.

Moreover, considering significant cross-session data variation, model adaptation methods are developed, by building a novel data-model mismatch metric without test labels and searching discriminative subspace for test data on

a manifold.

(i) Since session-to-session nonstationarity could be very significant, it is necessary for the computational model obtained from the calibration session to adapt to the data for BCI-based rehabilitation. The key challenge for adapting the computational model is how to construct a metric that measures the mismatch between test data and the model obtained from training data, especially when the labels of the test data are not available. To address this problem, in Chapter 5, we construct a metric that measures this data-model mismatch, which is used to guide the adaptation toward reducing the data-model mismatch. Since it is difficult to achieve the residual error minimization and the discrimination objective simultaneously, we propose a two-step approach where the residual error is estimated in the first step and then combined with the discrimination objective function in a regularized manner. Experimental results showed that the quantified mismatch was closely related to the classification accuracy, thus validating the proposed metric in measuring the data-model mismatch. The classification results also showed that the proposed adaptation framework reduced the feature distribution shift, increased the separability of the test features, and yielded higher classification accuracies compared with other regularization or adaptation methods.

(ii) As discussed in Section 2.1, the projection matrix in CSP consists of a whitening part and an orthogonal part, which can be deemed as the discriminative subspace for EEG data. There exist methods that adapt the projection matrix by re-estimating the whitening part, the

effectiveness of which is subject to whether the session-to-session data transformation of the two class is symmetric. In Chapter 6, we show that the effectiveness of the projection matrix closely relates to the consistence of discriminative subspace. Following the theoretical analysis, a discriminative subspace tracking method is introduced for model adaptation. In particular, the adaptation based on the searching on a manifold of orthogonal matrices is proposed to update the discriminative subspace, i.e., the orthogonal part in the projection matrix, so that the adaptation for a more general case, i.e., asymmetric transformation, could be addressed. To avoid possible problems arising from the semi-supervised learning, a cross-validation-based loss function is proposed to evaluate the adaption direction. Experimental results showed that compared to the normalization methods proposed in [104] the proposed adaptation method with discriminative subspace tracking could further enhance the classification results. Moreover, by analyzing the change in classification accuracy and loss function, the cross-validation-based loss function can stop the adaptation from wrong directions.

7.2 Limitations and Future Work

In this section, we discuss the limitations of our work and suggest topics for further investigation.

(i) The work in Chapter 3 focuses on modelling the motor imagery EEG using a convolutive model to describe the neuronal propagation dynamics, while the features extracted by the model are ERD/ERS features. However, the generation of the ERD/ERS in relation to the connectiv-

ity pattern is not fully explored. For example, it is not clear whether the ERD/ERS of single sources causes the propagation or the propagation involved in the oscillatory signals generates the ERD/ERS. By studying the role that connectivity plays in ERD/ERS, we could have a better understanding of the ERD/ERS generation for further enhancing the model generalization. Similarly, discriminative learning of connectivity related features could be investigated by studying the relationship between ERD/ERS effects and connectivity.

(ii) As shown in the numerical study in Chapter 6, after whitening, the change of the discriminative subspaces of two classes are related to each other. In other words, as the session-to-session shift of the discriminative subspaces is bias-like on manifolds, it could be learnt more efficiently in an unsupervised manner. The current semi-supervised adaptation strategies that treat each class independently fail to fully utilize this property. New methods should be developed to find a more effective way of variation tracking.

(iii) The proposed methods have addressed the feature extraction issue, as the feature extraction model is crucial to the classification error, e.g., Bayes error in [76]. However, the relationship between the nonstationarity in EEG and the classification error has not been established rigorously. In particular, discriminative subspaces could be used to model the distribution of the covariance matrices on manifolds. In the future work, it is necessary to perform rigorous theoretical analysis on the distribution of the discriminative subspaces on manifolds, and its relationship with the Bayes error of feature classification.

(iv) The proposed methods have addressed the nonstationarity problem from the perspective of computational models, while the generation of the nonstationarity has not been fully understood yet. With increasing interest of studying the resting state EEG, more neurophysiological knowledge of the nonstationarity inherent in EEG could be used as prior knowledge for computational model design. Thus, in the future work, experiments to understand nonstationarity from the neurophysiological perspective could be conducted as a basis of robust computational model design.

Bibliography

[1] J. Wolpaw, E. Wolpaw, BrainComputer Interfaces: Principles and Practice, Oxford University Press, New York, 2012.

[2] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, T. M. Vaughan, Brain-computer interfaces for communication and control, Clinical Neurophysiology 113 (6) (2002) 767–791.

[3] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, T. M. Vaughan, Brain-computer interfaces as new brain output pathways, The Journal of Physiology 579 (2007) 613–619.

[4] D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A.-S. LaMantia, J. O. McNamara, S. M. Williams, Neuroscience, 2nd edition, Sinauer Associates, Sunderland (MA), 2001.

[5] M. Hmaleinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, O. V. Lounasmaa, Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain, Reviews of Modern Physics 65 (1993) 413–497.

[6] P. L. Nunez, R. Srinivasan, The neurophysics of EEG, 2nd edition, Oxford University Press, New York, 2006.

[7] P. Shenoy, K. J. Miller, J. G. Ojemann, R. P. N. Rao, Generalized features for electrocorticographic BCIs, IEEE Transactions on Biomedical Engineering 55 (1) (2008) 273–280.

[8] C. M. Gray, P. E. Maldonado, M. Wilson, B. McNaughton, Tetrodes markedly improve the reliability and yield of multiple single-unit isolation from multi-unit recordings in cat striate cortex, Journal of Neuroscience Methods 63 (1-2) (1995) 43–54.

[9] M. M. Ter-Pogossian, M. E. Phelps, E. J. Hoffman, N. A. Mullani, A positron-emission transaxial tomograph for nuclear imaging (PETT), Radiology 114 (1) (1975) 89–98.

[10] Y. F. Tai, P. Piccini, Applications of positron emission tomography (PET) in neurology, Journal of Neurology Neurosurgery Psychiatry 75 (5) (2004) 669–676.

[11] S. Ogawa, T. M. Lee, A. R. Kay, D. W. Tank, Brain magnetic resonance imaging with contrast dependent on blood oxygenation, Proceedings of the National Academy of Sciences 87 (24) (1990) 9868–9872.

[12] J. C. W. Siero, N. Petridou, H. Hoogduin, P. R. Luijten, N. F. Ramsey, Cortical depth-dependent temporal dynamics of the bold response in the human brain, Journal of Cerebral Blood Flow and Metabolism 31 (4) (2011) 1999–2008.

[13] F. F. Jobsis, Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters, Science 198 (4323) (1977) 1264–1267.

[14] A. Villringer, J. Planck, C. Hock, L. Schleinkofer, U. Dirnagl, Near infrared spectroscopy (NIRS): A new tool to study hemodynamic changes during activation of brain function in human adults, Neuroscience Letters 154 (1-2) (1993) 101–104.

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

[16] S. M. Coyle, T. E. Ward, C. M. Markham, Brain-computer interface using a simplified functional near-infrared spectroscopy system, Journal of Neural Engineering 4 (3) (2007) 219–226.

[17] S. A. Huettel, A. W. Aong, G. McCarthy, Functional Magnetic Resonance Imaging, 2nd edition, Sinauer, Massachusetts, 2009.

[18] S. Ogawa, T. M. Lee, A. S. N. andP. Glynn, Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields, Magnetic Resonance in Medicine 14 (1) (1990) 68–78.

[19] J. Malmivuo, R. Plonsey, Bioelectromagnitism, Oxford University Press, New York, 1995.

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

[21] G. Gomez-Herrero, M. Atienza, K. Egiazarian, J. L. Cantero, Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials, Electroencephalography and Clinical Neurophysiology 70 (6) (1988) 497–508.

[22] C. C. Duncan, R. J. Barry, J. F. Connolly, C. Fischer, P. T. Michie, R. Naatanen, J. Polich, I. Reinvang, C. V. Petten, Event-related potentials in clinical research: guidelines for eliciting, recording, and quanti-

fying mismatch negativity, P300, and N400, Clinical Neurophysiology 120 (11) (2009) 1883–1908.

[23] E. Donchin, M. G. H. Coles, Is the P300 component a manifestation of context updating?, Behavioral and Brain Sciences 11 (3) (1988) 357–374.

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

[25] E. W. Sellers, E. Donchin, A P300-based brain-computer interface: initial tests by ALS patients, Clinical Neurophysiology 117 (3) (2006s) 538–548.

[26] U. Hoffmann, J. M. Vesin, T. Ebrahimi, K. Diserens, An efficient p300-based brain-computer interface for disabled subjects, Journal of Neuroscience Methods 167 (1) (2008) 115–125.

[27] F. Nijboer, E. W. Sellers, J. Mellinger, M. A. Jordan, T. Matuz, A. Furdea, S. Halder, U. Mochty, D. J. Krusienski, T. M. Vaughan, J. R. Wolpaw, N. Birbaumer, A. Kubler, A P300-based brain-computer interface for people with amyotrophic lateral sclerosis, Journal of Neuroscience Methods 119 (8) (2008) 1909–1916.

[28] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, J. R. Wolpaw, Toward enhanced P300 speller performance, Journal of Neuroscience Methods 167 (1) (2008) 15–21.

[29] G. Pfurtscheller, A. Aranibar, Evaluation of event-related desynchronization (ERD) preceding and following voluntary self-paced movements, Electroencephalography and Clinical Neurophysiology 46 (2) (1979) 138–146.

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

[31] G. Pfurtscheller, C. Neuper, D. Flotzinfger, M. Pregenzer, EEG-based discrimination between imagination of right and left hand movement, Electroencephalography and Clinical Neurophysiology 103 (6) (1997) 642–651.

[32] G. Pfurtscheller, Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest, Electroencephalography and Clinical Neurophysiology 83 (1) (1992) 62 – 69.

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

[34] C. Tangwiriyasakul, R. Verhagen, M. J. A. M. van Putten, W. L. C. Rutten, Importance of baseline in event-related desynchronization during a combination task of motor imagery and motor observation, Journal of Neural Engineering 10 (2) 026009.

[35] D. J. McFarland, L. A. Miner, T. M. Vaughan, J. R. Wolpaw, Mu and beta rhythm topographies during motor imagery and actual movements, Brain Topography 12 (2000) 177–186.

[36] G. Pfurtscheller, R. Leeb, C. Keinrath, D. Friedman, C. Neuper, C. Guger, M. Slater, Walking from thought, Brain Research 1071 (1) (2006) 145–152.

[37] M. Jeannerod, Mental imagery in the motor context, Neuropsychologia 33 (1995) 1419–1432.

[38] J. Munzert, B. Lorey, K. Zentgraf, Cognitive motor processes: The role of motor imagery in the study of motor representations, Brain Research Reviews 60 (2) (2009) 306–326.

[39] H. Chen, Q. Yang, W. Liao, Q. Gong, S. Shen, Evaluation of the effective connectivity of supplementary motor areas during motor imagery using granger causality mapping, NeuroImage 47 (4) (2009) 1844–1853.

[40] P. L. Jackson, M. F. Lafleur, F. Malouin, C. L. Richards, J. Doyon, Functional cerebral reorganization following motor sequence learning through mental practice with motor imagery, NeuroImage 20 (2) (2003) 1171–1180.

[41] J. Liepert, T. Hassa, O. Tuscher, R. Schmidt, Motor excitability during movement imagination and movement observation in psychogenic lower limb paresis, Journal of Psychosomatic Research 70 (1) (2011) 59–65.

[42] N. Sharma, V. M. Pomeroy, J.-C. Baron, Motor imagery: a backdoor to the motor system after stroke?, Stroke 56 (11) (2006) 1941–1952.

[43] K. K. Ang, C. Guan, Brain-Computer interface in stroke rehabilitation, IEEE International Joint Conference on Neural Networks and Computational Intelligence 7 (2) (2013) 139–146.

[44] K. K. Ang, C. Guan, K. S. G. Chua, B.-T. Ang, C. Kuah, C. Wang, K. S. Phua, Z. Y. Chin, H. Zhang, A clinical study of motor imagery-based brain-computer interface for upper limb robotic rehabilitation (2009) 5981–5984.

[45] E. Buch, C. Weber, L. G. Cohen, C. Braun, M. A. Dimyan, T. Ard, J. Mellinger, A. Caria, S. Soekadar, A. Fourkas, N. Birbaumer, Think to move: a neuromagnetic brain-computer interface (BCI) system for chronic stroke 39 (3) (2008) 910–917.

[46] G. Pfurtscheller, G. R. M 'thought'- control of functional electrical stimulation to restore hand grasp in a patient with tetraplegia, Neuroscience Letters 351 (1) (2003) 33–36.

[47] H. I. Krebs, J. J. Palazzolo, L. Dipietro, M. Ferraro, J. Krol, K. Rannekleiv, B. Volpe, N. Hogan, Rehabilitation robotics: Performance-based progressive robot-assisted therapy, Autonomous Robots 15 (2003) 7–20.

[48] S. Silvoni, A. Ramos-Murguialday, M. Cavinato, C. Volpato, G. Cisotto, A. Turolla, F. Piccione, N. Birbaumer, Brain-computer interface in stroke: a review of progress, clinical EEG and neuroscience 42 (4) (2011) 245.

[49] K. K. Ang, C. Guan, K. S. G. Chua, B. T. Ang, C. W. K. Kuah, C. Wang, K. S. Phua, Z. Y. Chin, H. Zhang, A large clinical study on the ability of stroke patients to use EEG-based motor imagery brain-computer interface, Clinical EEG and Neuroscience 42 (4) (2011) 253–258.

[50] P. von Bunau, F. C. Meinecke, F. J. Kiraly, K.-R. Müllerr, Finding stationary subspaces in multivariate time series, Physical Review Letters 103 (21) (2009) 214101.

[51] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, T. M. Vaughanr, Brain-computer interface technology: A review of the first international meeting, IEEE Transactions on Rehabilitaion Engineering 8 (2) (2000) 164–173.

[52] C. Tangwiriyasakul, V. Mocioiu, M. J. A. M. van Putten, W. L. C. Rutten, Classification of motor imagery performance in acute stroke, Journal of Neural Engineering 11 (3) 036001.

[53] B. Blankertz, R. Tomika, S. Lemm, M. Kawanabe, K.-R. Müller, Optimizing spatial filters for robust EEG single trial-trial analysis, IEEE Signal Processing Magazine 25 (1) (2008) 41–56.

[54] D. Choi, Y. Ryu, Y. Lee, M. Lee, Performance evaluation of a motor-imagery based EEG-brain computer interface using a combined cue with heterogeneous training data in BCI-naive subjects, BioMedical Engineering OnLine 10 (91).

[55] D. J. McFarland, L. M. McCane, S. V. David, J. R. Wolpaw, Spatial filter selection for EEG-based communication, Electroencephalography and Clinical Neurophysiology 103 (3) (1997) 386–394.

[56] J. R. Knott, F. A. Gibbs, C. E. Henry, Fourier transforms of the electroencephalogram during sleep, Journal of Experimental Psychology 31 (6) (1942) 465–477.

[57] N. Hazarika, J. Z. Chen, A. C. Tsoi, A. Sergejew, Classification of EEG signals using the wavelet transform, Signal Processing 59 (1) (1997) 61–72.

[58] A. W. Chiu, M. Derchansky, M. Cotic, P. L. Carlen, S. O. Turner, B. L. Bardakjian, Wavelet-based gaussian-mixture hidden markov model for the detection of multistage seizure dynamics: A proof-of-concept study, BioMedical Engineering OnLine 10 (29).

[59] C. Yeh, H. Chang, C. Wu, P. Lee, Extraction of single-trial cortical beta oscillatory activities in EEG signals using empirical mode decomposition, BioMedical Engineering OnLine 9 (25).

[60] A. Khorshidtalab, M. J. E. Salami, M. Hamedi, Robust classification of motor imagery EEG signals using statistical time-domain features, Physiological Measurement 34 (11) (2013) 1563.

[61] L. Gao, J. Wang, L. Chen, Event-related desynchronization and synchronization quantification in motor-related EEG by kolmogorov entropy, Journal of Neural Engineering 10 (3) (2013) 036023.

[62] L. Astolfi, F. Cincotti, D. Mattia, F. de Vico Fallani, S. Salinari, M. Ursino, M. Zavaglia, M. G. Marciani, F. Babiloni, Estimation of the cortical connectivity patterns during the intention of limb movements, IEEE Engineering in Medicine and Biology Magazine 25 (4) (2006) 32–38.

[63] M. L. Stavrinou, L. Moraru, L. Cimponeriu, S. D. Penna, A. Bezerianos, Evaluation of cortical connectivity during real and imagined rhythmic finger tapping, Brain Topography 19 (3) (2007) 137–145.

[64] A. Ewald, L. Marzetti, F. Zappasodi, F. C. Meinecke, G. Nolte, Estimating true brain connectivity from EEG/MEG data invariant to linear and static transformations in sensor space, NeuroImage 60 (1) (2012) 476–488.

[65] Q. Wei, Y. Wang, X. Gao, S. Gao, Amplitude and phase coupling measures for feature extraction in an EEG-based brain-ccomputer interface, Journal of Neural Engineering 4 (2007) 120–129.

[66] E. Gysels, P. Celka, Phase synchronization for the recognition of mental tasks in a braincomputer interface, IEEE Transactions on Rehablitation Engineering 12 (4) (2007) 406–415.

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

[68] K. K. Ang, Z. Y. Chin, H. Zhang, C. Guan, Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs, Pattern Recognition 45 (6) (2012) 2137–2144.

[69] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, B. Arnaldi, A review of classification algorithms for EEG-based brain-computer interfaces, Journal of Neural Engineering 4 (2) (2007) R1.

[70] S. Lemm, B. Blankertz, G. Curio, K.-R. Müller, Spatio-spectral filters for improving the classification of single trial EEG, IEEE Transactions on Biomedical Engineering 52 (9) (2005) 1541 –1548.

[71] W. Wu, X. Gao, B. Hong, S. Gao, Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL), IEEE Transactions on Biomedical Engineering 55 (6) (2008) 1733–1743.

[72] B. Blankertz, M. Kawanabe, R. Tomioka, F. U. Hohlefeld, V. Nikulin, K.-R. Müller, Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing, Advances in Neural Information Processing Systems 20 (2008) 113–120.

[73] Z. J. Koles, The quantitative extraction and toporraphic mapping of the abnormal components in the clinical EEG, Electroencephalography and Clinical Neurophysiology 79 (1991) 440–447.

[74] J. M. Gerkinga, G. Pfurtscheller, H. Flyvbjergc, Designing optimal spatial filters for single-trial EEG classiffication in a movement task, Clinical Neurophysiology 110 (1999) 787–798.

[75] W. Wu, Z. Chen, S. Gao, E. N. Brown, A hierarchical bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG, NeuroImage 56 (4) (2011) 1929–1945.

[76] H. Zhang, H. Yang, C. Guan, Bayesian learning for spatial filtering in an EEG-based brain-computer interface, IEEE Transactions on Neural Networks and Learning Systems 24 (7) (2013) 1049–1060.

[77] W. Samek, M. Kawanabe, K.-R. Müller, Divergence-based framework for common spatial patterns algorithms, IEEE Reviews in Biomedical Engineering 7 (2013) 50–72.

[78] G. Pfurtscheller, B. Graimann, J. E. Huggins, S. Levine, L. A. Schuh, Spatiotemporal patterns of beta desynchronization and gamma synchronization in corticographic data during self-paced movement, Clinical Neurophysiology 114 (7) (2003) 1226 – 1236.

[79] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, K.-R. Müller, Combined optimization of spatial and temporal filters for improving brain-computer interfacing, IEEE Transactions on Biomedical Engineering 53 (11) (2006) 2274–2281.

[80] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, H. Zhang, Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b, Frontiers in Neuroscience 6 (39).

[81] K. P. Thomas, C. Guan, C. T. Lau, A. P. Vinod, K. K. Ang, A new discriminative common spatial pattern method for motor imagery brain-computer interfaces, IEEE Transactions on Biomedical Engineering 56 (11) (2009) 2730 –2733.

[82] Q. Novi, C. Guan, T. H. Dat, P. Xue, Sub-band common spatial pattern (SBCSP) for brain-computer interface, in: the 3rd International IEEE/EMBS Conference on Neural Engineering, 2007, pp. 204–207.

[83] H. Zhang, Z. Y. Chin, K. K. Ang, C. Guan, C. Wang, Optimum spatio-spectral filtering network for brain-computer interface, IEEE Transactions on Neural Networks 22 (1) (2011) 52–63.

[84] H. Higashi, T. Tanaka, Simultaneous design of fir filter banks and spatial patterns for EEG signal classification, IEEE Transactions on Biomedical Engineering 60 (4) (2013) 1100–1110.

[85] E. Formaggio, S. F. Storti, R. Cerini, A. Fiaschi, P. Manganotti, Brain oscillatory activity during motor imagery in EEG-fMRI coregistration, Magnetic Resonance Imaging 28 (10) (2010) 1403–1412.

[86] Z. Chin, K. K. Ang, C. Guan, C. Wang, H. Zhang, Filter bank feature combination (FBFC) approach for brain-computer interface, The 2011 International Joint Conference on Neural Networks (IJCNN) (2011) 1352–1357.

[87] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. E. Raichle, Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior, Neuron 56 (2007) 171–184.

[88] F. de Pasquale, S. D. Penna, A. Z. Snyder, C. Lewis, D. Mantini, L. Marzetti, P. Belardinelli, L. Ciancetta, V. Pizzella, G. L. Romani, M. Corbetta, Temporal dynamics of spontaneous MEG activity in brain networks, Proceedings of the National Academy of Sciences 107 (2010) 6040–6045.

[89] P. von Bunau, F. C. Meinecke, F. J. Kiraly, K.-R. Müller, Finding stationary subspaces in multivariate time series, Phys. Rev. Lett. 103 (2009) 214101.

[90] P. von Bunau, F. C. Meinecke, S. Scholler, K.-R. Müller, Finding stationary brain sources in EEG data, 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2010) 2810–2813.

[91] F. Lotte, C. Guan, Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms, IEEE Transactions on Biomedical Engineering 58 (2) (2011) 355–362.

[92] W. Samek, C. Vidaurre, K.-R. Müller, M. Kawanabe, Stationary common spatial patterns for brain-omputer interfacing, Journal of Neural Engineering 9 (2) (2012) 026013.

[93] M. Arvaneh, C. Guan, K. K. Ang, C. Quek, Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain computer interface, IEEE Transactions on Neural Networks and Learning Systems 24 (4) (2013) 610–619.

[94] W. Samek, F. Meinecke, K.-R. Müller, Transferring subspaces between subjects in brain-computer interfacing, IEEE Transactions on Biomedical Engineering 55 (3) (2008) 902–913.

[95] H. Lu, H. Eng, C. Guan, K. N. Plataniotis, A. N. Venetsanopoulos, Regularized common spatial pattern with aggregation for EEG classification in small-sample setting, IEEE Transactions on Biomedical Engineering 57 (12) (2010) 2936–2946.

[96] M. Arvaneh, C. Guan, K. K. Ang, C. Quek, Optimizing the channel selection and classification accuracy in EEG-based BCI, IEEE Transactions on Biomedical Engineering 58 (6) (2011) 1865–1873.

[97] C. Vidaurre, M. Kawanabe, P. von Bunau, B. Blankertz, K.-R. Müller, Toward unsupervised adaptation of LDA for brain computer interfaces, IEEE Transactions on Biomedical Engineering 58 (3) (2011) 587–597.

[98] S. R. Liyanage, C. Guan, H. Zhang, K. K. Ang, J. Xu, T. H. Lee, Dynamically weighted ensemble classification for non-stationary EEG processing, Journal of Neural Engineering 10 (3) (2013) 036007.

[99] C. Vidaurre, A. Schlogl, R. Cabeza, R. Scherer, G. Pfurtscheller, Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces, IEEE Transactions on Biomedical Engineering 54 (3) (2007) 550–556.

[100] Y. Li, C. Guan, An extended EM algorithm for joint feature extraction and classification in brain-computer interfaces, Neural Computation 18 (2006) 2730–2761.

[101] A. Llera, V. Gomez, H. J. Kappen, Adaptive classification on brain-computer interfaces using reinforcement signals, Neural Computation 24 (2012) 2900–2923.

[102] A. Bamdadian, C. Guan, K. K. Ang, J. Xu, Online semi-supervised learning with KL distance weighting for motor imagery-based BCI, 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2012) 2732–2735.

[103] M. Arvaneh, C. Guan, K. K. Ang, C. Quek, EEG data space adaptation to reduce inter-session non-stationarity in brain-computer interface, Neural Computation 25 (8) (2013) 2146–2171.

[104] R. Tomioka, J. Hill, B. Blankertz, K. Aihara, Adapting spatial filtering methods for nonstationary BCIs, 2006 Workshop on Information-Based Induction Sciences (2006) 65–70.

[105] L. A. Baccala, K. Sameshima, Partial directed coherence: a new concept in neural structure determination, Biological Cybernetics 84 (2001) 463–474.

[106] M. Kaminski, K. Blinowska, A new method of the description of the information flow in the brain structures, Biological Cybernetics 65 (1991) 203–210.

[107] M. Kaminski, M. Ding, W. A. Truccolo, S. Bressle, Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance, Biological Cybernetics 85 (2001) 145–157.

[108] J. G. Jr, K. J. Blinowska, M. Kaminski, P. J. Durka, Phase and amplitude analysis in time-requency space-pplication to voluntary finger movement, Journal of Neuroscience Methods 110 (1-2) (2001) 113–124.

[109] A. Schlogl, G. Supp, Analyzing event-related EEG data with multivariate autoregressive parameters, Progress in Brain Research 159 (2006) 135–147.

[110] R. Kus, M. Kaminski, K. J. Blinowska, Determination of EEG activity propagation: pair-wise versus multichannel estimate, IEEE Transactions on Biomedical Engineering 51 (9) (2004) 1501–1510.

[111] G. Gomez-Herrero, M. Atienza, K. Egiazarian, J. L. Cantero, Measuring directional coupling between EEG sources, NeuroImage 43 (3) (2008) 497–508.

[112] M. Dyrholm, S. Makeig, L. K. Hansen, Convolutive ICA for spatio-temporal analysis of EEG, Neural Computation 19 (2007) 934–955.

[113] A. Bahramisharif, M. A. J. van Gerven, J. M. Schoffelen, Z. Ghahramani, T. Heskes, The dynamic beamformer, NIPS workshop on Machine Learning and Interpretation in Neuroimaging.

[114] M. Grosse-Wentrup, Understanding brain connectivity patterns during motor imagery for brain-computer interfacing, Conference on Advances in Neural Information Processing Systems (2009) 561–568.

[115] M. Mørup, K. H. Madsen, L. K. Hansen, Latent causal modelling of neuroimaging data, in: NIPS Workshop on Connectivity Inference in Neuroimaing, 2009.

[116] S. Haufe, R. Tomioka, G. Nolte, K.-R. Müller, M. Kawanabe, Modeling sparse connectivity between underlying brain sources for EEG/MEG, IEEE Transactions on Biomedical Engineering 57 (8) (2010) 1954 – 1963.

[117] L. Xu, P. Stoica, J. Li, S. L. Bressler, X. Shao, M. Ding, Aseo: A method for the simultaneous estimation of single-trial event-related potentials and ongoing brain activities, IEEE Transactions on Biomedical Engineering 56 (1) (2009) 111–121.

[118] K. K. Ang, C. Guan, C. Wang, K. S. Phua, A. H. G. Tan, Z. Y. Chin, Calibrating EEG-based motor imagery brain-computer interface from passive movement, 2011 Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, EMBC (2011) 4199–4202.

[119] T. Schneider, A. Neumaier, Algorithm 808: Arfit - a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models, ACM Transactions on Mathematical Software (TOMS) 6 (2001) 58–65.

[120] K. K. Ang, Z. Y. Chin, H. Zhang, C. Guan, Filter bank common spatial pattern (FBCSP) in brain-computer interface, IEEE International Joint Conference on Neural Networks and Computational Intelligence (2008) 2390–2397.

[121] C. Vidaurre, B. Blankertz, Towards a cure for BCI illiteracy, Brain Topography 23 (2) (2010) 194–198.

[122] N. Tomida, H. Higashi, T. Tanaka, A joint tensor diagonalization approach to active data selection for EEG classification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013), 2013, pp. 983–987.

[123] A. Cichocki, R. Zdunek, A. H. Phan, S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, Wiley, New York, 2009.

[124] W. Wu, Z. Chen, X. Gao, Y. Li, E. Brown, S. Gao, Probabilistic common spatial patterns for multichannel EEG analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence (2014) doi:10.1109/TPAMI.2014.2330598.

[125] K. M. M. Kawanabe, W. Samek, C. Vidaurre, Robust common spatial filters with a maxmin approach, Neural Computation 26 (2) (2014) 349–376.

[126] M. D. Plumbley, Geometrical methods for non-negative ica: Manifolds, lie groups and toral subalgebras, Neurocomputing 67 (2005) 161 – 197.

Appendix A

Appendix

A.1 Experiment Set-Up

EEGs from 27 channels were obtained using Nuamps EEG acquisition hardware with monopolar Ag/AgCl electrodes channels. The scalp map of the 27 channels being used is shown in Figure A.1. The sampling rate was 250 Hz with a resolution of 22 bits for the voltage range of ± 130 mV. A bandpass filter of 0.05 to 40 Hz was set in the acquisition hardware.

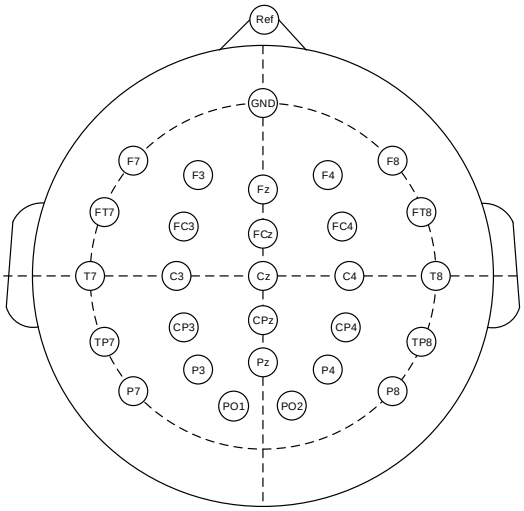


Figure A.1: Scalp map of the 27 channels.

The length of each trial was 12s, including 2s of preparatory segment, 4s of visual cue, and 6s of resting, which is shown in Figure A.2. During the

EEG recording process, the subjects were asked to avoid physical movement and minimize eye blinking.

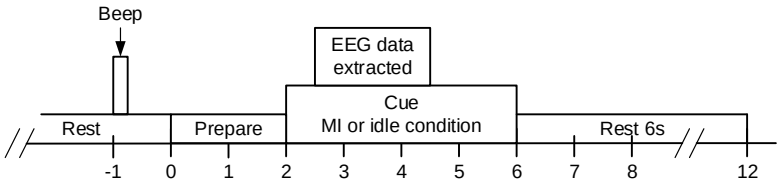


Figure A.2: Time segmentation of one trial.

A.2 Relations Between the Convolutive Model and the Instantaneous Model with Connected Sources

Based on the model in [116] and [111], $X(t)$ can be assumed to be generated as a linear instantaneous mixture of source signal $S(t)$, with the mixing matrix Φ_0, i.e.,

$$X(t) = \Phi_0 S(t) \tag{A.1}$$

Assume that $S(t)$ follows an MVAR model as below

$$S(t) = \sum_{\tau} B_s(\tau) S(t-\tau) + \epsilon(t) \tag{A.2}$$

where $B_s(\tau)$ is the coefficient matrix of the MVAR model and it represents the connectivity between sources [108, 109]. From (A.1), the innovation process $\epsilon(t)$ can be written as

$$\begin{aligned} \epsilon(t) &= \Phi_0^{-1} X(t) - \sum_{\tau} B_s(\tau) \Phi_0^{-1} X(t-\tau) \\ &= \sum_{\tau} \hat{B}_s(\tau) X(t-\tau) \end{aligned} \tag{A.3}$$

where

$$\hat{B}_s(\tau) = \begin{cases} \Phi_0^{-1}, & \tau = 0; \\ -B_s(\tau)\Phi_0^{-1}, & \tau > 0. \end{cases} \tag{A.4}$$

Equation (A.3) shows the equivalence between the MVAR model and the

convolutive model in [112, 115], with the innovation process $\epsilon(t)$ corresponding to the underlying convolutive sources. As the objective in [116] and [111] is connectivity analysis, the estimation of $B_s(\tau)$ and Φ_0 is based on the non-Gaussianity assumption of $\epsilon(t)$. In the proposed model, $S(t)$ represents the discriminative sources related to ERD/ERS, and thus the estimation of the FIR matrix $\hat{A}(\tau)$ in (3.10) and spatial filter $\mathbf{w}$ is based on maximizing the variance difference between the two classes. With the discriminative objective, it is preferable to apply the convolutive model to impose the variance difference as the prior information of the source. Moreover, since the two models are equivalent, it is also possible to build a discriminative model based on the instantaneous mixing model with connected sources in (A.1) and (A.2). In the future work, we would like to explore possible discriminative learning approaches to study the connectivity that contains class information.

A.3 Tensor-Related Notations and Basic Definitions

Definition 1. *Tensor: a tensor, also known as a Nth-order tensor, a multidimensional array, a N-way or a N-mode, is an element of the tensor product of N vector spaces, which is a higher-order generalization of a vector (first-order tensor) and a matrix (second-order tensor), denoted as $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, where N is the order of $\mathcal{A}$. An element of $\mathcal{A}$ is denoted by $a_{i_1,i_2,\ldots,i_N}$, $1 \leq i \leq I_n$, $n = 1, \ldots, N$.*

Definition 2. *Tensor Slice: a tensor slice is a two-dimensional section (fragment) of a tensor, obtained by fixing all indices except for two indices.*

Definition 3. *Unfolding: the n-mode unfolding of tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is denoted by $A_{(n)}$. More specifically, a tensor element $(i_1, i_2, \ldots, i_N)$ maps onto a matrix element (i_n, j), where*

$$
\begin{aligned}
j &= 1 + \sum_{p \neq n} (i_p - 1) J_p, \\
J_p &= \begin{cases} 1, & \text{if } p = 1 \text{ or} \\ & \text{if } p = 2 \text{ and } n = 1; \\ \prod_{m \neq n}^{p-1} I_m, & \text{otherwise.} \end{cases}
\end{aligned} \tag{A.5}
$$

Definition 4. *n-Mode Product: the n-mode product of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ and a matrix $U \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n U$, is a tensor in $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N}$*

given by

$$(\mathcal{A} \times_n U)_{i_1,i_2,...,i_{n-1},j_n,i_{n+1},...,i_N} = \sum_{i_n=1}^{I_n} a_{i_1,i_2,...,i_N}, u_{j_n,i_n} \tag{A.6}$$

Remark 1. *Given a tensor* $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_N}$*, and two matrices,* $F \in \mathbb{R}^{J_n \times I_n}$ *and* $G \in \mathbb{R}^{J_m \times I_m}$*, one has* $(\mathcal{A} \times_n F) \times_m G = (\mathcal{A} \times_m G) \times_n F = \mathcal{A} \times_n F \times_m G$*.*

Definition 5. *Khatri-Rao Product: For two matrices* $A = [a_1, a_2, ..., a_J] \in \mathbb{R}^{J_A \times J}$ *and* $B = [b_1, b_2, ..., b_J] \in \mathbb{R}^{J_B \times J}$ *with the same number of columns* J*, their Khatri-Rao product, denoted as* $\odot$*, performs the following operation:*

$$A \odot B = [\boldsymbol{vec}(b_1 a_1^T), ..., \boldsymbol{vec}(b_J a_J^T)] \in \mathbb{R}^{J_A J_B \times J} \tag{A.7}$$

Remark 2. *Given a tensor* $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_N}$ *and a sequence of matrices* $U^n \in \mathbb{R}^{I_n \times J_n}$*,* $n = 1, 2, ..., N$*, their multiplication* $\mathcal{A} \times_1 U^1 \times_2 U^2 ... \times_N U^N$ *satisfies*

$$\mathcal{A} \times_1 U^1 \times_2 U^2 ... \times_N U^N = U^n A_{(n)} [U^N \odot U^{N-1} ... U^{n+1} \odot U^{n-1} ... U^1] \tag{A.8}$$

A.4 Derivation of the Update Equations in Algorithm 3

Let $J_E = ||\mathcal{E}||_F^2$ and $E_{(3)}$ be the mode-3 unfolding of $\mathcal{E}$. Then, (5.3) becomes

$$E_{(3)} = R_{(3)} - \Lambda_d (V \odot V)^T \tag{A.9}$$

Substituting (A.9) into J_E, we have

$$J_E = \mathrm{tr}[R_{(3)} R_{(3)}^T - 2R_{(3)}(V \odot V)\Lambda_d^T + \Lambda_d (V \odot V)^T (V \odot V)\Lambda_d^T] \tag{A.10}$$

Differentiating (A.10) with respective to Λ_d^T, we obtain

$$\begin{aligned} \delta J_E &= \mathrm{tr}[-2R_{(3)}(V \odot V)\delta\Lambda_d^T + \delta\Lambda_d (V \odot V)^T (V \odot V)\Lambda_d^T \\ &\quad + \Lambda_d (V \odot V)^T (V \odot V)\delta\Lambda_d^T] \\ &= \mathrm{tr}[-2R_{(3)}(V \odot V)\delta\Lambda_d^T + 2\Lambda_d (V \odot V)^T (V \odot V)\delta\Lambda_d^T] \\ &= \mathrm{tr}[2(\Lambda_d (V \odot V)^T - R_{(3)})(V \odot V)\delta\Lambda_d^T] \end{aligned} \tag{A.11}$$

By setting $\delta J_E = 0$, we obtain

$$\Lambda_d = R_{(3)}\{(V \odot V)^T\}^\dagger \tag{A.12}$$

which is equivalent to (9) in Algorithm 3. Similarly, by substituting the mode-2 unfolding of $\mathcal{E}$ into J_E, we can obtain the update equation for V, i.e., (8) in Algorithm 3.

A.5 Comparison of Different "Flipping" Methods

As pointed out in [125, 77], the "flipping" method fails to capture relevant nonstationarity in certain cases, which is shown by the following example:

$$\begin{aligned} \bar{\Sigma}^{+} &= \begin{bmatrix} 0.9 & 0.15 \\ 0.15 & 0.1 \end{bmatrix}, \\ \Sigma^{+,1} &= \begin{bmatrix} 0.9 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}, \\ \Sigma^{+,2} &= \begin{bmatrix} 0.9 & 0.25 \\ 0.25 & 0.1 \end{bmatrix} \end{aligned} \tag{A.13}$$

Suppose that $\bar{\Sigma}^{+}$ is the average covariance matrix of class $+$, and $\Sigma^{+,1}$ and $\Sigma^{+,2}$ are covariance matrices of two trials. To extract the nonstationarity between trials, let $\Delta^i = \Sigma^i - \bar{\Sigma}$ and $\Delta^i \in \mathbb{R}^{M \times M}$. Then, the penalty matrix in sCSP with "flipping" is

$$\mathcal{F}(\Delta) = \frac{1}{2} \sum_{i=1}^{2} \mathcal{F}(\Sigma^{+,i} - \bar{\Sigma}^{+}) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \tag{A.14}$$

Thus, the nonstationarity of the off-diagonal elements cannot be penalized. To further investigate this problem, let the eigen-decomposition of Δ^i be

$$\Delta^i = U^i D^i {U^i}^T \tag{A.15}$$

where $U^i = [\mathbf{u}_1^i, \ldots, \mathbf{u}_M^i]$ are the eigenvectors and $D = \text{diag}(d_m)$, $m = 1, 2, ..., M$, is the diagonal matrix containing corresponding eigenvalues.

Then, the penalty term before "flipping" is

$$\begin{aligned} \mathbf{w}\Delta^i\mathbf{w}^T &= \mathbf{w}\left(\sum_{m=1}^{M} d_m^i \mathbf{u}_m^i (\mathbf{u}_m^i)^T\right)\mathbf{w}^T \\ &= \sum_{m=1}^{M} d_m^i \mathbf{w}\mathbf{u}_m^i (\mathbf{u}_m^i)^T \mathbf{w}^T \\ &= \sum_{m=1}^{M} d_m^i \sum_{p=1}^{M}\sum_{q=1}^{M} u_{mp}^i u_{mq}^i w_p w_q \end{aligned} \tag{A.16}$$

where u_{mp}^i or u_{mq}^i is the p-th or the q-th element in $\mathbf{u}_m^i$. The penalty term after "flipping" is

$$\mathbf{w}\mathcal{F}(\Delta^i)\mathbf{w}^T = \sum_{i=m}^{M} |d_m| \sum_{p=1}^{M}\sum_{q=1}^{M} u_{mp} u_{mq} w_p w_q \tag{A.17}$$

The reason why the "flipping" method fails to penalize relevant nonstatioary elements is that by only taking absolute value of eigenvalue d_m some coefficients $u_{mp}u_{mq}$ would cancel each other. In the example in (A.13), assume that $\Delta^1 = \Sigma^{+,1} - \bar{\Sigma}^+$ with $\Delta^1 = U^1 D^1 U^{1T}$, where

$$U^1 = \begin{bmatrix} -0.707 & -0.707 \\ -0.707 & 0.707 \end{bmatrix}, D^1 = \begin{bmatrix} -0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \tag{A.18}$$

Then, we have

$$\begin{aligned} \mathbf{w}\mathcal{F}(\Delta)\mathbf{w}^T = &| -0.1|(0.5w_1^2 + 0.1w_1w_2 + 0.5w_2^2) \\ &+|0.1|(0.5w_1^2 - 0.1w_1w_2 + 0.5w_2^2) \end{aligned} \tag{A.19}$$

where the coefficient of $w_1 w_2$ is 0 after taking absolute value of eigenvalues. To avoid this, $u_{ip}u_{iq}$ should be set to be positive if it is not, as below

$$\begin{aligned} \mathbf{w}\mathcal{F}^*(\Delta)\mathbf{w}^T &= \sum_{i=m}^{M} |d_m| \sum_{p=1}^{M}\sum_{q=1}^{M} |u_{mp}u_{mq}| w_p w_q \\ &\geq \mathbf{w}\mathcal{F}(\Delta)\mathbf{w}^T \\ &\geq |\mathbf{w}\Delta\mathbf{w}^T| \end{aligned} \tag{A.20}$$

which is equivalent to (5.13).

A.6 Rotation Matrix in 3D-Space

In 3D-space, the matrix for a rotation by an angle of θ about the axis in the direction of $u \in \mathbb{R}^{3\times1}$ is given by $R_t \in \mathbb{R}^{3\times1}$, i.e.,

$$R_t = \begin{bmatrix} R_{t,11} & R_{t,12} & R_{t,13} \\ R_{t,21} & R_{t,22} & R_{t,23} \\ R_{t,31} & R_{t,32} & R_{t,33} \end{bmatrix} \tag{A.21}$$

where

$$\begin{aligned}
&R_{t,11} = \cos\theta + u_1^2(1-\cos\theta), \quad R_{t,12} = u_1u_2(1-\cos\theta) - u_3\sin\theta, \\
&R_{t,13} = u_1u_3(1-\cos\theta) + u_2\sin\theta, \quad R_{t,23} = u_2u_1(1-\cos\theta) + u_3\sin\theta, \\
&R_{t,22} = \cos\theta + u_{2j}^2(1-\cos\theta), \quad R_{t,23} = u_2u_3(1-\cos\theta) - u_1\sin\theta, \\
&R_{t,31} = u_3u_1(1-\cos\theta) - u_2\sin\theta, \quad R_{t,32} = u_3u_2(1-\cos\theta) + u_1\sin\theta, \\
&R_{t,33} = \cos\theta + u_3^2(1-\cos\theta)
\end{aligned}$$

with $u_m,\ m \in \{1,2,3\}$ as the m-th element of u.