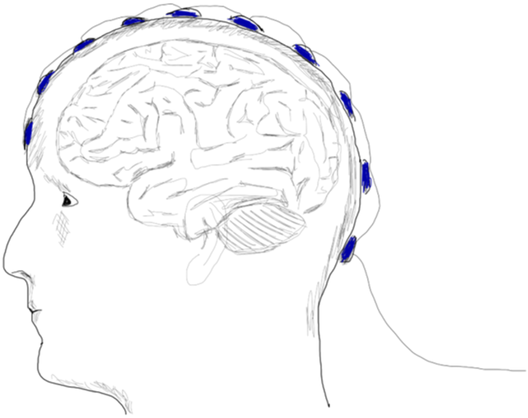
PhD Thesis

EEG Signal Processing in Motor Imagery Brain Computer Interfaces with Improved Covariance Estimators



Author:
F. Javier Olías Sánchez

Advisors:
Sergio Cruces
Rubén Martín Clemente

Teoría de la Señal y Comunicaciones
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

Sevilla, 2020

PhD Thesis

EEG Signal Processing in Motor Imagery Brain Computer Interfaces with Improved Covariance Estimators

Author:

F. Javier Olías Sánchez

Advisors:

Sergio Cruces
Rubén Martín Clemente

Teoría de la Señal y Comunicaciones
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

2020

Tesis Doctoral: EEG Signal Processing in Motor Imagery Brain Computer Interfaces with Improved Covariance Estimators

Autor: F. Javier Olías Sánchez
Tutores: Sergio Cruces, Rubén Martín Clemente

El tribunal nombrado para juzgar la Tesis arriba indicada, compuesto por los siguientes doctores:

Presidente:

Vocales:

Secretario:

acuerdan otorgarle la calificación de:

El Secretario del Tribunal

Fecha:

A mi familia.
A Carmen.

Agradecimientos

Me gustaría comenzar estos agradecimientos dándole las gracias a mi esposa Carmen por todos estos años a mi lado. Ella me ha apoyado emocionalmente, me ha cuidado, me ha mimado, me ha dado todo el amor que necesito y con ella crezco cada día…podría escribir otra tesis sobre todo lo que tengo que agradecerle. Pero, aun así, tengo que agradecerle especialmente la ayuda que me ha proporcionado en la redacción de este trabajo, ella también es ingeniera de telecomunicaciones, habla de forma nativa tanto español como inglés, y seguramente sin su ayuda esta tesis se habria escrito peor y en español. Por todo esto, gracias Carmen, este trabajo también es tuyo.

También quiero agradecerles a mis hermanas Macarena e Isabel todo su cariño y amor. Ellas son y serán mis compañeras en la vida, dándome siempre buenos momentos y muchas risas. Esta tesis no sería igual sin ellas. Además, durante estos años he contado con la ayuda de mis padres, Javier y Amparo, que me han guiado hasta aquí y me han dado la confianza y seguridad que necesito. Espero que se sientan orgullosos de mí.

No me quedaría tranquilo sin agradecer a mis suegros, Elena y Juan, que siempre me hayan tratado como a un hijo más y también por todos los tuppers de Elena que me han alimentado durante estos años.

Los fabulosos compañeros que he tenido en el departamento también se merecen un hueco en estos agradecimientos: Juan Antonio, Jose Carlos, Irene, Marta, Antonio, Paulina. Esta tesis también es vuestra porque de una manera u otra también habéis colaborado en ella.

Por último, me gustaría hacer mención a mis tutores Sergio y Rubén. Sin lugar a duda, este trabajo nunca se habría realizado sin ellos y solo puedo decir de ambos que son tutores ejemplares. Me gustaría agradecer especialmente a Sergio toda su dedicación y el tiempo que ha empleado en mí. Él me ha formado académicamente durante estos últimos años, enseñándome a ser minucioso y a prestar atención a los detalles, entre otras muchas cosas.

No puedo acabar estos agradecimientos sin expresar la gratitud que siento por haber tenido la oportunidad realizar esta tesis doctoral, la cual nunca habría ocurrido de no ser por la financiación que he recibido por parte del Ministerio de Ciencia e Innovación. Gracias, desde aquí, a todos los que hacen posible estas becas.

Javi,

Sevilla, 2020

Resumen

Desde hace unos años hasta la actualidad, el desarrollo en el campo de los interfaces cerebro ordenador ha ido aumentando. Este aumento viene motivado por una serie de factores distintos. A medida que aumenta el conocimiento acerca del cerebro humano y como funciona (del que aún se conoce relativamente poco), van surgiendo nuevos avances en los sistemas BCI que, a su vez, sirven de motivación para que se investigue más acerca de este órgano. Además, los sistemas BCI abren una puerta para que cualquier persona pueda interactuar con su entorno independientemente de la discapacidad física que pueda tener, simplemente haciendo uso de sus pensamientos.

Recientemente, la industria tecnológica ha comenzado a mostrar su interés por estos sistemas, motivados tanto por los avances con respecto a lo que conocemos del cerebro y como funciona, como por el uso constante que hacemos de la tecnología en la actualidad, ya sea a través de nuestros smartphones, tablets u ordenadores, entre otros muchos dispositivos. Esto motiva que compañías como Facebook inviertan en el desarrollo de sistemas BCI para que tanto personas sin discapacidad como aquellas que, si las tienen, puedan comunicarse con los móviles usando solo el cerebro.

El trabajo desarrollado en esta tesis se centra en los sistemas BCI basados en movimientos imaginarios. Esto significa que el usuario piensa en movimientos motores que son interpretados por un ordenador como comandos. Las señales cerebrales necesarias para traducir posteriormente a comandos se obtienen mediante un equipo de EEG que se coloca sobre el cuero cabelludo y que mide la actividad electromagnética producida por el cerebro. Trabajar con estas señales resulta complejo ya que son no estacionarias y, además, suelen estar muy contaminadas por ruido o artefactos.

Hemos abordado esta temática desde el punto de vista del procesado estadístico de la señal y mediante algoritmos de aprendizaje máquina. Para ello se ha descompuesto el sistema BCI en tres bloques: preprocesado de la señal, extracción de características y clasificación. Tras revisar el estado del arte de estos bloques, se ha resumido y adjuntado un conjunto de publicaciones que hemos realizado durante los últimos años, y en las cuales podemos encontrar las diferentes aportaciones que, desde nuestro punto de vista, mejoran cada uno de los bloques anteriormente mencionados. De manera muy resumida, para el bloque de preprocesado proponemos un método mediante el cual conseguimos normalizar las fuentes de las señales de EEG. Al igualar las fuentes efectivas conseguimos mejorar la estima de las matrices de covarianza. Con respecto al bloque de extracción de características, hemos conseguido extender el algoritmo CSP a casos no supervisados. Por último, en el bloque de clasificación también hemos conseguido realizar una separación de clases de manera no supervisada y, por otro lado, hemos observado una mejora cuando se regulariza el algoritmo LDA mediante un método específico para Gaussianas.

Abstract

The research and development in the field of Brain Computer Interfaces (BCI) has been growing during the last years, motivated by several factors. As the knowledge about how the human brain is and works (of which we still know very little) grows, new advances in BCI systems are emerging that, in turn, serve as motivation to do more research about this organ. In addition, BCI systems open a door for anyone to interact with their environment regardless of the physical disabilities they may have, by simply using their thoughts.

Recently, the technology industry has begun to show its interest in these systems, motivated both by the advances about what we know of the brain and how it works, and by the constant use we make of technology nowadays, whether it is by using our smartphones, tablets or computers, among many other devices. This motivates companies like Facebook to invest in the development of BCI systems so that people (with or without disabilities) can communicate with their devices using only their brain.

The work developed in this thesis focuses on BCI systems based on motor imagery movements. This means that the user thinks of certain motor movements that are interpreted by a computer as commands. The brain signals that we need to translate to commands are obtained by an EEG device that is placed on the scalp and measures the electromagnetic activity produced by the brain. Working with these signals is complex since they are non-stationary and, in addition, they are usually heavily contaminated by noise or artifacts.

We have approached this subject from the point of view of statistical signal processing and through machine learning algorithms. For this, the BCI system has been split into three blocks: preprocessing, feature extraction and classification. After reviewing the state of the art of these blocks, a set of publications that we have made in recent years has been summarized and attached. In these publications we can find the different contributions that, from our point of view, improve each one of the blocks previously mentioned. As a brief summary, for the preprocessing block we propose a method that lets us normalize the sources of the EEG signals. By equalizing the effective sources, we are able to improve the estimation of the covariance matrices. For the feature extraction block, we have managed to extend the CSP algorithm for unsupervised cases. Finally, in the classification block we have also managed to perform a separation of classes in an blind way and we have also observed an improvement when the LDA algorithm is regularized by a specific method for Gaussian distributions.

Short Contents

Contents

PART III: PUBLICATIONS

Notation

$\mathbb{R}$	Real Numbers.
$\|\|\mathbf{v}\|\|$	Norm of vector $\mathbf{v}$.
$\langle \mathbf{v}, \mathbf{w} \rangle$	Scalar product of the vectors $\mathbf{v}$ y $\mathbf{w}$.
$\det(\mathbf{A})$	Determinant of the square matrix $\mathbf{A}$.
$\mathbf{A}^\top$	Transpose of $\mathbf{A}$.
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$.
e	number e.
$\mathrm{P}(A)$	Probability of A.
$E[X]$	Expected value of the random variable X.
$var(X)$	Variance of the random variable X.
$\sim f_X(x)$	Distributed as the density of probability function $f_X(x)$.
$\mathscr{N}\left(m_X, \sigma_X^2\right)$	Gaussian distribution with mean m_X and variance σ_X^2.
$\mathbf{I}_n$	Identity matrix of dimension n.
$\mathrm{diag}(\mathbf{A})$	Diagonal vector of matrix $\mathbf{A}$.
$\mathrm{tr}(\mathbf{A})$	Trace operator of matrix $\mathbf{A}$. The trace operator is equal to the sum of the diagonal elements of $\mathbf{A}$.
$log(\mathbf{A})$	Logarithm of a SPD matrix $\mathbf{A}$. With $\mathbf{A}$ eigen decomposition such as $\mathbf{A} = \mathbf{U}\lambda\mathbf{U}^\top$ where $\mathbf{U}$ is the matrix containing the eigenvectors of $\mathbf{A}$ and λ the diagonal matrix containing its eigenvalues. The logarithm of $\mathbf{A}$ is defines as $log(\mathbf{A}) = \mathbf{U}\log(\lambda)\mathbf{U}^\top$. The logarithm of the determinant can be defines as: $log\|\mathbf{A}\| = \mathrm{tr}(\log(\lambda))$.
$\mathbb{E}[x]$	expectation operator x.

Specific definitions

$\mathbf{x}(t)$	EEG time signal. It can be combined with subscripts to indicate a certain trial or class. For example the trials τ from class k is represented by $x_{\tau,k}(t)$.
$\mathbf{\Sigma}$	Reference to a sample covariance matrix. When it is used with a subscript, it indicates the signal from which the covariance is estimated, i.e $\mathbf{\Sigma}_{\mathbf{X}}$ refers to the covariance of the signal $\mathbf{X}$ and if it is conditioned to the class k it is indicated as $\mathbf{\Sigma}_{\mathbf{X}\|k}$.
N_τ	Number of training trials. With the notation $N_{\tau\|k}$ we make reference to the number of training trials of the class k.
$\mathbf{S}$	Brain sources of an EEG signal.
$\mathbf{N}$	Additive Gaussian noise.
$\mathbf{A}$	Sources mixing matrix.
$\mathbf{X}$	Matrix form of $\mathbf{x}(t)$.
τ	Variable to indicate a generic training trial.
k	Variable to indicate a generic imagery movement class or label.
K	Number of imagery movement classes.
$\mathbf{u}$	Variable to refer to an eigenvector.
$\mathbf{U}$	Matrix of eigenvectors.
λ	Variable to refer to an eigenvalue.
$\mathbf{\Lambda}$	Diagonal matrix that holds a set of eigenvalues.
$\mathbf{w}$	Singular spatial filter. In general terms we refer with this variable to the spatial filter computed by the CSP algorithm or any of its variants.
$\mathbf{W}$	Matrix formed by spatial filters.
P	Number of spatial filters.
N_s	Number of EEG sensors.
T	Number of time samples in a trial.
K	Number of classes.
$\mathbf{y}$	Random variable which represent the set of features. To make reference to a set of features from the class k we use $\mathbf{y}_k$ and to reference a trial τ we use $\mathbf{y}_\tau$. It is usually used as a row vector.
$\boldsymbol{\mu}$	Mean vector of a random distribution that may be indicated by the subscript. For example $\boldsymbol{\mu}_{\mathbf{y}}$ represents the joint mean of the features $\mathbf{y}$.
$\mathbf{C}$	Variable to make reference to a true covariance covariance from a distribution. Only used in Chapter 6.

[*] There are variables that have been omitted from this notation table due to their sporadic use.

PART I

INTRODUCTION AND OBJECTIVES

Chapter 1

Motivation and introduction to the topic

This thesis started four years ago with the motivation of investigating and improving the Brain Computer Interface (BCI). One reason to choose this topic was to learn more about the brain, the organ that move us, but the main reason is the variety of possibilities that these systems enable.

The BCI systems are a group of communication systems through which a person is capable of transmitting information directly from her/his brain to a computer or machine without using any muscle. The information is usually in the form of commands that can allow the users to interact with their environment by just using their thoughts.

The development of the BCI systems is motivated by many factors but one of the most important or useful reasons is that these systems constitute a great tool for people with motor disabilities. The BCI systems open a new door with a set of opportunities for people with these kind of disabilities, allowing them for example, to move around with a wheel chair [1] or to control a computer [2] just through their thoughts, as we said before. Since the BCI systems enable a communication channel to the brain, the user does not need any other motor capability to interact with the system. In addition, BCI systems can also be used to help in rehabilitation processes [3], providing feedback about motor tasks or to understand how brain injuries are affecting the patient.

Not only can we find these systems useful in the medical field, the consumer industry is also starting to show interest in BCI systems. Some of the most notable examples of the development of BCI systems in applications for day to day consumers are being done by the company Neuralink, whose founder is Elon Musk, also the founder of the company Tesla. Neuralink is a company which objective is to build brain implants, for both medical uses and to give people new assistive powers in the form of a digital layer above the cortex. Another example that shows commercial purposes in the field of BCI is the Facebook BCI program which goal is "to build a non-invasive, wearable device that allows people to type by simply imagining themselves talking" [4].

As we can see the BCI systems are an active and top of the art research field and they count with a long history of almost half a century of development starting with the pioneer Jacques J. Vidal in 1973. In his article [5] he proposed to access the humans brain electrical signals to control computers, prosthetic devices and even spaceships. Since then, there has been a huge development in the scientific research field that he started and nowadays it is possible for people to control robotics arms [6] or artificial legs [7] using a brain implant. It is also possible to spell words [8] and there are even video-games based on BCI systems [9]. However, we are still far from being able to pilot a spaceship since BCI systems only understand a few set of commands and still have an important error rate that makes them unsuitable for tasks that require a high level of precision.

During all these years of development the general schema of these system have stayed the same. As exposed in the book [10], in the BCI systems we can find four parts as we describe below:

1. **Signal acquisition system**: Its objective is to measure and provide the next part of the system with the electromagnetic activity that the brain generates and which will control the system. There are two key aspects required from a good acquisition system: it is important to keep as much spatial resolution as possible while keeping the lowest signal to noise ratio possible. To measure the electromagnetic field of the brain the most popular technique is to place electrodes over the scalp, this is what is known as Electroencephalogram (EEG). There are also other acquisition system as for example the ECoG, which consists of placing electrodes over the brain and inside of the skull; or using brain scans provided by fMRI techniques.

2. **Signal processing module**: After acquiring the signals corresponding to different commands, the machine needs to be able to distinguish between them. This part of the system is the one in charge of extracting and classifying the features within the brain signals that will let us make a difference between the commands.

3. **Output device**: This is the part that we desire to control through the BCI system. Such as a computer screen, a wheel chair, a robotic arm or any other controlled device.

4. **Operating protocol**: The brain is a very complex organ, and to control the electrical signals it produces is a difficult task for the user. For this reason, the user usually needs training to learn how to generate the right stimuli for the system. There are external systems that help us produce or control the brain electrical signals by using visual, audio or any other stimuli that induces a waveform or activity in the brain. Nevertheless, the goal of a BCI system is that the user controls it willingly with their own thoughts. There are a few ways of controlling the brain electrical signal with the thoughts. We can control the activity of the brain by concentrating in our thoughts and their frecuency (for example, leaving the mind blank or thinking rapidly about something), or as we will see in more depth in this thesis, by thinking in different parts of the body which would generate electromagnetic activity in different parts of the brain. It is important to note that the different ways of controlling

the system (different methods of generating brain activity) establish the different BCI paradigms and define the whole system.

1.1 Overview on BCI

As it has been just said, the chosen operating protocol will define the architecture of the rest of the system. In that sense, there are several kinds of BCI systems that are based in different physiological processes of the brain. Researchers try to find patterns in the electrophysiology of the brain that can be measured and at the same time can be voluntary triggered by the users intention. For example, when you think about moving the right hand, the electromagnetic waves of the brain in the left hemisphere change. This event can be intentionally triggered by the user and can be measured by the EEG acquisition system. The decision of choosing a particular kind of system to control a device can depend of the device itself or the users limitations, making it necessary to design specific solutions for specific problems. For example, if the user is blind, he will not be able to use a screen, but he could use speakers or if the user is not able to write or read, he will not be able or it will be very difficult for him or her to use a speller machine. In the following sections we will review some of the most popular BCI paradigms.

1.1.1 P300 Speller

A very popular BCI system is the speller based on the P300 wave. This system is based in the brain activity that occurs when a person does not know when an event is going to happen but he or she is actively waiting for it. An example of the occurrence of this wave can be found in athlete brains when they wait for the shot or signal that indicates the start of the race. Under those circumstances and approximately 300 milliseconds after the occurrence of the event, the brain produces a powerful wave that can be measured with a EEG headset and which name is P300. The reasoning behind why the brain produces this wave is not yet very clear and it exists some controversy about what the explanation might be. For example, in [11] it is suggested that the goal of this wave is to improve decision making after the event happens, although there are other studies that suggest that this wave might occur to reset the neural system [12] as a preparation for taking action.

The P300 speller was first proposed in [13] and it consists on presenting a board or screen with the alphabet in form of a table as represented in figure 1.1. The user concentrates in a character, while rows and columns are randomly highlighted. When the row or column that contains the letter the user is focusing in is highlighted, the P300 wave is triggered. To determine that the person was thinking about a specific character, its row or column would have had to be highlighted in several occasions. This is necessary because a row or a column contains several characters and also to be able to make an average of responses for each character. This system is still being investigated and improved every year. For the interested reader, we point out to the following book chapter [14] where there is more information about this BCI system.

Choose one letter or command

A	G	M	S	Y	*
B	H	N	T	Z	*
C	I	O	U	*	TALK
D	J	P	V	FLN	SPAC
E	K	Q	W	*	BKSP
F	L	R	X	SPL	QUIT

Figure 1.1 Representation of the speller display used in [13].

1.1.2 Steady State Visual Evoked Potential

As the knowledge about the human brain increases, new BCI paradigms emerge. In 1982 it was discovered the Steady State Visual Evoked Potential (SSVEP) in monkeys [15]. The SSVEP refers to a electrophysiological phenomenon that occurs in the visual cortex of the brain. When a person looks to a flicking light in the range of 1-100Hz [16], that same frequency can be observed and detected in electromagnetic waves at the visual cortex. In the BCI system that Chen *et al.* presented in 2002 [17], they were able to estimate the telephone number that 8 of the 13 participants were trying to transmit by looking at the flicking panel. The SSVEP BCIs have kept evolving and nowadays by using these systems it is possible to drive a control remote car, a drone, or even a virtual vacuum [18, 19, 20].

1.1.3 Motor imagery Brain Computer Interface

The motor imagery movements paradigm consist in transmitting commands to the computer through the imagination of motor movements, being the most extended practice to transmit a set of two commands, one for each one of the hands.

the topic of this thesis is about this paradigm and we dive into the motor imagery paradigm in section 1.4.

1.2 Overview on human brain

To understand the motor imagery BCI system that this thesis focuses on, it is important to have basic notion of how the human brain works. Our brain can be seen as a distributed system in which each part of the brain is in charge of certain functions and depending on which BCI system is going to be used, different parts of the brain will be monitored.

We can divide the human brain in different areas, but to keep it simple and for the purpose of this thesis, we can start speaking of hemispheres. We can differentiate two hemispheres in the brain: the right and the left one. The right hemisphere controls the left side of the body and it is associated with mathematical and scientific thinking, sequential processing of information and it houses the abilities to write and speak. The left hemisphere controls the right side of the body and is associated with imagination, creativity, spatial orientation, emotions, intuition and multitasking. This said, we can see that each hemisphere controls the opposite part of the body.

At the same time, the brain is divided in four lobes. Each one of them holds symmetric parts of the two hemispheres:

- **Frontal lobe**: this is the lobe that manages action, controlling our behavior and it houses our creative abilities as well as our problem solving capacities. The frontal lobe is the biggest one and almost at the border with the parietal lobe, we can find the motor cortex. The motor cortex is the part which controls the motor movements and therefore, the part that we have to monitor in MI-BCI systems. The frontal lobe also plays an important role in the P300 speller BCI systems because it is one of the areas where the P300 wave is generated [21].
- **Parietal lobe**: this part of the brain is the one in charge of interpreting the reality we live in. It is where the language, reading and vision is processed. At this side of the border with the frontal lobe, the somatosensory cortex can be found. While the motor cortex is responsible for moving the muscles of our body, the somatosensory cortex is the part where the feelings about our bodies are processed, housing the pain and touch senses.
- **Occipital lobe**: this is the lobe where the information acquired with the eyes is projected and later distributed to other parts of the brain. This region is of interest for the BCI systems based on the paradigm SSVEP and in these systems they monitor the activity of this region.
- **Temporal lobe**: this is the part that holds the memories and also helps in speech and hearing tasks.

In figure 1.2 you can find a graphic representation of the four lobes and the most interesting cortices for the different BCI systems [22].

1.3 The EEG

As commented before there are others signal acquisition methods but in any case, the Electroencephalogram (EEG) is a very popular acquisition method for BCI systems [23]. The EEG method consists in placing electrodes over the scalp of the user to measure the electromagnetic field produced by the brain. Some of the reasons of its popularity reside in the price and its practicality: the equipment needed is very cheap in comparison with other acquisition methods, it can be carried around easily and no surgery is necessary for its usage. For those reasons the EEG is a great candidate for non-intrusive BCI systems. Nevertheless, we can also find a wide variety of EEG headsets with different prices, shapes and quality, some of them more portable while others require more equipment.

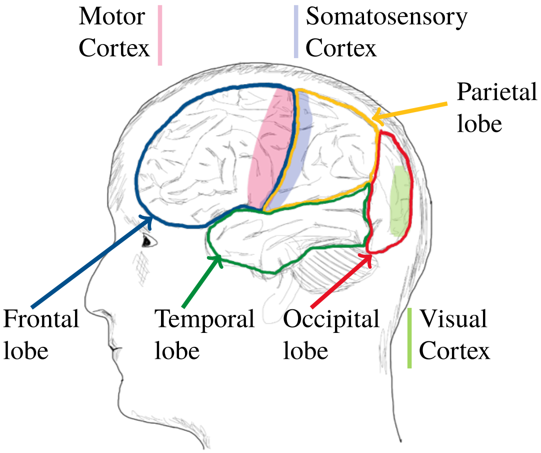


Figure 1.2 Representation of some parts of the human brain.

The electromagnetic field that the EEG headset measures is generated through the activity of the neurons. The neurons are specialized cells that process and transmit information. The human brain contains in average 86 billion neurons [24] that are connected one to another by their axons. The axons are long fibers that transmit electrical signals, called action potentials, and they work in an all-or-nothing way, resulting in a binary communication. Our brains are always working and the millions of action potentials that are constantly being fired up produce an electromagnetic field that can be measured. Ideally, the goal is to measure and locate the electromagnetic activity inside our brain with as much spatial resolution as possible, so we could know exactly which part of the brain is active. For example, the Electrocorticography (ECoG) measures this activity directly over the brain, placing the electrodes inside the subject's head using surgery and obtaining a very good signal quality. In the case of EEG, the signal is acquired from the surface of the scalp, avoiding surgery and with relatively cheap equipment.

The electrophysiological signals are very dependent of the patient metal state[25], they are very noisy and the spatial resolution is decremented, making it very difficult to work with. In addition, the EEG signals are usually contaminated with artifacts that also decreases the quality of the signal [26].

The artifact contamination is one of the most common contaminations that the EEG signal suffers, and we usually refer to this kind of contamination as the one produced by the subjects movements or actions. The most typical kind of artifacts in the EEG signals are the ones produced by blinking, swallowing or breathing. All those actions can have a major impact over the EEG signals and because of that there is a lot of research in the topic of avoiding them [27].

The EEG sensors can be dry or wet, which means that they might need a gel to adapt the impedance between the scalp and the electrode itself. Dry electrodes are relatively new and they perform as good as the wet ones in terms of Signal to Noise Ratio (SNR)

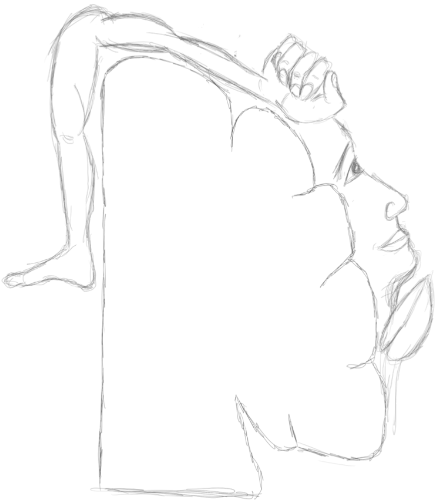


Figure 1.3 Representation of the mapping between the body and the motor cortex. Enumerating the body parts in a clockwise direction: Foot, leg, arm, hand, face and tongue.

but they are more sensible to movements or little variations [28]. Placing the electrodes in the same exact position in different sessions or in different users, is likely to be impossible and this means that usually the signals between sessions are very different. This is another handicap of the EEG, making seeking for spatial information and working across sessions extra hard. On top of that, due to little movements or the drying of the gel, the signal is non stationary within the same session.

We can conclude as a summary, that despite the advantages of the EEG for the BCI systems, the signals extracted with this method are very noisy, non-stationary, they are usually affected by artifacts perturbations and extrapolating the results across users or sessions is very difficult. All these reasons contribute to why the BCI systems based on EEG present such a challenge for signal processing research.

1.4 Introduction to Motor Imagery Brain Computer Interfaces

In the previous section we have introduced the Motor Imagery Brain Computer Interface (MI-BCI). The commands that are used in these systems are generated by the imagination of motor movements in a way that the movement of each part of the body is associated with a different command.

1.4.1 Physiological aspects of a MI-BCI

To differentiate between the user commands (imagery movements), the MI-BCI systems are based in two aspects of the motor cortex physiology:

- The first one is related to what we saw in the previous section 1.2: the brain works as a spatially distributed system that is divided in different regions and one of them is the motor cortex. This is the part of the brain that controls the voluntary movements of the body and, as the rest of the brain, it is spatially distributed in the sense that each part of the motor cortex controls a different part of the body. The motor cortex extends through both hemispheres of the brain: the part of the motor cortex which is in the right hemisphere controls the left side of the body and vice-versa. Apart from the hemispheres, it is difficult to establish the exact boundaries of the different parts within the motor cortex. However, in the figure 1.3 there is a sketch representation of the mapping between the body and the motor cortex.
- The second aspect that the MI-BCI systems rely on is the sensorimotor rhythms, that are a kind of waves that occur in the sensorimotor cortex [29] (that is how we refer to the somatosensory and motor cortex at the same time). This rhythms are concentrated in a bandwidth between 8-30Hz and the origin is not yet clear. They become of interest because when a person makes a motor movement or just imagines it, the sensorimotor rhythms become weaker and they could disappear completely from the sensorimotor cortex parts that are involved in the movement. This phenomenon is called Event Related with the Desynchronization (ERD). Similar to ERD, we can find another phenomenon that occurs in the opposite case. When there is a power increase in those rhythms, the phenomenon is called an Event Related with the Synchronization (ERS). The ERD and ERS can be detected in the EEG signals, and therefore, it is possible to discriminate between different motor movements by looking at the patterns in the EEG signal [30].

Finding the patterns that best discriminates between the imagery movements in the EEG signal becomes a very hard task and this thesis is dedicated to that purpose.

1.4.2 MI-BCI from the user point of view

In this section we talk about a case of use of a MI-BCI system and although there are many paradigms and variants, we will refer to the general case where the output device is a screen, there is a training phase in each session and the user tries to transmit commands that will be shown in the screen as feedback.

To use the MI-BCI system the headset is installed in the subject's head while the person is usually sitting down in a comfortable position where he or she can be still during the period of time that the system is in use. Each time a subject makes use of the system will be called a session.

MI-BCI are very dependent of the users and sensors position and because of that, in the general case, the system has to be trained every time it is used. The headset and position of sensors can not change within a session because, as we said before, these systems are based in spatial information and can be seriously affected by little variations in the position or behavior of the sensors.

To perform a command the user imagines a motor movement during a short period of time, usually between 3 and 6 seconds. After that, there is another short period of time during which the user can rest before performing the next command. Each time the user

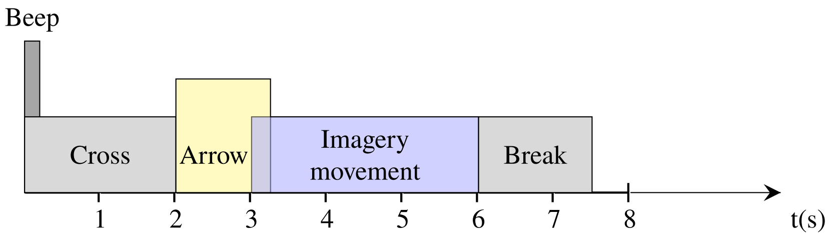


Figure 1.4 Time sequence representation of a trial from the BCI Competition IV dataset 2a.

performs a command, is called a trial and each trial has the label of the movement that was imagined.

The training phase consists in the acquisition of labeled trials that the computer will use to learn how to discriminate between the commands. Hence, the user is told what movement to think of in each one of the trials.

For the development of new variants or algorithms that improve the user experience we need some kind of objective measurement that enables the comparison between different proposals. The main measurement that is used is the percentage of trials in which the label was correctly estimated, although there are other measurements that are taken into account. For example, the required time needed to obtain a good performance (training phase time) or the complexity of the algorithms are important aspect of a MI-BCI system. Using the BCI system and evaluating its accuracy at the same time becomes a problem because while the system is in the usage phase, the user thinks of free will movements which label is unknown to the system or lost. Therefore, for researching purposes the user is always told which movement to perform and the usage phase is omitted in order to get more labeled trials to experiment with. Once a session has been recorded and labeled, there are many ways to split the data in training and test. For example, if we are running an online test, the first group of trials are used for training and the rest for test but it is usual to work with offline data and to perform what is called a Cross Validation (CV). The CV is technique which is used to select the correct parameter configuration of an algorithm and it consists in separating the training set into smaller sets to train and test the algorithm using the different parameter configurations and to select the configuration that perform the best over the training set.

An example of MI-BCI data that is used in this thesis and in many other researching works is the BCI competition IV dataset 2a [31]. This is a set of EEG recordings in which the trials follow the schema in the figure 1.4. Each trial starts with a beep to call the attention of the user, a cross appears in the screen to catch the users eye and after two seconds an arrow appears that tells the user which command to perform or think of. After four seconds the arrow disappears and the user rests.

Chapter 2

Objectives and structure of the thesis

2.1 Objectives

In a very general way, the objective of this thesis is to contribute to the Motor Imagery Brain Computer Interface field by applying and developing signal processing techniques that will increase the accuracy rate of the predictions, reduce the training period for each user and improve the overall experience when using these systems. To achieve this general goal, we have accomplished the following objectives which are indicated below:

- Study of the physiological aspect of the human brain that makes possible the existence of the MI-BCI systems and the timeline that is followed in the MI-BCI set up. By doing this we will gain the necessary knowledge and background to understand the experiments and the nature of the collected data.
- Review and research on the signal processing and machine learning algorithms that are currently being used in the MI-BCI field. The knowledge about these techniques and about their pros and cons will help in the development of new techniques.
- Investigate about proposals which are not framed in the MI-BCI field to find those that can be used or applied to these systems in order to improve them.
- Implement and use the state of the art techniques to gain experience in these systems and to be able to compare new possible proposals with the previous ones.
- Develop new proposals or modifications to existing algorithms that improve the current ones.

- Analyze and compare different cases scenarios of these new proposals to detect their disadvantages and perks.

2.2 Structure

This thesis is organized in three different parts, each containing chapters that offer a deeper understanding of the subject that is being studied.

In Part I we introduce the reader to the topic and justify the choice of it, as shown in the previous chapter. There, we introduced the reader to Brain Computer Interfaces by going over some BCI paradigms and the acquisition signal system. At last, we have introduced the reader to the motor imagery paradigm, establishing the basic framework to understand how they work. We then arrive to the current chapter, where we present the objectives of this thesis and review its structure.

Part II is divided in four different chapters, its objective is to review the state of the art in the field of MI-BCI, which we think is necessary to understand and contrast the contributions of our work. In figure 2.1 we represent the classic block diagram of a machine learning architecture, divided in: preprocessing, feature extraction and classification. We have chosen to apply this block diagram in this context since the MI-BCI system can also be considered a machine learning solution [32]. This decomposition is also useful to study the different blocks independently and, in practice, keeping these parts separated in the system lets us modify and troubleshoot each block by itself, without interfering with the rest of the system.

The different blocks represented in figure 2.1 are the study purpose of the first three chapters of Part II, that is, Chapter 3, Chapter 4 and Chapter 5. It is important to note that although the preprocessing part in a MI-BCI system is sometimes omitted and included as part of the feature extraction block (as for example in [33]), we have decided to dedicate a whole chapter to it since one of our most relevant contributions concerns to this part of the system. Therefore, Part II starts in Chapter 3 by studying the preprocessing block and also looking into more aspects about the EEG signal, like its frequency range and how it is transformed. In addition, we comment the problems that the artifacts pose. In Chapter 4 we go over the feature extraction block, explaining the CSP algorithm which is the most popular technique in the field of MI-BCI among some of its variants. We finish the chapter reviewing the Riemann framework approach, which has become more popular recently. In Chapter 5 we go through some of the state-of-the-art classification algorithms. In the last chapter of this part, Chapter 6, we study the covariances estimators which are used in the three previous chapters.

Part III is reserved to hold the publications that we have worked on during the PhD studies have taken place. Each publication is accompanied by a summary where we explicitly comment its contributions to the field.

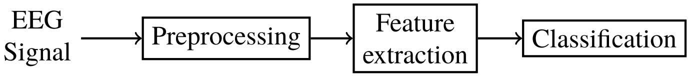


Figure 2.1 Block diagram of MI-BCI system.

In this document one can find a table with the general and specific notation that has been adopted, a list of figures and tables to help the reader navigate the document and a glossary with the acronyms that have been used throughout the text.

PART II

PRESENTATION AND DISCUSSION OF THE STATE-OF-THE-ART

Chapter 3

Preprocessing

As we have seen in section 2.2, the signal processing module can be divided in different blocks that can operate independently, being the first one of them the preprocessing of the EEG signal. Although we also saw that this block is sometimes overlooked by authors, we decided to dedicate a whole chapter of this thesis to it because it is a fundamental part in every machine learning system and we consider important to understand the basic concepts of this block. Also, we have developed some proposals that improve the preprocessing of the signal, making the whole MI-BCI system better. These proposals can be found in Publication A. In the following section we will introduce the basics on how the preprocessing of the EEG signal is done, and then we will center our attention on the artifacts and how to minimize their effect on the EEG signal.

3.1 Preprocessing of the signals

As explained in the previous section 1.4.1, the electromagnetic field that exists in the surface of our scalps is produced by the activity of the neurons that transmit electrical signals from one to another. We have billions of neurons that are constantly firing action potentials and each one of the electrical signals that are transmitted generate an electromagnetic field.

These electromagnetic fields are propagated through the brain and through the skull to the scalp where the headset measures the EEG signal. However, through the scalp we are only capable of measuring an overall activity of the brain and not the concrete activity of the different parts of the brain we saw in 1.2. These electrical signals proceding from the different parts of the brain are all mixed together, transformed through the propagation to the scalp and covered with noise and artifacts. So, despite all the sensors that are placed around the scalp, it becomes very difficult to identify the activity produced by the different parts of the brain.

Nevertheless, some spatial information is preserved and in that sense, the ERS and ERD that we introduced in section 1.4 are strong and synchronized waves that can represent important changes in the EEG signals [34]. Using the techniques that are developed within the BCI systems we can even estimate the location of those waves across a set of trials [30]. However, even though the occurrence of the ERD and ERS are the ones making possible that the MI-BCI systems work, finding these waves in each trial is not a reliable solution. Instead, we look for the variations that those waves produce on the EEG signals, without explicitly looking for them. This said, we can conclude saying that the main goal of the preprocessing block is to clean and adapt the signal in such a way that it becomes easier to find and extract the variations that we are looking for.

The headset provides us with a continuous stream of data per session and each session can last for more than an hour but as we have already commented in section 1.4.2, we will work with trials. This leads us to the first task of the preprocessing block, which is to cut the discrete signal provided by the headset into pieces containing each one of the trials. In the BCI competitions, as in others MI-BCI software acquisition, check [35] for example, we are provided with a sequence of events that mark the start of a trial among other events. Those marks may be used to extract the trials starting from the marked sample until the new event mark. It is also usual to discard the first part from each trial [36], because the ERD starts around 0.5 to 1 second after the user is told to start thinking of a certain movement [37]. In the experiments of the articles within this thesis we have discarded the first 0.5 seconds of each trial and we have worked with the following two seconds.

To prepare the EEG signals, one of the first steps is to filter them around the frequency band where the ERD and ERS occurs. In the reference [34] they establish the frequency range between 14 and 24 Hz, but for the MI-BCI systems, the range 8-30Hz is almost a standard [30, 33, 38]. There are some implementations of MI-BCI systems that do not use this frequency filtering right away, as for example the Filter Bank Common Spatial Patterns (FBCSP) or discriminative filter bank CSP [39, 40]. Those designs look for features that are in narrow frequency bands and that are covered with noise in wider frequency ranges. Those systems can improve the results obtained with a unique frequency band and in fact the FBCSP algorithm already counts with 4 patent citations.

However, the FBCSP design is difficult to reproduce, the training process is hard and time consuming and we can see that they do not always provide with better results than CSP [41, 40]. For these reasons we are not going to experiment with it and in the following, we consider that the EEG signals that the headset provide ($\mathbf{x}(t)$) are from a unique frequency band that covers the range 8-30Hz, using an eighth-order Butterworth filter:

$$\mathbf{x}(t) \leftarrow \underset{8-30Hz}{Band\ Pass\ Filter}\ \mathbf{x}(t). \tag{3.1}$$

In addition, we think it is worth noting that the acquisition system had a 50Hz notch filter, and that in our implementations, we filter the whole session before extracting the trials.

In classical approaches, the task of the preprocessing block would end here. Nevertheless, we will see in the following chapters that usually, in MI-BCI systems, all that we need for the next blocks are the covariances of the trials. Because of that, it makes sense to compute the covariances in the preprocessing block, offloading work and data volume of the rest of blocks:

$$\boldsymbol{\Sigma}_\tau = \frac{1}{T}\mathbf{x}_\tau(t)\mathbf{x}_\tau^\top(t), \quad \forall \tau \in N_\tau. \tag{3.2}$$

Where $\boldsymbol{\Sigma}_\tau$ represents the covariance of the training trial τ, T is the number of samples in each trial and N_τ is the number of training trials. We will see in further chapters that the covariances play a very important role in the field of MI-BCIs and we dedicate Chapter 6 to their estimators.

3.2 Dealing with the presence of artifacts

In the biomedical engineering field the term artifact is used to reference an undesired alteration in measurement that is produced by the patient or subject. In that sense the EEG signals are very sensitive to artifacts because while we are just interested in the electromagnetic field generated by the brain, any kind of movement of the subject can produce an alteration in the signal. Some typical artifacts are produced by blinking or other eye movements, which produce changes in the EEG signal amplitude that may be many times greater than the ones produced by brain activity [42]. Other common artifacts are produced by muscle movements, like swallowing or breathing.

As we concluded in the previous section, the MI-BCI systems use as main measure the covariance of the trials. In that sense, the artifacts can produce very important changes in the covariances of the trials, which in turn affect how the feature extraction and classification algorithms work, as they are trained with those covariances. Nevertheless, artifacts are not just a problem of the MI-BCI system but of any medical or BCI system based on EEG [42]. In fact, the artifact removal problem in the EEG signals is an active research field (a search on Scopus of works containing the keywords "EEG" and "artifact" generates 113 results during 2018).

Most of artifact removal algorithms can be categorized as preprocessing techniques that are applied before they can affect the system. There are tools for selecting and rejecting artifacts by visual inspection [43], but there are also many automatic algorithms, where the variants of the Independent Component Analysis (ICA) and Principal Component Analysis (PCA) algorithms are very popular [44][45]. Once a sample is labeled as an artifact, it is usual to erase it although there also techniques to reconstruct the signal [46].

In any case, it is also usual to ignore artifacts and instead use protection techniques that reduce the artifact effects as for example normalization techniques obtaining good results. For instance, the winning solution of a Kaggle competition in 2016 did not use any artifact rejection technique [47].

From this perspective, one popular option is to normalize the covariance matrices [48]:

$$\boldsymbol{\Sigma}_\tau \quad \leftarrow \quad \frac{\boldsymbol{\Sigma}_\tau}{\operatorname{tr}(\boldsymbol{\Sigma}_\tau)}, \quad \forall \tau \in N_\tau. \tag{3.3}$$

By doing this we equalize the power of all the trials and, at the same time, we get rid of the excess of power in the contaminated trials, although it does not dissolve all the effects caused by the artifacts.

3.3 Discussion

In this chapter we have seen the basic concepts of how the preprocessing of the signal block works and exposed the problems that artifacts cause in these signals. The importance of these problems explains why most of the development that has been done during these past years in the preprocessing field is related to palliate the effects of the artifacts.

On another note, and as a summary and guide to the practical user, the three basic concepts that can be extracted from this chapter related to the preprocessing block are the following:

1. The recommendation is to filter the whole EEG signal between 8-30Hz.
2. After filtering the signal, one can segment and extract the target trials.
3. The final step of the preprocessing is to compute the normalized covariance of the trials.

Chapter 4

Feature Extraction

The following chapter is dedicated to the study of the different methods and techniques that are used to transform the covariance matrices of the trials into a set of features that are more suitable to use for classification algorithms.

In the literature, this part of the MI-BCI system is more commonly known as Spatial Filtering because the most popular technique used to be the CSP algorithm, which performs a spatial filtering of the EEG signals and provides with a metric to choose the subspace projection which contains more class related information. In addition, other techniques that are used in this context such as ICA [49] would also perform a spatial filtering.

Nowadays, a new set of tools has been proposed and it has shown to be more flexible and powerful than CSP. The tangent space projection technique does not need to realize a spatial filtering of the signals. Instead, it projects the covariance matrices of the trials in a local Euclidean space which is tangent to the covariances space. In spite of the differences between the different techniques, they all have in common one thing: their goal is to extract a set of features that will be passed on to the classifiers.

This chapter will start with a deeper study of the MI-BCI systems, establishing the mathematical model on which the majority of feature extraction algorithms are based. After that, we explain and review the CSP algorithm and its variant for the multi-class paradigm, among others. Lastly, we will move on to the study of Riemann geometry techniques and tangent space projection.

4.1 Mathematical model of the observations in MI-BCI

We have already introduced some aspects about the MI-BCI systems in both Chapter 1 and Chapter 3, whereas in this section we explain the model that is used to process the EEG signals and transform them in commands.

In section 1.4.1 we explained that the electrical signals captured by the EEG headset come from the electrical impulses that are produced by the neurons in our brains. Taking into account that there are millions of neurons in our brains, to consider the contribution of each one of them to the EEG signal is not practical whatsoever. Instead, in the MI-BCI signal processing module we can consider that there are as many sources as EEG sensors, where each source represents the electromagnetic signals of an area of the brain. From all those sources, we are interested in those were we can find the sensorimotor rhythms. In that sense, one can say that the ERS are active sources and that the ERD are non-active sources. So, if we can detect which sources are active during a trial, we can discriminate between motor imagery movements, which is our main goal. This is not a simple task because the sources cannot be directly identified in the EEG signal since they are produced within the brain, and in the scalp we observe a distorted and contaminated version of those signals.

To express the EEG observations in a mathematical way, we will call $\mathbf{s}(t)$ to the source signals and $\mathbf{x}(t)$ to the EEG signals, or observations, which can be modeled as:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t), \tag{4.1}$$

Where the matrix $\mathbf{A}$ represents the mixing of the sources and the $\mathbf{n}(t)$ is a random Gaussian noise that represents the rest of variations that are not produced by the considered sources. In general it is also usual to use a matrix-wise notation

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}, \tag{4.2}$$

on which the time samples are concatenated in a column wise form: $\mathbf{X} = [\mathbf{x}(t_1), \cdots, \mathbf{x}(t_{end})]$, $\mathbf{S} = [\mathbf{s}(t_1), \cdots, \mathbf{s}(t_{end})]$, $\mathbf{N} = [\mathbf{N}(t_1), \cdots, \mathbf{N}(t_{end})]$.

The algorithm that we explain in this chapter, CSP, is based on this simple model and it is a well extended problem formulation.

4.2 Introducing the Common Spatial Patterns algorithm

The Common Spatial Patterns (CSP) algorithm [50] was first used to extract patterns from EEG signals in [51] and further on applied to MI-BCI in the article [30]. It is a method designed for those cases in which we have two classes, $K = 2$, and it transforms the EEG signals by applying a set of linear combinations between channels to reduce its dimensionality. That set of linear combinations is what we refer to as spatial filters.

In the previous chapter we commented that the main summary measure of each trial is its covariance. The CSP algorithm compute the average of these covariance matrices of each class. The resulting spatial filters simultaneously diagonalize the average covariance matrices, concentrating the whole power of the covariance in a diagonal matrix, whose elements are the variance of the filtered EEG signals. In addition, the transformation is made in such a way that for each CSP filter the variance of the resulting signal is maximized if the trial is from one class and minimized if it is from the other class.

CSP is a supervised algorithm, which means we need a set of labeled data to train the algorithm before it starts working with the unlabeled data. The training process of CSP begins by computing the two covariance matrices that represent the two classes. The matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ can be estimated from the training trials as the average covariance matrix of each class. Sometimes, we may find that there are trials that are more powerful, in the sense that they have more influence than others, specially the ones affected by the presence of artifacts. To avoid these trials from dominating the averaging covariance matrix of a class, and as we explained in section 3.2, these matrices are usually normalized by their trace and the computation of the class related covariance matrices result in

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{\tau:C_\tau=k}^{N_k} \frac{\mathbf{X}_{\tau,k}\mathbf{X}_{\tau,k}^\top}{\operatorname{tr}\left(\mathbf{X}_{\tau,k}\mathbf{X}_{\tau,k}^\top\right)}, \quad k = 1,2, \tag{4.3}$$

where the variable N_k is used to refer to the number of training trials in class k and C_τ refers to the class of the trial τ. Nevertheless, this normalization can be considered and applied during the preprocessing part as commented in section 3.2.

Once the class related covariances $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ have been computed, the following step consists it computing the global covariance of both classes as

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2. \tag{4.4}$$

Then $\boldsymbol{\Sigma}$ is factorized as $\boldsymbol{\Sigma} = \mathbf{U}^\top \boldsymbol{\Lambda} \mathbf{U}$, where $\mathbf{U}$ is the eigenvectors matrix and $\boldsymbol{\Lambda}$ is the diagonal matrix containing the sorted eigenvalues of $\boldsymbol{\Sigma}$. It is important to note that the eigenvalues in $\boldsymbol{\Lambda}$ should be sorted[*], either in a descending or ascending order. This done, the eigenvectors in $\mathbf{U}$ are sorted in the same way. In this thesis we have chosen to sort the eigenvalues in a descending order. Using the previous factorization, the next step is to define the whitening matrix $\mathbf{B} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}$, which will transform the class related covariance matrices as:

$$\mathbf{M_1} = \mathbf{B}^\top \boldsymbol{\Sigma}_1 \mathbf{B} \text{ and } \mathbf{M_2} = \mathbf{B}^\top \boldsymbol{\Sigma}_2 \mathbf{B} \tag{4.5}$$

After whitening the matrices $\mathbf{M_1}$ and $\mathbf{M_2}$, these matrices share the same eigenvectors $\mathbf{w}_i$ which concatenation form the matrix $\mathbf{W}$. Nevertheless, they are sorted in a different order since their eigenvalues are complimentary in the sense that the largest eigenvalue of $\mathbf{M_1}$ and the smallest eigenvalue of $\mathbf{M_2}$ sum 1. This can be expressed mathematically as:

$$\mathbf{M_1} = \mathbf{W}^\top \boldsymbol{\Lambda}_1 \mathbf{W} \;\Leftrightarrow\; \mathbf{M_2} = \mathbf{W}^\top \boldsymbol{\Lambda}_2 \mathbf{W} \tag{4.6}$$

and

$$\boldsymbol{\Lambda_1} + \boldsymbol{\Lambda_2} = \mathbf{I} \tag{4.7}$$

with $\boldsymbol{\Lambda}_1$ being the diagonal matrix containing the eigenvalues of $\mathbf{M_1}$ sorted in descending order, and $\boldsymbol{\Lambda}_2$ the diagonal matrix containing the eigenvalues of $\mathbf{M_2}$ sorted in ascending order.

The vectors $\mathbf{w}_i$ that form the matrix $\mathbf{W} = \left[\mathbf{w}_1, \cdots, \mathbf{w}_{N_s}\right]$ are the optimal filters that CSP looks for. The reason why they are optimal is due to the fact that all the vectors $\mathbf{w}_i$ are critic points of the functions $f_1(\mathbf{w}) = \mathbf{w}^\top \mathbf{M_1} \mathbf{w}$ and $f_2(\mathbf{w}) = \mathbf{w}^\top \mathbf{M_2} \mathbf{w}$, where the local maximums of $f_1(\mathbf{w})$ coincide with the local minimums of $f_2(\mathbf{w})$ and vice versa. Therefore,

[*] It is common that the algorithms used to compute the eigenvalues and the eigenvectors provide the results in a sorted way if the origin covariance matrices are symmetric. If they are not, the results may come out unsorted. Sometimes due to precision mistakes, the origin covariance matrices are almost symmetric but have some differences, resulting in non-sorted results from the algorithms. Therefore, our recommendation is to always make sure that these elements are sorted even though the algorithm does this in most of the cases.

the vectors $\mathbf{w}_i$ are the optimal filters that best separate the variances of the transformed EEG signals. In addition, they are also orthogonal to each other ($\mathbf{w}_i^\top \mathbf{w}_j = 0 \ \ \forall i \neq j$).

It is important to note that the only parameter that the original CSP algorithm has is the number of filters to be used. CSP can provide with as many spatial filters as EEG channels but because of the curse of dimensionality, a common practice is to reduce the dimensionality of the output by selecting the P filters that best discriminate between classes. Those P filters are the eigenvectors associated to the largest and smallest eigenvalues. Assuming P to be an even smaller number than N_s, the set of spatial filters is formed as:

$$\mathbf{W} \Leftarrow \left[\mathbf{w}_1, \ \cdots, \mathbf{w}_{P/2}, \mathbf{w}_{N_s - P/2}, \ \cdots, \ \mathbf{w}_{N_s}\right]. \tag{4.8}$$

4.2.1 Additional considerations about the CSP algorithm

One interesting property of the CSP algorithm is that it can be formulated by different optimization problems, of which the vectors $\mathbf{w}_i$ are solutions

$$\min_{\dim\{\mathscr{W}\}=N_s-i+1} \ \max_{w\in\mathscr{W}} \frac{var(\hat{\mathbf{w}}_i{}^\top \mathbf{X}_1)}{var(\hat{\mathbf{w}}_i{}^\top \mathbf{X}_2)}, \tag{4.9}$$

$$\min_{\dim\{\mathscr{W}\}=N_s-i+1} \ \max_{w\in\mathscr{W}} \frac{var(\hat{\mathbf{w}}_i{}^\top \mathbf{X}_1)}{var(\hat{\mathbf{w}}_i{}^\top \mathbf{X})}, \tag{4.10}$$

where $\mathbf{X}_k$ represents the column wise concatenation of all the training trials from class k, $\mathbf{X}$ is the concatenation of all the trials and the operators $\min_{\dim\{\mathscr{W}=N_s-i+1\}}$, $\max_{w\in\mathscr{W}}$ make reference to the application of the Courant–Fisher–Weyl minimax principle (check reference[52] in p.58 or reference [53] eq (19)-(21)). In the same way, the previous equations can be expressed as Rayleigh quotients. As an example, we will only express in this notation the equation 4.9 of the previous case:

$$\min_{\dim\{\mathscr{W}\}=N_s-i+1} \ \max_{w\in\mathscr{W}} \frac{\mathbf{w}_i^\top \mathbf{\Sigma}_1 \mathbf{w}_i}{\mathbf{w}_i^\top \mathbf{\Sigma} \mathbf{w}_i} \tag{4.11}$$

As we have already said, all the equations above have the same solutions which are given by the eigenvectors that form the matrix $\mathbf{W}$, but the values that those functions take for a given vector $\mathbf{w}_i$ change from one equation to another.

The spatial filters in $\mathbf{w}_i$ also inherit the orthogonality property of the eigenvector ($\mathbf{w}_i^\top \mathbf{w}_j \simeq 0 \ \ \forall i \neq j$). This is of interest because the variances are projected in the subspace form by the vectors $\mathbf{W}$ and since they are orthogonal to each other and of norm 1 they form an Euclidean space.

Another important consideration refers to the format of the data that CSP works with. We have seen that the output of the CSP algorithm over each trial is the variances of the different projection of the signals, that can be computed from the EEG signal of the trial τ:

$$\mathbf{y}_\tau = \left[var(\mathbf{w}_1^\top \mathbf{X}_\tau), \cdots, var(\mathbf{w}_P^\top \mathbf{X}_\tau)\right] \tag{4.12}$$

We can also choose to compute them using the diagonal of the transformed covariances:

$$\mathbf{y}_\tau = \operatorname{diag}\left(\mathbf{W}^\top \mathbf{\Sigma}_\tau \mathbf{W}\right). \tag{4.13}$$

Taking into consideration the format of the data is important because as we said before we can compute the covariance of the trials in the preprocessing part and forget about the EEG signal from that point on, or we can preprocess the signals and keep using them for feature extraction. Using the covariances has the advantage that they have less data than the EEG signals. For example, we have been working with 500 sample trials and 22 channels, while a trial covariance has 22x22 dimension. Using fewer data reduces the computation time and complexity, but keeping a higher abstraction of the data in the feature extraction block may reduce the options that we have to manipulate it.

4.3 CSP for the multi-class paradigm

Like we said in section 4.2, the CSP algorithm only works for two classes but in many BCI applications we are interested in using more than two commands. In that sense there are some proposals to extend the CSP algorithm to a multi-class paradigm [54, 55].

However, one of the most popular ways of doing this is based in the diagonalization that CSP performs. The diagonalization results as a direct consequence of equation (4.6), where we can see that the CSP filters simultaneously diagonalize the related covariance of both classes. Consequently, we could say that the CSP filters are the solution to a joint diagonalization problem for two classes. In that sense, the algorithm proposed in [54], which we refer to as Information Theory Feature Extraction (ITFE), uses the algorithm Joint Approximation Diagonalization of Eigen-matrices (JADE) to simultaneously diagonalize all the class related covariances and then select the filters that retain more class related information. The handicap of using JADE is that it is no longer an exact method but an approximation.

Since there is no analytic expression for the mutual information $I(C;\mathbf{w}_i^\top\mathbf{X})$ between the class C of each trial τ and the filtered signal $\mathbf{w}_i^\top\mathbf{X}_\tau$. They propose to use an approximation that they define as:

$$\begin{aligned} J_{ITFE}(\mathbf{w}_i) &= -\sum_{k=1}^{K} \mathrm{P}(C_k)\log\left(\sqrt{\mathbf{w}_i^\top\boldsymbol{\Sigma}_k\mathbf{w}_i}\right) - \frac{3}{16}\left(\sum_{k=1}^{K}\mathrm{P}(C_k)\left((\mathbf{w}_i^\top\boldsymbol{\Sigma}_k\mathbf{w}_i)^2-1\right)\right)^2 \\ &\approx I(C;\mathbf{w}_i^\top\mathbf{X}) \qquad (4.14)\end{aligned}$$

with K being the number of classes and $\boldsymbol{\Sigma}_k$ the class related covariance of the class k.

4.4 Some divergence-based criteria

The CSP algorithm is still an active research area because even though it provides us with good results in average, in the presence of artifacts it does not behave in the desired way [56, 57], and it is not robust to non-stationary changes. In this sense, the divergences interpretations have contributed to the CSP algorithm in several ways.

In the previous section (4.2) we have seen that by using CSP we assume that the EEG signals have zero mean and that both classes only differ in their covariance matrices. Consequently, we are establishing that the EEG signals are the result of a random Gaussian process with zero mean, which is unequivocally determined by its covariance matrix. This said, we can try to look at the dimensionality reduction problem from the point of view of statistics. Modeling the EEG signal as we just said

$$\mathbf{X}_1 \sim f_1 = \mathscr{N}(0, \boldsymbol{\Sigma}_1), \;\; \mathbf{X}_2 \sim f_2 = \mathscr{N}(0, \boldsymbol{\Sigma}_2). \tag{4.15}$$

Following this perspective the ratio in 4.11 can be seen as a measure of the dissimilarity between the two distributions.

In this section we will review the CSP algorithm from a divergence point of view, and we will also see some of the most popular divergence-based variants that have been proposed for the CSP algorithm. The divergences can be defined as set of statistic tools that measure the dissimilarities or distance between two distributions.

A good way of analyzing the resemblance between CSP and the divergences can be through the Kullback-Leibler divergence which is one of the most used ones. The Kullback-Leibler divergence between the two Gaussian distributions f_1 and f_2 is defined as:

$$D_{KL}(f_1 \| f_2) = \int_{-\infty}^{\infty} f_1(x) \log \frac{f_1(x)}{f_2(x)} dx. \tag{4.16}$$

For the case where f_1 and f_2 are defined as in (4.15), the Kullback-Leibler divergence can be simplified to:

$$\begin{aligned} D_{KL}(f_1 \| f_2) &= \int \mathscr{N}(0, \boldsymbol{\Sigma}_1) \log \frac{\mathscr{N}(0, \boldsymbol{\Sigma}_1)}{\mathscr{N}(0, \boldsymbol{\Sigma}_2)} \\ &= \frac{1}{2} \left[\log \frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)} - N_s + \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1\right) \right]. \end{aligned} \tag{4.17}$$

As one can see the KL is not a symmetric divergence, which means that

$$D_{KL}(f_1 \| f_2) \neq D_{KL}(f_2 \| f_1). \tag{4.18}$$

When measuring the divergence between two equally probable distributions, it does not make sense to use an asymmetric criterion, for this reason it is very common to use the symmetric version of the divergence, which is defined as:

$$D_{sKL}(f_1 \| f_2) = D_{KL}(f_1 \| f_2) + D_{KL}(f_2 \| f_1). \tag{4.19}$$

At this point, it is worth noting that the problem we are trying to solve is to find a set of spatial filters $\mathbf{W}$ of dimension $P < N_s$, that maximize the symmetric Kullback-Leibler divergence between the two classes. On the one hand this can be done from a hierarchical way resolving

$$\min_{\dim\{\mathscr{W}\} = N_s - i + 1} \; \max_{\mathbf{w}_i \in \mathscr{W}} \; D_{sKL}\left(\bar{f}_1 \| \bar{f}_2\right), \tag{4.20}$$

$$\bar{f}_k = \mathscr{N}(0, \mathbf{w}_i^\top \boldsymbol{\Sigma}_k \mathbf{w}_i), \;\; k \in (1,2) \;\; i \in (1, \cdots, N_s).$$

By doing it, we are looking for one optimal filter after the other. It is easy to see that both problems ((4.20), (4.11)) have equivalent set of solutions. It is also important to take into account that both solutions are equivalent but not necessary the exact same solution, in the sense that they span to the same subspace, but the ordering may change from one to another. This was shown in Samek's work [38], which we refer to the interested reader for a more detailed explanation.

Sometimes we are just interested in finding the subspace which maximize the separation between classes in terms of a divergence criterion. Therefore, we can also resolve the problem simultaneously, instead of looking for one filter after the other, we can look for the subspace that maximize the divergence between the two distributions, in which case we only compute the filters that maximize the following divergence

$$D_{sKL}\left(\bar{f}_1\|\bar{f}_2\right)=\frac{1}{2}\mathrm{tr}\left((\mathbf{W}^\top\boldsymbol{\Sigma}_2\mathbf{W})\left(\mathbf{W}^\top\boldsymbol{\Sigma}_1\mathbf{W}\right)^{-1}+(\mathbf{W}^\top\boldsymbol{\Sigma}_1\mathbf{W})(\mathbf{W}^\top\boldsymbol{\Sigma}_2\mathbf{W})^{-1}\right)-P. \tag{4.21}$$

In the CSP criterion there was a min max expression, meanwhile in this case there is only a maximization problem but note that the expression in (4.21) has the form $z+{}^1\!/_z$, which is maximized when z goes to either infinity or zero.

The use of the divergence criterion provides us with some advantages because it allows us to optimize different objective functions. For example, in the works [38, 58], it is also proposed to look for a projection that minimizes the within class distance at the same time that the distance between classes is maximized. To do that they use the following objective function:

$$\mathscr{L}_{sKL}(\mathbf{W})=(1-\phi)D_{sKL}(\mathbf{W}^\top\boldsymbol{\Sigma}_1\mathbf{W}\|\mathbf{W}^\top\boldsymbol{\Sigma}_2\mathbf{W})-\phi\Delta(\mathbf{W}), \tag{4.22}$$

where $0\leq\phi<1$ and:

$$\Delta(\mathbf{W})=\frac{1}{2L}\sum_{k=1}^{K}\sum_{\tau:C_\tau=k}^{N_k}D_{KL}\left(\mathscr{N}(0,\mathbf{W}^\top\boldsymbol{\Sigma}_\tau\mathbf{W})\|\mathscr{N}(0,\mathbf{W}^\top\boldsymbol{\Sigma}_k\mathbf{W})\right). \tag{4.23}$$

It is interesting to note that in equation 4.23, the Kullback-Leibler divergence is chosen instead of the symmetric version. The reason for this is that in this equation the comparison is done with a trial covariance to the class related covariance. The trial covariance may be ill-conditioned and it makes sense to avoid computing the inverse of these matrices.

The Kullback-Leibler divergence can be generalized through the Beta divergence. In [59] it is shown that the CSP solutions can also be obtained through the Beta divergence. In [60] is suggested a new objective function to maximize the divergence between trials covariance pairs from different classes

$$\bar{D}_{s\beta}(\mathbf{W})=\sum_{\tau:C_\tau=1}^{N_1}D_{s\beta}(\mathbf{W}^\top\boldsymbol{\Sigma}_{\tau,C_\tau=1}\mathbf{W}\|\mathbf{W}^\top\boldsymbol{\Sigma}_{|\tau,C_\tau=2}\mathbf{W}),\ N_1=N_2\ .$$

Where $D_{s\beta}$ represents the symmetric Beta divergence [60]. One reason to use this objective function is based on the argument that small values of β penalizes sporadic variations of trials while large values of β penalizes small variations among trials.

We cannot end this section without talking about another generalization of the two previous divergences: the Alpha-Beta Log-Det divergence (AB-LD). For two Gaussian distributions with zero mean, it takes the following value:

$$D_{LD}^{(\alpha,\beta)}(\mathbf{\Sigma}_1\|\mathbf{\Sigma}_2) = \frac{1}{\alpha\beta}\log\left|\frac{\alpha(\mathbf{\Sigma}_2^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-\frac{1}{2}})^{\beta}+\beta(\mathbf{\Sigma}_2^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-\frac{1}{2}})^{-\alpha}}{\alpha+\beta}\right|_{+} \tag{4.24}$$

$$\text{for } \alpha \neq 0,\ \beta \neq 0,\quad \alpha+\beta \neq 0,$$

where:

$$|x|_+ = \begin{cases} x & x \geq 0, \\ 0, & x < 0, \end{cases}$$

denotes the non-negative truncation operator. For the singular cases, the definition becomes:

$$\begin{aligned} &D_{LD}^{(\alpha,\beta)}(\mathbf{\Sigma}_1\|\mathbf{\Sigma}_2) \\ &= \begin{cases} \frac{1}{\alpha^2}\left[\mathrm{tr}\left((\mathbf{\Sigma}_2^{\frac{1}{2}}\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2^{\frac{1}{2}})^{\alpha}-\mathbf{I}\right)-\alpha\log|\mathbf{\Sigma}_2^{\frac{1}{2}}\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2^{\frac{1}{2}}|\right] & \text{for } \alpha\neq 0, \beta=0 \\ \frac{1}{\beta^2}\left[\mathrm{tr}\left((\mathbf{\Sigma}_2^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-\frac{1}{2}})^{\beta}-\mathbf{I}\right)-\beta\log|\mathbf{\Sigma}_2^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-\frac{1}{2}}|\right] & \text{for } \alpha=0,\ \beta\neq 0 \\ \frac{1}{\alpha^2}\log\left|(\mathbf{\Sigma}_2^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-\frac{1}{2}})^{\alpha}(\mathbf{I}+\log(\mathbf{\Sigma}_2^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-\frac{1}{2}})^{-\alpha})\right|_{+} & \text{for } \alpha=-\beta \\ \frac{1}{2}||\log(\mathbf{\Sigma}_2^{\frac{1}{2}}\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2^{\frac{1}{2}})||_F^2 & \text{for } \alpha,\ \beta=0 \end{cases} \end{aligned} \tag{4.25}$$

The interest in the AB-LD divergence [61] is that it also reproduces the CSP solution for any value of α and β through the maximization of the function

$$D_{LD}^{(\alpha,\beta)}\left(\mathbf{W}^{\top}\mathbf{\Sigma}_1\mathbf{W}\|\mathbf{W}^{\top}\mathbf{\Sigma}_2\mathbf{W}\right) \tag{4.26}$$

as it has been proven in our previous work [53]. As we show in the figure 4.1, the AB-LD generalizes other existing divergences as for example the Riemann distance or the Kullback-Leibler divergence. This divergence may also be useful in configurations like (4.22) since the parameters α and β provide with greater freedom.

To finish this section, we show in figure 4.2 the value of several divergences as a spatial filter vary in a 2-D schema ($P = 1$), where the value of the spatial filters $\mathbf{W}$ only depend on the parameter θ which denotes the angle of the $\mathbf{w}$ vector in a 2-D schema. The angle of the CSP solutions are marked and the coincide with the extreme points of all the divergences. Note that in this schema we are just using one spatial filter and therefore, the hierarchical and the simultaneously way of resolving the divergence criterion meet. In this figure one can also see that the CSP solution matches the critic points of the set of divergences.

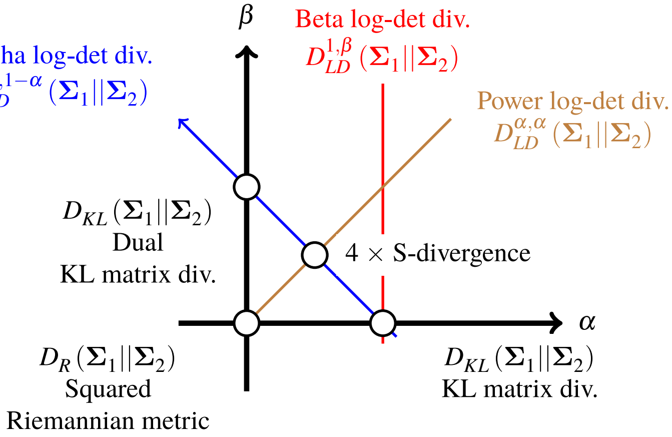

β
Beta log-det div.
$D_{LD}^{1,\beta}(\Sigma_1||\Sigma_2)$
Alpha log-det div.
$D_{LD}^{\alpha,1-\alpha}(\Sigma_1||\Sigma_2)$
Power log-det div.
$D_{LD}^{\alpha,\alpha}(\Sigma_1||\Sigma_2)$
$D_{KL}(\Sigma_1||\Sigma_2)$
Dual
KL matrix div.
4 × S-divergence
α
$D_R(\Sigma_1||\Sigma_2)$
Squared
Riemannian metric
$D_{KL}(\Sigma_1||\Sigma_2)$
KL matrix div.

Figure 4.1 Map of the AB-LD divergence $D_{LD}^{(\alpha,\beta)}(\Sigma_1||\Sigma_2)$ in the (α,β)-plane. Where the position of each divergence is specified by the value of (α,β).

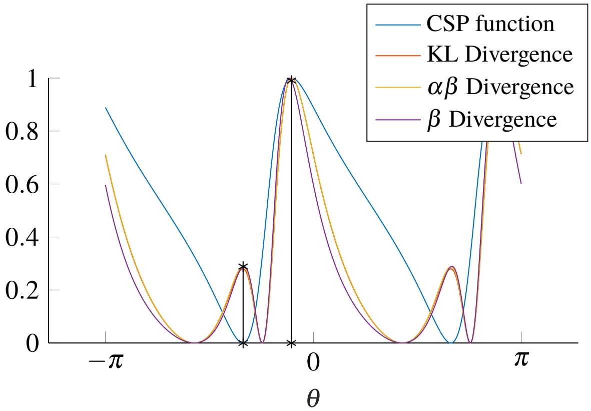


Figure 4.2 This figure shows the evolution of several divergences and the CSP criterion for two random cavariance matrices. With one spatial filter ($P = 1$), its components vary with respect θ as: $\mathbf{w} = [\cos\theta, \sin\theta]$. To show a result that we can appreciate, and since we are only interested in the extreme points. The function were normalized between zero and one.

4.5 Feature extraction based on Riemannian geometry

It has already been pointed out how the divergences are a great metric to work with covariance matrices and how the CSP filters can be found through the maximization of the divergence between two distributions. In that sense, Riemann's distance, which matches the Alpha-Beta divergence for α, $\beta = 0$ as shown in the figure 4.1, are a great set of tools. One of the reasons is that the Riemann distance has no hyper-parameter and it is a symmetric divergence by itself. Also, the Riemann distance has been used to compare and measure covariance matrices for years [62] and has been linked to CSP in multiple references, such as [63].

As seen in [64], the Riemann distance has also been successfully used to classify MI-BCI trials covariances without using any dimensionality reduction algorithms. The classification is performed by comparing the Riemann distance between the test trial covariance and the center of each class, choosing as estimated class the one with the closest center. In [64], it is also proposed to project the covariance matrices into a Riemann space which is tangent to the trials mean covariance matrix, as represented in figure 4.3. In our experiments, this algorithm obtained the higher accuracy results and it is for this reason that we have chosen to study it the present section.

To see how to project the covariance matrices into the tangent space we start referencing [62], where we can see that the SPD subspace where the covariance matrices are, is a differentiable manifold. This means that the tangent vectors in a given point vary smoothly from one point to another close point. We will denote $\mathscr{P}$ to the subspace where the SPD matrices are and $\mathscr{T}_\int$ to the tangent space, $\mathscr{P}, \mathscr{T}_\int \in \mathbb{R}^{N_s(N_s+1)/2}$. The reason why this tangent space is useful is because most of the classification algorithms need to work in Euclidean spaces, and the Riemann distance between two SPD matrices can be approximated by the Euclidean distance between the two projection of the SPD matrices in the

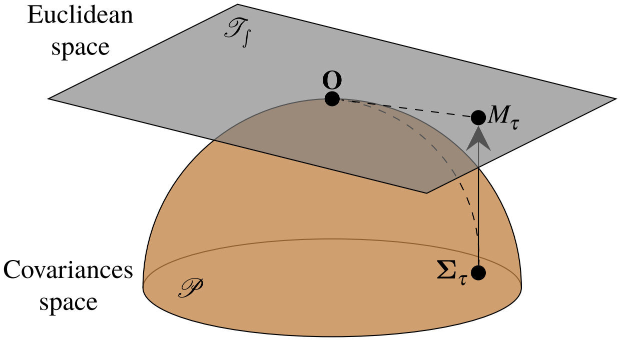


Figure 4.3 Representation of a covariance matrix Σ_τ in a given covariance space, and its projection M_τ in the Euclidean space which is tangent to the covariance space at the center O.

$\mathscr{T}_{\int}$ space.

To work with the matrices as in an Euclidean space, it is necessary to define a vectorization of the matrices which is given by:

$$vec(\mathbf{C}) = \left[c_{1,1},\, \sqrt{2}c_{1,2},\, c_{2,2},\, \sqrt{2}c_{1,3},\, \sqrt{2}c_{2,3},\, c_{3,3}, \cdots, c_{j,j},\right]. \tag{4.27}$$

The reason why the $\sqrt{2}$ coefficient appears is to maintain the norm $\|\mathbf{C}\|_F = \|vec(\mathbf{C})\|_2$. This said, projecting the covariance matrices using 4.27 is not advised because this projection does not respect the Riemann properties between SPD matrices.

Since the main goal is to transform a set of covariance matrices ($\mathbf{\Sigma}_\tau$, $\tau \in [1,N_\tau]$) in such a way that standard classifiers can be used, in [64] it is proposed to project the covariance matrices into the Euclidean tangent space of the center of the set. We can refer to the center of the set as $\mathbf{O}$. The center $\mathbf{O}$ is usually computed as the Riemann geometric mean of the set of matrices [65], although in our experiments there was not a significant improvement in comparison to computing the Euclidean mean of the set of matrices as the class center is computed in equation (4.3).

After computing the mean covariance matrix of the set using any of the two methods commented previously, the covariance matrices of the trials are projected in the tangent space by:

$$\mathbf{M}_{\tau,\mathbf{O}} = \mathbf{O}^{1/2} \log\left(\mathbf{O}^{-1/2}\mathbf{\Sigma}_\tau\mathbf{O}^{-1/2}\right)\mathbf{O}^{1/2}, \tag{4.28}$$

where $\mathbf{M}_{\tau,\mathbf{O}}$ is the projection of $\mathbf{\Sigma}_\tau$ on the Euclidean tangent space of $\mathbf{O}$. We are not interested in the projection itself, but in transforming the covariances in such a way that they can be passed on to a classification module. Following this idea, we are interested in a transformation that can satisfy the following:

$$\delta_{\mathscr{R}}(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2) \simeq \|\mathbf{y}_1 - \mathbf{y}_2\|_2, \tag{4.29}$$

where $\delta_{\mathscr{R}}$ represents the Riemann distance and $\mathbf{y}_1, \mathbf{y}_2$ are the vectorized forms of $\mathbf{\Sigma}_1,\mathbf{\Sigma}_2$. The norm of the difference can be expressed through scalar products that in the new space are defined as $\langle \mathbf{y}_1,\mathbf{y}_2\rangle_{\mathbf{O}} = \mathrm{tr}\left(\mathbf{y}_1\mathbf{O}^{-1}\mathbf{y}_2\mathbf{O}^{-1}\right)$. For these reasons and to satisfy (4.29), the transformation that is used is:

$$\mathbf{y}_{\tau,\mathbf{O}} = vec\left(\mathbf{O}^{-1/2}\mathbf{M}_{\tau,\mathbf{O}}\mathbf{O}^{-1/2}\right), \tag{4.30}$$

$$= vec\left(\log\left(\mathbf{O}^{-1/2}\mathbf{\Sigma}_\tau\mathbf{O}^{-1/2}\right)\right). \tag{4.31}$$

It is important to note that the test trials have to go through the same process as the training trials, therefore they need to be projected using the same center. It is also worth noting that the dimension of these projections may be too high in comparison with the number of training trials and, for this reason, it is usual to reduce the dimension with the use of any standard classification algorithm as for example the Fisher LDA algorithm (explained in section 5.1). The tangent space technique can also be combined with CSP to first reduce the dimension of the covariance matrices and later compute their projections.

However, we can find a complete and useful review on Riemannian approaches that extend this explanation in [66].

4.6 Discussion

In this chapter we have studied the CSP algorithm and some of its variants which are related with the processing of the information which is present in the data. We have also chosen to review the divergence approaches because of our experience using those techniques but there are other CSP variants that have shown very good results. In this sense, we highlight the FBCSP algorithm [39] and the regularized versions of CSP shown in [67, 68]. Even though all these variants perform quite well, they all need hyper-parameters that need to be configured through CV which slow down the training process. In addition, if there is not enough training data, it is advised to use as fewer hyper-parameters as possible, since fewer data can lead to overfitting. Apart from the hyper-parameter configuration through CV, we still have to choose the ranges on which they will vary. Because of all these reasons, it becomes very difficult to compare and experiment with those algorithms in a wide range of datasets and users, so for many applications and research works the extra load of applying CV does not always compensate the improvements that they may contribute.

Through the study of the CSP method we have seen that by its definition it only works for two classes. So, to extend it to a multi-class paradigm, we have to use tricks as "divide and conquer", "one versus rest" [39] or we can also approximate the diagonalization that is given as a result of the JADE algorithm [54].

In contrast with the CSP variants, we can find Riemann geometry techniques that allow us to directly classify the covariance matrices using the Riemann distance to the class related covariances, or to project the covariances in the tangent space on which we can use any classifier available in the literature. The Riemann techniques work for the binary case as well as for the multi-class problem, in the latter, they compose a more precise tool than other techniques based in approximations. In addition, since these techniques work with covariances, they can be used in combination with the CSP algorithm after it has reduced the dimensionality of the signals.

Chapter 5

Popular classification techniques in BCI

The purpose of this chapter is to do a brief review of the state-of-the-art classifiers that are used in the field of MI-BCI systems nowadays. For the interested reader, we refer him to the following book, where he will find a complete guide about these classifiers [69].

We are going to concentrate on two classifiers that have been widely used in MI-BCI, which are Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM). We will also go over another algorithm that can result of interest, which is the the Logistic Regression classifier. This classifier is not as popular as the previous in MI-BCI, but when it is used in combination with the Tangent Space technique, it provides us with the best accuracy results that we have been able to achieve.

In addition, we will review two popular clustering algorithms, the k-means and the Gaussian mixture model. These algorithms are not supervised and can be used when the training data is unlabeled. They are of importance to us since in this thesis we demostrate that CSP can be performed in an unsupervised manner.

5.1 Linear Discriminant Analysis

The LDA classifier, first developed by Fisher in 1936 [70], can be considered the first algorithm for pattern recognition [71]. Although there are non-linear versions of this algorithm, in this work we will only focus in the case that assumes that the within class covariance matrix is the same one for all the classes and the border between classes is linear. This is a supervised algorithm that assumes that all the classes' features are distributed as Gaussian variables, and that the features from one class differ from the others only by their mean values. In other words, the features extracted from a trial τ with class k is modeled as Gaussian distribution

$$f_{\tau|C=k} \sim \mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{\mathbf{y}}\right) \quad \text{for } k \in (1,2) \tag{5.1}$$

where μ_k represent the unique mean values of the class k and $\boldsymbol{\Sigma}_{\mathbf{y}}$ is the common between-class covariance matrix.

The usage of the CSP algorithm explained in section 4.2 is usually followed by the usage of the LDA classifier. This may be due to historical reasons since the first proposal for the use of CSP in MI-BCI systems, was in combination with the LDA classifier[30]. This said and leaving aside historical reasons, the combined used of both of them makes sense, since the assumptions made by both algorithms are the same ones. In this sense, we can recall that CSP assumes that the signal from both classes follow a Gaussian distribution with zero bias but different covariance matrices. If CSP is followed by LDA, the latter assumes that the distribution of the sample variances around the center of each class is identical and that the logarithm of the samples variances also follow a Gaussian distribution.

When there are only two classes with the same prior probability, the classification border can be easily found. Given the density probability function of a multivariate normal distribution:

$$\mathrm{P}(\mathbf{y}|c=k) = \frac{1}{(2\pi)^{P/2}\sqrt{|\boldsymbol{\Sigma}_{\mathbf{y}}|}} \exp\left(\frac{-1}{2}(\mathbf{y}-\boldsymbol{\mu}_k)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(\mathbf{y}-\boldsymbol{\mu}_k)^\top\right) \quad \text{for } k \in (1,2) \tag{5.2}$$

and using the Bayes theorem [69], we can compute the border by resolving the equation resulting of applying $\mathrm{P}(\mathbf{y}|c=1) = \mathrm{P}(\mathbf{y}|c=2)$ to equation (5.2). We can see that equation (5.2) can be simplified into the following expression:

$$0 = \mathbf{y}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top - \frac{1}{2}\boldsymbol{\mu}_2\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\mu}_2^\top + \frac{1}{2}\boldsymbol{\mu}_1\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\mu}_1^\top \tag{5.3}$$

which coincides with the equation of an hyperplane with a term dependent of the variable $\mathbf{y}$ and an independent term. Therefore, we can rewrite the previous expression as:

$$\begin{aligned} \hat{C} &= sign(\mathbf{y}\mathbf{d}+b)) \\ \mathbf{d} &= \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top \\ b &= -\frac{1}{2}\boldsymbol{\mu}_2\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\mu}_2^\top + \frac{1}{2}\boldsymbol{\mu}_1\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\mu}_1^\top = \frac{-1}{2}(\boldsymbol{\mu}_2+\boldsymbol{\mu}_1)\mathbf{d}, \end{aligned} \tag{5.4}$$

where the estimated class ($\hat{C}$) of a vector of features ($\mathbf{y}_\tau$) can be easily computed using the equation (5.4).

When there are more than two classes the border between classes can not be so easily computed and it may not be lineal anymore. Nevertheless, we can always compute the posterior probability of each class using Bayes theorem and equation 5.2, selecting the class with higher probability.

At this point it is important to note that there is a variation of LDA, in which the covariance $\boldsymbol{\Sigma}_{\mathbf{y}}$ is computed using a shrinkage covariance estimator. This topic is discussed in Chapter 6 and it leads to what is known as Shrinkage Linear Discriminant Analysis (sLDA).

5.1.1 LDA in dimensionality reduction

In the field of MI-BCI, the LDA algorithm can also be used as an algorithm for dimensionality reduction after applying a tangent space projection [64].

Performing a dimensionality reduction with Fisher LDA consists in making the same assumptions that are made in LDA classification algorithm (identical Gaussian distributions around their centers) and maximizing the ratio between within class covariance matrix ($\mathbf{\Sigma_y}$) and between class covariance matrix ($\mathbf{\Sigma}_B$). The $\mathbf{\Sigma}_B$ matrix is computed as:

$$\mathbf{\Sigma}_B = \frac{1}{N_C}\sum_{k=1}^{N_C}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\mathbf{y})^\top(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\mathbf{y}), \tag{5.5}$$

where $\boldsymbol{\mu}_\mathbf{y}$ is the overall features mean.

Afterwards, the SVD of the matrix $\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{\Sigma}_B$ is computed. The pair of eigenvalues and eigenvectors are sorted in descending order and the P first eigenvectors are selected as filters to reduce the dimensionality. The Fisher LDA algorithm is indeed the solution to the following maximization problem where each eigenvector is a solution and a local maximum, making each eigenvalue the value that the following maximization ratio takes for each eigenvector:

$$\max_w \frac{w^\top \mathbf{\Sigma}_B w}{w^\top \mathbf{\Sigma}_\mathbf{y} w} \tag{5.6}$$

By selecting the eigenvectors associated to the highest eigenvalues, we are maximizing the between class variance at the same time that we are minimizing the within class variance. On top of this, each eigenvector is orthogonal to the others and by combining them we create an Euclidean space which suits other classifiers.

5.2 Support Vector Machines

The SVM algorithm is one of the most popular algorithms in the field of computer science and engineering. In fact, there are more than 100.000 works that are related to SVM[*] and one of the most popular libraries[72] for the usage of SVM already counts with more than 18.000 cites in Scopus.

In the field of MI-BCI it has been widely used [73] because of its versatility and also because libraries like [72], make it very easy to use and apply to any classification problem.

The problem that SVM solves is related to the computation of "the widest street" or maximum gap between two groups of points. The class of each point can be determined by the variable c and each class can be represented by one or minus one ($c_\tau = -1, 1$). To understand how it works, we start by making the assumption that the two classes of points can be linearly separated by an hyperplane. This hyperplane would match the midway of the "widest street".

[*] Number of results obtained with the search –"SVM" OR "Support Vector Machine"– made the 17th of September of 2019.

The hyperplane can be defined by its perpendicular vector **d** and an offset b. Under this definition, the class of an unclassified point **u** can be determined by:

$$\hat{c} = sign(\mathbf{ud} + b). \tag{5.7}$$

To compute the values of **d** and b, we impose the following restriction:

$$c_\tau(\mathbf{yd} + b) \geq 1, \tag{5.8}$$

for $\tau \in (1, \ldots, N_\tau)$, where the $\mathbf{y}_\tau$ are the training points or features.

Equation (5.8) imposes that all the points of a class are in one side of the hyperplane while the other class points are in the other side of the hyperplane, leaving a gap completely empty between the two classes, as represented in figure 5.1. The Support Vectors (SV) are those for which the equality is satisfied:

$$c_\tau(\mathbf{y}_\tau \mathbf{d} + b) = 1 \ \ \forall \tau \in (SV). \tag{5.9}$$

As the SV are the closest ones to the border, and any two SP of different classes ($\mathbf{y}_+$, $\mathbf{y}_-$) can define the gap size as the vector difference projected in the direction of **d**:

$$\text{Gap} = (\mathbf{y}_+ - \mathbf{y}_-)\frac{\mathbf{d}}{||\mathbf{d}||} = \frac{2}{||\mathbf{d}||} \tag{5.10}$$

where the result $^2/_{||\mathbf{d}||}$ can be obtained by plugging (5.9) in $\mathbf{y}_+$ and $\mathbf{y}_-$. In figure 5.1 we represent the SVM geometry that we just explained.

Returning to our objective, the goal is to obtain a gap as wide as possible and, for that reason, the optimization problem changes to:

$$\max_{\mathbf{d}} \frac{2}{||\mathbf{d}||} = \min_{\mathbf{d}} \frac{1}{2}||\mathbf{d}||^2, \tag{5.11}$$

subject to the restriction in (5.8). We can solve the previous optimization problem by applying the Lagrange multipliers (α_τ), obtaining the following expression:

$$\min_{\mathbf{d},b} \ L(\mathbf{d},b,\alpha_\tau) = \min_{\mathbf{d},b} \ \left[\frac{1}{2}||\mathbf{d}||^2 - \sum_{\tau=1}^{N_\tau} \alpha_\tau \left[c_\tau(\mathbf{y}_\tau \mathbf{d} + b) - 1\right]\right] \tag{5.12}$$

We can easily arrive to the dual problem by taking the derivatives and making them equal to zero:

$$\frac{\partial L(\mathbf{d},b,\alpha_\tau)}{\partial \mathbf{d}} = \mathbf{d} - \sum_{\tau=1}^{N_\tau} \alpha_\tau c_\tau \mathbf{y}_\tau = 0 \implies \mathbf{d} = \sum_{\tau=1}^{N_\tau} \alpha_\tau c_\tau \mathbf{y}_\tau \tag{5.13}$$

$$\frac{\partial L(\mathbf{d},b,\alpha_\tau)}{\partial b} = -\sum_{\tau=1}^{N_\tau} \alpha_\tau c_\tau = 0 \implies \sum_{\tau=1}^{N_\tau} \alpha_\tau c_\tau = 0. \tag{5.14}$$

By plugging (5.13) and (5.14) in $L(\mathbf{d},b,\alpha_\tau)$, and doing some algebra we obtain:

$$L(\alpha_1, \cdots, \alpha_\tau) = \sum_{\tau=1}^{N_\tau} \alpha_\tau - \frac{1}{2}\sum_{i=1}^{N_\tau}\sum_{j}^{N_\tau} \alpha_i \alpha_j c_i c_j \mathbf{y}_i \mathbf{y}_j^\top, \tag{5.15}$$

where we can see that the whole function depends on the product $\mathbf{y}_i\mathbf{y}_j^\top$. Furthermore, if we plug (5.13) in the decision boundary (5.7), it results in:

$$\hat{c} = sign\left(\sum_{\tau=1}^{N_\tau} \alpha_\tau c_\tau \mathbf{y}_\tau \mathbf{u}^\top + b\right) \tag{5.16}$$

which also depends on scalar products.

To minimize the function $L(\alpha_1,\cdots,\alpha_\tau)$ in equation (5.15), a quadratic optimization algorithm that is out of the scope of this work is used.

Even though, we have only explained SVM performing linear classification, the fact that the whole classifier only depends on scalar products makes possible what is known as the kernel trick, which allows us to operate in other spaces drawing nonlinear borders in the current space without projecting the points. This is useful because sometimes it is not possible to draw a straight line that separate the two classes and it is necessary to use other types of separations which are linear borders in a certain and higher dimension space. For example, in the previous chapter, in section 4.5 we saw how to project the covariance matrices onto the tangent space. From now on, we will see how to obtain a kernel function that provides with the scalar product of two matrices in the tangent space.

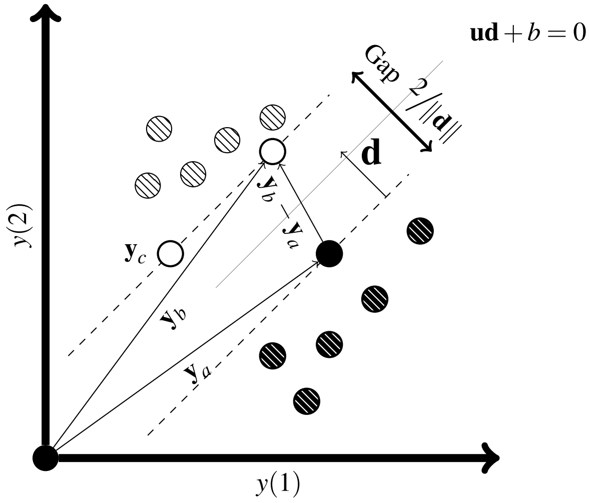


Figure 5.1 Representation of how a SVM works. In the figure we represent points of two classes, the classification border ($\mathbf{ud}+b=0$), the gap represented by the slashed lines and the gap size ($2/\|\mathbf{d}\|$), which is the result of projecting the difference between two SP over the direction of the vector $\mathbf{d}$.

The transformation from the covariance space to the Euclidean plane in equation (4.30) can be denoted as:

$$\phi(\boldsymbol{\Sigma}_\tau) = \mathbf{y}_\tau = vec\left(\log\left(\mathbf{O}^{-1/2}\boldsymbol{\Sigma}_\tau\mathbf{O}^{-1/2}\right)\right). \tag{5.17}$$

Since we are only interested in the result of the product between the projections, we can define a function that provides us with those results:

$$K(\boldsymbol{\Sigma}_a,\boldsymbol{\Sigma}_b) = \phi(\boldsymbol{\Sigma}_a)^\top \phi(\boldsymbol{\Sigma}_b). \tag{5.18}$$

This kernel, which defines the scalar products, can be plugged in the minimization and decision functions ((5.15), (5.16)):

$$L(\alpha_1,\cdots,\alpha_\tau) = \sum_{\tau=1}^{N_\tau} \alpha_\tau - \frac{1}{2}\sum_{i=1}^{N_\tau}\sum_{j}^{N_\tau} \alpha_i\alpha_j c_i c_j K\left(\boldsymbol{\Sigma}_i,\boldsymbol{\Sigma}_j\right) \tag{5.19}$$

$$\hat{c} = sign\left(\sum_{\tau=1}^{N_\tau} \alpha_\tau c_\tau K\left(\boldsymbol{\Sigma}_\tau,\boldsymbol{\Sigma}_u\right) + b\right). \tag{5.20}$$

Apart from this kernel which was defined in [74], there are many other kernels, of which we highlight the popular polynomial and Gaussian kernels.

5.3 Logistic Regression

The Logistic Regression (LR) is a widely used technique, making it easy to find numerous tutorials and implementations of it in different languages, for example in Python [75], or in Matlab with the function "mnrfit".

The LR is a predictive statistical model. In contrast with other classifiers like LDA or SVM, that try to draw a border that separates two classes, the LR is a model that tries to fit the logistic function into a set of points from two classes. The logistic function is defined as:

$$f_{LR}(y) = \frac{1}{1+e^{-(b+\mathbf{y}\mathbf{d})}}, \tag{5.21}$$

where the variables $\mathbf{d}$ and b are the parameters of this function.

The logistic function can be seen as a continuous approximation to the the step function. In its general form ($b = 0$), its output converges to zero for negative inputs, while it converges to one for positive inputs, having a smooth step shape for inputs around zero. In some models the value of the function is interpreted as the probability of each point of being from class one [76]. The shape of the logistic function is represented in figure 5.2, where we can see that its shape makes it perfect to fit in a constellation of points from two classes (zero and one). For this reason it can be used as a classifier. In the case that the output value is higher than 0.5 , its input is classified as class one, while if the output is less than 0.5, the assigned class would be zero.

As a regression problem, we try to minimize the error between the logistic function and the points within the target constellation. However, if we try to solve this optimization problem using Minimum Square Error (MSE), we would find that it is not a convex

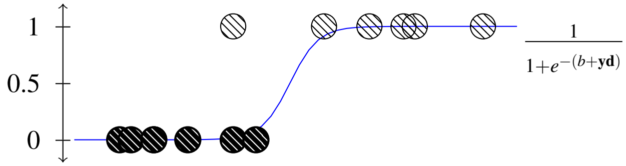


Figure 5.2 Sketch of a logistic regressor fitting in a two class problem.

function and it has more than one local minimum. For this reason the optimization problem is transformed into the following one:

$$\min_{\mathbf{d},b} \sum_{\tau=1}^{N_\tau} c_\tau \log\left(f_{LR}(\mathbf{d},b)\right) + (1-c_\tau)\log\left(1-f_{LR}(\mathbf{d},b)\right), \tag{5.22}$$

where c_τ are the class of each training trial. Equation (5.22) is know as "categorical cross entropy" function, and it is a convex function with only one minimum. The values of the parameter can be computed using any optimization algorithm like, for example, gradient descent.

5.4 Unsupervised classification algorithms

In the field of machine learning, there is a clear division between supervised and unsupervised techniques. Until now, we have only studied supervised algorithms, this is, algorithms in which the training set is already classified, having a label that distinguishes the class of each training point. In these kind of problems, the classification algorithm learns how to classify new and unlabeled data from the training data provided.

On the other hand, there are many situations where the main goal is to separate a set of points that share alike and unknown features in different groups. This group of algorithms are known as Clustering algorithms because they try to form clusters of points within a set of data. In a more general way, we can also refer to the unsupervised machine learning techniques as blind methods.

Most BCI systems fall in the supervised category, because during the training phase the user is told the class of the action he or she has to perform. However, blind algorithms are also of interest because they give us tools to work with mislabeled data or with applications where the class label is unavailable.

5.4.1 k-means

The k-means algorithm is the most popular clustering algorithm and it has many variants, as for example the one defined in [77], which introduces us to the algoritm and how to use it with the family $\alpha\beta$-divergences.

It is called k-means because it starts by setting a number k of clusters and tries to classify the training set in such a way that the square distant of each point to its cluster center is minimized:

$$\min_{\mathbf{S}} \sum_{i=1}^{N_c} \sum_{\mathbf{y}_j \in \mathbf{S}_i} \|\mathbf{y}_j - \boldsymbol{\mu}_i\|^2. \tag{5.23}$$

To obtain this minimum, the Lloyd's algorithm [78] is used. This algorithm consist in initializing the centroids with some arbitrary points (usually within the set), and it iterates over the two following steps until convergence. Each point is assigned to the closest centroid; the centroid of each clusters is computed.

This method is very dependent on the initialization and it can very easily fall into a local minima. Therefore, it is usual and recommended to try different initializations and to choose the best solution. In addition, it is considered a hard method because it supposes k number of clusters and that each point can only be from an unique cluster, ignoring the probabilities of a point being from one cluster in particular.

5.4.2 Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a soft algorithm that tries to fit a given number of Gaussian distributions into a set of data, maximizing the probability of each point of being from one of the clusters.

When the number of clusters is defined and with the assumption that they are independent Gaussian distributions, the probability of each one of the points is given by:

$$\mathrm{P}(\mathbf{y}) = \sum_{k=1}^{N_c} \pi_k \mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{y|k}\right), \tag{5.24}$$

which is the equation of a Guassian Mixture. Variable π_k is the occurrence probability of the class k.

Since we do not only have an isolated point but a set of them, we can build a matrix of features, where each row correspond to the features of a trial, and it has as many rows as training trials:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{N_\tau} \end{bmatrix}. \tag{5.25}$$

Because we consider each trial as an independent random process, the probability of $\mathbf{Y}$ is given by the product of the individual probabilities:

$$\mathrm{P}(\mathbf{Y}) = \prod_{\tau=1}^{N_\tau} \mathrm{P}(\mathbf{y}_\tau). \tag{5.26}$$

We would like to find the ML estimator parameter (π_k,$\boldsymbol{\mu}_k$,$\boldsymbol{\Sigma}_{y|k}$) for equation (5.26), but the sum in the equation (5.24) makes it a nuisance. For this reason, we rather find the ML estimator of the conditional probability:

$$\mathrm{P}\left(\mathbf{Y}|z_{\tau,k}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{y|k}\right) = \prod_{\tau=1}^{N_\tau} \prod_{k=1}^{N_k} \pi_k^{z_{\tau,k}} \mathcal{N}\left(\mathbf{y}_\tau|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{y|k}\right)^{z_{\tau,k}}, \tag{5.27}$$

where the variables $z_{\tau,k}$ represent binary variables which take the value 1 if the features $\mathbf{y}_\tau$ come from the cluster k and 0 if trial τ is from another class. Each trial's features can only come from a class or cluster $\sum_{k=1}^{N_k} z_{\tau,k}$.

For the sake of simplicity, we can express the maximization problem as:

$$\max_{z_{\tau,k},\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_{y|k}} \ln \mathrm{P}\left(\mathbf{Y}|z_{\tau,k},\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_{y|k}\right), \tag{5.28}$$

which can be solved through gradient methods. These gradient methods need to be initialized and it is usual to initialize GMM algorithm with the result obtained after applying k-means.

For a more detailed explanation, we recommend again the book [69], and for a tutorial and a Python implementation one can check [79].

5.5 Discussion

In this chapter we have studied a variety of classification algorithms that are used in multiple applications and fields of knowledge, not only in MI-BCI systems. Since they are widely used, it is easy to find coded implementations in different libraries throughout the internet and the documentation on this subject. This is useful to us since we don't need to study these algorithms in depth to apply them to an MI-BCI system, we can just make use of them in a high level understanding, always keeping in mind their hypothesis. This said, we consider it important to discuss the different configurations that have been used in the MI-BCI field and which have been proved to obtain good accuracy results.

In that sense, it is very common to use the CSP algorithm in combination with LDA, although nowadays it is more recommended to use sLDA (see Chapter 6). Nevertheless, the SVM classifier is also a good choice after applying CSP, but this algorithm demands more computation power and choosing this option instead of LDA should be somehow justified. However, when the tangent space technique is used, the three supervised classification algorithms have been used and they all provide with very good results but we have obtained the best results when combining the tangent space feature extraction with the logistic regression classifier. Combining tangent space and SVM also results in very good results. An alternative approach is to set up different solutions, combining several weak classifiers which probably would yield different results and include a final step to select the best or more robust combination of them. This is known as ensemble methods ans we can find an overview about this technique in [80].

To finish this chapter, it its important to talk about the multiclass scenario since we have only sutided the case of having two classes for the supervised algorithms. For the LDA algorithm we have developed the multiclass case by selecting the class with a higher probability. This is possible because it can be considered a soft classifier. On the other hand, for SVM and LR, it is often used the one versus the rest technique or the one-vs-one decomposition strategies.

Chapter 6

Covariance matrix estimators

As we have seen through previous chapters, the covariances are present in different parts of the MI-BCI systems, playing a very important role in each one of the blocks showed in Figure 2.1. For this reason it is necessary to review in this chapter some important estimators and considerations regarding the covariance matrices.

Until now, we have been using $\boldsymbol{\Sigma}$ to refer to the sample covariances. However, in this chapter we also need to refer to the true covariance of the distributions, which will be denoted as $\mathbf{C}$, this notation differs from the one that we used in the publications. The covariance matrices are important because they provide us with a measure about the correlation between a set of random variables. We will start by defining $(c_{i,j})$, which is the element of the matrix $\mathbf{C}$ in the column i and row j and it measures the correlation between the random variables i and j. The elements in the diagonal $(c_{i,i})$ coincide with the variance of the variable i. Therefore, they are symmetric matrices $(c_{i,j} = c_{j,i})$ where each element is given by:

$$C_{i,j} = \mathbb{E}\left[(x_i - \mu_i)(x_j - \mu_j)\right], \tag{6.1}$$

with x_i representing the random variable i and μ_i its mean.

There are a few reasons to use the covariance matrices of the trials, one of them being that the dimension of the data is reduced. Another reason is that the variance of the sources is a very good feature to check whether a source is active or not, keeping in mind that the ERS can be associated with the activation of a source, and the ERS to a non active sources, as we explained before.

A third reason to use the covariance matrices is that the expectation operator automatically reduces the noise influence under the assumption of independent Gaussian noise $(\mathbf{N} \sim \mathscr{N}(0, \boldsymbol{\Sigma}_N))$. We can recall the observations model in equation (4.2) as

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}, \tag{6.2}$$

which covariance results in

$$\mathbf{C}_{\mathbf{X}} \simeq \left\langle \mathbb{E}\left[\mathbf{X}\mathbf{X}^{\top}\right]\right\rangle_{N_{\tau},T} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{S}}\mathbf{A}^{\top} + \boldsymbol{\Sigma}_{N}, \tag{6.3}$$

where $\mathbf{C}_{\mathbf{X}}$ is the true covariance matrix of $\mathbf{X}$. We can expect that the power of the signal part $\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{S}}\mathbf{A}^{\top}$ dominates the noise contribution.

Although working with the covariances is the recommended and most extended procedure, it is important to take into account some considerations about them. The covariances are Symmetric Positive Definite (SPD) matrices, which means that they are in a special space that is not Euclidean. We can see this in figure 6.1, where the space for a two dimension square matrix is represented. The space of larger matrices with more than three dimensions, escape from our imagination. What we just explained is the reason why trying to classify covariance matrices with Euclidean distances or any Euclidean algorithm is not a good option. To solve this issue and to be able to use Euclidean algorithms, we make use of the feature extraction part in MI-BCI systems.

The sample covariances are not only used as features of the trials. Depending on which algorithms are used, there are usually other covariances estimations that are part of the algorithms used to extract the features (CSP) and to classify them (LDA).

The most popular covariance estimator is the Maximum Likelihood (ML), which is implemented in most of programming languages and it is taught in many science degrees.

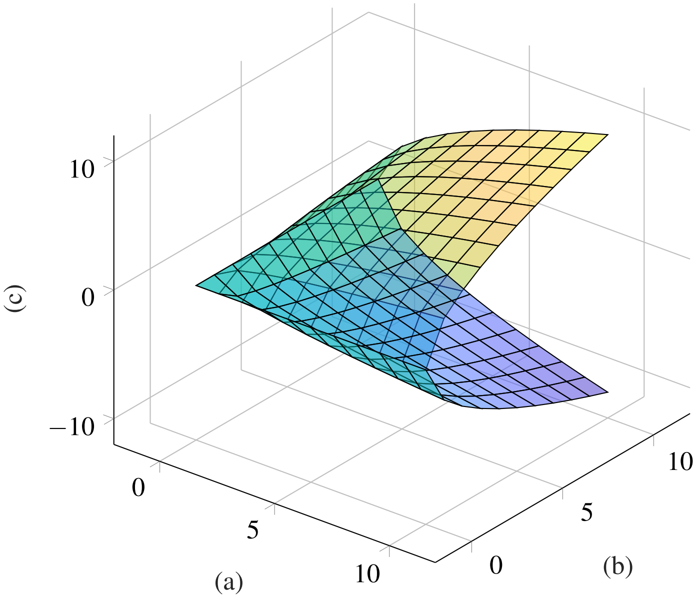


Figure 6.1 This surface with the shape of a cone represents the borders that hold the subspace of SPD matrices. Where $\left[\begin{smallmatrix} a & c \\ c & b \end{smallmatrix}\right]$ represents the values of a matrix.

This method provides with a simple formula that computes the estimation of the covariance from the multivariate sample data:

$$\boldsymbol{\Sigma} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{x}_t - \bar{\boldsymbol{\mu}})(\mathbf{x}_t - \bar{\boldsymbol{\mu}})^\top, \tag{6.4}$$

where, T is the number of data samples and $\bar{\boldsymbol{\mu}}$ is the estimated mean of the samples. This is a general purpose estimator that can be used independently of the data distribution. The $(N_\tau - 1)$ normalization factor is used because the estimation of the mean is used, but, in a case where we use a Gaussian distribution for example, the normalization factor becomes T.

As a ML estimator, one of its properties is that the estimation tends to the true covariance as the number of samples goes to infinity, and this estimator works just fine when the number of samples is high enough in comparison with the number of variables. But regardless of its popularity, the estimator from (6.4) needs the number of samples to be many times greater than the number of variables.

When the number of samples is relatively low, the sample covariance matrix is ill conditioned, having eigenvalues that tend to zero as the number of samples decrease. When the number of samples is lower than the number of variables, the covariance estimation will have at least one eigenvalue equal to zero. Having fewer samples also increase the large eigenvalues of the sample covariance matrix, so we can say that the dispersion of the eigenvalues is related to the number of samples [81].

The implication of having eigenvalues close to zero can be easily studied by decomposing the covariance matrix in its Singular Value Decomposition (SVD). As it is a SPD matrix, its decomposition can be found through its eigenvalues and eigenvectors:

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top, \tag{6.5}$$

where $\mathbf{U}$ is the matrix that holds the eigenvectors in its columns, and $\boldsymbol{\Lambda}$ is the matrix that contains the eigenvalues in its diagonal, having zero values in the rest of the positions. The inverse can be calculated through its SVD as:

$$\boldsymbol{\Sigma}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^\top. \tag{6.6}$$

In this equation one can see that if the sample covariance is not well conditioned and has eigenvalues close to zero, then those values will dominate the result of the inverse. This can be a real issue, specially in classifiers that use the inverse of the sample covariance matrix as saw in Chapter 5.

6.1 Shrinkage of the covariance and Ledoit and Wolf implementation

We have finished the introduction of the chapter establishing why the lack of samples to estimate the covariance becomes a problem, especially when it comes to compute its inverse. During years there has been new proposals to estimate the covariance with fewer samples to try to solve this problem among others [82, 83]. However, in 2004, Ledoit and

Wolf published their shrinkage method in their article [81], that regularizes the sample covariance matrix and which has become a standard in many applications [84, 85, 67].

This method consists in computing the estimated covariance matrix as the pondered addition of the sample covariance matrix and an identity matrix. By adding the identity matrix, the dispersion of the eigenvalues shrink (making this the reason why these kind of methods are called covariance matrix shrinkage algorithms):

$$\mathbf{C} \sim \tilde{\boldsymbol{\Sigma}} = \rho_1 \boldsymbol{\Sigma} + \rho_2 \mathbf{I}, \tag{6.7}$$

where the tilde denotes the shrieked covariance ($\tilde{\cdot}$).

In [81] we can find a formula that provides with the optimal value ρ_1 and ρ_2 without any previous assumption about the distribution from where the samples come from, and just using a finite number of samples.

This said, we can simplify equation (6.7) into the following equation where we only have a parameter:

$$\tilde{\boldsymbol{\Sigma}} = (1-\rho)\boldsymbol{\Sigma} + \rho \frac{\operatorname{tr}(\boldsymbol{\Sigma})}{P} \mathbf{I}. \tag{6.8}$$

From this equation, we can see that we have a shrinkage intensity (ρ) and a shrinkage target $\left({}^{\operatorname{tr}(\boldsymbol{\Sigma})}/_{P}\mathbf{I}\right)$.

To obtain the optimal shrinkage intensity, we need to solve the following MSE problem:

$$\min_{\rho} E\left[||\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_F^2 \right], \tag{6.9}$$

where the operator $||\cdot||_F^2$ refers to the square of the Frobenius norm.

In the work [86], Ledoit and Wolf proved that the optimal value for the parameter ρ was:

$$\rho_O = \frac{E\left[(\operatorname{tr}(\mathbf{C} - \boldsymbol{\Sigma})) \left(\frac{\operatorname{tr}(\boldsymbol{\Sigma})}{P}\mathbf{I} - \boldsymbol{\Sigma} \right) \right]}{E\left[\left|\left| \boldsymbol{\Sigma} - \frac{\operatorname{tr}(\boldsymbol{\Sigma})}{P}\mathbf{I} \right|\right|_F^2 \right]}. \tag{6.10}$$

However, this formula does not help us find the optimal value of ρ when we do have previous information about the distribution where the data comes from. This is because ρ_O depends of the true covariance of the data $\boldsymbol{\Sigma}$, and for this reason the estimator is called an oracle estimator.

Ledoit and Wolf also showed that the optimal value of ρ is between 0 and 1, and they finally obtained a well-conditioned estimator of the ρ parameter for large-dimensional covariance matrices in their work [81]:

$$\hat{\rho}_{\text{LW}} = \min\left\{ \frac{\sum_{\tau=1}^{N_\tau} ||(\mathbf{y}_\tau - \boldsymbol{\mu}_k)(\mathbf{y}_\tau - \boldsymbol{\mu}_k)^T - \boldsymbol{\Sigma}||_F^2}{N_\tau^2 ||\boldsymbol{\Sigma} - (\operatorname{tr}(\boldsymbol{\Sigma})/P)\, \boldsymbol{I}||_F^2}, 1 \right\}, \tag{6.11}$$

where the $\hat{}$ denotes that it makes reference to an estimation. We have also changed the notation since in the MI-BCI systems the shrinkage methods are usually used to compute the classifiers precision matrix (inverse of the covariance matrix). Therefore, to keep the consistency with Chapter 5, the samples are denoted as the features ($\mathbf{y}_\tau$), the number of samples is equal to the number of trials (N_τ) and the number of variables is equal to the number of features (P).

However, to obtain this result they had to establish an asymptotic framework where they considered that both the number of samples (N_τ) and their dimension (P), tend to infinity. However, to estimate the optimal shrinkage intensity it is only required that the distribution has fourth order finite moments that are estimated within the formula (6.11).

6.2 Improved covariance shrinkage for Gaussian distributions

Although the shrinkage proposed by Ledoit and Wolf is very effective, other authors have followed this research line proposing new shrinkage methods. In that sense, Chen et al. showed in [87] a specific shrinkage estimator for Gaussian distributions.

One of the perks of assuming Gaussianity is that these distributions are completely characterized by their covariance and mean, so a specific estimator can be designed. Taking this into account, we can evaluate the expectations of equation (6.10) to obtain that:

$$\hat{\rho}_{\mathrm{O}} = \frac{\left(1 - \frac{2}{p}\right) \operatorname{tr}\left(\mathbf{C}^2\right) + \operatorname{tr}^2(\mathbf{C})}{\left(N_\tau + 1 - \frac{2}{p}\right) \operatorname{tr}(\mathbf{C}^2) - \left(1 - \frac{N_\tau}{p}\right) \operatorname{tr}^2(\mathbf{C})}. \tag{6.12}$$

We can see that this formula still depends on the true covariance of the distribution, which we don't have access to. This expression can be approximated using the sample covariance matrix through the following expression that the authors called Oracle Approximating Shrinkage (OAS), and that only differs from the previous in the coefficients:

$$\hat{\rho}_{\mathrm{OAS}} = \min\left\{ \frac{\left(\frac{1-2}{p}\right) \operatorname{tr}\left(\mathbf{C}^2\right) + \operatorname{tr}^2(\mathbf{C})}{\left(\frac{N_\tau+1-2}{p}\right) \left[\operatorname{tr}(\mathbf{C}^2) - \frac{\operatorname{tr}^2(\mathbf{C})}{p}\right]}, 1 \right\}, \tag{6.13}$$

With this technique, the benefits of assuming a Gaussian distribution are several:

- The computational efficiency of equation (6.13) is higher than the computational efficiency of equation (6.11).
- While Ledoit and Wolf had to establish an asymptotic framework, this is not needed for this estimator and, in comparison, it works especially good when there are fewer samples.
- When the distribution is indeed Gaussian, the OAS estimator is always more precise that the one suggested by Ledoit and Wolf.

6.3 The Maronna-Tyler estimator of the scatter matrix

The scatter matrix can be seen as a covariance matrix that has no scale and it is common to compute the scatter matrix by normalizing the covariance matrix by the trace [88]. This matrix is of relevance because it is a parameter of a family of random variables distributions that are characterized by having a density function with the shape of an ellipse. Of course, the Gaussian distribution is part of this family but there are other famous ones like the $Student-T$ or logistic distributions that are part of this family too. In the MI-BCI field the scatter matrix also has importance since, as we commented in section 3.2, sample covariance matrices are often normalized as the scatter matrix [30, 89] is.

These kind of distributions are characterized by three parameters: the mean vector, the scatter matrix and the generator or tail function. The generalized form of their density function is:

$$f(x) = |\mathbf{C}^{-1/2}| g\left((x-\boldsymbol{\mu})\mathbf{C}^{-1}(x-\boldsymbol{\mu})^{\top}\right), \tag{6.14}$$

where the function $g(\cdot)$ is a non negative function which does not depend on either the mean ($\boldsymbol{\mu}$) or the scatter matrix $\boldsymbol{\Sigma}$.

As we have already established, the scatter matrix can be estimated using the formula (6.4) to estimate the covariance matrix and later normalize it. But this estimation does not provide with a reliable result when the tails of the distribution are heavy, as it happens in many elliptical distributions, where some samples may dominate the result. In 1976, Maronna proposed an optimization algorithm in his work [90] to compute robust estimators, but he could not provide a close form. After a decade, Tyler published a new estimator in his work [91], which became the most robust estimator of the scatter matrix. This estimator consists of an iterative algorithm that normalizes the samples using the precision matrix. The precision matrix can be initialized using (6.4) or the identity matrix, and its formula yields:

$$\boldsymbol{\Sigma}_{k+1} = \frac{N_s}{T}\sum_{t=0}^{T}\frac{\mathbf{x}_t\mathbf{x}_t^{\top}}{\mathbf{x}_t^{\top}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_t}. \tag{6.15}$$

This estimator counts with many applications such as radar [92] or financial engineering. However, in the field of BCI it has been used in [93] for the steady state visually evoked potential paradigm.

6.4 Generalized means of covariances

We have commented before that the CSP algorithm is one of the most used algorithms in MI-BCI, and that it depends on the sample mean covariance matrix within each class, although we can also consider that it depends on the overall mean covariance of both classes. In the same way, the tangent space technique also depends on the overall mean covariance. We have also noted that those mean covariances are usually estimated by computing the arithmetic mean of the covariances from all the trials. In the case that all the trials have mean zero, this estimation technique is equivalent to concatenate all the trials and compute the covariance of all of them at the same time. If the trials are biased, the difference of mean in each trial would affect the estimation in the concatenation technique. Nevertheless, in the MI-BCI field the signals are filtered in the 8-30Hz range, so they should not be biased.

Considering the case of computing the sample class related covariance (as well as the overall sample covariance) as the mean of all the trials, there are various alternatives (check [94]). On the one hand, a popular technique is to compute it as the geometric mean instead of using the arithmetic mean. This is done by computing the Riemann mean that we reviewed in section 4.5. Among the advantages of the geometric mean, we highlight that it has no hyper-parameters, the existence of a geometric center is guaranteed and it is more robust against outliers [64]. Despite of the existence of a unique center, there is not an exact formula to compute it and we have to rely in the following optimization problem:

$$\min_{\mathbf{O}} \sum_{\tau=1}^{N_\tau} \delta_{\mathscr{R}}(\mathbf{O},\mathbf{\Sigma}_\tau) \;\; \text{s.t.} \;\; \mathbf{O} \in \text{SPD} \tag{6.16}$$

where we try to minimize the sum of Riemann distances between the center and each trial. This problem can be addressed using any gradient descend method.

Another proposal that goes in the same direction can be found in the work of Samek [95]. Here, it is suggested to compute the center that minimizes the β-divergence from the center itself to each one of the trials within the class. That may be appropriate because the parameter beta can control the outliers interference. However, we find this method hard to implement due to the fact that the optimization function that is used contains local minimums, therefore it needs several initializations and in addition, the beta parameter adds a degree of freedom and may need of CV to choose a correct value.

6.5 Discussion

Through the chapters previous to this one, we have learned the importance that the estimations of the covariances have in a MI-BCI system, making them an important object of study in this field. Because of it, and because improving these techniques is a very effective way to improve the MI-BCI systems, in this chapter we have reviewed some methods to estimate the covariance of a group of random variables from a set of its samples.

Despite the different methods we have seen in this chapter, we can conclude that the simplest approach is usually the best. In that sense, the simplest covariance estimator is the maximum likelihood (ML) estimator which we can find in equation 6.4. This estimator is widely used and very easy to implement. However, when the precision matrix is needed, using a shrinkage method is advised, and nowadays it is common to find that new libraries already include them.

PART III

PUBLICATIONS

General notes

As commented in section 2.2, we have decided to include in this part of the document the publications on which this work is based and that contain the contributions made to the MI-BCI field during these past years. These publications have been done while the author has been coursing his PhD studies in the University of Seville. They consist of four publications, particularly three journal papers (one of them containing supplementary material) and a conference paper.

As a personal note, the author of the thesis would like to establish a chronological order or timeline, so the reader can gain some perspective about the work done to complete this thesis. Javier Olías started his studies in May of 2016, helping with the simulations and revisions of the paper [53], which has not been included in the present work because it is part of another thesis. During the following year 2017, the author worked in the review paper (publication D) which helped him gain more perspective and experience in the MI-BCI systems. Later, during the fall of 2018 he realized a stay of three month in Berlin in the group of Wojchiech Samek, this stay helped him to gain knoledge about Python coding, this knoledge and the work on the previous publications made possible for him to develop the work in Publication A, published in 2019. Also during 2019 he wrote and presented in the URSI congress the conference paper (publication C). As to publication E, which uses the preprocessing technique explained in Publication A, he worked on it also during the past year and was published a few months after Publication A.

In the following sections we will present the details of each publication and a summary with the contributions of each one of them, so the reader can obtain a previous knowledge about each topic before diving into the papers. At the end of each section we also include the original paper or publication, in case the interested reader wants to study them in more depth.

In this document the publications have not been sorted in a chronological order but according to the topics they are about. We start with publication A, which is the cornerstone of this thesis. In it we explain a new technique to normalize the EEG signal by accessing the power of the hidden EEG sources. The publication B is related to this topic and in it we show that with this normalization the effects of the artifacts in the EEG signals are considerably reduced. Afterwards, it is attached Publication C, a review paper in which the reader can find a complete review on CSP and its more important variants from the point of view of the information theory. Following this review we have attached the program of an oral exposition in an international workshop. Lastly we attach Publication D, which we consider to be of mayor importance too because it demonstrates how it is possible to perform the CSP algorithm in a blind context.

Publication A

EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimators

- **Title:** EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimator.
- **Authors:** Javier Olias; Rubén Martín-Clemente; M. Auxiliadora Sarmiento-Vega; Sergio Cruces.
- **DOI:** 10.1109/TNSRE.2019.2905894.
- **Published in:** IEEE Transactions on Neural Systems and Rehabilitation Engineering (Volume: 27 , Issue: 5).
- **Impact factor:** 3.972;
- **Quartile:** Q1 (Engineering – Biomedical, Rehabilitation)
- **Date:** May 2019.
- **Pages:** 895 - 904.
- **Publisher:** IEEE.

- **Abstract:** In brain-computer interfaces (BCIs), the typical models of the EEG observations usually lead to a poor estimation of the trial covariance matrices, given the high non-stationarity of the EEG sources. We propose the application of two

techniques that significantly improve the accuracy of these estimations and can be combined with a wide range of motor imagery BCI (MI-BCI) methods. The first one scales the observations in such a way that implicitly normalizes the common temporal strength of the source activities. When the scaling applies independently to the trials of the observations, the procedure justifies and improves the classical preprocessing for the EEG data. In addition, when the scaling is instantaneous and independent for each sample, the procedure particularizes to Tyler's method in statistics for obtaining a distribution-free estimate of scattering. In this case, the proposal provides an original interpretation of this existing method as a technique that pursuits an implicit instantaneous power-normalization of the underlying source processes. The second technique applies to the classifier and improves its performance through a convenient regularization of the features covariance matrix. Experimental tests reveal that a combination of the proposed techniques with the state-of-the-art algorithms for motor-imagery classification provides a significant improvement in the classification results.

Summary of Publication A

The paper "EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimators" was published in the journal with the highest impact factor and recognition in the field of BCI. This paper represents a relevant contribution to the MI-BCI algorithms because it proposes a method to improve many algorithms in a very simple way. In the following we expose the contributions and novelties that were introduced in this manuscript.

Contribution 1: *Proposal of an over-complete formulation for the EEG model.*

In section IV of the paper, we propose to transform the general EEG model ($\mathbf{X} = \mathbf{AS} + \mathbf{N}$) that has been studied in section 4.1, to a model that includes the noise as another EEG source, introducing the variables $\mathbf{A}'$ and $\mathbf{S}'$. That is:

$$\mathbf{X} = (\mathbf{A}\ \mathbf{I}) \begin{pmatrix} \mathbf{S} \\ \mathbf{N} \end{pmatrix} = \mathbf{A}'\mathbf{S}'. \tag{A.17}$$

This model is equivalent to the previous one but this notation allows us to simplify the model into two matrices: the sources matrix and the mixing matrix. However, using this notation introduces a linear indeterminacy between $\mathbf{A}'$ and $\mathbf{S}'$. We resolve this indeterminacy by assuming that the covariance matrix of the sources is equal to the identity matrix and therefore, the global covariance matrix of the observations is equal to:

$$\frac{1}{TN_\tau}\mathbf{S}'\mathbf{S}'^\top = \mathbf{I}, \tag{A.18}$$

$$\Sigma_\mathbf{X} = \frac{1}{TN_\tau}\mathbf{X}\mathbf{X}^\top = \frac{1}{TN_\tau}\mathbf{A}'\mathbf{S}'\mathbf{S}'^\top\mathbf{A}'^\top = \mathbf{A}'\mathbf{A}'^\top \tag{A.19}$$

where N_τ is the number of trials where each one of them contains T samples. In equation (A.17), we introduce a notation where the mixing matrix $\mathbf{A}'$ is wide and the sources matrix $\mathbf{S}'$ is tall. Therefore we will have orthogonal and aligned components to the mixing matrix: the orthogonal components will not contribute to the EEG observations while the components that are aligned will contribute to the EEG observations. We will refer to the latter as the effective sources $\tilde{\mathbf{S}}$ to rewrite the previous equation as

$$\mathbf{X} = \mathbf{A}'\tilde{\mathbf{S}}. \tag{A.20}$$

In the same way we can define the global power of the effective as

$$P_{\tilde{\mathbf{S}}} = \frac{1}{TN_\tau} \operatorname{tr}\left(\tilde{\mathbf{S}}\tilde{\mathbf{S}}^\top\right) = \operatorname{tr}(\mathbf{I}) = N_s \tag{A.21}$$

Nevertheless, for a more detailed explanation we recommend to check *Section IV* of the publication A.

Contribution 2: *Normalization by the power of the EEG effective sources.*

In Chapter 3 it was shown how the trials were usually normalized as in the equation (3.3) which we rewrite for the convenience of the reader

$$\mathbf{\Sigma}_\tau \quad \leftarrow \quad \frac{\mathbf{\Sigma}_\tau}{\operatorname{tr}(\mathbf{\Sigma}_\tau)}, \quad \forall \tau \in N_\tau. \tag{A.22}$$

We will be referring to this as the classic normalization. However, in the *section V* of the paper, we show that it is possible to access the power of the sources. To see this, we use equation (A.21) where we established that the global covariance of the sources is equal to the identity matrix and therefore, the power of the sources matches the number of sensors. Nevertheless, the singular trials do not need to satisfy this imposition since the power of the sources might fluctuate from a trial to another decreasing the precision of the covariance estimators. Therefore, in the same way as before, we can define the power of the sources in a trial as:

$$P_{\tilde{\mathbf{S}}_\tau} = \operatorname{tr}(\mathbf{\Sigma}_\tau) = \frac{1}{T} \operatorname{tr}\left(\tilde{\mathbf{S}}_\tau \tilde{\mathbf{S}}_\tau^\top\right). \tag{A.23}$$

This is important because in Chapter 3 we explained that a extended practice is to normalize the trials using the power of the observations. But in *subsection V.A* of this article, we suggest that this kind of normalization is useful but sub-optimal, since it is possible to normalize the trials using the power of the sources even though we cannot access the sources covariance since it is hidden by the mixing matrix.

To see how this is possible, we use the model in (A.20) to extract the sources from a trial and, keeping in mind that the matrix $\mathbf{A}'$ is wide, we have that:

$$\tilde{\mathbf{S}}_\tau = (\mathbf{A}')^\top \left(\mathbf{A}'\mathbf{A}'^\top\right)^{-1} \mathbf{X}_\tau. \tag{A.24}$$

Since we are only interested in $\operatorname{tr}\left(\tilde{\mathbf{S}}_\tau \tilde{\mathbf{S}}_\tau^\top\right)$ which is equal to $\operatorname{tr}\left(\tilde{\mathbf{S}}_\tau^\top \tilde{\mathbf{S}}_\tau\right)$, if we apply it to the previous equation we obtain:

$$P_{\tilde{\mathbf{S}}_\tau} = \operatorname{tr}\left(\tilde{\mathbf{S}}_\tau^\top \tilde{\mathbf{S}}_\tau\right) = \operatorname{tr}\left(\mathbf{X}_\tau^\top \left(\mathbf{A}'\mathbf{A}'^\top\right)^{-1} \mathbf{X}_\tau\right). \tag{A.25}$$

Using equation (A.19) we can conclude without any approximation that:

$$P_{\tilde{\mathbf{S}}_\tau} = \mathrm{tr}\left(\mathbf{X}_\tau^\top \left(\boldsymbol{\Sigma}_\mathbf{X}\right)^{-1} \mathbf{X}_\tau\right); \tag{A.26}$$

and instead of normalizing using the power of the observations, we can normalize using the power of the sources as follows:

$$\boldsymbol{\Sigma}_\tau \quad \leftarrow \quad \frac{\boldsymbol{\Sigma}_\tau}{P_{\tilde{\mathbf{S}}_\tau}}, \quad \forall \tau \in N_\tau. \tag{A.27}$$

Taking into account that the normalization parameter depends on the global covariance matrix, which at the same time is computed as the mean covariance among all the trials, we can see that they both depend on each other and, consequently, this can be implemented as an iterative algorithm, where we denote the global covariance matrix at iteration k as $\boldsymbol{\Sigma}_\mathbf{X}^{(k)}$.

Furthermore, we can perform this normalization sample by sample (instantaneous normalization) and express the iterative algorithm as:

1. Compute:

$$\boldsymbol{\Sigma}_\mathbf{X}^{(0)} = \sum_{\tau=1}^{N_\tau} \frac{\boldsymbol{\Sigma}_\tau}{\mathrm{tr}\left(\boldsymbol{\Sigma}_\tau\right)}$$

2. Repeat:

$$\boldsymbol{\Sigma}_\tau \quad \leftarrow \quad \sum_{t=1}^{T} \frac{\mathbf{x}_t \mathbf{x}_t^\top}{\mathbf{x}_t^\top \left(\boldsymbol{\Sigma}_\mathbf{X}^{(k-1)}\right)^{-1} \mathbf{x}_t} \tag{A.28}$$

$$\boldsymbol{\Sigma}_\mathbf{X}^{(k)} = \frac{1}{N_\tau} \sum_{\tau=1}^{N_\tau} \boldsymbol{\Sigma}_\tau.$$

Again, a more detailed explanation can be found in *publication A*.

Contribution 3: *Proof that instantaneous normalization leads to the Tyler scatter matrix estimator in the case of using one trial.*

In the case of just having a singular trial the proposed technique leads to the Tyler scatter matrix estimator. However, on the one hand Tyler addressed the problem from a statistical point of view and we address it from a source mixture model. On the other hand having a singular trial is never an option since it is needed at least one trial from each class to training the models and in any case we also normalize the test trials which does not fit with Tyler approach. The reader can find the explanation about the link between the two approaches in *section VI*.

Contribution 4: *Application of a novel shrinkage method for the covariances in the LDA classifier.*

In *section VII* of the paper we propose to apply the Chen shrinkage method as an alternative to the widely used Ledoit and Wolf shrinkage, since the Chen method has proven to be more accurate than the Ledoit and Wolf under the Gaussianity assumption which is made by LDA.

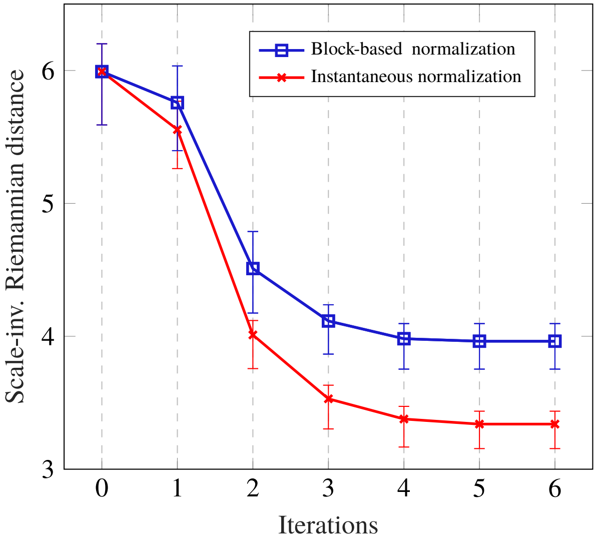


Figure A.1 Variations of the scale-invariant Riemannian distance (between the reference $\mathbf{C}_{C_k}$ and estimated $\mathbf{\Sigma}_{C_k}$ covariance matrices) with respect to the number of iterations of the proposed power-normalization procedures. The solid lines represent average distances while the bars represent the 25% and 75% percentiles. Iteration 0 refers to the absence of normalization, iteration 1 coincides with the standard trace-based normalization used in CSP, while the remaining iterations are instances of the proposed normalization.

Contribution 5: *Method to create artificial EEG data to experiment with.*

There are many circumstances in which it is very helpful to create artificial data to experiment with. We detail a methodology to create artificial data using information from real data in the *Apendix C* of the article.

Using this artificial data we perform a test that shows how the classical normalization is a suboptimal approach in comparison with the proposed normalization method. For the reader convenience in figure A.1 we reproduce the *figure 1* of the paper that shows the experiment results.

Contribution 6: *Proof that the new normalization technique improves the classical one.*

In *section VIII* we show the results obtained for a set of simulations that we performed over artificial and real data. In *table III* we show the average accuracy for each user using the same number of trials for training and testing, and in *figure 2* we plot the average result improvement as the number of training trials change. Also, in *figure 3* we show the results as the number of CSP filters change, proving an overall improvement in accuracy results. When looking at those figures and tables, it is important to note that even though the LDA and TSLR approaches use the same number of dimensionality reduction filters, the LDA variants use considerably less features since they only extract the diagonal elements of the matrix and the TSLR combination uses all the elements within the filtered covariance matrix.

Table A.1 Expected user accuracy for the *binary* MI classification problem in each of the considered datasets. The best performances are marked in bold. One can observe that in the majority of the cases the improvements obtained when combining the state-of-the-art methods with the proposed techniques can be regarded as statistically significant ($p\text{-}value < 5e\text{-}02$) (McNemar's).

		Classic	**State-of-the-art: CSP+...**			**Proposed: nCSP+...**		
Dataset	**User**	**CSP+LDA**	**sLDA**	**RMDM**	**TSLR**	**gLDA** (p-value)	**RMDM** (p-value)	**TSLR** (p-value)
III-3a	k3b	94.31	95.09	94.90	96.13	95.38 (1.7e-02)	94.97 (1.2e-01)	**96.31** (2.8e-02)
	k6b	75.77	77.77	77.23	78.94	78.78 (7.1e-04)	78.15 (7.7e-05)	**79.69** (1.0e-03)
	l1b	86.73	88.18	88.46	89.25	89.48 (1.6e-07)	89.90 (1.4e-13)	**90.29** (2.7e-08)
	– – –							
	mean	85.60	87.01	86.87	88.11	87.88 (4.0e-09)	87.68 (6.9e-13)	**88.76** (4.3e-09)
III-4a	aa	67.68	68.56	63.12	**72.37**	68.0 (8.6e-01)	63.62 (1.3e-01)	70.31 (1.0e+00)
	al	**96.81**	96.5	96.25	96.56	96.18 (9.3e-01)	95.75 (1.0e+00)	96.31 (9.8e-01)
	av	62.12	62.25	58.31	66.25	63.06 (8.7e-02)	59.75 (4.7e-03)	**67.87** (5.1e-03)
	aw	85.12	83.37	80.62	87.62	82.93 (8.7e-01)	80.75 (1.9e-01)	**87.75** (1.9e-01)
	ay	87.68	90.87	87.62	91.0	89.31 (1.0e+00)	87.68 (2.0e-01)	**91.62** (3.7e-02)
	– – –							
	mean	79.88	80.31	77.18	82.76	79.89 (9.4e-01)	77.51 (5.8e-02)	**82.77** (2.3e-01)
IV-2a	A01	86.97	88.18	88.15	89.02	89.0 (1.3e-06)	88.74 (1.0e-05)	**89.23** (3.2e-02)
	A02	73.15	74.63	76.20	**77.65**	75.20 (7.3e-03)	74.78 (1.0e+00)	76.15 (1.0e+00)
	A03	87.34	88.53	88.30	89.75	89.78 (1.1e-13)	90.08 (3.3e-28)	**90.60** (6.1e-11)
	A04	68.77	69.99	70.63	71.36	70.95 (3.0e-05)	70.93 (4.7e-02)	**71.38** (2.3e-01)
	A05	56.89	58.61	58.78	59.27	60.07 (4.9e-07)	**60.19** (5.5e-08)	59.82 (1.1e-02)
	A06	61.41	62.36	62.41	**63.32**	63.12 (2.6e-03)	62.81 (2.5e-02)	63.26 (8.1e-01)
	A07	88.75	89.92	90.44	91.09	91.20 (5.4e-17)	91.39 (5.5e-09)	**91.70** (9.0e-06)
	A08	86.40	87.65	86.33	88.32	88.73 (2.0e-10)	88.72 (8.2e-48)	**89.18** (2.1e-09)
	A09	82.70	83.95	82.64	84.25	84.67 (6.7e-05)	84.59 (7.5e-27)	**85.26** (2.5e-09)
	– – –							
	mean	76.93	78.20	78.21	79.34	79.19 (8.8e-39)	79.14 (6.0e-41)	**79.62** (6.4e-06)

Contribution 7: *We use the McNemar's test to check whether a result has significant improvements over a set of simulations.*

We have used one-sided test of hypothesis for paired data: McNemar's tests of hypothesis, paired Student's t-test and Wilcoxon signed-rank test, obtaining equivalent results with the three options. In the manuscript we have reported the p-values of McNemar's test (more specifically the mid-p values) because they have low type error I (i.e., small number of false positives) and are one of the preferred choices for the comparison of the statistical performance between classifiers [96].

Although in the article we only show the results of the McNemar's test (which is copied in here for the reader convenience), in this document we reproduce *table III* but using the Wilcoxon test and the one side T-test, so the reader can compare the three results.

Table A.2 Expected user accuracy for the binary MI classification problem in each of the considered datasets. The best performances are marked in bold. One can observe that in the majority of the cases the improvements obtained when combining the state-of-the-art methods with the proposed techniques can be regarded as statistically significant ($p\text{-}value < 5e\text{-}02$) (T-test).

		Classic		**State-of-the-art**			**Proposed**	
Dataset	**User**	**CSP+LDA**	**CSP+sLDA**	**CSP+RMDM**	**CSP+TSLR**	**nCSP+gLDA** (p-value)	**nCSP+RMDM** (p-value)	**nCSP+TSLR** (p-value)
III-3a	k3b	94.31	95.09	94.90	96.13	95.38 (3.9e-02)	94.97 (2.8e-01)	**96.31** (6.7e-02)
	k6b	75.77	77.77	77.23	78.94	78.78 (1.6e-03)	78.15 (1.8e-04)	**79.69** (2.3e-03)
	l1b	86.73	88.18	88.46	89.25	89.48 (4.2e-07)	89.90 (6.8e-13)	**90.29** (8.5e-08)
	——— mean	85.60	87.01	86.87	88.11	87.88 (9.4e-09)	87.68 (2.0e-12)	**88.76** (4.3e-09)
III-4a	aa	67.68	68.56	63.12	**72.37**	68.0 (7.1e-01)	63.62 (2.9e-01)	70.31 (9.9e-01)
	al	**96.81**	96.5	96.25	96.56	96.18 (8.2e-01)	95.75 (9.8e-01)	96.31 (9.2e-01)
	av	62.12	62.25	58.31	66.25	63.06 (1.9e-01)	59.75 (1.2e-02)	**67.87** (1.3e-02)
	aw	85.12	83.37	80.62	87.62	82.93 (7.2e-01)	80.75 (4.2e-01)	**87.75** (4.2e-01)
	ay	87.68	90.87	87.62	91.0	89.31 (9.9e-01)	87.68 (4.5e-01)	**91.62** (9.5e-02)
	——— mean	79.88	80.31	77.18	82.76	79.89 (8.8e-01)	77.51 (1.3e-01)	**82.77** (4.8e-01)
IV-2a	A01	86.97	88.18	88.15	89.02	89.0 (3.1e-06)	88.74 (2.5e-05)	**89.23** (6.9e-02)
	A02	73.15	74.63	76.20	**77.65**	75.20 (1.5e-02)	74.78 (1.0e+00)	76.15 (1.0e+00)
	A03	87.34	88.53	88.30	89.75	89.78 (3.3e-13)	90.08 (1.0e-99)	**90.60** (1.9e-10)
	A04	68.77	69.99	70.63	71.36	70.95 (6.5e-05)	70.93 (9.7e-02)	**71.38** (4.7e-01)
	A05	56.89	58.61	58.78	59.27	60.07 (1.1e-06)	**60.19** (1.2e-07)	59.82 (2.3e-02)
	A06	61.41	62.36	62.41	**63.32**	63.12 (5.4e-03)	62.81 (5.1e-02)	63.26 (6.0e-01)
	A07	88.75	89.92	90.44	91.09	91.20 (2.2e-16)	91.39 (1.4e-08)	**91.70** (2.1e-05)
	A08	86.40	87.65	86.33	88.32	88.73 (5.2e-10)	88.72 (1.0e-99)	**89.18** (5.7e-09)
	A09	82.70	83.95	82.64	84.25	84.67 (1.5e-04)	84.59 (1.0e-99)	**85.26** (6.4e-09)
	——— mean	76.93	78.20	78.21	79.34	79.19 (1.0e-99)	79.14 (1.0e-99)	**79.62** (1.3e-05)

Table A.3 Expected user accuracy for the binary MI classification problem in each of the considered datasets. The best performances are marked in bold. One can observe that in the majority of the cases the improvements obtained when combining the state-of-the-art methods with the proposed techniques can be regarded as statistically significant ($p\text{-}value < 5e\text{-}02$) (Wilcoxon).

		Classic		**State-of-the-art**			**Proposed**	
Dataset	**User**	**CSP+LDA**	**CSP+sLDA**	**CSP+RMDM**	**CSP+TSLR**	**nCSP+gLDA** (p-value)	**nCSP+RMDM** (p-value)	**nCSP+TSLR** (p-value)
III-3a	k3b	94.31	95.09	94.90	96.13	95.38 (7.9e-02)	94.97 (4.5e-01)	**96.31** (1.3e-01)
	k6b	75.77	77.77	77.23	78.94	78.78 (3.1e-03)	78.15 (3.7e-04)	**79.69** (4.6e-03)
	l1b	86.73	88.18	88.46	89.25	89.48 (8.6e-07)	89.90 (1.4e-12)	**90.29** (1.7e-07)
	——— mean	85.60	87.01	86.87	88.11	87.88 (1.9e-08)	87.68 (4.1e-12)	**88.76** (4.3e-09)
III-4a	aa	67.68	68.56	63.12	**72.37**	68.0 (5.9e-01)	63.62 (4.2e-01)	70.31 (9.8e-01)
	al	**96.81**	96.5	96.25	96.56	96.18 (6.3e-01)	95.75 (9.7e-01)	96.31 (8.4e-01)
	av	62.12	62.25	58.31	66.25	63.06 (3.8e-01)	59.75 (2.5e-02)	**67.87** (2.6e-02)
	aw	85.12	83.37	80.62	87.62	82.93 (5.6e-01)	80.75 (1.5e-01)	**87.75** (1.5e-01)
	ay	87.68	90.87	87.62	91.0	89.31 (9.8e-01)	87.68 (9.3e-02)	**91.62** (1.9e-01)
	——— mean	79.88	80.31	77.18	82.76	79.89 (7.6e-01)	77.51 (2.5e-01)	**82.77** (3.5e-02)
IV-2a	A01	86.97	88.18	88.15	89.02	89.0 (6.3e-06)	88.74 (4.9e-05)	**89.23** (1.4e-01)
	A02	73.15	74.63	76.20	**77.65**	75.20 (3.1e-02)	74.78 (1.0e+00)	76.15 (1.0e+00)
	A03	87.34	88.53	88.30	89.75	89.78 (6.8e-13)	90.08 (5.9e-27)	**90.60** (3.8e-10)
	A04	68.77	69.99	70.63	71.36	70.95 (1.3e-04)	70.93 (1.9e-01)	**71.38** (5.6e-02)
	A05	56.89	58.61	58.78	59.27	60.07 (2.1e-06)	**60.19** (2.5e-07)	59.82 (4.7e-02)
	A06	61.41	62.36	62.41	**63.32**	63.12 (1.1e-02)	62.81 (1.0e-01)	63.26 (7.9e-01)
	A07	88.75	89.92	90.44	91.09	91.20 (4.3e-16)	91.39 (2.8e-08)	**91.70** (4.3e-05)
	A08	86.40	87.65	86.33	88.32	88.73 (1.0e-09)	88.72 (1.7e-45)	**89.18** (1.1e-08)
	A09	82.70	83.95	82.64	84.25	84.67 (3.0e-04)	84.59 (9.2e-26)	**85.26** (1.3e-08)
	——— mean	76.93	78.20	78.21	79.34	79.19 (4.4e-38)	79.14 (3.2e-40)	**79.62** (2.7e-05)

EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimators

Javier Olias, Rubén Martín-Clemente, M^{a} Auxiliadora Sarmiento-Vega and Sergio Cruces

***Abstract*—In brain-computer interfaces the typical models of the EEG observations usually lead to a poor estimation of the trial covariance matrices, given the high non-stationarity of the EEG sources. We propose the application of two techniques that significantly improve the accuracy of these estimations and can be combined with a wide range of motor imagery BCI methods. The first one scales the observations in such a way that implicitly normalizes the common temporal strength of the source activities. When the scaling applies independently to the trials of the observations the procedure justifies and improves the classical preprocessing for the EEG data. Additionally, when the scaling is instantaneous and independent for each sample, the procedure particularizes to Tyler's method in statistics for obtaining a distribution-free estimate of scattering. In this case, the proposal provides an original interpretation of this existing method as a technique that pursuits an implicit instantaneous power-normalization of the underlying source processes. The second technique applies to the classifier and improves its performance through a convenient regularization of the features covariance matrix. Experimental tests reveal that a combination of the proposed techniques with state-of-the-art algorithms for motor-imagery classification provides a significant improvement in the classification results.**

***Index Terms*—Common spatial pattern, brain-computer interfaces, motor-imagery classification, covariance matrix estimation.**

I. INTRODUCTION

Brain-computer interfaces (BCI) have a great potential for enabling the communication between machine and humans by means of the analysis of the electroencephalographic activity. Nowadays, almost all the Motor Imagery BCI (MI-BCI) systems summarize most of the relevant information about the measurements in two kinds of covariance matrices: the covariance matrices of the filtered observations (employed for dimensionality reduction) and the covariance matrices of the features (which are required for classification). In the dimensionality reduction stage one tries to select those subspaces of the observations that retain most of the discriminative power, for instance, using the technique of Common Spatial Patterns (CSP) [1]. After that, the features are usually chosen as a non-linear transformation of the band-power statistics of the projected observations onto the previously selected subspaces [2]. The covariance matrices of these features (together with their class-conditional expectations) play a relevant role in the classification stage of MI-BCI [3]. Although CSP was only suitable for two-class classification problems, some later alternatives have been also proposed for multi-class settings (see, for instance, [4], [5]).

There are several sources of difficulty in the processing of EEG signals. Among them, we may cite: the inevitable presence of noise and interference at the sensors, the low spatial resolution of the BCI headsets [1], the possible presence of outliers in the measurements [6], the difficulty in gathering sufficient data trials for training [7], the need to determine the suitable number of features in those method that apply to dimensionality reduction [8], and the non-stationarity of the EEG signals [9]–[11].

The non-stationary can happen at different levels. The classical inter-subject and inter-session variabilities have been frequently addressed in the literature [11]. In this work, we will shift our attention to the less studied variabilities that happen between trials, and also within samples of the same trial. The signals generated by the brain are non-stationary in power at the trial and sample levels. We will show later that this power variability hinders the correct estimation of the covariance matrices of the trials, which are the most used statistics in the existing MI-BCI implementations. Our experimental results avail the hypothesis that the correction of this EEG signal variability leads to improved covariance matrix estimates, which allow transversal improvements in accuracy for the tested classification algorithms.

The main contributions of the article are the following:

- We show that the standard power normalization of the observations, which is widely used in the preprocessing of the EEG data for MI-BCI, is useful but suboptimal.
- We propose the power-normalization of the effective EEG source activities. This normalization has no hyper-parameters and, in general, improves the quality of the covariance matrix estimates during training and testing.
- The shrinkage of the feature covariance matrices in MI-BCI was shown to be beneficial when the number of training trials is small [12]. We propose the application of an alternative shrinkage estimate (gLDA) that is based on the Gaussianity of the features [13].

Our experimental results confirm that the proposed power-normalization and gLDA implementation lead to a transversal improvement in the performance of the existing MI-BCI algorithms. In addition, the proposal seems to be much less sensitive with respect to the number of features employed in the dimensionality reduction stage.

The article is organized as follows. Section II introduces the basic model of the EEG measurements and section III discusses some classical and state-of-the-art approaches for

All the authors are with the Department of Teoría de la Señal y Comunicaciones, Universidad de Sevilla, Camino de los Descubrimientos s/n, Seville 41092, Spain. E-mails: {folias, ruben, sarmiento, sergio}@us.es

This work has been led by the corresponding author: Dr. Cruces.

This work was supported in part by the Spanish Government under MINECO grant TEC2017-82807-P.

MI-BCI. Section IV presents an overcomplete model of the observations and defines the effective sources of the mixture. Section V describes the proposal for the normalization in power of the EEG sources and also analyzes its links with the standard preprocessing of the observations. This method is extended in section VI to the case of the instantaneous power-normalization of the effective sources. Section VII presents some variations in the implementations of the classifier using shrinkage estimates of the feature covariance matrices. The experimental results are provided and discussed in section VIII, while section IX is devoted to the conclusions.

II. Basic model of the EEG observations

The EEG headset is based on an array of sensors that measures the electromagnetic activity on the scalp. At time t, the variations of the activities of the sensors are measured with respect to a given referential system (see EEG referencing in [1]) and passband filtered to retain the 8Hz-32Hz band. After that, they are centered at the origin by subtracting the estimated mean of each trial and collected into the observation vector $\boldsymbol{x}(t) = [x_1(t), \ldots, x_{N_x}(t)]^T \in \mathbb{R}^{N_x}$.

The physiological nature of the problem allows one to model the i^{th}-element of the observation vector as a superposition of contributions from: some desired latent EEG source activities $s_j(t),\ j = 1, \ldots, N_s$, and some filtered additive interference or noise component which we denote by $n_j(t),\ j = 1, \ldots, N_x$. We will not assume any specific value for N_s which, depending on the experiment, could be greater or lower than N_x. The contribution of i^{th}-source $s_j(t)$ to the j^{th}-observation $x_j(t)$ is modeled as $a_{ij}s_j(t)$, where the factor a_{ij} refers to the attenuation of the almost instantaneous propagation of the source activity to the sensor position. In vector form, the filtered observations are known to follow the linear instantaneous mixing model [14]

$$\boldsymbol{x}(t) \quad = \quad \mathbf{A}\mathbf{s}(t) + \boldsymbol{n}(t)\ , \tag{1}$$

where $\mathbf{A} = [a_{ij}]_{ij} \in \mathbb{R}^{N_x \times N_s}$ refers to the mixing matrix.

In those cases where we would like to make explicit the trial to which the observations belong to, we will use the notation $\boldsymbol{x}_\tau(t)$ that refers to the vector of observations of the trial τ at time t. The global power of the non-stationary process of filtered observations is defined by

$$P_{\boldsymbol{x}} \equiv \langle E[\|\boldsymbol{x}_\tau(t)\|^2] \rangle_{t,\tau} = \frac{1}{N_\tau T} \sum_{t=1}^{T} \sum_{\tau=1}^{N_\tau} E[\|\boldsymbol{x}_\tau(t)\|^2]. \tag{2}$$

A column-wise concatenation of the observed vector samples from a trial results in the matrix model of the observations

$$\mathbf{X}_\tau \quad = \quad \mathbf{A}\mathbf{S}_\tau + \mathbf{N}_\tau\ , \tag{3}$$

where $\mathbf{X}_\tau, \mathbf{N}_\tau \in \mathbb{R}^{N_x \times T}$ and $\mathbf{S}_\tau \in \mathbb{R}^{N_s \times T}$. In the following, the class of a trial τ will be denoted by $c(\tau) \in \{c_1, \ldots, c_K\}$.

III. The common spatial patterns and other successful approaches for MI-BCI

The Common Spatial Patterns (CSP) is a method designed for the case of having two classes ($K = 2$) [15]. Let the class-conditional covariance matrices of the classes be $\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_1}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_2}}$. The CSP algorithm (see Table I) tries to reduce the dimensionality of the observations by finding a p-dimensional subspace for which the two classes are maximally separated in a certain divergence sense [6], [16]. This goal is achieved by setting the $p < N_x$ spatial filters $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ (for the sake of simplicity p is assumed to be even) equal to the $p/2$ principal and $p/2$ minor eigenvectors of the following generalized eigenvalue problem

$$\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_1}}\boldsymbol{w} = \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_2}}\boldsymbol{w}\ \lambda\ . \tag{4}$$

The selected eigenvectors are grouped in the matrix of spatial filters $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p]$, which is used to perform the dimensionality reduction of the observations

$$\mathbf{Y}_\tau = \mathbf{W}^T\mathbf{X}_\tau \in \mathbb{R}^{p \times T}. \tag{5}$$

There are several possible extension of CSP to multi-class ($K > 2$) scenarios. Some are based on the joint approximated diagonalization of the covariances matrices of the observations for each class [14]

$$\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_k}}\mathbf{W} = \mathbf{D}_{c_k} \quad k = 1, \ldots, K, \tag{6}$$

where $\mathbf{D}_{c_k}$ refers to an approximately diagonal matrix. The one proposed in [4] combines this approximated diagonalization with a method to choose the most relevant filters based on an Information Theoretic Feature Extraction criterion (ITFE).

Although the dimensionality reduction stage (implemented by CSP and ITFE) sometimes is omitted, in general, as we will see in the simulations, it is a recommended procedure for datasets with moderate or relatively large number of sensors.

After the dimensionality reduction, some basic linear classification results can be obtained using Fisher's Linear Discriminant Analysis (LDA). However, other state-of-the-art proposals are nowadays preferable. This is the case of sLDA [7] a shrinkage variant of LDA and also of the classifiers that exploit the Riemmanian geometry of the manifold of symmetric and positive definite (SPD) matrices. Among these classifiers, we can mention the Riemannian Minimum Distance to Mean (RMDM) [5], which is based on the minimization of the Riemmanian distance between the sample covariance matrices of the test trials and the Riemmanian mean of the classes. Other improved classification methods are obtained by using as features the projection of the sample covariance matrices onto the tangent space (of the Riemmanian SPD manifold) at the referential Riemmanian mean of the set of covariance matrices. In this way, an LDA classifier applied to tangent space (TS) features give rise to a TSLDA implementation. Similarly, the logistic regression (LR) classification of TS features leads to a TSLR implementation [17]. The interested reader in Riemmanian approaches for Brain-Computer Interfaces can find in [17] and [18] respective tutorial reviews on this topic.

A. Classical estimation of the class covariance matrices

As the observations have been already centered, the EEG spatial covariance matrix of trial τ is given by

$$\mathbf{C}^{(0)}_{\mathbf{X}_\tau} \quad = \quad \frac{1}{T}\mathbf{X}_\tau\mathbf{X}_\tau^T. \tag{7}$$

The notation $\mathbf{C}^{(i)}_{\mathbf{X}_\tau}$ is adopted in this paper in order to allow the possibility to refine this estimate through additional iterations. Then, since the trials may have unequal power, the standard CSP implementation [1] normalizes the EEG covariance matrices as

$$\mathbf{C}^{(1)}_{\mathbf{X}_\tau} = \frac{\mathbf{C}^{(0)}_{\mathbf{X}_\tau}}{\mathrm{Tr}\{\mathbf{C}^{(0)}_{\mathbf{X}_\tau}\}/N_x} \equiv \frac{\mathbf{X}_\tau \mathbf{X}_\tau^T}{\frac{1}{N_x}\mathrm{Tr}\{\mathbf{X}_\tau \mathbf{X}_\tau^T\}}. \qquad (8)$$

One may note that this definition only differs from the classical normalization in the following irrelevant $\frac{1}{N_x}$ scaling term, which is mainly adopted here for notational convenience.

Finally, the class-conditional covariance matrices are usually estimated by means of the arithmetic mean of the trials

$$\hat{\mathbf{\Sigma}}^{(1)}_{\mathbf{x}|\mathbf{c_k}} = \frac{1}{N_{c_k}} \sum_{\tau : c(\tau)=c_k} \mathbf{C}^{(1)}_{\mathbf{X}_\tau} \qquad k = 1, \ldots, K. \qquad (9)$$

In the following sections, we propose an alternative normalization for the training and test covariance matrices. It has no additional hyperparameters and, in general, outperforms the standard one considered in (8). In particular, we will show that the standard normalization can be regarded as a first approximation to the proposed approach.

At this point, it is worth to comment other estimators for $\mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}$ which have been suggested for MI-BCI applications according to various strategies. The adaptation with respect to differences between the training and testing distributions of the data has been considered in [19], which suggests the weighting of the samples according to their estimated importance. In [20], class covariance matrices estimators of minimum β-divergence for a Wishart model have been proposed to ensure the robustness with respect to data outliers. The solution, which is based on an iteratively weighting of the trial covariance matrices of each class, uses cross-validation (CV) for the determination of the hyper-parameter of the divergence. The use of CV is also required in [21], which proposed several regularized covariance matrix estimates with the aim to avoid overfitting. One regularized estimate, which has the remarkable advantage of avoiding CV, was proposed by Ledoit and Wolf in [22].

IV. Overcomplete model of the observations and effective component of the sources

The linear mixing model of equation (1) provides an overcomplete representation of the observations. This can be seen by integrating the noise/interference contribution into an extended sources vector $\mathbf{s}'_\tau(t)$ to obtain

$$\boldsymbol{x}_\tau(t) = (\mathbf{A}\ \mathbf{I}) \begin{pmatrix} \mathbf{s}_\tau(t) \\ \boldsymbol{n}_\tau(t) \end{pmatrix} = \mathbf{A}'\, \mathbf{s}'_\tau(t). \qquad (10)$$

Moreover, there is an inherent linear indeterminacy between the sources and the columns of the mixing matrix. In this sense, note that, for any arbitrary invertible matrix $\mathbf{M} \in \mathbb{R}^{(N_s+N_x)\times(N_s+N_x)}$, the model satisfies

$$\boldsymbol{x}_\tau(t) = \mathbf{A}'\, \mathbf{s}'_\tau(t) = (\mathbf{A}'\mathbf{M}^{-1})\,(\mathbf{M}\mathbf{s}'_\tau(t)). \qquad (11)$$

We avoid this indeterminacy by assuming, from here on, that the global covariance matrix of the source signal process is equal to the identity matrix. As we initially considered the centering of the observations, this matrix is then given by $\mathbf{\Sigma}'_\mathbf{s} = \langle E[\mathbf{s}'_\tau(t)(\mathbf{s}'_\tau(t))^T]\rangle_{t,\tau} = \mathbf{I}$, and the global covariance matrix of the observations is

$$\mathbf{\Sigma}_\mathbf{x} = \langle E[\boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T]\rangle_{t,\tau} = \mathbf{A}'\mathbf{A}'^T\ . \qquad (12)$$

The fact that the resulting mixing matrix $\mathbf{A}' \in \mathbb{R}^{N_x\times(N_s+N_x)}$ is wide and of rank N_x, implies that not all the components of the extended vector of sources $\mathbf{s}'(t)$ will contribute to the observations. Only the component of the sources that is aligned with the range space of the rows of $\mathbf{A}'$ will have an effective contribution, while the orthogonal component to this subspace will be discarded. To see this, consider the orthogonal decomposition of the extended sources

$$\mathbf{s}'_\tau(t) = \mathbf{\Pi}_{\mathbf{A}'^T}\, \mathbf{s}'_\tau(t) + \mathbf{\Pi}^\perp_{\mathbf{A}'^T}\, \mathbf{s}'_\tau(t), \qquad (13)$$

where the proyection matrix onto the rows of the extended mixing matrix is $\mathbf{\Pi}_{\mathbf{A}'^T} = \mathbf{A}'^T(\mathbf{A}'\mathbf{A}'^T)^{-1}\mathbf{A}'$ and the orthogonal projection matrix is given by $\mathbf{\Pi}^\perp_{\mathbf{A}'^T} = \mathbf{I} - \mathbf{\Pi}_{\mathbf{A}'^T}$. Since these projection matrices satisfy $\mathbf{A}'\mathbf{\Pi}_{\mathbf{A}'^T} = \mathbf{A}'$ and $\mathbf{A}'\mathbf{\Pi}^\perp_{\mathbf{A}'^T} = \mathbf{0}$, it is easily observed that

$$\boldsymbol{x}_\tau(t) = \mathbf{A}'\, \mathbf{s}'_\tau(t) = \mathbf{A}'\, \tilde{\mathbf{s}}_\tau(t) \qquad (14)$$

where $\tilde{\mathbf{s}}_\tau(t) = \mathbf{\Pi}_{\mathbf{A}'^T}\mathbf{s}'_\tau(t)$ represents the *effective sources*, i.e., the component of the extended sources with a non-negligible influence in the value of the observations $\boldsymbol{x}_\tau(t)$. Moreover, it is straightforward to check that the global covariance matrix of the effective sources coincides with the following projection matrix $\mathbf{\Sigma}_{\tilde{\mathbf{s}}} = \Pi_{\mathbf{A}'^T}$, which is unitary similar to the identity matrix of dimension N_x. Hence, the global power of the effective sources is

$$P_{\tilde{\mathbf{s}}} = \mathrm{Tr}\{\mathbf{\Sigma}_{\tilde{\mathbf{s}}}\} = \mathrm{Tr}\{\mathbf{I}_{N_x}\} = N_x. \qquad (15)$$

V. Normalization of the power of the sources

Let's define the power of the effective sources for trial τ as

$$P_{\tilde{\mathbf{S}}_\tau} = \mathrm{Tr}\{\mathbf{C}_{\tilde{\mathbf{S}}_\tau}\} = \frac{1}{T}\,\mathrm{Tr}\{\tilde{\mathbf{S}}_\tau \tilde{\mathbf{S}}_\tau^T\}. \qquad (16)$$

When $\tilde{\mathbf{S}}_\tau$ for $\tau = 1, \ldots, N_\tau$ have dissimilar powers, their contribution to the class-conditional covariance matrices is not homogeneous. In this situation, a fraction of the trials may dominate the estimation, implying a higher variance in the estimates.

The covariance normalization by the power of the observations $P^{(0)}_{\mathbf{X}_\tau} = \mathrm{Tr}\{\mathbf{C}^{(0)}_{\mathbf{X}_\tau}\}/N_x$ in (8) only partially alleviates the previous effect, since, due to the equivalence $P^{(0)}_{\mathbf{X}_\tau} = \mathrm{Tr}\{\mathbf{A}\mathbf{C}^{(0)}_{\tilde{\mathbf{S}}_\tau}\mathbf{A}^T\}$, it depends on the interaction between the mixing matrix and the trial covariance of the sources. Instead, we propose to equalize the power of the effective sources in each trial in such a way that they all coincide with the global power of the process, which was defined in (15). Although we don't have direct access to $\tilde{\mathbf{S}}_\tau$, we explain in the sequel a method that allows us to iteratively equalize its power, contributing in this way to obtain more reliable estimates of the covariance matrices.

TABLE I
PSEUDO-CODE OF CSP+LDA ALGORITHM FOR MI-BCI.

```
PREPROCESSING FOR TRAINING & TESTING
Freq. filtering & centering of the data $\forall\tau$.
Compute $\mathbf{C}^{(0)}_{\mathbf{X}_\tau}$ and $\mathbf{C}^{(1)}_{\mathbf{X}_\tau}$, $\forall\tau$, using (7)-(8).
Determine $\hat{\mathbf{\Sigma}}^{(1)}_{\mathbf{x}|\mathbf{c_1}}$ and $\hat{\mathbf{\Sigma}}^{(1)}_{\mathbf{x}|\mathbf{c_2}}$ with (9).

METHOD FOR DIM. REDUCTION (STANDARD-CSP)
% Obtain the spatial filters solving (4)
  $[\mathbf{V},\mathbf{D}] = eig(\hat{\mathbf{\Sigma}}^{(1)}_{\mathbf{x}|\mathbf{c_1}}, \hat{\mathbf{\Sigma}}^{(1)}_{\mathbf{x}|\mathbf{c_2}})$
% Sort the $N_x$ solutions
  $[\sim, ind] = sort(diag(\mathbf{D}))$, $\mathbf{V} = \mathbf{V}(:, ind)$,
% Select the extreme $p$ eigenvectors
  $\mathbf{W} = [\mathbf{V}(:, 1:p/2), \mathbf{V}(:, N_x - p/2) + 1 : N_x]$
% Spatial filtering
  $\mathbf{C}_{\mathbf{Y}_\tau} = \mathbf{W}^T \mathbf{C}^{(0)}_{\mathbf{X}_\tau} \mathbf{W}, \quad \tau = 1, \ldots, N_\tau.$
% Transf. for obtaining normal-like features
  $\mathbf{f}_\tau = \log(\mathrm{diag}(\mathbf{C}_{\mathbf{Y}_\tau})/\mathtt{sum}(\mathrm{diag}(\mathbf{C}_{\mathbf{Y}_\tau})))$        (F1)

LEARNING THE LDA (BINARY) CLASSIFIER
% Using the training pairs $(c(\tau), \mathbf{f}_\tau)$
  $\boldsymbol{\mu}_k = \frac{1}{N_{c_k}} \sum_{\tau:c(\tau)=c_k} \mathbf{f}_\tau$ for $k = 1, 2.$
  $\hat{\mathbf{\Sigma}}_{\mathbf{f}|\mathbf{c_k}} = \frac{1}{N_{c_k}} \sum_{\tau:c(\tau)=c_k} (\mathbf{f}_\tau - \boldsymbol{\mu}_k)(\mathbf{f}_\tau - \boldsymbol{\mu}_k)^T, \quad k = 1, 2.$
  $p(c_k) = N_{c_k}/(N_{c_1} + N_{c_2}), \quad k = 1, 2.$
  $\hat{\mathbf{\Sigma}}_{\mathbf{f}} = p(c_1)\hat{\mathbf{\Sigma}}_{\mathbf{f}|\mathbf{c_1}} + p(c_2)\hat{\mathbf{\Sigma}}_{\mathbf{f}|\mathbf{c_2}}$
  $\boldsymbol{\alpha} \ = \hat{\mathbf{\Sigma}}_{\mathbf{f}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$
  $\boldsymbol{\beta} \ = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{\log p(c_1) - \log p(c_2)}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \hat{\mathbf{\Sigma}}_{\mathbf{f}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

CLASSIFICATION OF TEST DATA
% Implementation using LDA with equal cov.
  $\mathbf{C}_{\mathbf{Y}_\tau} = \mathbf{W}^T \mathbf{C}^{(0)}_{\mathbf{X}_\tau} \mathbf{W}, \quad \tau \in Set\ of\ test\ trials.$
  Evaluate $\mathbf{f}_\tau$ using the same formula as in (F1)
  $\hat{c}(\tau) = c_k \ \ where \ \ k = 1.5 + 0.5\,\mathrm{sign}(\boldsymbol{\alpha}^T(\mathbf{f}_\tau - \boldsymbol{\beta})).$
```

Consider the notation for the inner product between two symmetric positive definite matrices of dimension N_x,

$$\langle \mathbf{C}_{\mathbf{X}_\tau} \ , \ \mathbf{\Sigma}_{\mathbf{x}}^{-1} \rangle = \frac{1}{N_x} \mathrm{Tr}\{\mathbf{C}_{\mathbf{X}_\tau} \mathbf{\Sigma}_{\mathbf{x}}^{-1}\}. \tag{17}$$

Lemma 1 (Power of the effective sources): The power of the effective sources for each trial τ is given by the scaled inner product between the covariance matrix of the trial and the inverse of the global covariance matrix of the observations

$$P_{\tilde{\mathbf{S}}_\tau} = N_x \ \langle \mathbf{C}_{\mathbf{X}_\tau} \ , \ \mathbf{\Sigma}_{\mathbf{x}}^{-1} \rangle. \tag{18}$$

This lemma, which provides an exact formula for the evaluation of the power of the effective sources, is proved in Appendix A. However, the determination of $\mathbf{\Sigma}_{\mathbf{x}} = \langle E[\boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T] \rangle_{t,\tau}$ involves an expectation operation and, as a consequence, is not feasible. Instead, we can estimate it from the available samples at a given iteration $i - 1$.

Under a Gaussian mixture model for the observations, a natural estimate of $\mathbf{\Sigma}_{\mathbf{x}}$ (built from a combination of maximum likelihood estimates) is given by the arithmetic mean of the covariances of the trials

$$\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)} = \langle \mathbf{C}^{(i-1)}_{\mathbf{X}_\tau} \rangle_\tau = \frac{1}{N_\tau} \sum_{\tau=1}^{N_\tau} \mathbf{C}^{(i-1)}_{\mathbf{X}_\tau}. \tag{19}$$

After substituting $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)}$ for $\mathbf{\Sigma}_{\mathbf{x}}$ in (18), the estimated power of the effective sources at iteration $i - 1$ is $\hat{P}^{(i-1)}_{\tilde{\mathbf{S}}_\tau}$. The ratio between the power of the effective sources at iteration $(i - 1)$ and their global power (obtained in (15)) is given by

$$(\hat{\sigma}_\tau^{(i-1)})^2 \equiv \frac{1}{N_x} \hat{P}^{(i-1)}_{\tilde{\mathbf{S}}_\tau} = \langle \mathbf{C}^{(0)}_{\mathbf{X}_\tau} \ , \ (\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)})^{-1} \rangle. \tag{20}$$

In order to equalize the power across trials at the i^{th} iteration, we should normalize the observations as

$$\mathbf{X}_\tau^{(i)} \ = \ \mathbf{X}_\tau / \hat{\sigma}_\tau^{(i-1)}, \tag{21}$$

since this scaling replaces the estimated power $\hat{P}^{(i-1)}_{\tilde{\mathbf{S}}_\tau}$ of the effective sources in each trial by the global average power $P_{\tilde{\mathbf{S}}} = N_x$. After that, the scaled observations $\mathbf{X}_\tau^{(i)}$ lead to normalized estimates of the trial covariance matrices

$$\mathbf{C}^{(i)}_{\mathbf{X}_\tau} = \frac{1}{T} \mathbf{X}_\tau^{(i)} (\mathbf{X}_\tau^{(i)})^T = (\hat{\sigma}_\tau^{(i-1)})^{-2} \, \mathbf{C}^{(0)}_{\mathbf{X}_\tau} \quad \forall\tau \tag{22}$$

and to an improved estimate of the global covariance matrix

$$\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)} = \langle \mathbf{C}^{(i)}_{\mathbf{X}_\tau} \rangle_\tau = \frac{1}{N_\tau} \sum_{\tau=1}^{N_\tau} \mathbf{C}^{(i)}_{\mathbf{X}_\tau}. \tag{23}$$

This new estimate can still help in improving the normalization of the sources, so the estimation procedure can continue in a recursive manner until the relative variation in the estimate of the global covariance matrix falls below a tolerance threshold ϵ. For instance, by continuing with the iteration until the following condition is met: $\|\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)} - \hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)}\|_F / \|\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)}\|_F < \epsilon$. After the convergence of the iteration, the following average covariance matrices of each class are used as inputs to the method of dimensionality reduction

$$\hat{\mathbf{\Sigma}}^{(i)}_{\mathbf{x}|\mathbf{c_k}} = \frac{1}{N_{c_k}} \sum_{\tau:c(\tau)=c_k} \mathbf{C}^{(i)}_{\mathbf{X}_\tau} \qquad k = 1, \ldots, K. \tag{24}$$

A. Expliciting the link with the standard preprocessing of the EEG observations

At iteration $i = 0$, before having access to the observed data, we may consider an initial isotropic estimate for the covariance matrix of the observations $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(0)} = \mathbf{I}$. Hence, the estimates of the covariance matrices in (22) are, for $i = 1$, equal to

$$\mathbf{C}^{(1)}_{\mathbf{X}_\tau} = (\hat{\sigma}_\tau^{(0)})^{-2} \, \mathbf{C}^{(0)}_{\mathbf{X}_\tau} = \frac{\mathbf{X}_\tau \mathbf{X}_\tau^T}{\frac{1}{N_x} \mathrm{Tr}\{\mathbf{X}_\tau \mathbf{X}_\tau^T\}} \quad \forall\tau, \tag{25}$$

which exactly coincide with those provided by the standard normalization of the trials in (8). Next, the global covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(1)}$ is evaluated using (23) and used, in another iteration ($i = 2$), to improve the normalization of the observations in each trial. Then, the new trial covariance matrices are

$$\mathbf{C}^{(2)}_{\mathbf{X}_\tau} = (\hat{\sigma}_\tau^{(1)})^{-2} \, \mathbf{C}^{(0)}_{\mathbf{X}_\tau} = \frac{\mathbf{C}^{(0)}_{\mathbf{X}_\tau}}{\langle \mathbf{C}^{(0)}_{\mathbf{X}_\tau} \ , \ (\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(1)})^{-1} \rangle} \quad \forall\tau \tag{26}$$

and again the new global covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(2)}$ is evaluated. One can continue with the iterations of the procedure until it convergences. In the section of simulations, we will later illustrate with a controlled experiment (see Figure 1) the improvement of the estimates of the trial covariance matrices with respect to the number of iterations.

Although we have previously suggested the initialization of the iteration with $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(0)} = \mathbf{I}$ for revealing the link between the proposal and the classical preprocessing of CSP, in practice, it is better to choose as initial estimate the sample covariance matrix of the trials $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(0)} =< \mathbf{C}_{\mathbf{X}_\tau}^{(0)} >_\tau$. This latter estimate is more informative than the identity matrix, which contributes to a faster convergence of the iteration.

VI. Instantaneous power normalization leads to an existing estimator of scatter

Until now, in order to illustrate the links of the proposed power-normalization with the preprocessing used in classical CSP, we have only addressed the equalization of the power across trials. However, the technique is easily extended for equalizing the power of the sources over temporal juxtaposed (or overlapped) windows of arbitrary length. For signals like the EEG sources, which are highly non-stationary, one can improve the estimates of the covariance matrices by equalizing the power across samples, i.e., considering windows of one sample length.

Let us consider the instantaneous correlation matrix estimate of the observations at the trial τ and time t

$$\mathbf{C}_{\boldsymbol{x}_\tau(t)}^{(0)} = \boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T, \tag{27}$$

which is based on a single sample. In similarity with (20), given $\hat{\mathbf{\Sigma}}_{\mathbf{x}}$ at iteration $(i-1)$, we obtain the power ratio for each trial and time sample

$$\begin{aligned}
(\hat{\sigma}_{\tau,t}^{(i-1)})^2 &\equiv \langle \mathbf{C}_{\boldsymbol{x}_\tau(t)}^{(0)} \, , \, (\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)})^{-1} \rangle & (28)\\
&= \frac{1}{N_x} \operatorname{Tr}\{\boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T(\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)})^{-1}\} & (29)\\
&= \frac{1}{N_x} \operatorname{Tr}\{(\boldsymbol{x}_\tau(t))^T(\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)})^{-1}\boldsymbol{x}_\tau(t)\}. & (30)
\end{aligned}$$

Its evaluation with (30), is recommended in the instantaneous case because of the computational advantages over (29).

The instantaneous power-normalization of $\tilde{\mathrm{s}}_\tau(t)$ is simply obtained by scaling the observations

$$\boldsymbol{x}_\tau^{(i)}(t) \quad = \quad \boldsymbol{x}_\tau(t)/\hat{\sigma}_{\tau,t}^{(i-1)} \qquad \forall\ \tau, t. \tag{31}$$

Therefore, the covariance matrices estimates of each trial

$$\mathbf{C}_{\mathbf{X}_\tau}^{(i)} \quad = \quad \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_\tau^{(i)}(t)(\boldsymbol{x}_\tau^{(i)}(t))^T \tag{32}$$

are, in general, more reliable and contribute, using (23), to an improved estimation of the averaged covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)}$. The whole iteration over the set of training trials is summarized in the top part of Table II. The procedure for the estimation of the covariance matrix of the test trials, which is shown in the second part of Table II, is coherent with the updates performed in the last iteration for the training trials.

TABLE II

Pseudocode of the instantaneous power-normalization, which particularizes to a version of Tyler's method in statistics for obtaining a robust estimator of scatter.

```
PREPROCESSING FOR TRAINING TRIALS
Freq. filtering & centering of the data $\forall \tau$.
Set   i=0, $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(0)} =< \mathbf{C}_{\mathbf{X}_\tau}^{(0)} >_\tau$
Repeat
  i=i+1
  $(\hat{\sigma}_{\tau,t}^{(i-1)})^2 = \frac{1}{N_x} \operatorname{Tr}\{(\boldsymbol{x}_\tau(t))^T(\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)})^{-1}\boldsymbol{x}_\tau(t)\}\ \forall \tau, t$
  $\mathbf{C}_{\mathbf{X}_\tau}^{(i)} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\sigma}_{\tau,t}^{(i-1)})^{-2}\ \boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T\ \forall \tau$
  $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)} = \frac{1}{N_\tau}\sum_{\tau=1}^{N_\tau} \mathbf{C}_{\mathbf{X}_\tau}^{(i)}$
Until $\|\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)} - \hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)}\|_F / \|\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)}\|_F < \epsilon$

Return $\mathbf{C}_{\mathbf{X}_\tau}^{(i)}$ $\forall \tau$ and
  $\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}^{(i)} = \frac{1}{N_{c_k}}\sum_{\tau:c(\tau)=c_k} \mathbf{C}_{\mathbf{X}_\tau}^{(i)} \quad k = 1, \dots, K.$

PREPROCESSING FOR A TESTING TRIAL $\tau$
Freq. filtering & centering of the trial.
Given the last used estimate $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)}$ in training...
Evaluate $(\hat{\sigma}_{\tau,t}^{(i-1)})^2 = \frac{1}{N_x} \operatorname{Tr}\{(\boldsymbol{x}_\tau(t))^T(\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)})^{-1}\boldsymbol{x}_\tau(t)\}\ \forall t$
Return $\mathbf{C}_{\mathbf{X}_\tau}^{(i)} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\sigma}_{\tau,t}^{(i-1)})^{-2}\ \boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T\ \forall \tau$

*Note: after this preprocessing the evaluation of
 the features no longer needs normalization, i.e.,
 $\mathbf{f}_\tau = \log(\operatorname{diag}(\mathbf{C}_{\mathbf{Y}_\tau}))$                                (F2)
```

The combination of the instantaneous power-normalization with CSP will be referred, hereinafter, as nCSP. However, since the proposed normalization aims to recover the stationarity in power of the effective sources vector, it is unnecessary to apply any additional normalization of the features. Hence, the recommended evaluation of the features is simply given by formula (F2) of Table II, i.e., $\mathbf{f}_\tau = \log(\operatorname{diag}(\mathbf{C}_{\mathbf{Y}_\tau}))$, which replaces all the instances of formula (F1) in Table I.

In Appendix B, we discuss the link between the instantaneous power-normalization iteration and the method proposed by Tyler in [23] for obtaining a distribution-free estimator of scatter within the class of elliptically distributed data. Both techniques use complementary arguments that arrive at a similar final result. However, Tyler's method assumes that the observations are drawn from an elliptical distribution, a hypothesis that may be true for a single trial ($N_\tau = 1$) and a unique class ($K = 1$). For multiple classes, the previous hypothesis can no longer be true, whereas the proposal based on the power-normalization of the effective sources still provides an admissible statistical interpretation for the iteration.

VII. Gaussian Shrinkage LDA

Under the hypotheses of p-dimensional Gaussian features for each class, with means $\boldsymbol{\mu}_k$, $k = 1, 2$, and homoscedastic covariance matrices $\mathbf{\Sigma}_{\mathbf{f}|c_1} = \mathbf{\Sigma}_{\mathbf{f}|c_2} = \mathbf{\Sigma}_{\mathbf{f}}$, the LDA classifier considered in the Table I implements the maximum a posteriori (MAP) Bayesian classification [24]. However, when the number of feature vectors $\mathbf{f}_\tau$ for training is not sufficiently large with respect to their dimension p, this method can be prone to overfitting. Moreover, the implementation of the classifier uses the within-class precision matrix of the features $\hat{\mathbf{\Sigma}}_{\mathbf{f}}^{-1}$, which in this situation may be poorly conditioned.

To address this problem we should resort to some form of regularization of the averaged within-class covariance

$$\hat{\mathbf{\Sigma}}_{\mathbf{f}} = p(c_1)\hat{\mathbf{\Sigma}}_{\mathbf{f}|c_1} + p(c_2)\hat{\mathbf{\Sigma}}_{\mathbf{f}|c_2}. \tag{33}$$

Regularized Discriminant Analysis [25] considered the projection of the sample covariance estimate (in our case, the $\hat{\mathbf{\Sigma}}_{\mathbf{f}}$ defined previously) onto the identity matrix to obtain $\langle\ \hat{\mathbf{\Sigma}}_{\mathbf{f}}\ ,\ \mathbf{I}\ \rangle\ \mathbf{I} \equiv v\ \mathbf{I}$, and then estimate the true covariance matrix $\mathbf{\Sigma}_{\mathbf{f}}$ with the convex combination

$$\tilde{\mathbf{\Sigma}}_{\mathbf{f}} = (1-\rho)\hat{\mathbf{\Sigma}}_{\mathbf{f}} + \rho(v\mathbf{I}). \tag{34}$$

The shrinkage of the sample covariance matrix towards the projection can improve the matrix conditioning and provide a closer estimate to the true covariance matrix for a carefully chosen parameter ρ. The problem consists in finding the optimal value for ρ. Ledoit and Wolf in [22] studied how to automatically determine it by approximately minimizing the minimum quadratic error between the unknown covariance matrix $\mathbf{\Sigma}_{\mathbf{f}}$ and its shrunken estimation $\tilde{\mathbf{\Sigma}}_{\mathbf{f}}$

$$\min_{\rho} E\left\{\left\|\mathbf{\Sigma}_{\mathbf{f}} - \tilde{\mathbf{\Sigma}}_{\mathbf{f}}\right\|_F^2\right\} \quad \text{s.t.} \quad \tilde{\mathbf{\Sigma}}_{\mathbf{f}} = (1-\rho)\hat{\mathbf{\Sigma}}_{\mathbf{f}} + \rho(v\mathbf{I}). \tag{35}$$

The estimator of ρ obtained by Ledoit and Wolf is given by

$$\hat{\rho}_{\text{LW}} = \min\left\{\frac{\sum_{\tau=1}^{N_\tau}\|(\mathbf{f}_\tau - \boldsymbol{\mu}_{k_\tau})(\mathbf{f}_\tau - \boldsymbol{\mu}_{k_\tau})^T - \hat{\mathbf{\Sigma}}_{\mathbf{f}}\|_F^2}{N_\tau^2\|\hat{\mathbf{\Sigma}}_{\mathbf{f}} - (\text{Tr}\{\hat{\mathbf{\Sigma}}_{\mathbf{f}}\}/p)\ \mathbf{I}\|_F^2}, 1\right\}, \tag{36}$$

where $\boldsymbol{\mu}_{k_\tau}$ refers to the mean of the class to which the feature $\mathbf{f}_\tau$ belongs. This choice for the estimate guarantees an asymptotically optimal combination of the sample covariance matrix and the identity matrix, is asymptotically consistent and makes no assumption over the data distribution.

In the context of MI-BCI, Lotte considered in [7] the Shrunken LDA (sLDA) classification. This method replaces the sample covariance matrix of the features $\hat{\mathbf{\Sigma}}_{\mathbf{f}}$ in Linear Discriminant Analysis with the Ledoit and Wolf regularized covariance matrix $\tilde{\mathbf{\Sigma}}_{\mathbf{f}}$ for $\rho = \hat{\rho}_{\text{LW}}$. sLDA obtained significant accuracy improvements over standard LDA so its use was highly recommended [3]. Note, however, that the LDA classifier assumes conditional Gaussian classes and, under this assumption, the Ledoit and Wolf regularization technique usually does not provide the best possible mean-square error for finite samples.

Chen *et al.* recognized in [13] that $\hat{\rho}_{\text{LW}}$ uses statistics of the features up to order four, while under Gaussian hypothesis the mean and covariance condense all the relevant information. They developed an Oracle Approximate Shrinkage (OAS) procedure for small samples that exploits the Gaussian hypothesis. The estimator of ρ provided by the OAS method is

$$\hat{\rho}_{\text{OAS}} = \min\left\{\frac{\left(\frac{1-2}{p}\right)\text{Tr}(\hat{\mathbf{\Sigma}}_{\mathbf{f}}^2) + \text{Tr}^2(\hat{\mathbf{\Sigma}}_{\mathbf{f}})}{\left(\frac{N_\tau+1-2}{p}\right)\left[\text{Tr}(\hat{\mathbf{\Sigma}}_{\mathbf{f}}^2) - \frac{\text{Tr}^2(\hat{\mathbf{\Sigma}}_{\mathbf{f}})}{p}\right]}\ ,\ 1\right\}, \tag{37}$$

and it was shown to attain a better mean square error in simulations than $\hat{\rho}_{\text{LW}}$ and other alternatives.

In what follows we denote by gLDA the implementation of the LDA classifier in combination with the Oracle Approximate Shrinkage estimator of the feature covariance matrix, which is obtained from equation (34) with $\rho = \hat{\rho}_{\text{OAS}}$. The added “g” refers to the Gaussian hypothesis of the centered features.

According to our experiments in MI-BCI, the classification with gLDA provides relevant gains in accuracy with respect to both the standard LDA and sLDA techniques.

VIII. Experimental results

In this section, we will try to corroborate through illustrative simulations the good performance of the proposed covariance estimators that form part of the proposals nCSP and gLDA. The first simulation reveals the expected improvement in the estimation of the covariances with a set of synthetic data since, for its evaluation, the true underlying covariance matrices of the classes have to be known. The remaining simulations consider real datasets from the BCI competitions and test the possible combination of the proposals with state-of-the-art techniques.

A. Testing the improvement in the estimation of $\mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}$

In this experiment, we design a synthetic simulation for corroborating the improvement that can be obtained with the proposed estimation method for the class covariance means. The centroids for the right-hand and left-hand classes have been set equal to the estimated class covariances of user A01 from the dataset IV-2a [29]. We used 25 training trials per class, each with 22 sensors and a length of 500 samples. The samples $\boldsymbol{x}_\tau(t)$ of each trial τ were drawn from a multidimensional Gaussian density $\mathcal{N}(\mathbf{0}, \tilde{\mathbf{C}}_\tau)$, where $\tilde{\mathbf{C}}_\tau$ was generated from a local perturbation of the conditional-class mean $\mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}$ of the trial. The details of the procedure for the generation of local random covariance matrices in the neighborhood of its centroids are described in Appendix C.

The proposed power-normalization technique does not help to guess the absolute scales of the underlying covariance matrix centroids, because these scales are subordinated to the objective of equalizing the power of the effective sources. Fortunately, it is well known that they are irrelevant in the evaluation of the common principal directions. Hence, a good measure of similarity between the true and estimated covariance centroids should be invariant with respect to the scaling of the compared arguments. A natural measure of dissimilarity between covariance matrices with arbitrary scaling is the scale-invariant version of the Riemmanian distance

$$\begin{aligned} D_R(\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}\ ,\ \mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}) &= \min_{\alpha} \delta_R(\alpha\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}\ ,\ \mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}) && (38) \\ &= \min_{\alpha}\left(\sum_{i=1}^{N_x}\log^2\frac{\lambda_i}{\alpha}\right)^{\frac{1}{2}} && (39) \\ &= \left(\sum_{i=1}^{N_x}\log^2\frac{\lambda_i}{e^{\frac{1}{N_x}\sum_{j=1}^{N_x}\log\lambda_j}}\right)^{\frac{1}{2}} && (40) \end{aligned}$$

where $\delta_R(\cdot\,,\cdot)$ denotes the standard Riemmanian distance and $\lambda_i,\ i = 1, \ldots, N_x$, refers to the eigenvalues of $\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}^{-1}\mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}$.

Dataset	User	Classic	State-of-the-art: CSP+...			Proposed: nCSP+...		
		CSP+LDA	sLDA	RMDM	TSLR	gLDA (p-value)	RMDM (p-value)	TSLR (p-value)
III-3a	k3b	94.31	95.09	94.90	96.13	95.38 (1.7e-02)	94.97 (1.2e-01)	**96.31** (2.8e-02)
	k6b	75.77	77.77	77.23	78.94	78.78 (7.1e-04)	78.15 (7.7e-05)	**79.69** (1.0e-03)
	l1b	86.73	88.18	88.46	89.25	89.48 (1.6e-07)	89.90 (1.4e-13)	**90.29** (2.7e-08)
	mean	85.60	87.01	86.87	88.11	87.88 (4.0e-09)	87.68 (6.9e-13)	**88.76** (4.3e-09)
III-4a	aa	67.68	68.56	63.12	**72.37**	68.0 (8.6e-01)	63.62 (1.3e-01)	70.31 (1.0e+00)
	al	**96.81**	96.5	96.25	96.56	96.18 (9.3e-01)	95.75 (1.0e+00)	96.31 (9.8e-01)
	av	62.12	62.25	58.31	66.25	63.06 (8.7e-02)	59.75 (4.7e-03)	**67.87** (5.1e-03)
	aw	85.12	83.37	80.62	87.62	82.93 (8.7e-01)	80.75 (1.9e-01)	**87.75** (1.9e-01)
	ay	87.68	90.87	87.62	91.0	89.31 (1.0e+00)	87.68 (2.0e-01)	**91.62** (3.7e-02)
	mean	79.88	80.31	77.18	82.76	79.89 (9.4e-01)	77.51 (5.8e-02)	**82.77** (2.3e-01)
IV-2a	A01	86.97	88.18	88.15	89.02	89.0 (1.3e-06)	88.74 (1.0e-05)	**89.23** (3.2e-02)
	A02	73.15	74.63	76.20	**77.65**	75.20 (7.3e-03)	74.78 (1.0e+00)	76.15 (1.0e+00)
	A03	87.34	88.53	88.30	89.75	89.78 (1.1e-13)	90.08 (3.3e-28)	**90.60** (6.1e-11)
	A04	68.77	69.99	70.63	71.36	70.95 (3.0e-05)	70.93 (4.7e-02)	**71.38** (2.3e-01)
	A05	56.89	58.61	58.78	59.27	60.07 (4.9e-07)	**60.19** (5.5e-08)	59.82 (1.1e-02)
	A06	61.41	62.36	62.41	**63.32**	63.12 (2.6e-03)	62.81 (2.5e-02)	63.26 (8.1e-01)
	A07	88.75	89.92	90.44	91.09	91.20 (5.4e-17)	91.39 (5.5e-09)	**91.70** (9.0e-06)
	A08	86.40	87.65	86.33	88.32	88.73 (2.0e-10)	88.72 (8.2e-48)	**89.18** (2.1e-09)
	A09	82.70	83.95	82.64	84.25	84.67 (6.7e-05)	84.59 (7.5e-27)	**85.26** (2.5e-09)
	mean	76.93	78.20	78.21	79.34	79.19 (8.8e-39)	79.14 (6.0e-41)	**79.62** (6.4e-06)

TABLE III

EXPECTED USER ACCURACY FOR THE *binary* MI CLASSIFICATION PROBLEM IN EACH OF THE CONSIDERED DATASETS. THE BEST PERFORMANCES ARE MARKED IN BOLD. ONE CAN OBSERVE THAT IN THE MAJORITY OF THE CASES THE IMPROVEMENTS OBTAINED WHEN COMBINING THE STATE-OF-THE-ART METHODS WITH THE PROPOSED TECHNIQUES CAN BE REGARDED AS STATISTICALLY SIGNIFICANT ($p\text{-}value < 5e\text{-}02$).

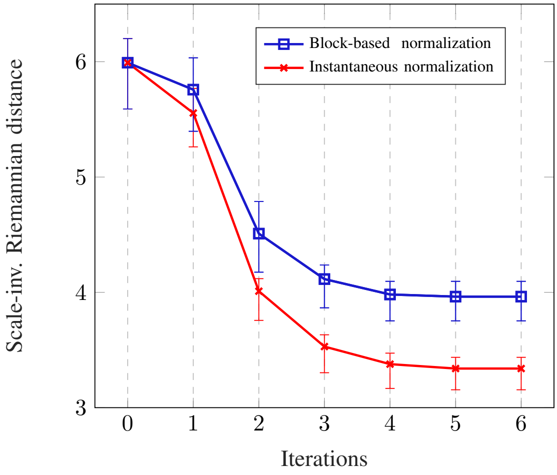


Fig. 1. Variations of the scale-invariant Riemannian distance (between the reference $\mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}$ and estimated $\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}$ covariance matrices) with respect to the number of iterations of the proposed power-normalization procedures. The solid lines represent average distances while the bars represent the 25% and 75% percentiles. Iteration 0 refers to the absence of normalization, iteration 1 coincides with the standard trace-based normalization used in CSP, while the remaining iterations are instances of the proposed normalization.

Figure 1 illustrates the improvement in the estimation of the class-conditional covariance matrix means for the block (Section V) and instantaneous (Section VI) power-normalization procedures, when both share the initialization $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(0)} = \mathbf{I}$. The x-axis represents the iteration i at which the covariance matrix estimate $\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}^{(i)}$ is evaluated, whereas the y-axis represents the average across classes of the scale-invariant Riemannian distances between $\mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}^{(i)}$.

The simulation results confirm the expected improvement of these normalizations with respect to the classical one, which corresponds with the result obtained for iteration 1. Being, in this case, the power-instantaneous equalization method slightly more precise than the block based-implementation.

B. Experiments using the BCI competitions datasets

This subsection is devoted to the experimental comparison of the proposals on real BCI datasets. The normalization scheme for the estimation of the class-conditional covariance matrices, proposed in Section VI, can be combined with a variety of MI-BCI techniques to improve their accuracy. In particular, we compare the differences of performance, between classical CSP and nCSP (our proposal), when they are used in combination with the following classifiers: LDA, its shrinkage variants sLDA and gLDA, RMDM and TSLR. The Python code for the RMDM and tangent space (TS) implementations can be downloaded from [5]. For TSLR, the Logistic Regression (LR) classifier was implemented according to version 0.19.2 of [26] with its default parameters.

The experiments in this section have been carried out using three datasets from BCI competitions. Dataset 3a from BCI competition III [27], which contains 60 EEG channels, three users and four classes of motor imagery movements (MIM); dataset 4a from BCI competition III [28] with 108 EEG channels, four users and two classes of MIM; finally, dataset 2a from BCI competition IV [29] has 22 EEG channels, nine users with two sessions per user and four classes of MIM.

Each experiment consists of 40 Monte-Carlo simulations where the whole set of available trials for each session, user and pair of movements is randomly split into testing and training groups. After that, the averaged performance over the test trials is reported. The simulations report the average classification accuracy over all the possible confrontations of pairs of classes ($K = 2$) for each user. By default, the number of training and testing trials is set to 40 and the number of spatial filters is set 8, except for those cases where a range of these values is specified.

In Table III, we show the accuracy results for each subject in each of the three datasets. We also report the *mid-p values* of one-sided McNemar's tests of hypotheses [30] for paired data that allows to check whether the proposals have significant advantages in accuracy with respect to their respective state-of-the-art approaches. One can observe in Table III that for two of the datasets the proposal nCSP leads to significant improvements in expected accuracy with respect to classical CSP, whereas, its performance remains equivalent for the dataset III-4a.

We also compare the algorithms when the number of training trials varies from 4 to 80, while the number of testing trials remains equal to the default value of 40. For this purpose, we have employed the dataset IV-2a. Figures 2(a) and 2(b) represent the improvement of nCSP+gLDA, nCSP+RMDM and nCSP+TSLR with respect to their respective baselines: CSP+sLDA, CSP+RMDM and CSP+TSLR. In both figures, the best performance over the whole range of training trials is obtained for the proposed normalization. Figure 2(b) reveals that the use of nCSP instead of CSP progressively increases the improvement with the number of training trials. In Figure 2(a), the combination of nCSP with gLDA sustains the improvement across the number of training trials. A disaggregated analysis reveals that gLDA improves greatly over sLDA for a small number of training trials.

In the last experiment, we analyze the sensitivity of the methods with respect to the chosen number of spatial filters p for dimensionality reduction. Figure 3 enables us to compare the accuracy of the proposals nCSP+gLDA and nCSP+TSLR with respect the existing approaches, for the datasets IV-2a and III-3a. These figures reveal that the standard method CSP+LDA (orange dashed-line) is quite sensitive to the choice p. Its performance attains a maximum at a relatively small value of p and greatly decreases as this number increases. This finding supports the necessity of employing automatic selection techniques to determine the right number of spatial filters for each user [8]. Although, use of Ledoit and Wolf covariance shrinkage estimates (green dashed-line) partially alleviates the previous drawback, the accuracy for the proposed nCSP+gLDA (green solid-line) is more robust with respect to a misspecification of the optimum number of spatial filters.

In our simulations, the best performance was obtained for the nCSP procedure in combination with the Tangent Space Logistic Regression (TSLR) classifier (blue continuous-line). This method has outperformed CSP+TSLR (blue dashed-line) in expected user accuracy over all the range of the number of spatial filters and training trials.

Similar results have been obtained for multiclass scenarios. We refer the interested reader to the supplementary material [31] that accompanies this manuscript and includes an illustrative Python demo.

IX. Conclusion

In this work, we have studied the problem of obtaining improved covariance matrix estimators for the processing of the MI-BCI signals. We have proposed the application of two techniques that improve the accuracy of these estimations. To counter the inter and intra-trial non-stationarity that hinders the correct estimation of the trial covariance matrices, we propose a power normalization of the EEG source activities. When this is implemented across trials, it improves the classical normalization of the observations used for the EEG trials. Furthermore, the instantaneous power-normalization of the sample source vector seems to enable superior classification results. In this latter case, the proposal extends Tyler's method (for obtaining an estimate of scatter) to the context of heterogeneous trial observations. The second technique refers to a convenient regularization of the feature covariance matrix of the classifiers. Both proposals are transversal, in the sense that they can be easily combined with the existing MI-BCI algorithms to boost their performance. Experimental tests on several BCI competition datasets reveal that a combination of the proposed techniques with state-of-the-art algorithms for motor-imagery classification provides a significant improvement in the classification results.

Appendix

A. Proof of the formula for the power of the effective sources

We start by noting that there is a one to one correspondence between $\mathbf{X}_\tau$ and $\tilde{\mathbf{S}}_\tau$, which is given by

$$\tilde{\mathbf{S}}_\tau = \mathbf{\Pi}_{\mathbf{A}'^T}\tilde{\mathbf{S}}_\tau = \mathbf{A}'^T(\mathbf{A}'\mathbf{A}'^T)^{-1}\mathbf{X}_\tau. \quad (41)$$

Recalling the invariance of the trace of the product of compatible matrices with respect to cyclic permutations in the matrix positions, i.e., $\mathrm{Tr}\{\tilde{\mathbf{S}}_\tau\tilde{\mathbf{S}}_\tau^T\} = \mathrm{Tr}\{\tilde{\mathbf{S}}_\tau^T\tilde{\mathbf{S}}_\tau\}$, and using (41) to substitute the value of $\tilde{\mathbf{S}}_\tau$ in (16), we obtain

$$P_{\tilde{\mathbf{S}}_\tau} \;=\; \frac{1}{T}\,\mathrm{Tr}\{\mathbf{X}_\tau^T(\mathbf{A}'\mathbf{A}'^T)^{-1}\mathbf{X}_\tau\}. \quad (42)$$

As we have seen in equation (12), $\mathbf{A}'\mathbf{A}'^T$ coincides with the global average covariance matrix of the observations $\mathbf{\Sigma}_\mathbf{x}$, hence, we can write without any approximations that

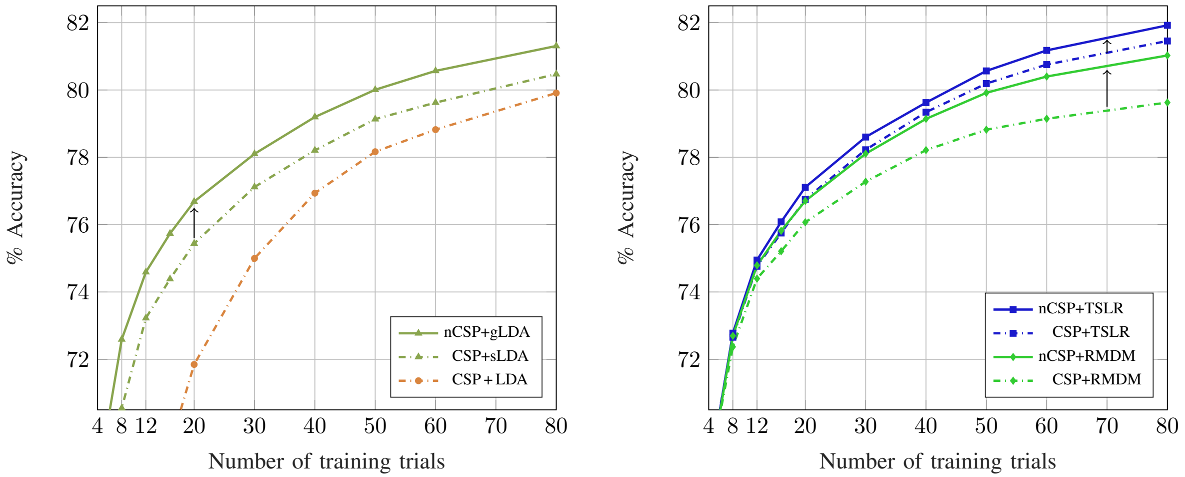
$$P_{\tilde{\mathbf{S}}_\tau} = \frac{1}{T}\,\mathrm{Tr}\{\mathbf{X}_\tau^T\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{X}_\tau\} = \mathrm{Tr}\{\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{C}_{\mathbf{X}_\tau}^{(0)}\}. \quad (43)$$

B. Equivalence with Tyler's method for estimation of scatter

The algorithmic solution provided by the proposed instantaneous power-normalization technique may be regarded as a variation of Tyler's method used in statistics for obtaining a robust m-estimator of scatter [23]. As it will be shown, for a single trial ($N_\tau = 1$) and a single class ($K = 1$), both techniques use complementary arguments to arrive by different paths to a similar final result. To trace back the equivalence, we review the problem considered by Maronna in [32], where he studied how to obtain robust affine-invariant estimates of mean and scatter from a set $\{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)\}$ of multivariate i.i.d. samples, drawn from an elliptical distribution. Let the density of $\boldsymbol{x}(t)$ for a given scatter matrix $\mathbf{C}_{\boldsymbol{x}}$ be

$$p(\boldsymbol{x}(t);\mathbf{C}_{\boldsymbol{x}}) = \kappa|\mathbf{C}_{\boldsymbol{x}}|^{-1/2}\phi((\boldsymbol{x}(t)-\boldsymbol{\mu}_x)^T\mathbf{C}_{\boldsymbol{x}}^{-1}(\boldsymbol{x}(t)-\boldsymbol{\mu}_x)) \quad (44)$$

where $\phi(\cdot)$ is an integrable and non-negative function with domain $\mathbb{R}^+$ and κ is the normalization constant. For simplicity,



(a) Improvement of nCSP+gLDA over the baselines.

(b) Improvement of nCSP+TSLR and nCSP-RMDM over the baselines.

Fig. 2. This experiment shows the accuracy of the binary classification methods with respect to the number of training trials for dataset IV-2a. The arrows in Subfigures (a) and (b) represent the improvement in performance of nCSP+gLDA, nCSP+RMDM and nCSP+TSLR with respect to their baselines.

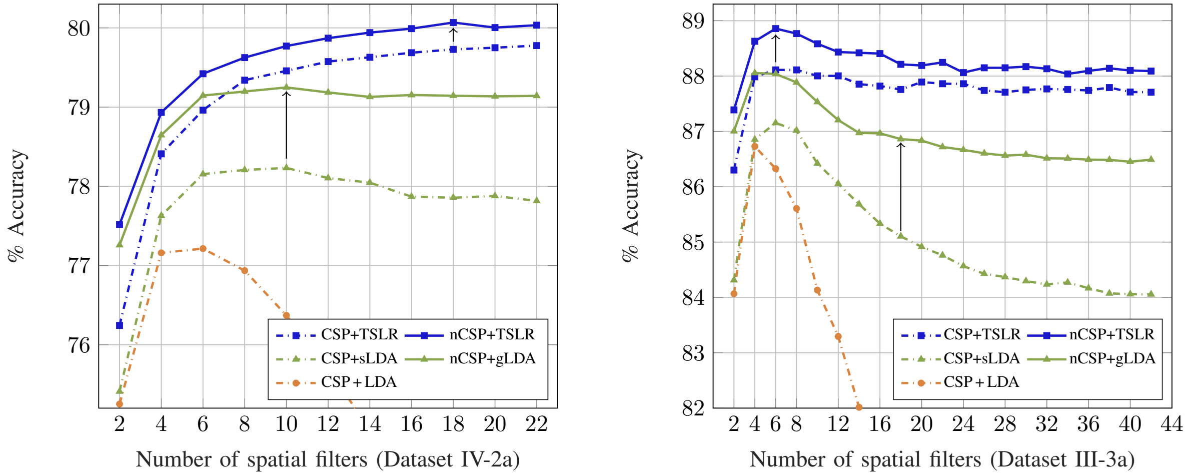


Fig. 3. Variations in performance of the MI-BCI binary classification methods with respect to p, the number of spatial filters. The results confirm the advantages of using nCSP in combination with the state-of-the-art classifiers to improve the expected user accuracy.

we assume in this exposition that the mean $\boldsymbol{\mu}_x$ is known (or can be reasonably estimated from the data) and focus on the steps for the estimation of $\mathbf{C}_{\boldsymbol{x}}$. The normalized log-likelihood of the observations is

$$\langle \log p(\boldsymbol{x}(t);\mathbf{C}_{\boldsymbol{x}})\rangle_t \quad = \quad \log\kappa - \frac{1}{2}\log|\mathbf{C}_{\boldsymbol{x}}| + \langle \log\phi(\alpha_t)\rangle_t$$

where $\alpha_t = (\boldsymbol{x}(t)-\boldsymbol{\mu}_x)^T\mathbf{C}_{\boldsymbol{x}}^{-1}(\boldsymbol{x}(t)-\boldsymbol{\mu}_x)$. In [32] the maximization of the log-likelihood leads to an M-estimator of scatter $\hat{\mathbf{C}}_{\boldsymbol{x}}$ that satisfies the estimating equation

$$\hat{\mathbf{C}}_{\boldsymbol{x}} - \left\langle u(\hat{\alpha}_t)(\boldsymbol{x}(t)-\boldsymbol{\mu}_x)(\boldsymbol{x}(t)-\boldsymbol{\mu}_x)^T\right\rangle_t \quad = \quad \mathbf{0} \quad (45)$$

where $u(\alpha_t) = -2\frac{d\log\phi(\alpha_t)}{d\alpha_t}$.

Although there is no close-form solution to this equation because of the coupling between $\hat{\alpha}_t$ and $\mathbf{C}_{\mathbf{X}}$, there is a general set of conditions that guarantees its uniqueness (see [32]). Years later, Tyler considered in [23] the same problem. He studied the properties of the specific weighting function $u(\alpha_t) = N_x/\alpha_t$ and showed that this choice gives the "most robust estimator of the scatter matrix of an elliptical distribution in the sense of minimizing the maximum asymptotic variance". He also proposed to iteratively solve the estimation equation through a fixed point iteration.

In our particular case, $\boldsymbol{x}_\tau(t) \equiv \boldsymbol{x}(t) - \boldsymbol{\mu}_x$ and Tyler's iteration for the estimation of the trial covariance matrices

$\hat{\mathbf{C}}_{\boldsymbol{x}} \equiv \mathbf{C}_{\mathbf{X}_\tau}$ is given by

$$\mathbf{C}^{(i)}_{\mathbf{X}_\tau} = \frac{N_x}{T}\sum_{t=1}^{T} \frac{\boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T}{(\boldsymbol{x}_\tau(t))^T(\mathbf{C}^{(i-1)}_{\mathbf{X}_\tau})^{-1}\boldsymbol{x}_\tau(t)}. \tag{46}$$

In fact, this estimate of the covariance matrices of the trials has been recently considered for SSVEP-BCI in [33], however, we are not aware of its previous use in MI-BCI applications. Our proposed instantaneous power-normalization in (32), i.e.,

$$\mathbf{C}^{(i)}_{\mathbf{X}_\tau} = \frac{N_x}{T}\sum_{t=1}^{T} \frac{\boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T}{(\boldsymbol{x}_\tau(t))^T(\hat{\boldsymbol{\Sigma}}^{(i-1)}_{\mathbf{x}})^{-1}\boldsymbol{x}_\tau(t)} \tag{47}$$

simplifies to (46) in the specific case of having a unique class and a single trial. This is a straightforward consequence of the fact that $\hat{\boldsymbol{\Sigma}}^{(i-1)}_{\mathbf{x}} = \mathbf{C}^{(i-1)}_{\mathbf{X}_\tau}$ for $N_\tau = 1$, see (19). However, for the case of heterogeneous trials and classes, this last proposal will eventually improve the obtained performance results.

C. Procedure for generating random and locally perturbed covariance matrices

The procedure for its generation has been the following. Initially, for each trial, a symmetric random perturbation $\mathbf{H}$ is built on the tangent space of the matrix mean $\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_k}}$. This can be done with the help of the following MatLab commands:

$$\begin{aligned} \mathbf{G} &= \text{randn}(N_x);\ \mathbf{H}_0 = (\mathbf{G}+\mathbf{G}^T)/2 && (48)\\ \mathbf{H} &= \text{rand}(1)\ (2.5\sqrt{N_x})\ \mathbf{H}_0/\|\mathbf{H}_0\|_{\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_k}}} && (49)\end{aligned}$$

where the natural norm in the tangent space is given by

$$\|\mathbf{H}_0\|_{\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_k}}} = \sqrt{\text{Tr}\{\mathbf{H}_0\boldsymbol{\Sigma}^{-1}_{\mathbf{x}|\mathbf{c_k}}\mathbf{H}_0\boldsymbol{\Sigma}^{-1}_{\mathbf{x}|\mathbf{c_k}}\}}. \tag{50}$$

After that, the retraction of $\mathbf{H}$ onto the covariance matrix manifold results in the randomly perturbed covariance matrix

$$\tilde{\mathbf{C}}_\tau = \boldsymbol{\Sigma}^{1/2}_{\mathbf{x}|\mathbf{c_k}}\text{expm}(\boldsymbol{\Sigma}^{-1/2}_{\mathbf{x}|\mathbf{c_k}}\mathbf{H}\boldsymbol{\Sigma}^{-1/2}_{\mathbf{x}|\mathbf{c_k}})\boldsymbol{\Sigma}^{1/2}_{\mathbf{x}|\mathbf{c_k}}. \tag{51}$$

Lastly, the samples of the trial $\boldsymbol{x}_\tau(t)$ are drawn according to the Gaussian density $\mathcal{N}(\mathbf{0}, \tilde{\mathbf{C}}_\tau)$.

References

[1] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE Trans. on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[2] R. Martín-Clemente, J. Olias, D. B. Thiyam, A. Cichocki, and S. Cruces, "Information theoretic approaches for motor-imagery bci systems: Review and experimental comparison," *Entropy*, vol. 20, no. 1, 2018.

[3] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005.

[4] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. on Biomedical Engineering*, vol. 55, no. 8, pp. 1991–2000, 2008.

[5] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain-computer interface classification by riemannian geometry," *IEEE Trans. on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2012. [Online]. Available: https://github.com/alexandrebarachant/covariancetoolbox

[6] W. Samek, M. Kawanabe, and K. R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[7] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 871–890, June 2015.

[8] Y. Yang, S. Chevallier, J. Wiart, and I. Bloch, "Automatic selection of the number of spatial filters for motor-imagery bci," in *Proceedings of the 20th European Symposium on Artificial Neural Networks (ESANN)*, 2012, pp. 109–114.

[9] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-spectral filters for improving the classification of single trial eeg," *IEEE Trans. on Biomedical Engineering*, vol. 52, no. 9, pp. 1541–1548, 2005.

[10] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain-computer interfacing," *Journal of Neural Engineering*, vol. 9, no. 2, 2012.

[11] W. Samek, F. C. Meinecke, and K. Müller, "Transferring subspaces between subjects in brain–computer interfacing," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[12] F. Lotte and C. Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2010, pp. 614–617.

[13] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for mmse covariance estimation," *IEEE Trans. on Signal Processing*, vol. 58, no. 10, pp. 5016–5029, Oct 2010.

[14] M. Congedo, C. Gouy, C. Jutten, "On the blind source separation of human electroencephalogram by approx joint diagonalization of second order statistics", *Clinical Neurophysiology* 119, pp. 2677-2686, 2008.

[15] K. Fukunaga and W. Koontz, "Application of the Karhunen-Loeve expansion to feature selection and ordering," *IEEE Trans. on Computers*, vol. C-19, pp. 311 – 318, 1970.

[16] D. Thiyam, S. Cruces, J. Olias, and A. Cichocki, "Optimization of alpha-beta log-det divergences and their application in the spatial filtering of two class motor imagery movements," *Entropy*, vol. 19, no. 3, 2017.

[17] M. Congedo, A. Barachant, R. Bhatia, "Riemannian Geometry for EEG-based Brain-Computer Interfaces; a Primer and a Review", *Brain-Computer Interfaces* 4(3), pp. 155-174, 2017.

[18] F. Yger, M. Berar, F. Lotte, "Riemannian Approaches in Brain-Computer-Interfaces: A Review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1753-1762, 2017.

[19] A. Balzia, F. Yger, M. Sugiyama, "Importance-weighted covariance estimation for robust common spatial pattern," *Pattern Recognition Letters* 68, pp. 139–145, 2015.

[20] W. Samek, S. Nakajima, M. Kawanabe, KR. Müller, "On robust parameter estimation in brain–computer interfacing," *J. Neural Eng.* 14(6):061001, 2017.

[21] F. Lotte, C. Guan, "Regularizing Common Spatial Patterns to Improve BCI Designs: Unified Theory and New Algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355-362, 2011.

[22] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365 – 411, 2004.

[23] D. E. Tyler, "A distribution-free m-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. New York, NY, USA: Wiley-Interscience, 2000.

[25] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–174, 1989.

[26] F. Pedregosa *et al.* "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* 12, pp. 2825–2830, 2011.

[27] D. A. Schlögl, "Dataset IIIa: 4-class EEG data," http://www.bbci.de/competition/iii/desc_IIIa.pdf, 2004, [Online; accessed 25-May-2018].

[28] K.-R. Müller and B. Blankertz, "Data set IVa - motor imagery, small training sets ," http://www.bbci.de/competition/iii/desc_IVa.html, 2004, [Online; accessed 25-May-2018].

[29] C. Brunner, R. Leeb, G. R. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI Competition 2008 - Graz data set A," http://www.bbci.de/competition/iv/desc_2a.pdf, 2008, [Online; accessed 25-May-2018].

[30] T.G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.* 10, 7, pp. 1895-1923, 1998.

[31] J. Olias, R. Martin-Clemente, M.A. Sarmiento-Vega, S. Cruces, "Supplementary Material for EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimators", *IEEE Dataport*, 2019. [Online]. Available: http://dx.doi.org/10.21227/a8yg-4f68. Accessed: Mar. 14, 2019.

[32] R. A. Maronna, "Robust m-estimators of multivariate location and scatter," *The Annals of Statistics*, vol. 4, no. 1, pp. 51–67, 1976.

[33] S. Chevallier, E. Kalunga, Q. Barthélemy, and F. Yger, "Riemannian Classification for SSVEP-Based BCI: Offline versus Online Implementations," in *Brain-Computer Interfaces Handbook : Technological and Theoretical Advances*, 2018, pp. 372–398.

Supplementary material

- **Title:** Supplementary Material for “EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimators” .
- **Authors:** Javier Olias; Rubén Martín-Clemente; M. Auxiliadora Sarmiento-Vega; Sergio Cruces.
- **DOI:** 10.21227/a8yg-4f68.
- **Published in**: IEEE Dataport.
- **Date:** March 2019.

Summary of supplementary material

In this supplementary material of the previous article we extend the results and provide with the codes to reproduce a mock example. Therefore the contributions of this publication are:

Contribution 1: *Extension to the multi-class paradigm.*

The preprocessing that was proposed in *Publication A* is not dependent of the classes, so it can be used in a multi-class paradigm without any modification. In this supplementary material we extend the results in *Publication A*, that address the binary class paradigm to a multi-class paradigm using the ITFE algorithm that we explained in 4.3 and the Riemannian geometry approaches that also work in multi-class scenarios. The multi-class classification is performed as One versus the rest when the LR classifier is used. In the case of LDA classifiers, the class which maximize the equation (5.2) is selected. We recall the equation for the reader convenience:

$$p(\mathbf{y}|C_k) = \frac{1}{(2\pi)^{p/2}\sqrt{|\mathbf{\Sigma_y}|}} \exp\left(\frac{-1}{2}(\mathbf{y}-\mu_k)\mathbf{\Sigma}_{\mathbf{y}}^{-1}(\mathbf{y}-\mu_k)^{\top}\right) \tag{A.29}$$

In that sense, in *table I* of the supplementary material we show the accuracy results of each user comparing the use of the proposed normalization with the classical one, and providing with McNemar’s test to check whether a result has significant improvements or not. The *figure 1* and *figure 2* of the supplementary material show the evolution of the overall accuracy as the number of training trials and the number of CSP filters vary respectively.

Contribution 2: *Publication of the necessary codes to implement the instantaneous normalization technique.*

We also made public the implementation in Phyton language of the instantaneous normalization using simulated data. Making it easier to reproduce the results and the extension of the technique.

Supplementary Material for “EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimators”

Javier Olias, Rubén Martín-Clemente, M^a Auxiliadora Sarmiento-Vega and Sergio Cruces

IEEE Transactions on Neural Systems and Rehabilitation Engineering

(accepted for publication in March 2019)

I. Simulations for the multiclass paradigm

In this supplementary file, we report the results of the simulations for the multiclass paradigm, which complements the binary case already presented in the main manuscript.

Only datasets III-3a [27] and IV-2a [29] provide four classes of MIM that correspond to the left hand, right hand, tongue, and feet. For the dimensionality reduction stage, we have used the multi-class version of CSP, which is implemented by the Information Theoretic Feature Extraction criterion (ITFE) [4], with 12 spatial filters.

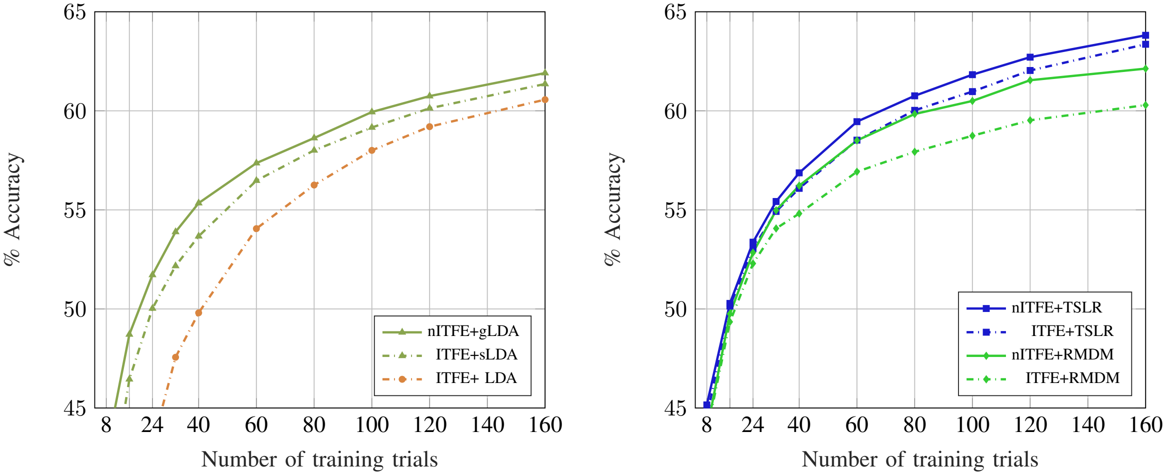
The experiments compare the classic ITFE dimensionality reduction with the improved version nITFE (which includes the proposed normalization) in combination with LDA, sLDA, gLDA, RMDM and TSLR classifiers.

In Table I, we present the expected user accuracies and *p-values* that reveal the degree of significance for the improvements which are obtained when nITFE is used. This experiment averaged the results obtained for 40 Monte Carlo runs, considering 80 random trials for training and 80 trials for testing.

Figure 1 presents the expected accuracy results, for the whole dataset IV-2a, when the number of training trials varies. Similarly, Figure 2 represents the expected accuracy results versus the number of spatial filters for dimensionality reduction. All these results confirm that the advantages of the proposals, for the binary case, are also extended to the multiclass paradigm.

		Classic	**State-of-the-art**			**Proposed**		
Dataset	**User**	**ITFE+LDA**	**ITFE+sLDA**	**ITFE+RMDM**	**ITFE+TSLR**	**nITFE+gLDA** (p-val.)	**nITFE+RMDM** (p-val.)	**nITFE+TSLR** (p-val.)
III-3a	k3b	88.75	89.31	84.40	88.75	86.71 (1.0e+00)	86.06 (1.8e-05)	**90.21** (3.5e-05)
	k6b	50.21	51.56	51.90	54.65	57.99 (5.3e-13)	58.34 (4.2e-17)	**61.84** (1.8e-15)
	l1b	70.31	75.09	71.31	71.90	**75.68** (1.1e-01)	75.25 (2.3e-10)	75.09 (3.4e-06)
	– – –							
	mean	69.76	71.98	69.20	71.77	73.46 (1.9e-04)	73.21 (1.9e-28)	**75.71** (5.6e-22)
IV-2a	A01	65.39	67.09	68.09	70.39	67.95 (2.9e-02)	69.39 (8.1e-04)	**70.98** (4.6e-02)
	A02	50.53	52.32	54.81	**55.73**	51.73 (9.2e-01)	50.92 (1.0e+00)	53.70 (1.0e+00)
	A03	72.07	74.68	73.81	77.32	75.31 (3.7e-02)	77.14 (1.5e-16)	**78.84** (3.0e-05)
	A04	43.85	45.14	45.06	46.89	45.65 (9.1e-02)	47.68 (7.3e-07)	**48.34** (3.8e-03)
	A05	31.76	33.67	33.96	**35.37**	34.35 (6.1e-02)	34.93 (1.9e-02)	35.01 (8.7e-01)
	A06	35.57	36.68	36.09	37.39	36.40 (8.4e-01)	37.37 (4.5e-03)	**38.46** (1.7e-02)
	A07	73.71	75.5	74.54	77.81	76.81 (1.3e-03)	78.46 (1.9e-15)	**78.93** (2.9e-03)
	A08	69.92	71.84	70.93	73.45	72.42 (5.2e-02)	**75.15** (1.1e-22)	75.09 (3.2e-05)
	A09	63.43	65.09	64.06	65.84	66.98 (4.0e-05)	**67.46** (4.3e-14)	67.42 (1.9e-04)
	– – –							
	mean	56.25	58.00	57.93	60.02	58.62 (1.4e-04)	59.83 (4.0e-30)	**60.75** (5.4e-06)

TABLE I
User accuracy for the *four-class* MI-BCI classification problem in each of the considered datasets. Again, the results confirm that state-of-the-art methods for the multi-class problem can be improved when combined with the proposed techniques.



(a) Improvement of nITFE+gLDA over the baselines (4-class problem).

(b) Improvement of nITFE+TSLR over the baselines (4-class problem).

Fig. 1. This experiment considers the dataset IV-2a and shows the variations in accuracy of the four-class classification methods with respect to the number of training trials. Subfigures (a) and (b) present the improvement in performance of nITFE+gLDA, nITFE+RMDM and nITFE+TSLR with respect their baselines.

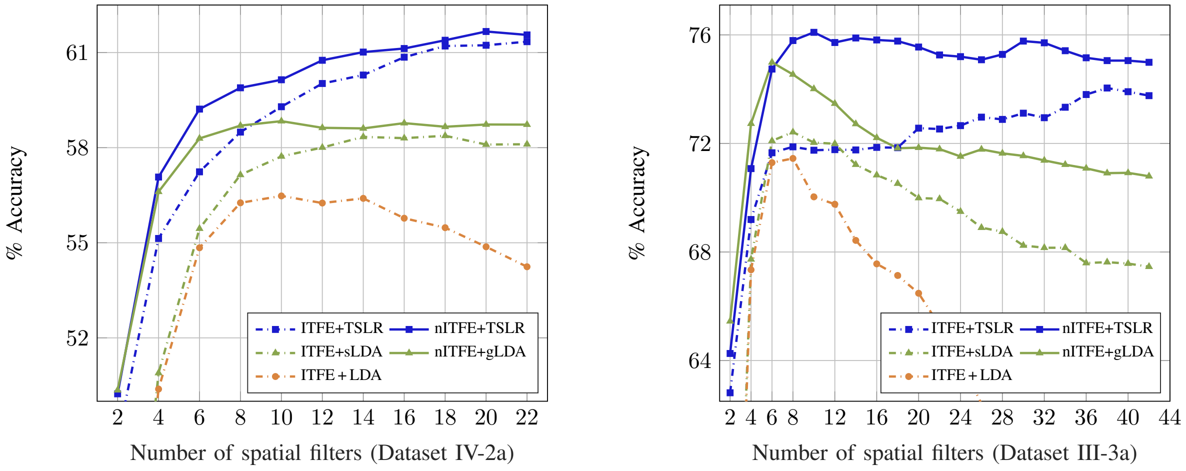


Fig. 2. Variations in performance of the MI-BCI classification methods with respect to p, the number of spatial filters. The results confirm the advantages of using nITFE in combination with the state-of-the-art classifiers to improve the expected user accuracy.

Publication B

A Technique for Artifact Attenuation in Motor-Imagery BCI

- **Title:** A Technique for Artifact Attenuation in Motor-Imagery BCI.
- **Authors:** Javier Olias; Rubén Martín-Clemente; M. Auxiliadora Sarmiento-Vega; Sergio Cruces.
- **Available at:** researchgate.net.
- **Published in:** URSI Spain 2019.
- **Date:** September 2019.

- **Abstract:** Artifact is a term used in the biomedical field to make reference to a kind of signal distortion that is produced by non-biological process and that are not inherent of the system. In Brain Computer Interfaces (BCI) they are usually produced by a person's movement like blinking or breathing. Although most of the artifacts are of a short duration and sporadic, their effects can disrupt the whole system. We have found that the new method that we have recently presented in the field of Motor-Imagery BCI (MI-BCI) to estimate the trial covariance matrices, is a very powerful tool to palliate the effects of these artifacts. This algorithm normalizes the underlying EEG sources by weighting the samples of each trial before computing the covariance matrix. In this work we explain this technique and we show how by using it, we are also establishing a robust penalty to the samples that were contaminated by artifacts, without using any extra parameters or specific training.

Summary of Publication B

In this work we analyze the benefits of normalizing using the EEG sources instead of the classical normalization. We do this by studying the effects of the sample by sample EEG sources normalization in the presence of artifacts.

Contribution 1: *Inclusion of the artifacts in the mathematical model of the observations.*

We consider any of the two models that we have explained before to describe the EEG signals in equation (4.2) or in equation (A.20). Here, we will use (4.2) to keep the consistency with the paper where we use this model as it is more extended,

$$\begin{aligned} \mathbf{X}_{\nu} &= \mathbf{AS}+\mathbf{N}+\nu && \text{(B.30)} \\ &= \ddot{\mathbf{X}}+\nu,\ \ \ddot{\mathbf{X}},\nu \in \mathbb{R}^{n_s\times T}, && \text{(B.31)} \end{aligned}$$

where the diaeresis (̈) denotes that the variables are statistically the same random variables that we study in the previous model (4.2) but in this case they are going to be mixed with artifact contamination (ν). As we can see in equation (B.31), we model a contaminated trial as a clean one with the addition of the artifact.

Contribution 2: *Demonstration of why the instantaneous power normalization attenuates the samples which are contaminated with artifacts.*

To demonstrate this we rely on the sporadic occurrence of the artifacts. Having this in mind, we can see that their effects over the global covariance matrix ($\mathbf{\Sigma_X}$) are considered to be negligible. Although one can argue that trials contaminated with artifacts usually have more power than the rest of trials and, consequently, they can be more present in the mean covariances, the EEG sources normalization is performed sample by sample, equalizing the power of each sample. In addition, we will perform an iterative algorithm in which the first iteration might be still contaminated but as the algorithm keeps iterating, the presence of the artifacts in the global covariance matrix will keep decreasing as the power of the rest of the samples that are not related to the set. However, for our analysis we consider just the case of the last iteration, and as we said before, we consider that the artifact presence in the global covariance matrix at this point is negligible.

Before diving into how the artifacts are affected by the normalization, we will first show the case of a clean sample. To do this, we use equation (A.28), from which we extract the normalization parameter which we define as γ^2:

$$\gamma_t^2 = (\mathbf{x}_t+\nu)^{\top}(\mathbf{\Sigma_X})^{-1}(\mathbf{x}_t+\nu)\,. \tag{B.32}$$

We can transform the product by decomposing the global precision matrix $(\mathbf{\Sigma_X})^{-1}$ in its eigenvalues λ_i and eigenvectors $\mathbf{u}_i$.

$$\gamma_c^2 \rightarrow (\mathbf{x}_t+\nu)^{\top}\mathbf{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x}_t+\nu) = \sum_{i=1}^{n_s}\langle \mathbf{u}_i,(\mathbf{x}_t+\nu)\rangle^2\lambda_i^{-1}.$$

Where we can see that the normalization value depends on the scalar products of a sample vector $(\mathbf{x}_t+,\nu)$ with each eigenvectors $\mathbf{u}_i$ of the global covariance matrix

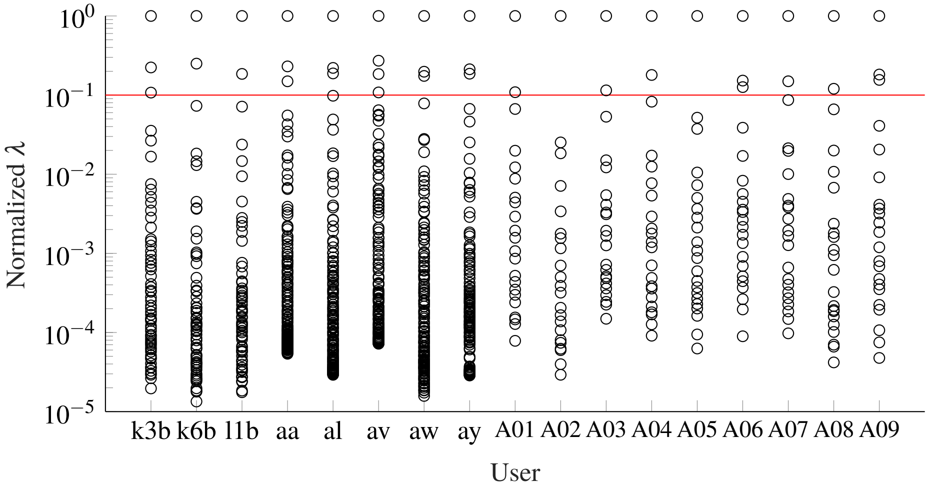


Figure B.1 Representation of the normalized eigenvalues of the global matrix of a left hand versus right hand session of each user. Each covariance has as many eigenvalues as channels has the EEG session and there are three data-sets: Data-set III-3a (k3b, k6b, l1b) with 60 channels; Data-set III-4a (aa, al, etc...) with 108 channels; Data-set IV-2a (A01,...,A09) with 22 channels. The normalization is perform by the highest eigenvalue. The red line mark the value 0.1 from which we consider the eigenvalues to be close to zero. We can see that only a few eigenvalues of each session have a significant value.

and the inverse of the eigenvalue associated to it λ_i^{-1}. To follow with the analysis it is important to understand that most of the eigenvalues λ_i of $\Sigma_{\mathbf{X}}$ are close to zero [97]. This is shown in B.1, where a representation of those eigenvalues is plotted and one can see that there usually is a dominant eigenvalue and only a few more which values are not negligible. Therefore, the inverse of the eigenvalues (λ_i^{-1}) that are close to zero would have a very high value and the samples that are not aligned with a principal direction of the global covariance matrix will be heavily attenuated. In any case, the samples that are aligned are just normalized.

To finish with the analysis, we recall that the artifact signals are by definition those that are not related to brain activity and therefore it is very unlikely that an artifact sample would be in one of the few main direction of the EEG signal, and consequently those samples are attenuated.

Contribution 3: *To support the previous theoretical contribution, we provide with test results over simulated and real data.*

In *figure 1* of the *publication B*, we find a synthetic trial with controlled artifacts in it to see how they are attenuated. Later, we find in *figure 2* a real trial that was marked by an expert as a contaminated trial, and thanks to the normalization parameter we can see very clearly the contaminated part. Finally, in the *table* (that we reproduce in the following) we can see improvements in terms of accuracy of

Table B.1 Accuracy results (%) comparing the standard (Std.) method against the use of the proposed normalization over clean and contaminated trials. To contrast this numbers can be useful to note that each user has 288 total trials.

	Artifact	Clean trials		Contaminated trials	
User	count	Std.	Proposed	Std.	Proposed
A01	22	88.40	89.23	77.43	84.62
A02	23	73.78	73.20	77.22	78.88
A03	33	86.09	89.41	82.51	89.51
A04	86	68.88	70.25	63.11	65.90
A05	38	57.86	59.74	62.93	62.93
A06	142	61.12	61.74	63.17	64.75
A07	28	90.31	91.62	89.25	91.59
A08	71	86.24	89.02	87.76	89.85
A09	45	82.23	84.17	76.61	82.54
Mean	Accuracy	77.21	78.71	75.56	78.95

the normalization over the contaminated trials in comparison with the clean trials.

A Technique for Artifact Attenuation in Motor-Imagery BCI

Javier Olias, Rubén Martín-Clemente, M^{a} Auxiliadora Sarmiento-Vega and Sergio Cruces
folias@us.es, ruben@us.es, sarmiento@us.es and sergio@us.es
All the authors are with the Departamento de Teoría de la Señal y Comunicaciones, Universidad de Sevilla.

Abstract—**Artifact is a term used in the biomedical field to make reference to a kind of signal distortion that is produced by non-biological process and that are not inherent of the system. In Brain Computer Interfaces (BCI) they are usually produced by a person's movement like blinking or breathing. Although most of the artifacts are of a short duration and sporadic, their effects can disrupt the whole system. We have found that the new method that we have recently presented in the field of Motor-Imagery BCI (MI-BCI) to estimate the trial covariance matrices, is a very powerful tool to palliate the effects of these artifacts. This algorithm normalizes the underlaying EEG sources by weighting the samples of each trial before computing the covariance matrix. In this work we explain this technique and we show how by using it, we are also establishing a robust penalty to the samples that were contaminated by artifacts, without using any extra parameters or specific training.**

I. Introduction

The term BCI makes reference to the communication systems in which the user is capable of transmitting information directly from her or his brain to a computer or machine without the intervention of any other organs or muscles [1]. To establish this communication, we need to measure the electromagnetic signals that our brain is constantly producing. In this manuscript we focus on the case where this activity is measured by an electroencephalogram (EEG) headset, placing electrodes over the scalp, using relatively cheap and non intrusive techniques.

The BCI systems based on EEG are an engineering challenge because the signal to noise ratio of the EEG signals is very low in comparison with other techniques, since they are usually affected by interferences and by common artifacts as the ones produced by little movements like blinking or swallowing. These artifacts produce important variations in the signals that disturb the whole system. In addition, the EEG signals suffer from other common issues as non-stationary signals, or the necessity of specific training for each user and session.

It is common for artifacts to be more powerful than the EEG signal, which is one of the reasons why they provoke such detrimental effects. Therefore, if they affect a trial of the training phase, the trial can be cataloged as an outlier and discarded or we can try to clean or remove the artifacts from the contaminated trial. Anyhow, in the field of MI-BCI there are many proposals to detect and erase artifacts. One way to address this issue is to separate them using Independent Component Analysis (ICA) [2], [3]. There are also other approaches such as [4] that identify contaminated trials using other metrics, like for example the Riemannian Distance in this case. There are also other techniques that try to reduce the impact of artifacts by, for example, looking for projections in subspaces that are more robust to artifacts or non-stationarities [5], [6] but then we need to find the right balance between those subspaces that retain more relevant information and those that discard interferences. This is the reason why most of these techniques need extra parameters that drastically increase the complexity of the system because in order to find the appropriate value or combination of values, there is a Cross Validation (CV) process needed for each hyper-parameter. The CV makes the implementation slower and it may result in overfitting. In the case where the artifacts are automatically detected and discarded, it is also possible to misclassify uncontaminated trials, resulting in fewer data for the training phase.

In the next section of this manuscript we introduce the interested reader to the basic concepts of MI-BCI systems that are needed to understand how the algorithm proposed in [7] and later explained briefly in section III normalizes the underlying EEG sources. Section IV is dedicated to the main contribution of this manuscript. Unlike in [7], here we provide with the theoretical explanation of why this method is useful to resolve the problems caused by artifacts. Section V shows some experiments that we have considered relevant to prove that this technique truly palliates the effects of artifacts and how good it works over real data and to finish, in section VI we go over the conclussions that can be drawn with the usage of this method.

II. Basics on MI-BCI systems

To explain the technique developed in this manuscript, it's important to explain how a person can communicate directly through his or her brain with MI-BCI systems by measuring the electromagnetic signals of the brain. We also explain the dimensionality reduction stage that is widely used in MI-BCI and the classification stage that is always needed to determine the intentions of the system users.

The MI-BCI systems rely on the brain processes that occur during the movement of a certain part of the body. When a person is relaxed and not thinking actively of anything, electromagnetic waves are generated in the brain. When these oscillations occur in a certain part of the brain, we refer to them as an Event Related Synchronization (ERS). When this person starts moving a part of the body or just imagining its movement, the ERS disappears from the motor-cortex part of the brain associated to the body part of the imagined movement. This is called an Event Related Desynchronization (ERD) [8]. By looking at the brain parts where the ERDs and ERSs happen, we can decode which body part the person is thinking about. Therefore we can decode a pre-established command associated to the imagination of a specific movement.

From the point of view of the users, they imagine a movement and then the system is capable of discerning between the movements they have thought about, relying on two different phases: training and test phase. To begin, we need to establish time windows called trials. To train the system the user is told to imagine a certain movement during each trial. Once the computer has collected enough labeled trials, the user can start using the BCI system by thinking about the movement associated to the action he or she wants to transmit.

We need to express this biological process in a mathematical way. We can start by defining $\mathbf{X}$ as the matrix notation of the signals that are recorded by the EEG sensors. In the same way, we can define $\mathbf{S}$ which corresponds with the sources that produce the ERS when they are active and the ERD when they are not. The EEG sensors record the brain signals that are transmitted from the sources inside the brain. We will refer to the mixing matrix that projects the sources on the sensors as $\mathbf{A}$. According to this notation it is common to find in the literature the following equation where $\mathbf{N}$ represents an additive Gaussian noise:

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}, \quad \mathbf{X},\ \mathbf{S},\ \mathbf{N} \in \mathbb{R}^{n_s \times n_m} \tag{1}$$

where we refer to n_s and n_m as the number of sensors and the number of samples per trial respectively. The EEG signals provided by the headset are filtered between 8-30Hz which is the spectrum range where the ERD and ERS occurs.

For example, a way to determine whether a source is active or not during a trial is to compute the covariance of the signals. If the source is active, the sensors that are closer to the source will record a signal with higher variance. In fact, a simple MI-BCI system consists in computing the averaged covariance matrix corresponding to the imagination of each one of the movements during the training phase. Then, during the test phase, each test trial covariance is compared with the mean covariance of each movement to determine to which one it is more alike. That is in essence what the method proposed by Barachant in [9] does.

To obtain a good spatial resolution a minimum number of EEG sensors is needed, but then the dimensionality of the covariances might be too high to work with. To solve this problem the Common Spatial Patterns (CSP) algorithm is used. This technique, which was proposed in [10], provides orthogonal spatial filters ($\boldsymbol{w}$) that maximize the ratio between the two averaged covariance matrices of each movement ($\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_1}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_2}}$) and at the same time diagonalizes them. The solution to find $\boldsymbol{w}$ is obtained by the maximization problems:

$$\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_1}} \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_2}} \boldsymbol{w}}, \quad \max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_2}} \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_1}} \boldsymbol{w}}. \tag{2}$$

The two maximization problems of the equation 2 have equivalent solutions since the vector that maximizes one of them, minimizes the other. Both problems can be solved by a unique Reighley quotient which solution is given by the generalized eigen-problem:

$$\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_1}} \boldsymbol{w} = \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{c_2}} \boldsymbol{w}\ \lambda \tag{3}$$

where each pair of eigenvector and eigenvalue gives a different solution. Each eigenvector is a projection on which the variance of the signal corresponding to the imagination of a movement is maximized with respect to the other, and each eigenvalue is the ratio between the two variances. By selecting the P eigenvectors associated to the $P/2$ greatest and to the $P/2$ lowest eigenvalues, the P projection that best separates the two signals variance are selected. The selected filters are coupled to form the matrix $\mathbf{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_{P/2}, \boldsymbol{w}_{n_s - P/2}, ..., \boldsymbol{w}_{n_s}]$.

The features used to classify can be the whole covariance matrix although it is more common to just use the variance of the filtered signal:

$$f_\tau = diag(\mathbf{W}^\top \boldsymbol{\Sigma}_{\mathbf{X}_\tau} \mathbf{W}) \tag{4}$$

with $diag()$ being the diagonal operator and $\boldsymbol{\Sigma}_{\mathbf{X}_\tau}$ the covariance of the trial τ. To classify these features, any algorithm in the literature can be used, as, for instance the Linear Discriminant Analysis or Support Vector Machine.

III. Power normalization of the underlying sources

The power of the received signals $\mathbf{X}$ is a mixture of the power of $\mathbf{S}$ and the mixing matrix $\mathbf{A}$, plus the power of noise. In our work [7] we propose a method to estimate the power of $\mathbf{S}$ from the observations $\mathbf{X}$ using an iterative algorithm. To do this, the first step is to compute the averaged covariance of the training trials. This can be done with any estimator, in this case we use the arithmetic mean of the training trials covariances:

$$\boldsymbol{\Sigma}_{\boldsymbol{x}}^{(0)} = \frac{1}{T} \sum_{\tau=1}^{T} \boldsymbol{\Sigma}_{\mathbf{X}_\tau}^{(0)} \tag{5}$$

where T represents the total number of training trial without distinguishing the classes and the upper index $^{(0)}$ denotes the estimation at iteration zero. Then the normalization factor, at iteration k of each trial sample ($\gamma_\tau^{(k)}(t)$) can be computed as in [7]:

$$\gamma_\tau^{(k)}(t) = \sqrt{\boldsymbol{x}_\tau(t)^\top \left(\boldsymbol{\Sigma}_{\boldsymbol{x}}^{(k-1)}\right)^{-1} \boldsymbol{x}_\tau(t)} \tag{6}$$

where $x_\tau(t)$ is the sample at time t of the τ trial. To simplify, in the following we will omit the dependency with time.

Then, each trial covariance at the iteration k is computed as:

$$\begin{aligned} \bar{\boldsymbol{x}}_\tau^{(k)} &= \boldsymbol{x}_\tau / \gamma_\tau^{(k)}, \\ \boldsymbol{\Sigma}_{\mathbf{X}_\tau}^{(k)} &= \frac{1}{n_m - 1} \sum_{i=1}^{n_m} \left(\bar{\boldsymbol{x}}_\tau^{(k)}\right)^\top \bar{\boldsymbol{x}}_\tau^{(k)} \end{aligned} \tag{7}$$

by plugging the previous equation in 5 we iterate until convergence. A small fixed number of iterations is usually enough. For example, for the experimental results in section V we use 3 iterations.

IV. How the power normalization of the sources attenuates contaminated samples

In this section we are going to see why the normalization of the samples using the method described in [7] is not only useful because it helps to equalize the power of the underlying sources, but also because it helps by applying a greater attenuation factor to those samples that were contaminated by artifacts.

To see this, we can define an artifact as any signal variation that does not come from the brain. In that sense, an artifact can be any kind of added noise or just a scale that attenuates or amplifies the signal $\mathbf{X}$. To prevent scale distortions, it may suffice with normalizing each sample by its power but, by doing that, we could amplify very noisy samples that had very little power.

In a mathematical way, we can see an artifact as an extra term in the equation (1) that is not usual in $\mathbf{X}$ and that has a duration of a_l (artifact length) samples:

$$\begin{aligned}\mathbf{X}_{\boldsymbol{\nu}} &= \ddot{\mathbf{A}}\ddot{\mathbf{S}} + \ddot{\mathbf{N}} + \boldsymbol{\nu} \\ &= \ddot{\mathbf{X}} + \boldsymbol{\nu}, \quad \ddot{\mathbf{X}}, \boldsymbol{\nu} \in \mathbb{R}^{n_s \times a_l}\end{aligned}$$

where the $(\ddot{\cdot})$ represent the different signal parts of contaminated samples, $\boldsymbol{\nu}$ makes reference to the artifact, which is not correlated to the signal $\mathbf{X}$ and we establish that $\ddot{\mathbf{X}} \simeq \mathbf{SA}$. We can assume that the artifact contribution in the covariance matrix of the whole set of data $(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ is barely appreciable, and therefore:

$$\mathbb{E}\left[\ddot{\mathbf{X}} \times \ddot{\mathbf{X}}^\top\right] \simeq \boldsymbol{\Sigma}_{\boldsymbol{x}}. \tag{8}$$

As the normalization is applied sample to sample, in the following, we will work with independent time samples. Therefore, we proceed by computing the value of γ_c^2 from a contaminated sample. According to the definition in (6), calling $\ddot{\boldsymbol{x}}$ to a time sample of $\ddot{\mathbf{X}}$ and using the previous assumptions:

$$\begin{aligned}\gamma_c^2 &= \left(\ddot{\boldsymbol{x}}^\top + \boldsymbol{\nu}^\top\right)\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\left(\ddot{\boldsymbol{x}} + \boldsymbol{\nu}\right) \\ &= \ddot{\boldsymbol{x}}^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\ddot{\boldsymbol{x}} + 2\ddot{\boldsymbol{x}}^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\nu} + \boldsymbol{\nu}^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\nu}. \end{aligned}\tag{9}$$

From the previous equation we study the two extreme cases: first, when the power of the signal is dominant and therefore it can be considered a clean sample; second, when the power of the artifact is dominant. In the rest of the cases, the cross term $2\ddot{\boldsymbol{x}}^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\nu}$ plays an important role and the attenuation factor is somewhere in the middle of the two previous cases.

When the power of the signal is dominant it means that $\ddot{\boldsymbol{x}}^\top\ddot{\boldsymbol{x}} \gg \boldsymbol{\nu}^\top\boldsymbol{\nu}$ and, consequently, $\gamma_c^2 \simeq \ddot{\boldsymbol{x}}^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\ddot{\boldsymbol{x}}$. This means that we would be equalizing the power of the underlying sources as we described in [7].

When the power of the artifact dominates the sample we have that $\ddot{\boldsymbol{x}}^\top\ddot{\boldsymbol{x}} \ll \boldsymbol{\nu}^\top\boldsymbol{\nu}$ and, in the same way as before: $\gamma_c^2 \simeq \boldsymbol{\nu}^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\nu}$. This case was not studied in our previous work and we proceed its examination in the following.

We apply the Singular Value Decomposition (SVD) to the averaged covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$. Because $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is a covariance matrix, $\mathbf{U}$ coincides with its eigenvectors and $\boldsymbol{\Lambda}^{-1}$ is a diagonal matrix formed with its eigenvalues $\boldsymbol{\lambda}_i^{-1} > 0, \ \forall i \in n_s$. The inverse of the covariance matrix SVD is:

$$\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^\top$$

Therefore, the value of the normalization term when the sample is dominated by an artifact, is similar to:

$$\gamma_c^2 \to \boldsymbol{\nu}^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\nu} = \sum_{i=1}^{n_s}\langle \boldsymbol{u}_i, \boldsymbol{\nu}\rangle\lambda_i^{-1}$$

Where the $\langle\cdot,\cdot\rangle$ represents the scalar product and $\boldsymbol{u}_i$, $\boldsymbol{\lambda}_i$ are the $i-$th eigenvector and eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ respectively. During motor imagery tasks, EEG signals from different sensors tend to be correlated [11], which means that most of their power relies on a few directions. By looking at equation (10) we can see that the value of γ_c^2 depends on the inverse of each eigenvalue and the result of the scalar product of the artifact sample with the corresponding eigenvector. Therefore we can say that if the artifact is aligned with the signal, it means that the sample would not be drastically attenuated or that it could even be amplified (in case the eigenvalue is greater than one). This is not usually expected, because artifacts are by definition not correlated to the signal and it is more common that artifacts are randomly distributed among all directions. That means that most part of their power is aligned with some of the directions with lowest eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. The weakest eigenvalues are close to zero and their inverse $(\boldsymbol{\lambda}_i^{-1})$ are usually quite high. When there is power in a none signal direction, the normalization factor (γ_c^2) increases rapidly and therefore, it attenuates those samples with little representation of the signal of interest.

V. Experimental Results

In this section we present the results of the simulations, first over synthetic data and later over real data.

For the synthetic simulation it has been generated a unique trial with five channels in which the covariance matrix eigenvalues are: (3.8, 0.46, 0.42, 0.16, 0.12). Three artifacts have been placed in the samples (30, 50, 70). The first artifact is aligned with the weakest eigenvector and has more power than the average of samples. The second artifact is aligned with the main direction of the signal and it is the most powerful sample. The last artifact is generated from Gaussian noise in a random direction and has the signal average power. The generated signal is represented in the top plot of figure 1. The middle plot shows the proposed normalization factor in a logarithmic representation. The bottom plot, shows the result of the normalized signal.

It can be seen how the artifact that had the direction of the weakest eigenvector has the most powerful normalization term. In contrast, the artifact that was placed in the main direction of the signal has a normalization term that is not as big as its power and that the random artifact also has a strong normalization term.

To test how the normalization term affects real data, we will use the Dataset 2a from the IV BCI competition [12]. This competition provides with a label that marks trials that were contaminated by artifacts. Although we can not access to the exact samples of the artifacts, by having a look at the normalization parameter we can guess where they occurred. For example, in the figure 2, it seems that in this trial there were two important artifacts centered in sample 125 and in sample 360 and between the two artifacts, the signal seems to be very noisy. This trial was selected because it is a good example of how before using the normalization the classification is not right but after doing so, we obtain the correct label.

To provide with a more general result, we present the table I where the accuracy result for clean trials and for contaminated trials is represented. Here we can see how the improvement of

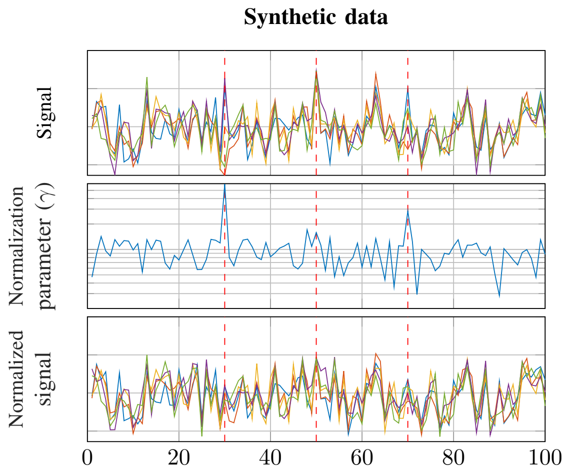


Fig. 1. Synthetic simulation with three artifacts in samples 30, 50 and 70 (red lines). The first artifact is an artifact aligned with the weakest eigenvector of the covariance of data, the second is aligned with the strongest eigenvector and the third is an arbitrary Gaussian noise.

User	Artifact count	Clean trials Std.	Clean trials Proposed	Contaminated trials Std.	Contaminated trials Proposed
A01	22	88.40	89.23	77.43	84.62
A02	23	73.78	73.20	77.22	78.88
A03	33	86.09	89.41	82.51	89.51
A04	86	68.88	70.25	63.11	65.90
A05	38	57.86	59.74	62.93	62.93
A06	142	61.12	61.74	63.17	64.75
A07	28	90.31	91.62	89.25	91.59
A08	71	86.24	89.02	87.76	89.85
A09	45	82.23	84.17	76.61	82.54
Mean	Accuracy	77.21	78.71	75.56	78.95

TABLE I
ACCURACY RESULTS (%) COMPARING THE STANDARD (STD.) METHOD AGAINST THE USE OF THE PROPOSED NORMALIZATION OVER CLEAN AND CONTAMINATED TRIALS. EACH USER HAS 288 TOTAL TRIALS.

the proposed technique over contaminated trials is remarkable in comparison to the improvement over clean trials, but that in any case it is convenient to use it. To obtain those results 10 Monte-Carlo simulations were performed. In each one of them 40 random trials were selected for training and 40 for testing. This was repeated for each pair of movements of each session and user, using 6 spatial filters of CSP in combination with the classifier Riemanian Distance to Mean. [9].

VI. Conclusions

In this work we have seen how to palliate the effects of the artifacts in a very robust way without using any extra parameter. This new algorithm is applied to all trials because it benefits most of the clear trials and it specially benefits trials that were contaminated by artifacts. Although it is an iterative algorithm, it is not very computationally demanding because it converges with a few iterations and because it does not need any hyper parameters. As a consequence of this, over-fitting is not an issue. Furthermore, since it only tries to improve the covariance of trials and do not look for artifacts in a explicit way, it avoids the problem of mislabeling clear samples as artifacts.

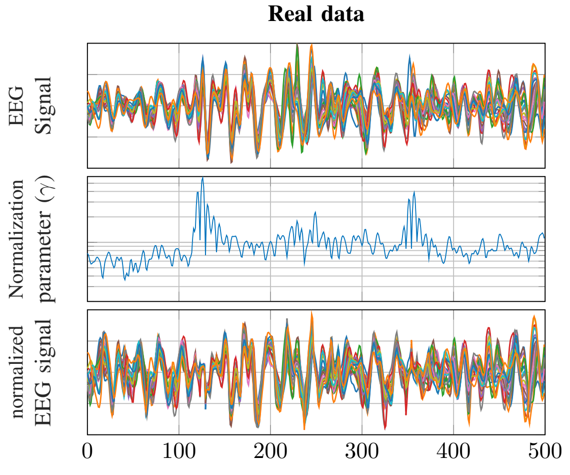


Fig. 2. Real contaminated trial from user A01. The normalization term that is shown was computed using the averaged covariance matrix for this user.

References

[1] R. Martín-Clemente, J. Olias, D. Beeta, A. Cichocki, and S. Cruces, “Information Theoretic Approaches for Motor-Imagery BCI Systems: Review and Experimental Comparison,” *Entropy*, vol. 20, no. 7, 2018.

[2] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski, “Removing Electroencephalographic Artifacts by Blind Source Separation,” *Psychophysiology*, vol. 37, no. 2, 2000.

[3] I. Dowding, S. Haufe, and M. Tangermann, “Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals,” *Behavioral and brain functions : BBF*, vol. 7, 2011.

[4] A. Barachant, A. Andreev, and M. Congedo, “The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry,” *TOBI Workshop lV*, 2013.

[5] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, “Stationary Common Spatial Patterns for Brain–Computer Interfacing,” *Journal of Neural Engineering*, vol. 9, no. 2, 2012.

[6] W. Samek, S. Nakajima, M. Kawanabe, and K.-R. Müller, “On Robust Parameter Estimation in Brain-Computer Interfacing,” *Journal of Neural Engineering*, vol. 14, no. 6, nov 2017.

[7] J. Olias, R. Martín-Clemente, M. A. Sarmiento-Vega, and S. Cruces., “EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimators,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2019, in print.

[8] H. Aboul Ella and A. Ahmad Taher, *Brain-Computer Interfaces: current trends and applications*. Springer International Publishing, 2014.

[9] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, “Multiclass Brain-Computer Interface Classification by Riemannian Geometry,” *IEEE transactions on bio-medical engineering*, vol. 59, 10 2011.

[10] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, “Optimal Spatial Filtering of Single Trial EEG During Imagined Hand Movement,” *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, 2000.

[11] R. Bhavsar, Y. Sun, N. Helian, N. Davey, D. Mayor, and T. Steffert, “The Correlation Between EEG Signals as Measured in Different Positions on Scalp Varying with Distance,” *Procedia Computer Science*, vol. 123, 2018.

[12] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, “Dataset IIa,” *BCI Competition 2008 – Graz data set A*, 2008.

Publication C

Information Theoretic Approaches for Motor-Imagery BCI Systems

- **Title:** Information Theoretic Approaches for Motor-Imagery BCI Systems: Review and Experimental Comparison.
- **Authors:** Rubén Martín-Clemente; Javier Olias; Deepa Beeta Thiyam; Andrzej Cichocki and Sergio Cruces.
- **DOI:** 10.3390/e20010007.
- **Published in:** Entropy (Volume: 20 , Issue: 1).
- **Impact factor:** 2.303;
- **Quartile:** Q2 (Physics)
- **Date:** January 2018.
- **Article:** 7.
- **Publisher:** MDPI.
- **Abstract:** Brain computer interfaces (BCIs) have been attracting a great interest in recent years. The common spatial patterns (CSP) technique is a well-established approach to the spatial filtering of the electroencephalogram (EEG) data in BCI applications. Even though CSP was originally proposed from a heuristic viewpoint, it can be also built on very strong foundations using information theory. This paper reviews the relationship between CSP and several information-theoretic approaches,

including the Kullback–Leibler divergence, the Beta divergence and the Alpha-Beta log-det (AB-LD)divergence. We also revise other approaches based on the idea of selecting those features that are maximally informative about the class labels. The performance of all the methods will be also compared via experiments.

Summary of Publication C

In this review paper we gather different CSP variant, we revisit their theoretical background and later we compare their performance under different tests. We pay special attention to those variants that are based on a information theory framework and more specifically to the divergence interpretations of the CSP algorithm. However, we compare the information theory based algorithm with the classical CSP and other variants which have not a information theory background. In that sense, we study the Regularized Tikhonov-CSP (RTCSP) and FBCSP algorithms.

Contribution 1: *Theoretical and real data comparison of different CSP variants.*

In the article we examine those methods from a theoretical point of view and we also perform a set of experimental comparison.

We compare the methods DivCSP, developed by Wojciech Samek and based in the beta divergence, the Sub-LD which is based in the alpha-beta divergence, and rest of the variants that we have commented (CSP, FBCSP and RTCSP) all of them in combination with the standard LDA classification algorithm. The FBCSP method was lightly modified with respect to the original proposal that uses 4Hz-bands while we used the defined bands (α,β,θ). Recently, we have been aware that this configuration has been chosen by other works [98]. We compare the execution time in c.I, in figure c.6 we compare de overall accuracy and in figure c.7 we compare the accuracy per pairs of movements. We also analyze the accuracy per user in figure c.8 on which we can see the p-value of each method in comparison with CSP, showing that none of the method seems to be superior to the rest in an overall context, while for some users, it might seem to be a method superior to the rest. However the best method varies from one user to another and therefore we can not conclude that any of the method is superior to the rest. In the figures c.9 and c.10 we show the histograms of the selected values for the divergences methods and we difference the cases on which it performed better than CSP or worse, with the hope that the reader would be able to draw his/her own conclusions. Finally we also perform a couple of test related to the robustness of the algorithms and show the results in c.11 and c.12.

Contribution 2: *Modification of a previous technique to select the number of CSP filters.*

We select the number of filters of the CSP algorithm using a novel modification of the method described in [99]. While Yang suggested to use the regular two sides t-test, we used the right side t-test because as we are only interested in checking if there is an improvement in accuracy when increasing the number of filters we are only interested in one side of the t-Student distribution. By doing it, we select the optimal number of spatial filter for each user and we provide with figure c.5 that justify the selection of this method over a fixed number of filters. In addition, in figure c.5, we also plot an histogram showing the probability of choosing a given number of filters.

Contribution 3: *We conclude that the classic CSP algorithm is still one of the best options.*

In the light of the results obtained we can see that the considered variants of the CSP algorithm can only outperform the classical approach in specific cases and at the cost of using CV techniques that slow down the training process and which require more training data.

Review

Information Theoretic Approaches for Motor-Imagery BCI Systems: Review and Experimental Comparison

Rubén Martín-Clemente [1,*], Javier Olias [1], Deepa Beeta Thiyam [1,2], Andrzej Cichocki [3,4,5] and Sergio Cruces [1,*]

[1] Departamento de Teoría de la Señal y Comunicaciones, Universidad de Sevilla, Camino de los Descubrimientos s/n, Seville 41092, Spain; folias@us.es (J.O.)
[2] Department of Sensor and Biomedical Technology, School of Electronics Engineering, VIT University, Vellore, Tamil Nadu 632014, India; thiyamdeepa@gmail.com (D.B.T.)
[3] Skolkovo Institute of Science and Technology (Skoltech), Moscow 143026, Russia; A.Cichocki@skoltech.ru
[4] Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan; a.cichocki@riken.jp (A.C.)
[5] Systems Research Institute, Polish Academy of Sciences, Warsaw 01-447, Poland.
* Correspondence: ruben@us.es (R.M.-C.); sergio@us.es (S.C.); Tel.: +34-954-487-475

Received: 31 October 2017; Accepted: 19 December 2017; Published: date

Abstract: Brain computer interfaces (BCIs) have been attracting a great interest in recent years. The common spatial patterns (CSP) technique is a well-established approach to the spatial filtering of the electroencephalogram (EEG) data in BCI applications. Even though CSP was originally proposed from a heuristic viewpoint, it can be also built on very strong foundations using information theory. This paper reviews the relationship between CSP and several information-theoretic approaches, including the Kullback–Leibler divergence, the Beta divergence and the Alpha-Beta Log-Det divergence. We also revise other approaches based on the idea of selecting those features that are maximally informative about the class labels. The performance of all the methods will be also compared via experiments.

Keywords: common spatial patterns; generalized divergences; brain computer interfaces

1. Introduction

The electroencephalogram (EEG) is a record over time of the differences of potential that exist between different locations on the surface of the head [1,2]. It originates from the summation of the synchronous electrical activity of millions of neurons distributed within the cortex. In recent years, there has been a growing interest in using the EEG as a new communication channel between humans and computers. Brain-computer interfaces (BCIs) are computer-based systems that enable us to control a device with the mind, without any muscular intervention [3,5–7]. This technology, though not yet mature, has a number of therapeutic applications, such as the control of wheelchairs by persons with severe disabilities, but also finds use in fields as diverse as gaming, art or access control.

There are several possible approaches for designing a BCI [1,5,8]. Among them, motor imagery (MI)-based BCI systems seem to be the most promising option [7,9–11]. In MI-based BCI systems, the subject is asked to imagine the movement of different parts of his or her body, such as the hands or the feet. The imagined actions are then translated into different device commands (e.g., when the subject imagines the motion of the left hand, the wheelchair is instructed to turn to the left). What makes this possible is that the spatial distribution of the EEG differs between different imagined movements. More precisely, since each brain hemisphere mainly controls the opposite side of the body, the imagination of right and left limb movements produces a change of power over the contralateral left and right brain motor areas. These fluctuations, which are due to a pair of phenomena known as event-related desynchronization (ERD) or power decrease and event-related synchronization (ERS) or power increase [12,13], can be detected and converted into numerical features. By repeating the imagined actions several times, a classifier can be trained to determine which kind of motion the subject is imagining (see [14] for a review). In practice, three classes of MI are used in BCIs, namely the movements of the hands, the feet and the tongue. Left hand movement imagery is more prominent in the vicinity of the electrode C4 (see Figure 1), while right hand imagined actions are detected around electrode C3 [15]. The imagery of feet movements appears in the electrode Cz and its surrounding area; nevertheless, it is not usually possible to distinguish between left foot or right foot motor imagery because the corresponding activation areas are too close in the cortex [12,15]. Finally, imagery of tongue movements can be detected on the primary motor cortex and the premotor cortex [16]. One of the inherent difficulties of designing a BCI is that the EEG features are highly non-stationary and vary over sessions. To cope with this problem, the background state of the subject (i.e., his or her motivation, fatigue, etcetera) and the context of the experiment can be both modeled as latent variables, whose parameters can be estimated using the expectation-maximization (EM) algorithm [17,18]. Overall, current BCI approaches achieve success rates of over 90%, although much depends on the person from whom the EEG data are recorded [15].

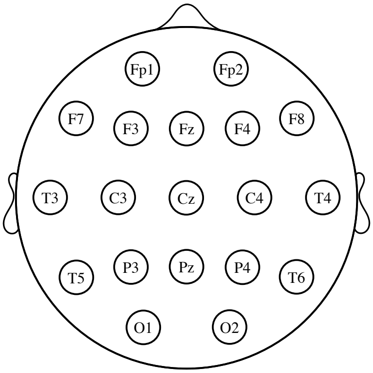


Figure 1. Electrode locations of the international 10–20 system for EEG recording. The letters "F", "T", "C", "P" and "O" stand for frontal, temporal, central, parietal and occipital lobes, respectively. Even numbers correspond to electrodes placed on the right hemisphere, whereas odd numbers refer to those on the left hemisphere. The "z" refers to electrodes placed in the midline.

The common spatial patterns (CSP) method [4,5,19–22] is a method of dimensionality reduction that is widely used in BCI systems as a preprocessing step. Basically, assuming two classes of MI-EEG signals (e.g., left hand and right hand MI tasks), CSP projects the EEG signals onto a low-dimensional subspace, which captures the variability of one of the classes while, at the same time, trying to minimize the variance in the other class. The goal is to enhance the ability of the BCI to discriminate between the different MI tasks, and it has been shown that CSP is able to reduce the dimension of the data significantly without decreasing the classification rate. It is noteworthy that CSP admits an interesting probabilistic interpretation. Under the assumption of Gaussian distributed data, CSP is equivalent to maximizing the symmetric Kullback–Leibler (KL) divergence between the probability distributions of the two classes after the projection onto the low dimensional space [23,24]. As a generalization of this idea in the context of BCI, it is interesting to investigate the dimensionality reduction ability of other different divergence-based criteria, which is drawing a lot of interest among the computational neuroscience community.

The present manuscript is a review of the state of the art of information theoretic approaches for motor imagery BCI systems. The article is written as a guideline for researchers and developers both in the fields of information theory and BCI, and the goal is to simplify and organize the ideas.

We will present a number of approaches based on Kullback–Leibler divergence, Beta divergence (which is a generalization of Kullback–Leibler's) and Alpha-Beta Log-Det divergence (which include as special cases Stein's loss, the S-divergence or the Riemannian metric), as well as their relation to CSP. We will also review a technique based on the idea of selecting those features that are maximally informative about the class labels. Complementarily, for the purpose of comparison, several non-information theoretic variants of CSP and their different regularization schemes are revised in the paper. The performance of all approaches will be evaluated and compared through simulations using both real and synthetic datasets.

The paper is organized as follows: The CSP algorithm is introduced in Section 2. Section 3 introduces the main characteristics of the Kullback–Leibler divergence, the Beta divergence and the Alpha-Beta Log-Det divergence, respectively, as well as their application to the problem of designing MI-BCI systems and the algorithms used to optimize them. Section 4 reviews an information-theoretic feature extraction framework. Section 5 presents, as has been said before, several extensions of CSP not based on information-theoretic principles. Finally, Section 6 presents the results of some experiments in which the performances of the above criteria are tested, in terms of their accuracy, computational burden and robustness against errors.

EEG Measurement and Preprocessing

For measuring the EEG, several different standardized electrode placement configurations exist. The most common among them is the International 10–20 system, which uses a set of electrodes placed at locations defined relative to certain anatomical landmarks (see Figure 1). The ground reference electrode is usually positioned at the ears or at the mastoid. To obtain a reference-free system, it is common practice to calculate the average of all the electrode potentials and subtract it from the measurements [1,2].

The EEG is usually contaminated by several types of noise and artifacts. Eye blinks, for example, elicit a large potential difference between the cornea and the retina that can be several orders of magnitude greater than the EEG. In the rest of the paper, it is assumed that the signals have already been pre-processed to remove noise and interferences. To this end, several techniques [25], such as autoregressive modeling [26], the more complex independent component analysis (ICA) [27], or the signal space projection (SSP) method [28], have shown good or excellent results (see also [29] and the references therein). Signal preprocessing includes also the division of the EEG into several frequency bands that are separately analyzed [30,31]. The "mu" band (8–15 Hz) and the "beta" band (16–31 Hz) are particularly useful in BCIs, as they originate from the sensorimotor cortex, i.e., the area that controls voluntary movements [2].

2. The Common Spatial Pattern Criterion

In this section, we present the common spatial patterns (CSP) method [4,5,19–22,32,33]. Consider a two-class classification problem, where the EEG signals belong to exactly one of two classes or conditions (e.g., left-/right-hand movement imagination).

To fix notation, let $\mathbf{X}_{i,k} \in \mathbb{R}^{D\times T}$ be the matrix that contains the EEG data of class $i \in \{1,2\}$ in the k-th trial or experiment, where D is the number of channels and T the number of samples in a trial. The corresponding sample covariance estimator is defined by:

$$\mathbf{\Sigma}_{i,k} = \frac{1}{T-1}\mathbf{X}_{i,k}\mathbf{X}_{i,k}^{\top}, \tag{1}$$

where $(\cdot)^\top$ denotes "transpose". Here, the EEG signals are assumed to have zero-mean, which is fulfilled as they are band-pass filtered (see the previous section). If L trials per class are performed, the spatial covariance matrix for class i is usually calculated by averaging the trial covariance matrices as:

$$\mathbf{\Sigma}_i = \frac{1}{L}\sum_{k=1}^{L}\mathbf{\Sigma}_{i,k} \tag{2}$$

In practice, these covariance matrices are often normalized in power with the help of the following transformation:

$$\mathbf{\Sigma}_i \leftarrow \mathbf{\Sigma}_i / \operatorname{tr}\left(\mathbf{\Sigma}_i\right), \tag{3}$$

where $\operatorname{tr}(\cdot)$ denotes the trace operator.

After the BCI training phase, in which matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are estimated using training data, suppose that a new, not previously observed, data matrix $\mathbf{X} \in \mathbb{R}^{D\times T}$ of imagined action is captured. The problem that arises is to develop a rule to allocate these new data to one class or the other. A useful approach is to define a weight vector $\boldsymbol{w} \in \mathbb{R}^D$ (also known as a 'spatial filter') and allocate $\mathbf{X}$ to one class if the variance of $\boldsymbol{w}^\top\mathbf{X}$ exceeds a certain predefined threshold and to the other if not; this relates to the fact that event-related desynchronizations and event-related synchronizations, i.e., the phenomena underlying the MI responses, are associated with power decreases/increases of the ongoing EEG activity [13].

Of course, not just any spatial filter is of value. To enhance the discrimination of the MI tasks, CSP proposes using spatial filters that maximize the variance of the band-pass filtered EEG signals in one class while, simultaneously, minimizing it for the other class. Mathematically, CSP aims at maximizing an objective function based on the the following Rayleigh quotient:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top\mathbf{\Sigma}_1\boldsymbol{w}}{\boldsymbol{w}^\top\mathbf{\Sigma}_2\boldsymbol{w}} = \frac{\sigma_1^2}{\sigma_2^2}, \tag{4}$$

where σ_i^2 is the variance of the i-th projected class and $\mathbf{\Sigma}_i$ is the covariance matrix of the i-th class.

It is a straightforward derivation to obtain that the spatial filters that hierarchically maximize (4) can be computed by solving the generalized eigenvalue problem:

$$\mathbf{\Sigma}_1\boldsymbol{w} = \lambda\mathbf{\Sigma}_2\boldsymbol{w}. \tag{5}$$

Each eigenvector $\boldsymbol{w}_i$ gives a different solution. Observe that:

$$\boldsymbol{w}_i^\top\mathbf{\Sigma}_1\boldsymbol{w}_i = \lambda_i\boldsymbol{w}_i^\top\mathbf{\Sigma}_2\boldsymbol{w}_i \rightarrow \lambda_i = \frac{\boldsymbol{w}_i^\top\mathbf{\Sigma}_1\boldsymbol{w}_i}{\boldsymbol{w}_i^\top\mathbf{\Sigma}_2\boldsymbol{w}_i} = J(\boldsymbol{w}_i),$$

where λ_i is the generalized eigenvalue corresponding to $\boldsymbol{w}_i$. Therefore, the larger (or smaller) the eigenvalue, the larger the ratio between the variances of the two classes and the better the discrimination accuracy of the filter.

The latter readily suggests selecting the spatial filters among the principal and the minor eigenvectors (i.e., the eigenvectors associated with the largest and smallest eigenvalues, respectively). Let:

$$\mathbf{W}_{CSP} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d] \in \mathbb{R}^{D\times d} \tag{6}$$

be the matrix that collects these $d \leq D$ top (i.e., most discriminating) spatial filters. Given a data matrix $\mathbf{X} \in \mathbb{R}^{D \times T}$ of observations, all of the same class, the outputs of the spatial filters are defined as:

$$\boldsymbol{y}_i = \boldsymbol{w}_i^\top \mathbf{X}, \qquad i = 1, \dots, d, \qquad (d \leq D) \tag{7}$$

which can be gathered at the $d \times T$ output matrix $\mathbf{Y} = \mathbf{W}_{CSP}^\top \mathbf{X}$. Denoting by $\boldsymbol{\Sigma}$ the sample covariance matrix of $\mathbf{X}$, it follows that the covariance matrix of the outputs is given by $\mathbf{W}_{CSP}^\top \boldsymbol{\Sigma} \, \mathbf{W}_{CSP}$, while the variance of the output of the i-th spatial filter is equal to $\boldsymbol{w}_i^\top \boldsymbol{\Sigma} \boldsymbol{w}_i$. Finally, not the sample variances, but the log transformed sample variances of the outputs, i.e.,

$$F_i = \log(\boldsymbol{w}_i^\top \boldsymbol{\Sigma} \boldsymbol{w}_i), \qquad i = 1, \dots, d, \tag{8}$$

are used as features for the classification of the imagined movements. Observe that, as long as $d < D$, the dimensionality of the data is reduced.

CSP admits an interesting neurological interpretation. First note that the scalp EEG electrodes measure the addition of numerous sources of neural activity, which are spread over large areas of the neocortical surface, and this does not always allow a reliable localization of the cortical generators of the electrical potentials. It has been suggested that CSP linearly combines the EEG signals so that the sources of interest are enhanced while the others are suppressed [34].

Another interpretation of (4) may be as follows: the basic theory of principal component analysis (PCA) states that maximizing $\boldsymbol{w}^\top \boldsymbol{\Sigma}_i \boldsymbol{w}$ finds the direction vector that best fits, in the least-squares sense, the data of class i in the D-dimensional space. Similarly, minimizing this ratio obtains the opposite effect. Thus, we can interpret that CSP seeks directions that fit well with the data in one class, but are not representative of the data in the other class. By projecting the EEG data onto them, a significant reduction of the variance of one of the classes, while preserving the information content of the other, can be thus obtained.

An interesting generative model perspective has been proposed in [35,36]. Here, the above data matrices are assumed to be generated by a latent variable model:

$$\mathbf{X}_i(:, k) = \mathbf{A} \, \mathbf{Y}_i(k) + \mathbf{N}_i(k),$$

where we have used the notation $\mathbf{X}_i(:, k) \in \mathbb{R}^D$ for the k-th column of the data matrix $\mathbf{X}_i$, i.e., it is the observation vector at time k for class i, $i = 1, 2$; $\mathbf{A} \in \mathbb{R}^{D \times s}$ is a mixing matrix, the same for both classes; $\mathbf{Y}_i(k) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_i)$ is an s-dimensional column vector of latent variables (s has to be estimated from the data) and $\mathbf{N}_i(k) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta}_i)$ is a D-dimensional vector of noise, independent of the data. Here, the covariance matrices $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Delta}_i$ are assumed to be diagonal matrices, implying that the latent factors are also independent of each other. Under this model, the columns of matrix $\mathbf{A}$ can be regarded as the 'spatial patterns' that explain how the EEG data are formed at each electrode location, where the latent variables represent the degree to which each 'spatial pattern' appears in the data. Under the assumptions that the noise is negligible and matrix $\mathbf{A}$ is square, it is noteworthy that the CSP spatial filters are precisely the columns of the matrix $\mathbf{A}^{-\top}$ [35].

3. Divergence-Based Criteria

CSP produces quite good results in general, but also suffers from various shortcomings: e.g., it is sensitive to artifacts [37,38] and its performance is degraded for non-stationary data [39]. For these reasons, CSP is still an active line of research, and a number of variants have been proposed in the literature. In particular, in this paper, we are interested in reviewing CSP-variants based on an information-theoretic framework.

There is a common assumption in the literature that the classes can be modeled by multivariate Gaussian distributions with zero-means and different covariance matrices. This assumption is based on the principle of maximum entropy, not in actual measures of EEG data. By projecting the data onto the principal generalized eigenvectors, CSP transforms them onto a lower dimensional space where the variance of Class 1 is maximized, while the variance for Class 2 is minimized. Conversely, the projection onto the minor generalized eigenvectors has the opposite effect. Since a zero-mean univariate normal variable is completely determined by its variance, we can understand the ratio (4) as a measure of how much the distributions of the projected classes differ from each other (the larger the ratio between the variances, the more different the distributions). By accepting this viewpoint, it is interesting to investigate the ability of other measures of dissimilarity between statistical distributions, rather than the ratio of the corresponding variances, to help in discriminating between the classes. In fact, the most interesting features for classification often belong to those subspaces where there is a large dissimilarity between the conditional densities of the considered classes, which is another justification for proposing a divergence maximization framework in the context of MI-BCI.

In the following sections, we review the main information-theoretic-based approaches.

3.1. Criterion Based on the Symmetric Kullback–Leibler Divergence

Divergences are functions that measure the dissimilarity or separation between two statistical distributions. Given two univariate Gaussian densities $\mathcal{N}_1(0,\sigma_1)$ and $\mathcal{N}_2(0,\sigma_2)$, their Kullback–Leibler divergence (the KL divergence between two distributions f_1 and f_2 is defined as $Div_{KL}\left(f_1 \| f_2\right) = \int_{-\infty}^{\infty} f_1(x) \log \frac{f_1(x)}{f_2(x)} dx$) is easily found to be:

$$Div_{KL}\left(\mathcal{N}_1(0,\sigma_1) \| \mathcal{N}_2(0,\sigma_2)\right) = \frac{1}{2}\left(\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right).$$

If the densities have interchangeable roles, it is reasonable to consider the use of a symmetrized measure like the one provided by the symmetrized Kullback–Leibler (sKL) divergence. This is defined simply as:

$$\begin{aligned} sDiv_{KL}\left(\mathcal{N}_1 \| \mathcal{N}_2\right) &= Div_{KL}\left(\mathcal{N}_1 \| \mathcal{N}_2\right) + Div_{KL}\left(\mathcal{N}_2 \| \mathcal{N}_1\right) \\ &= \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2}\right) - 1. \end{aligned} \tag{9}$$

The resemblance to the CSP criterion (4) is quite obvious, as was already noted, e.g., in [24]. In particular, note that, since $z + \frac{1}{z}$ increases when z goes to either infinity or zero, (9) is maximized by either maximizing or minimizing the ratio of the variances σ_1 and σ_2.

The generalization to multivariate data is straightforward. Let $\mathbf{Y} = \mathbf{W}^\top\mathbf{X}$, where $\mathbf{X}$ is the observed data matrix and $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d] \in \mathbb{R}^{D\times d}$ denotes an arbitrary matrix of spatial filters with $1 \leq d \leq D$. Under the assumption that the EEG data are conditionally Gaussian distributed for each class $c_k \in \{1,2\}$, i.e., $\mathbf{X}|c_k \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma}_k)$, the spatially-filtered data are also from a normal distribution, i.e., $\mathbf{Y}|c_k \sim \mathcal{N}(\mathbf{0},\bar{\mathbf{\Sigma}}_k)$, where:

$$\bar{\mathbf{\Sigma}}_k = \mathbf{W}^\top\mathbf{\Sigma}_k\mathbf{W} \in \mathbb{R}^{d\times d},$$

$k = 1,2$. The KL divergence between two d-dimensional multivariate Gaussian densities $f_1 = \mathcal{N}_1(\mathbf{0},\bar{\mathbf{\Sigma}}_1)$ and $f_2 = \mathcal{N}_2(\mathbf{0},\bar{\mathbf{\Sigma}}_2)$, that is,

$$Div_{KL}\left(f_1 \| f_2\right) = \int \mathcal{N}_1(\mathbf{0},\bar{\mathbf{\Sigma}}_1) \log\frac{\mathcal{N}_1(\mathbf{0},\bar{\mathbf{\Sigma}}_1)}{\mathcal{N}_2(\mathbf{0},\bar{\mathbf{\Sigma}}_2)} d\boldsymbol{y},$$

can be shown to be (after some algebra):

$$Div_{KL}\left(f_1 \| f_2\right) = \frac{1}{2}\left[\log\frac{|\bar{\mathbf{\Sigma}}_2|}{|\bar{\mathbf{\Sigma}}_1|} - d + \mathrm{trace}(\bar{\mathbf{\Sigma}}_2)^{-1}(\bar{\mathbf{\Sigma}}_1)\right], \tag{10}$$

where $|\cdot|$ stands for "determinant". The symmetrized Kullback–Leibler (sKL) divergence between the probability distributions of the two classes is now defined as:

$$\begin{aligned} Div_{sKL}\left(f_1 \| f_2\right) &= Div_{KL}\left(f_1 \| f_2\right) + Div_{KL}\left(f_2 \| f_1\right) \\ &= \frac{1}{2}\operatorname{trace}\Big((\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}) \\ &\quad +(\mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W})\Big) - d, \end{aligned}$$

where we show again the explicit dependency on $\mathbf{W}$.

We can naturally extend this formula to define the equivalent sKL matrix divergence:

$$\begin{aligned} D_{sKL}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W} \| \mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}) &= \frac{1}{2}\operatorname{trace}\Big((\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}) \\ &\quad +(\mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W})\Big) - d. \end{aligned} \tag{11}$$

It has been shown in [23] that the subspace of the filters that maximize the sKL matrix divergence,

$$\mathbf{W}_{sKL} \quad = \arg\max_{\mathbf{W}} D_{sKL}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W} \| \mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}), \tag{12}$$

coincides with the subspace of those that maximize the CSP criterion, in the sense that the columns of $\mathbf{W}_{sKL}$ and $\mathbf{W}_{CSP}$ span the same subspace:

$$\operatorname{span}(\mathbf{W}_{sKL}) \quad = \quad \operatorname{span}(\mathbf{W}_{CSP}), \tag{13}$$

that is, every column of $\mathbf{W}_{sKL}$ is a combination of the top spatial filters of $\mathbf{W}_{CSP}$ and vice versa.

In practice, $\mathbf{W}_{sKL}$ is first used to project the data onto a lower dimensional subspace, and then, $\mathbf{W}_{CSP}$ is determined by applying CSP to the projected data. Some advantage can be gained, compared to using CSP only, if in the first step the optimization of the sKL matrix divergence is also combined with some suitable regularization scheme. For example, to fight against issues caused by the non-stationarity of the EEG data, it has been proposed to maximize the regularized objective function [23]:

$$\mathcal{L}_{sKL}(\mathbf{W}) = (1-\phi)D_{sKL}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W} \| \mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}) - \phi\Delta(\mathbf{W}), \tag{14}$$

where $0 \leq \phi < 1$ and:

$$\Delta(\mathbf{W}) = \frac{1}{2L}\sum_{i=1}^{2}\sum_{k=1}^{L} Div_{KL}\left(\mathcal{N}(0, \mathbf{W}^\top \boldsymbol{\Sigma}_{i,k} \mathbf{W}) \| \mathcal{N}(0, \mathbf{W}^\top \boldsymbol{\Sigma}_i \mathbf{W})\right) \tag{15}$$

is a regularization term, where we have assumed that L trials per class have been performed and $\boldsymbol{\Sigma}_{c,k}$ is the covariance matrix in the k-th trial of class $c \in \{1,2\}$. This proposed regularization term enforces the transformed data in all the trials to have the same statistical distribution. Other ideas have been proposed in [23], and a related approach can be found in [40]. Observe also that (15) is defined on the basis of the KL divergence, not on its symmetrized version. The KL divergence is calculated by a formula similar to (10), giving:

$$Div_{KL}\left(\mathcal{N}(0, \mathbf{W}^\top \boldsymbol{\Sigma}_{i,k} \mathbf{W}) \| \mathcal{N}(0, \mathbf{W}^\top \boldsymbol{\Sigma}_i \mathbf{W})\right) \quad = \quad \frac{1}{2}\left[\log\frac{|\mathbf{W}^\top \boldsymbol{\Sigma}_i \mathbf{W}|}{|\mathbf{W}^\top \boldsymbol{\Sigma}_{i,k} \mathbf{W}|} - d + \operatorname{trace}(\mathbf{W}^\top \boldsymbol{\Sigma}_i \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_{i,k} \mathbf{W})\right]. \tag{16}$$

The inverse of $\boldsymbol{\Sigma}_{i,k}$ does not appear in (16), which makes sense if this matrix is ill-conditioned due to insufficient sample size. For this reason, the KL divergence is preferred to its symmetric counterpart. In addition, the logarithm in (16) downweights the effect of $|\mathbf{W}^\top \boldsymbol{\Sigma}_{i,k} \mathbf{W}|^{-1}$ in case $\boldsymbol{\Sigma}_{i,k}$ is nearly singular.

3.2. Criterion Based on the Beta Divergence

The beta divergence, which is a generalization of the Kullback–Leibler's, seems to be an obvious alternative measure of discrepancy between Gaussians. Given two zero-mean multivariate probability density functions $f_1(\boldsymbol{y})$ and $f_2(\boldsymbol{y})$, the beta divergence is defined for $\beta > 0$ as:

$$Div_\beta(f_1(\boldsymbol{y})\|f_2(\boldsymbol{y})) = \frac{1}{\beta}\int\left(f_1^\beta(\boldsymbol{y}) - f_2^\beta(\boldsymbol{y})\right)f_1(\boldsymbol{y})d\boldsymbol{y} - \frac{1}{\beta+1}\int\left(f_1^{\beta+1}(\boldsymbol{y}) - f_2^{\beta+1}(\boldsymbol{y})\right)d\boldsymbol{y}.$$

As $\lim_{\beta\to 0}\frac{f_1^\beta - f_2^\beta}{\beta} = \log\left(\frac{f_1}{f_2}\right)$, it can be shown that the beta divergence converges to the KL divergence for $\beta \to 0$.

Let $f_1 = \mathcal{N}(0, \bar{\boldsymbol{\Sigma}}_1)$ and $f_2 = \mathcal{N}(0, \bar{\boldsymbol{\Sigma}}_2)$, with $\bar{\boldsymbol{\Sigma}}_i = \mathbf{W}^\top\boldsymbol{\Sigma}_i\mathbf{W} \in \mathbb{R}^{d\times d}$, $i = 1, 2$, be the zero-mean Gaussian distributions of the spatially-filtered data. In this case, the symmetric beta divergence between them yields the following closed form formula [41]:

$$D_{s\beta}(\mathbf{W}^\top\boldsymbol{\Sigma}_1\mathbf{W}\|\mathbf{W}^\top\boldsymbol{\Sigma}_2\mathbf{W}) = \gamma\left(|\bar{\boldsymbol{\Sigma}}_1|^{-\beta/2} + |\bar{\boldsymbol{\Sigma}}_2|^{-\beta/2} - (\beta+1)^{d/2}\left(\frac{|\bar{\boldsymbol{\Sigma}}_2|^{\frac{1-\beta}{2}}}{|\beta\bar{\boldsymbol{\Sigma}}_1 + \bar{\boldsymbol{\Sigma}}_2|^{1/2}} + \frac{|\bar{\boldsymbol{\Sigma}}_1|^{\frac{1-\beta}{2}}}{|\beta\bar{\boldsymbol{\Sigma}}_2 + \bar{\boldsymbol{\Sigma}}_1|^{1/2}}\right)\right), \quad (17)$$

where $\gamma = \frac{1}{\beta}\sqrt{\frac{1}{(2\pi)^{\beta d}(\beta+1)^d}}$. Observe that $D_{s\beta}$ is somewhat protected against possible large increases in the elements of $\boldsymbol{\Sigma}_1$ or $\boldsymbol{\Sigma}_2$ caused by outliers or estimation errors. For example, if $\boldsymbol{\Sigma}_i$ (resp. $\bar{\boldsymbol{\Sigma}}_i$) grows, $i \in \{1, 2\}$, then the contribution of all the terms containing $\boldsymbol{\Sigma}_i$ (resp. $\bar{\boldsymbol{\Sigma}}_i$) in (17) tends to vanish. Compared with the previous case, if $\boldsymbol{\Sigma}_1$ (for example) increases, then the term:

$$\text{trace}\left((\mathbf{W}^\top\boldsymbol{\Sigma}_2\mathbf{W})^{-1}(\mathbf{W}^\top\boldsymbol{\Sigma}_1\mathbf{W})\right)$$

may dominate (11).

With the necessary changes of divergences being made, the regularizing framework previously defined by Equations (14) and (15) can be easily adapted to the present case [23]. It has been argued in [23] that small values of β penalize abrupt changes in the covariance matrices caused by single extreme events, such as artifacts, whereas a large β is more suitable to penalize the gradual changes over the dataset from trial to trial.

Alternatively, supposing that L trials per class are performed, it has been also proposed in [41] to use as the objective function the sum of trial-wise divergences:

$$\bar{D}_{s\beta}(\mathbf{W}) = \sum_{i=1}^{L} D_{s\beta}(\mathbf{W}^\top\boldsymbol{\Sigma}_{1,i}\mathbf{W}\|\mathbf{W}^\top\boldsymbol{\Sigma}_{2,i}\mathbf{W}),$$

where $\boldsymbol{\Sigma}_{1,i}$ and $\boldsymbol{\Sigma}_{2,i}$ are the covariance matrices in the i-th trial of Class 1 and Class 2, respectively.

3.3. Criterion Based on the Alpha-Beta Log Det Divergence

Given the covariance matrices of each class, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, an extension of the Kullback–Leibler symmetric matrix divergence given in Equation (11) is the Alpha-Beta log-det divergence (AB-LD), defined as [42,43]:

$$D_{LD}^{(\alpha,\beta)}(\boldsymbol{\Sigma}_1\|\boldsymbol{\Sigma}_2) = \frac{1}{\alpha\beta}\log\left|\frac{\alpha(\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-\frac{1}{2}})^\beta + \beta(\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-\frac{1}{2}})^{-\alpha}}{\alpha+\beta}\right|_+ \quad (18)$$

$$\text{for } \alpha \neq 0, \quad \beta \neq 0, \quad \alpha+\beta \neq 0,$$

where:

$$|x|_+ = \begin{cases} x & x \geq 0, \\ 0, & x < 0, \end{cases}$$

denotes the non-negative truncation operator. For the singular cases, the definition becomes:

$$D_{LD}^{(\alpha,\beta)}(\boldsymbol{\Sigma}_1 \| \boldsymbol{\Sigma}_2) = \begin{cases} \frac{1}{\alpha^2}\left[\mathrm{tr}\left((\boldsymbol{\Sigma}_2^{\frac{1}{2}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2^{\frac{1}{2}})^{\alpha} - \mathbf{I}\right) - \alpha \log|\boldsymbol{\Sigma}_2^{\frac{1}{2}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2^{\frac{1}{2}}|\right] & \text{for} \quad \alpha \neq 0,\ \beta = 0, \\ \frac{1}{\beta^2}\left[\mathrm{tr}\left((\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-\frac{1}{2}})^{\beta} - \mathbf{I}\right) - \beta \log|\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-\frac{1}{2}}|\right] & \text{for} \quad \alpha = 0,\ \beta \neq 0, \\ \frac{1}{\alpha^2}\log\left|(\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-\frac{1}{2}})^{\alpha}(\mathbf{I} + \log(\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-\frac{1}{2}})^{-\alpha})\right|_+ & \text{for } \alpha = -\beta, \\ \frac{1}{2}||\log(\boldsymbol{\Sigma}_2^{\frac{1}{2}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2^{\frac{1}{2}})||_F^2 & \text{for} \quad \alpha,\ \beta = 0. \end{cases} \tag{19}$$

It can be easily checked that $D_{LD}^{(\alpha,\beta)}(\boldsymbol{\Sigma}_1 \| \boldsymbol{\Sigma}_2) = 0$ iff $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. The interest in the AB-LD divergence is motivated by the fact that, as can be observed in Figure 2, it generalizes several existing log-det matrix divergences, such as the Stein's loss (the Kullback–Leibler matrix divergence), the *S*-divergence, the Alpha and Beta log-det families of divergences and the geodesic distance between covariance matrices (the squared Riemannian metric), among others [43].

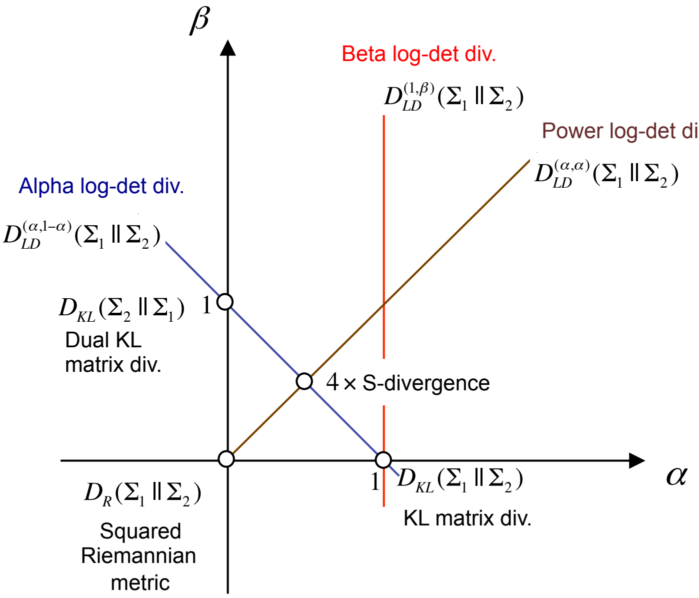

β
Beta log-det div.
$D_{LD}^{(1,\beta)}(\Sigma_1 \| \Sigma_2)$
Power log-det div.
$D_{LD}^{(\alpha,\alpha)}(\Sigma_1 \| \Sigma_2)$
Alpha log-det div.
$D_{LD}^{(\alpha,1-\alpha)}(\Sigma_1 \| \Sigma_2)$
$D_{KL}(\Sigma_2 \| \Sigma_1)$
1
Dual KL
matrix div.
4 × S-divergence
1
$D_{KL}(\Sigma_1 \| \Sigma_2)$
α
KL matrix div.
$D_R(\Sigma_1 \| \Sigma_2)$
Squared
Riemannian
metric

Figure 2. Illustration of the Alpha-Beta Log-Det divergence (AB-LD) divergence $D_{LD}^{(\alpha,\beta)}(\boldsymbol{\Sigma}_1 \| \boldsymbol{\Sigma}_2)$ in the (α, β)-plane. Note that the position of each divergence is specified by the value of the hyperparameters (α, β). This parameterization smoothly connects several positive definite matrix divergences, such as the squared Riemannian metric ($\alpha = 0, \beta = 0$), the KL matrix divergence or Stein's loss ($\alpha = 1, \beta = 0$), the dual KL matrix divergence ($\alpha = 0, \beta = 1$) and the *S*-divergence ($\alpha = \frac{1}{2}, \beta = \frac{1}{2}$), among others.

There is a close relationship between the AB-LD divergence criterion and CSP: it has been shown [42] that the sequence of Courant-like minimax divergence optimization problems [42]

$$\boldsymbol{w}_{\pi_i} = \arg \min_{\dim\{\mathcal{W}\}=D-i+1} \max_{\boldsymbol{w}\in\{\mathcal{W}\}} D_{LD}^{(\alpha,\beta)}(\boldsymbol{w}^\top \boldsymbol{\Sigma}_1 \boldsymbol{w} \parallel \boldsymbol{w}^\top \boldsymbol{\Sigma}_2 \boldsymbol{w}), \quad i = 1, \ldots, D, \tag{20}$$

yields spatial filters $\boldsymbol{w}_{\pi_i}$ that essentially coincide (i.e., up to a permutation π_i in the order) with the CSP spatial filters $\boldsymbol{w}_i$, i.e., with the generalized eigenvectors defined by (5). The permutation ambiguity can be actually avoided if we introduce a suitable scaling $\kappa \in \mathbb{R}^+$ in one of the arguments of the divergence, so (20) becomes

$$\boldsymbol{w}_i = \arg \min_{\dim\{\mathcal{W}\}=D-i+1} \max_{\boldsymbol{w}\in\{\mathcal{W}\}} D_{LD}^{(\alpha,\beta)}(\boldsymbol{w}^\top \boldsymbol{\Sigma}_1 \boldsymbol{w} \parallel \kappa\, \boldsymbol{w}^\top \boldsymbol{\Sigma}_2 \boldsymbol{w}), \quad i = 1, \ldots, D, \tag{21}$$

where κ is typically close to the unity.

For $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d] \in \mathbb{R}^{D\times d}$ with $1 \leq d \leq D$, a criterion based on the AB log-det divergence takes the following form [42]

$$\mathcal{L}_{LD}(\mathbf{W}) = D_{LD}^{(\alpha,\beta)}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W} \| \, \kappa\, \mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}) - \eta \left(P(c_1)\mathbf{R}_1 + P(c_2)\mathbf{R}_2 \right) \;, \tag{22}$$

where $P(c_1)$ and $P(c_2)$ are the prior probabilities of Class 1 and Class 2,

$$\mathbf{R}_1 = \frac{1}{L}\sum_{i=1}^{L} D_{LD}^{(\alpha,\beta)}(\mathbf{W}^\top \boldsymbol{\Sigma}_{1,i} \mathbf{W} \| \mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W}), \tag{23}$$

$$\mathbf{R}_2 = \frac{1}{L}\sum_{i=1}^{L} D_{LD}^{(\alpha,\beta)}(\mathbf{W}^\top \boldsymbol{\Sigma}_{2,i} \mathbf{W} \| \mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}), \tag{24}$$

where L is the number of trials per class and $\boldsymbol{\Sigma}_{1,i}$ and $\boldsymbol{\Sigma}_{2,i}$ are the covariance matrices in the i-th trial of Class 1 and Class 2, respectively.

The regularization term:

$$P(c_1)\mathbf{R}_1 + P(c_2)\mathbf{R}_2$$

may be interpreted as a sort of within-class scatter measure, which is reminiscent of that used in Fisher's linear discriminant analysis. The parameter η thus controls the balance between the maximization of the between-class scatter and the minimization of the within-class scatter. Observe that when both classes are equiprobable, $P(c_1) = P(c_2) = 1/2$, this regularization term is the equivalent of the one defined in Equation (15).

3.4. Algorithms for Maximizing the Divergence-Based Criteria

To give some idea of how the objective functions are, Figure 3 depicts the divergences defined in Sections 3.1, 3.2 and 3.3 assuming two-dimensional data in the particular case $d = 1$ (so that the projected data are one dimensional). These divergence-based criteria can be optimized in several ways. In practice, a two-step procedure seems convenient, in which a first "whitening" of the observed EEG data is followed by maximization where the search space is the set of the orthogonal matrices.

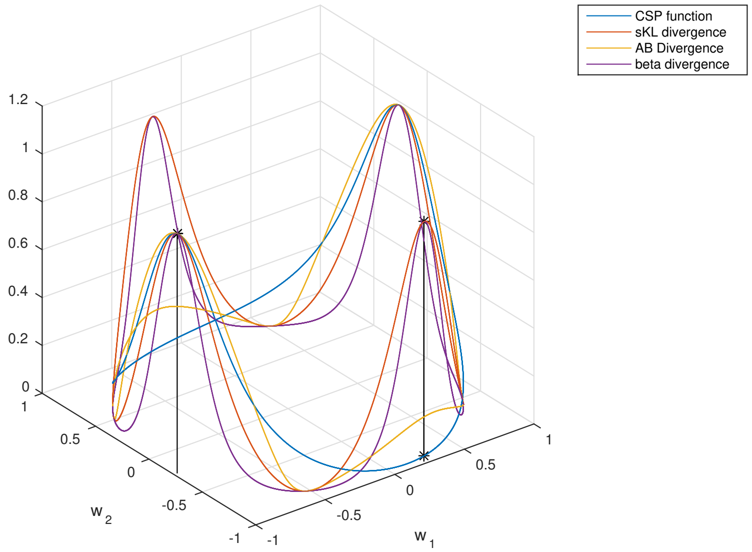


Figure 3. This figure shows the evolution of the common spatial patterns (CSP) criterion function (in blue line), the symmetrized Kullback–Leibler divergence (sKL) (in red line), the symmetrized beta divergence (in purple line) and the AB-LD divergence (in yellow line), all of them as a function of the components of the spatial filter $\boldsymbol{w} = [w_1, w_2]$ in the two-dimensional case, where it is assumed that $\|\boldsymbol{w}\|_2^2 = w_1^2 + w_2^2 = 1$. All the divergences are normalized with respect to their maximum values, and no regularization has been applied. Observe the coincidence of all the critical points. The covariance matrices were generated at random in this experiment.

The rationale is as follows. Observe first that the CSP filters, i.e., the solutions to Equation (5), which is rewritten next for the reader's convenience,

$$\boldsymbol{\Sigma}_1 \boldsymbol{w} = \lambda \boldsymbol{\Sigma}_2 \boldsymbol{w} \rightarrow \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1 \boldsymbol{w} = \lambda \boldsymbol{w},$$

are also the eigenvectors of the matrix $\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1$. Since this matrix is not necessarily symmetric, it follows that these eigenvectors do not form an orthogonal set. A well-posed problem can be obtained by transforming the covariance matrices $\boldsymbol{\Sigma}_i$ into $\hat{\boldsymbol{\Sigma}}_i \equiv \mathbf{P}\boldsymbol{\Sigma}_i\mathbf{P}^\top$, where $\mathbf{P} \in \mathbb{R}^D$ is chosen in such a way to ensure the whitening of the sum of the expected sample observations, i.e.,

$$\mathbf{P}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{P}^\top = \mathbf{I}.$$

Let $\mathbf{W}$ be the matrix that contains the eigenvectors of $\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1$ in its columns, and let $\mathbf{V}$ be the matrix with the eigenvectors of $\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\Sigma}}_1$. It can be shown that matrix $\mathbf{V}$ is orthogonal. Furthermore,

$$\mathbf{W} = \mathbf{P}^\top \mathbf{V} \boldsymbol{\Lambda} \rightarrow \mathbf{W}^\top = \boldsymbol{\Lambda}^\top \mathbf{V}^\top \mathbf{P},$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix (up to elementary column operations) that contains scale factors. In practice, since only the directions of the spatial filters (i.e., not the magnitude) are of interest, we can ignore the above-defined scale matrix $\boldsymbol{\Lambda}$. Then, when only $d \leq D$ filters are retained, it can be assumed that $\mathbf{W}^\top$ can be decomposed into two components $\mathbf{W}^\top = \tilde{\mathbf{R}}\mathbf{P}$ that successively transform the observations. The first matrix $\mathbf{P} \in \mathbb{R}^D$ is chosen in such a way to ensure the whitening of the sum of the expected sample observations, i.e., $\mathbf{P}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{P}^\top = \mathbf{I}$, as was previously explained. The second transformation $\tilde{\mathbf{R}} \in \mathbb{R}^{d \times D}$ is performed by a semi-orthogonal projection matrix, which rotates and reflects the whitened observations and projects this result onto a reduced d-dimensional subspace. This is better seen through the decomposition $\tilde{\mathbf{R}} = \mathbf{I}_d\mathbf{R}$, where $\mathbf{R}$ is a full rank orthogonal matrix ($\mathbf{R}\mathbf{R}^\top = \mathbf{I}$) and $\mathbf{I}_d \in \mathbb{R}^{d\times D}$ is the identity matrix truncated to have only the first d rows.

Let $D(\cdot||\cdot)$ denote any of the previously-studied divergences. The above discussion suggests maximizing the criterion:

$$\begin{aligned} D(\mathbf{W}^\top\mathbf{\Sigma}_1\mathbf{W}\|\mathbf{W}^\top\mathbf{\Sigma}_2\mathbf{W}) &= D(\tilde{\mathbf{R}}\tilde{\mathbf{\Sigma}}_1\tilde{\mathbf{R}}^\top\|\tilde{\mathbf{R}}\tilde{\mathbf{\Sigma}}_2\tilde{\mathbf{R}}^\top) \\ &= D(\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_1\mathbf{R}^\top\mathbf{I}_d\|\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_2\mathbf{R}^\top\mathbf{I}_d) \\ &\equiv J(\mathbf{R}) \end{aligned} \tag{25}$$

under the constraint that $\mathbf{R}$ is an orthogonal matrix, where $\tilde{\mathbf{\Sigma}}_i = \mathbf{P}\mathbf{\Sigma}_i\mathbf{P}^\top$.

Now, we face the problem of optimizing $J(\mathbf{R})$ under the orthogonality constraint $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$. This problem can be addressed in several ways, and here, we review two particularly significant approaches.

3.4.1. Tangent Methods

First of all, it has been shown that the gradient of J at $\mathbf{R}$ on the group of orthogonal matrices is given by [44,45]:

$$\nabla J(\mathbf{R}) = \partial J(\mathbf{R}) - \mathbf{R}(\partial J(\mathbf{R}))^\top\mathbf{R}, \tag{26}$$

where $\partial J(\mathbf{R})$ is the matrix of partial derivatives of J with respect to the elements of $\mathbf{R}$, i.e.,

$$(\partial J(\mathbf{R}))_{ij} = \frac{\partial J(\mathbf{R})}{\partial r_{ij}}, \tag{27}$$

where r_{ij} is the $(i,j)th$ entry of matrix $\mathbf{R}$. Therefore, for steepest ascent search, consider small deviations of $\mathbf{R}$ in the direction $\nabla J(\mathbf{R})$ as follows:

$$\mathbf{R} \to \bar{\mathbf{R}} = \mathbf{R} + \mu\nabla J(\mathbf{R}), \tag{28}$$

with $\mu > 0$. If $\mathbf{R}$ is orthogonal, this update direction maintains the orthogonality condition, in the sense that $\bar{\mathbf{R}}\bar{\mathbf{R}}^\top = \mathbf{I} + o(\mu^2)$. Furthermore, since the first order Taylor expansion of $J(\mathbf{R})$ is:

$$J(\mathbf{R}+\Delta\mathbf{R}) = J(\mathbf{R}) + < \partial J(\mathbf{R})|\Delta\mathbf{R} > + o(\Delta\mathbf{R}), \tag{29}$$

where $< \mathbf{A}|\mathbf{B} >= \text{trace}\left(\mathbf{A}^\top\mathbf{B}\right)$ represents the inner product of two matrices, if $\mathbf{R}$ is modified into $\bar{\mathbf{R}}$, it follows that:

$$J(\bar{\mathbf{R}}) = J(\mathbf{R}) + \mu < \partial J(\mathbf{R})|\nabla J(\mathbf{R}) > + o(\mu). \tag{30}$$

Some algebra shows that:

$$< \partial J(\mathbf{R})|\nabla J(\mathbf{R}) > \;=\; \frac{1}{2} < \nabla J(\mathbf{R})|\nabla J(\mathbf{R}) > \tag{31}$$

which is always positive, and therefore, J always increases. The steepest ascent method thus becomes:

$$\mathbf{R}_{t+1} = \mathbf{R}_t + \mu\nabla J(\mathbf{R}) = \left[\mathbf{I} + \mu H(\mathbf{R}_t)\right]\mathbf{R}_t, \tag{32}$$

where:

$$H(\mathbf{R}_t) = \partial J(\mathbf{R}_t)\mathbf{R}_t^\top - \mathbf{R}_t\partial J(\mathbf{R}_t)^\top. \tag{33}$$

A drawback of this approach is that, as the algorithm iterates, the orthogonality constraint may be no longer satisfied. One possible solution is to re-impose the constraint from time to time by projecting **R** back to the constraint surface, which may be performed using an orthogonalization method such as the Gram–Schmidt technique. This approach has been used, e.g., in [42].

3.4.2. Optimization on the Lie Algebra

Alternatively, **R** can be forced to remain always on the constraint surface using an iteration of the form [44]:

$$\mathbf{R}_{t+1} = \mathbf{Q}_t \mathbf{R}_t, \tag{34}$$

where $\mathbf{Q}_t = \exp(\mathbf{M}_t)$ and $\mathbf{M}_t$ is skew symmetric, i.e., $\mathbf{M}_t = -\mathbf{M}_t^\top$. As the exponential of a skew symmetric matrix is always orthogonal, we ensure that $\mathbf{R}_{t+1}$ is orthogonal, as well, supposing $\mathbf{R}_t$ to be. Technically speaking, the set of the skew symmetric matrices is called a Lie algebra, and the idea is to optimize J moving along it. As the update rule for **R** given in (34) may be also considered as an update for **M** from the zero matrix to its actual value $\mathbf{M}_t$, the algorithm is as follows:

1. Start at the zero matrix **0**.
2. Move from **0** to

$$\mathbf{M}_t = \mu \nabla_{\mathbf{M}} J|_{\mathbf{M}=\mathbf{0}}, \tag{35}$$

where $\nabla_{\mathbf{M}} J$ is the gradient of J with respect to **M** in the Lie algebra:

$$\nabla_{\mathbf{M}} J = \partial J(\mathbf{R})\mathbf{R}^\top - \mathbf{R}\partial J(\mathbf{R})^\top. \tag{36}$$

3. Define $\mathbf{Q}_t = \exp(\mathbf{M}_t)$, and use it to come back into the space of the orthogonal matrices.
4. Update $\mathbf{R}_{t+1} = \mathbf{Q}_t \mathbf{R}_t$.

Note that, for small enough μ, we have that $\exp(\mathbf{M}) = \exp(\mu\nabla_{\mathbf{M}} J) \approx \mathbf{I} + \mu\nabla_{\mathbf{M}} J$, so that (34) coincides with (32). From this viewpoint, it may seem that (34), which is used in [23,41], is superior to (32), in the sense that includes (32) as a particular case. Nevertheless, the main drawback of (34) is that it is necessary to calculate the exponential of a matrix, which is a somewhat "tricky" operation [46]. In both approaches, the optimal value of μ can be chosen by a line search along the direction of the gradient.

More advanced optimization techniques, like the standard quasi-Newton algorithms based on the Broyden–Fletcher–Goldfarb–Shannon (BFGS) method [24] have been recently extended to work on Riemannian manifolds [47]. The algorithm used in Section 6 for the optimization of the AB-LD divergence criterion [42], which we will denote in this paper as the Sub-LD algorithm, is based on the BFGS implementation on the Stiefel manifold of semi-orthogonal matrices [48]. Finally, note that spatial filters can be computed all at a once, yielding the so-called subspace approach, or one after the other by a sequential procedure, which is called the deflation approach. In the latter case, the problem is repeatedly solved for $d = 1$, and a projection mechanism is used to prevent the algorithms from converging to previously found solutions [23].

3.4.3. Post-Processing

Finally, it has to be pointed out that, by maximizing any divergence, we may not obtain the CSP filters, i.e, the vectors $\boldsymbol{w}_i$ computed by the CSP method, but a linear combination of them [23,42]. The filters are actually determined by applying CSP to the projected data in a final step.

4. The Information Theoretic Feature Extraction Framework

Information theory can play a key role in the dimensionality reduction step that extracts the relevant subspaces for classification. Inspired by some other papers in machine learning, the authors of [49] adopted an information theoretic feature extraction (ITFE) framework based on the idea of selecting those features, which are maximally informative about the class labels. Let $\mathcal{X}$ be the D-dimensional random variable describing the observed EEG data. In this way, the desired spatial filters are the ones that maximize the mutual information between the output random variable $\mathcal{Y} = \boldsymbol{w}^\top \mathcal{X}$ and a class random variable C that represents the true intention of the BCI user, i.e.,

$$\boldsymbol{w}_* \quad = \quad \arg\max_{\boldsymbol{w}} I(C\,;\,\boldsymbol{w}^\top \mathcal{X}). \tag{37}$$

As was noted in [49], this criterion can be also linked with the minimization of an upper-bound on the probability of classification error. Consider the entropy $H(C)$ and a function:

$$U(\gamma) = 1 - 2^{-(H(C)-\gamma)}, \tag{38}$$

which was used in [50] to obtain an upper-bound for the probability of error:

$$P_e \quad \leq \quad U(I(C\,;\,\mathcal{Y})). \tag{39}$$

Since $U(\gamma)$ is a strictly monotonous descending function, the minimization of the upper-bound of P_e is simply obtained through the maximization of the mutual information criterion:

$$J_{ITFE}(\boldsymbol{w}) = I(C\,;\,\boldsymbol{w}^\top \mathcal{X}). \tag{40}$$

Although the samples in each class are assumed to be conditionally Gaussian distributed, the evaluation of this criterion also requires one to obtain $h(\boldsymbol{w}^\top \mathcal{X})$, the differential entropy of the output of the spatial filter, which is non-trivial to evaluate, and therefore, it has to be approximated. The procedure starts by choosing the scale of the filter that normalizes the random variable $\boldsymbol{w}^\top \mathcal{X}$ to unit variance. Assuming that $\boldsymbol{w}^\top \mathcal{X}$ is nearly Gaussian distributed, the differential entropy of this variable is approximated with the help of a truncated version of the Edgeworth expansion for a symmetric density [51]:

$$h(\boldsymbol{w}^\top \mathbf{X}) \approx h_g(\boldsymbol{w}^\top \mathcal{X}) - \tfrac{1}{48}\left(k_4(\boldsymbol{w}^\top \mathcal{X})\right)^2, \tag{41}$$

where $h_g(\boldsymbol{w}^\top \mathcal{X})$ denotes the entropy of a Gaussian random variable with power $E[|\boldsymbol{w}^\top \mathcal{X}|^2] = 1$ and kurtosis $k_4(\boldsymbol{w}^\top \mathcal{X})$. By expressing the value of the kurtosis of a mixture of conditional Gaussian densities in terms of the conditional variances of the output for each class, after substituting these values in (41), the authors of [49] arrive to the approximated mutual information criterion that they propose to maximize:

$$\begin{aligned} \tilde{J}_{ITFE}(\boldsymbol{w}) \quad &\equiv \quad -\tfrac{1}{2}\sum_{k=1}^{n_c} P(c_k)\log_2\left(\boldsymbol{w}^\top \boldsymbol{\Sigma}_k \boldsymbol{w}\right) - \tfrac{3}{16}\left(\sum_{k=1}^{n_c} P(c_k)\left((\boldsymbol{w}^\top \boldsymbol{\Sigma}_k \boldsymbol{w})^2 - 1\right)\right)^2 \\ &\approx \quad J_{ITFE}(\boldsymbol{w}), \end{aligned} \tag{42}$$

where n_c is the number of classes and $\boldsymbol{\Sigma}_k$ denotes the conditional covariance matrix of the k-th class.

On the one hand, for only two classes ($n_c = 2$), the exact solution of the ITFE criterion can be shown to coincide with the one of CSP. On the other hand, for multiclass scenarios ($n_c > 2$), it is proposed to use a Joint Approximate Diagonalization (JAD) procedure (which is no longer exact) for obtaining the independent sources of the observations and then retain only those sources that maximize the approximated mutual information with the class labels.

5. Non-Information-Theoretic Variants of CSP

In this section we review, for the purposes of comparison, some variants of CSP that are not based on information-theoretic principles. Although CSP is considered to be the most effective algorithm for the discrimination of motor imagery movements, it is also sensitive to outliers. Several approaches have been proposed to improve the robustness of the algorithm.

Using the sample estimates of the covariance matrices, the CSP criterion (4) can be rewritten as:

$$\hat{J}(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_1 \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_2 \boldsymbol{w}} = \frac{\boldsymbol{w}^\top \mathbf{X}_1 \mathbf{X}_1^\top \boldsymbol{w}}{\boldsymbol{w}^\top \mathbf{X}_2 \mathbf{X}_2^\top \boldsymbol{w}} = \frac{\|\boldsymbol{w}^\top \mathbf{X}_1\|_2^2}{\|\boldsymbol{w}^\top \mathbf{X}_2\|_2^2}, \tag{43}$$

where $\mathbf{X}_i$ denotes the data matrix of class i. Therefore, CSP is not a robust criterion as large outliers are favored over small data values by the square in Equation (43). To fix this problem, some approaches use robust techniques for the estimation of the covariance matrices [37]. Alternatively, as presented in [52], a natural extension of CSP that eliminates the square operation, having it replaced by the absolute value, is given by:

$$\hat{J}_1(\boldsymbol{w}) = \frac{\|\boldsymbol{w}^\top \mathbf{X}_1\|_1}{\|\boldsymbol{w}^\top \mathbf{X}_2\|_1}. \tag{44}$$

This l_1-norm-based CSP criterion is more robust against outliers than the original l_2-norm-based formula (43). However, l_1-norm CSP does not explicitly consider the effects of other types of noise, such as those caused by ocular movements, eye blinks or muscular activity, supposing that they are not completely removed in the preprocessing step [53,54]. To take them into account, [55] added a penalty term in the denominator of the CSP-l_2 objective function, obtaining:

$$\hat{J}_{1r}(\boldsymbol{w}) = \frac{\|\boldsymbol{w}^\top \mathbf{X}_1\|_2^2}{\|\boldsymbol{w}^\top \mathbf{X}_2\|_2^2 + \rho R(\boldsymbol{w})}, \tag{45}$$

where $R(\boldsymbol{w})$ is some measure of the intraclass scattering of the filtered data in each of the classes, so the maximization of $\hat{J}_{1r}(\boldsymbol{w})$ encourages the minimization of $R(\boldsymbol{w})$, and ρ is a positive tuning parameter. Finally, a generalization of the l_1-norm-based approach has been proposed in [56,57], which explores the use of l_p norms through the following criterion:

$$\hat{J}_{1p}(\boldsymbol{w}) = \frac{\|\boldsymbol{w}^\top \mathbf{X}_1\|_p^{1/p}}{\|\boldsymbol{w}^\top \mathbf{X}_2\|_p^{1/p}}. \tag{46}$$

Other approaches for regularizing the original l_2-norm based CSP algorithm include performing a robust estimation of the covariance matrices $\boldsymbol{\Sigma}_i$ or adding a penalty term $\boldsymbol{\Delta}$ in the objective function. With regard to the first approach, [58] proposes the use of information from various subjects as a regularization term, so the sample covariance matrices $\boldsymbol{\Sigma}$ are substituted in the formulas for:

$$\tilde{\boldsymbol{\Sigma}} = (1-\psi)\boldsymbol{\Sigma} + \psi \frac{1}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} \boldsymbol{\Sigma}^k,$$

where $\mathcal{S}$ is a set of subjects whose data have been previously recorded, $\boldsymbol{\Sigma}^k$ is the sample covariance matrix of the k-th subject and $\psi \in (0,1)$ is a regularization parameter. Related approaches can be found in [21,37,38,59–63]. Finally, in [8] the covariance matrices are estimated using data originating from specific regions of interest within the brain.

The second regularization approach consists of including a penalty term in the CSP objective function [64]. The regularized CSP objective functions can be represented as:

$$\tilde{J}_1(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_1 \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_2 \boldsymbol{w} + \alpha \Delta(\boldsymbol{w})} \tag{47}$$

$$\tilde{J}_2(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_2 \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_1 \boldsymbol{w} + \alpha \Delta(\boldsymbol{w})} \tag{48}$$

where α is the regularization parameter. The regularized Tikhonov-CSP approach (RTCSP) penalizes the solutions with large weights by using a penalty term $\boldsymbol{\Delta}(\boldsymbol{w})$ of the form:

$$\Delta(\boldsymbol{w}) = \|\boldsymbol{w}\|.$$

The filters $\boldsymbol{w}$ can computed by solving an eigenvalue problem similar to that of the standard CSP algorithm. Specifically, the stationary points of $\tilde{J}_1(\boldsymbol{w})$ verify [64]:

$$(\boldsymbol{\Sigma}_2 + \alpha \mathbf{I})^{-1} \boldsymbol{\Sigma}_1 \boldsymbol{w} = \lambda \boldsymbol{w}.$$

Similarly, the stationary points of $\tilde{J}_2(\boldsymbol{w})$ are the eigenvectors of matrix $(\boldsymbol{\Sigma}_1 + \alpha \mathbf{I})^{-1} \boldsymbol{\Sigma}_2$. Observe that it is necessary to optimize both objective functions, as the stationary points of any of them alone maximize the variance of one class, but do not minimize the variance of the other class.

Finally, all the previous approaches admit the following generalization: in traditional CSP, the EEG data is usually band-pass pre-filtered using one single filter between 8 and 30 Hz, which is a range that covers the so-called "alpha", "beta" and "mu" EEG bands. An straightforward extension, known as the filter bank CSP (FBCSP) technique, was proposed in [30], where the input MI-EEG signals are bandpass filtered between different bands of frequency ((4–8 Hz), (8–12 Hz), . . ., (36–40 Hz)) and the CSP algorithm, or any of its variants, is applied to each band for the computation of the spatial filters. The results of all analyses are then combined to form the final response (see Figure 4). Similar approaches have been proposed in [11,65,66]. An extension to the multiclass problem can be found in [67]. Since the optimal frequency bands can vary from subject to subject, several alternative approaches have been proposed that combine the time-frequency characteristics of the EEG data [68,69] for improving the classification accuracy and reducing the number of electrodes [70].

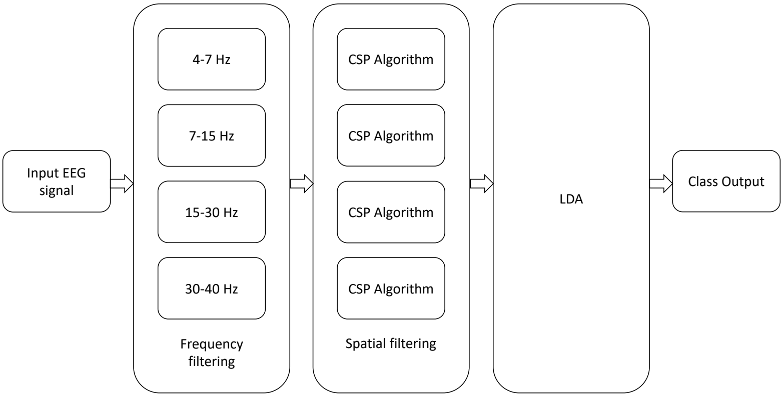


Figure 4. Architecture of filter bank CSP. LDA is shorthand for Linear Discriminant Analysis.

6. Experimental Results

Initially, we will test the algorithms using real datasets obtained from the BCI competition III (dataset 3a) and BCI competition IV (dataset 2a), which are publicly available at [71]. On the one hand, the dataset 3a from BCI competition III consists of EEG data acquired from three subjects (k3b, k6b and l1b) at a sampling frequency of 250 Hz using a 60-channel EEG system. In each trial, an arrow to the left, right, up or down was shown on a display for a few seconds, and in response to the stimulus, the subject was asked to respectively perform left hand, right hand, tongue and foot MI movements. The dataset consists of 90 trials per class for Subject k3b and 60 trials per class for Subjects k6b and l1b. On the other hand, the dataset 2a from BCI competition IV was acquired by using 22 channels from nine subjects (A01–A09) while also performing left hand, right hand, tongue and foot MI movements following a similar procedure. The signals were also sampled at 250 Hz and were recorded in two sessions on different days, each of them with 72 trials per each class.

For a total of four possible motor-imagery (MI) movements, $\binom{4}{2} = 6$ different combinations of pairs of MI movements (i.e., left hand-right hand, left hand-foot, left hand-tongue, right hand-foot, right hand-tongue, foot-tongue) can be formed. The experiments below consider all possible combinations: since 12 users are available and for nine of them we have recordings performed on two different days, this makes a total of $3 \times 6 + 9 \times 6 \times 2 = 126$ different experiments. We repeated eight times each of these 126 possible experiments, and results were averaged. For each repetition, 60 trials were selected at random from each MI movement, which were split into 40 trials for training and 20 trials for testing. Additionally, in the case of the BCI competition IV, we averaged over the two sessions conducted for each user to avoid biasing the statistical tests. As a result, $3 \times 6 + 9 \times 6 \times 2/2 = 72$ averaged performance measures are finally available for each algorithm. The data have been initially bandpass filtered between the cut-off frequencies of 8–30 Hz, except before using the FBCSP method, which as we explained in Section 5, considers four bands for covering the frequency range between 4 and 40 Hz. The information of the classes in each trial is summarized by their respective covariance matrices. These matrices are estimated, normalized by their trace and used as input to the algorithms that carry out the calculation of the spatial filters prior to the MI classification, which is performed by using linear discriminant analysis (LDA).

The only parameter of the CSP algorithm is the number of spatial filters that one would like to consider. Although, this number d is usually fixed a priori for each dataset, it is advantageous to estimate automatically the best number of spatial filters for each user by using the combination of cross-validation and hypothesis testing proposed in [73]. Figure 5a illustrates this fact. The figure represents the scatter plot of the accuracies, expressed as a percentage, that have been respectively obtained by the CSP algorithm for a fixed value of $d = 8$ (x-axis) and for the estimation of the best value of d (y-axis). These estimated accuracies have been obtained by averaging eight test samples, as explained above. The accuracies obtained for different individuals or for different pairs of conditions can be reasonably considered approximately independent and nearly Gaussian. Under this hypothesis, a one-sided paired t-test of statistical significance can be used to compare the results obtained by both alternatives. Let $\delta f(m) = f_y(m) - f_x(m)$ be the paired differences of accuracy ((y-axis value) vs. (x-axis value)) for $m = 1, \ldots, M$, where $M = 72$ is the number of samples. Then, the averaged difference is:

$$\overline{\Delta f} = \frac{1}{M} \sum_{m=1}^{M} \delta f(m) \tag{49}$$

and the unbiased estimate of its variance is:

$$s^2 = \frac{s_{\Delta f}^2}{M}, \tag{50}$$

where $s^2_{\Delta f} = \frac{1}{M-1}\sum_{m=1}^{M}(\delta f(m) - \overline{\Delta f})^2$. Under the null hypothesis ($H0$) that the expected performance values coincide, i.e., $E[f_y(m)] = E[f_x(m)]$, the t-statistic:

$$T - STAT = \frac{\overline{\Delta f}}{s_{\Delta f}/\sqrt{M}}. \tag{51}$$

follows a Student's t distribution with $M - 1$ degrees of freedom. Thus, the probability that the null hypothesis can generate a t-statistic larger than $T - STAT$ gives the p-value of the right-sided test:

$$P - VAL = Prob(t > T - STAT|H0). \tag{52}$$

The more positive is $T - STAT$, the smaller is the $P - VAL$, and the probability of observing a t-statistics larger than $T - STAT$ decreases under the null hypothesis. When the p-value falls below the 0.05 threshold of significance, the hypothesis of not having a performance improvement when using the alternative procedure can be rejected, because this would correspond to a quite improbable situation. On the contrary, if the p-value of the right-sided test is above 0.05, the null hypothesis cannot be rejected.

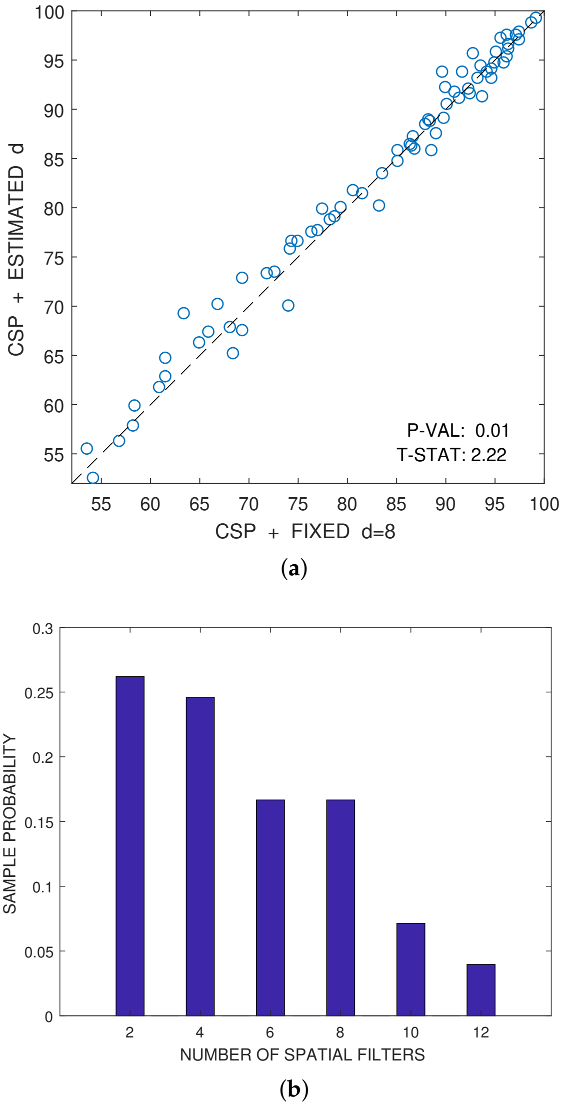
In this particular case, the p-value of the test in Figure 5a is below 0.05; therefore, one can reject the hypothesis that the automatic estimation of d does not improve the results over the method that a priori selects $d = 8$ filters.

We briefly name and describe below some of the implementations that optimize the already mentioned criteria for dimensionality reduction in MI-BCIs. Because of the substantially higher computational complexity of most of the alternatives to CSP (see Table 1), it is not practical to develop a specific automatic estimation procedure of the number of spatial filters for each of them. For this reason, we will consider in their implementations the same number of spatial filters that was automatically estimated for CSP.

- CSP (see Section 2) and ITFE (see Section 4): apart from the number of spatial filters, these two methods do not have hyper-parameters to tune. Their respective algorithms have been implemented according to the specifications given in [4] and [49].
- RTCSP (see Section 5): RTCSP has a regularization parameter, which has been selected by five-fold cross-validation in $\{0, 0.1, 0.2, \ldots, 1\}$. The MATLAB implementation of this algorithm has been obtained from [72].
- FBCSP (see Section 5): In this case, we have used a variation of the algorithm in [30]. The selected frequency bands correspond to the brainwaves *theta* (4–7 Hz), *alpha* (8–15 Hz), *beta* (16–31 Hz) and *low gamma* (32–40 Hz), where five-fold cross-validation has been used to select the best combination of these frequency bands. We extract d features from each band, where d is selected using the method in [73].
- DivCSP (see Sections 3.2 and 3.4). The values of β and ϕ (the regularization parameter) have been selected by five-fold cross-validation, $\beta \in [0, 1]$, $\phi \in [0, 0.5]$. This divergence includes the KL divergence as a particular case when $\beta = 0$. MATLAB code of the algorithm has been downloaded from [74] and used without any modification. Optimization has been performed using the so-called subspace method (see Section 3.4).
- Sub-LD (sub-space Log-Det): this algorithm, which also belongs to the class of the subspace methods, is based on the criterion in [42] to maximize the Alpha-Beta Log-Det divergence (see Sections 3.3 and 3.4). In this paper, the implementation of the algorithm is based on the BFGS method on the Stiefel manifold of semi-orthogonal matrices and takes as the initialization point the solution obtained by the CSP algorithm. The regularization parameter η has been chosen by five-fold cross-validation in the range of values $(-0.2, 0.2)$, which are not far from zero. The negative values of η favor the expansion of the clusters, while the positive values favor their contraction. For η close to zero, the solution of this criterion should not be far from that of CSP,

which improves the convergence time of the algorithm and reduces the impact of the values of α, β in the results, so both parameters have been fixed to 0.5.

Table 1 shows the typical execution time of a single run of each algorithm, programmed in MATLAB language, in a PC with Intel I7-6700 CPU @ 3.4-GHz processor and 16 GB of RAM. The algorithms that use cross-validation for selecting the hyper-parameters need more iterations, hence the run time has to be multiplied by the number of the hyper-parameters combinations that are evaluated.



(**a**)

(**b**)

Figure 5. Illustration of the advantages in performance of using an automatic cross-validation method to estimate the best even number of features d with respect to using an a priori fixed value of d. The automatic method relies on the technique proposed in [73], which was implemented here using one-sided t-tests of significance instead of the original two-sided tests. (**a**) Scatter plot comparison of the accuracies (in percentage) obtained by the CSP algorithm for fixed $d = 8$ (x-axis) and for the automatic estimation of d (y-axis); (**b**) histogram of the estimated best even number of features d.

Table 1. Computational burden of the considered algorithms, which are sorted in increasing value of their respective execution times without using cross-validation. FBCSP, filter bank CSP; ITFE, information theoretic feature extraction.

Algorithm	Time (s)
CSP	0.0017
FBCSP	0.0050
ITFE	0.3070
Sub-LD	1.0538
DivCSP	4.6696

Figure 6 represent the boxplot of the accuracy of the algorithms, considering together all the combinations of the motor imagery movements from all subjects in datasets III 3a and IV 2a. The p-values and t-statistics shown below the box-plots of Figure 6 are above the 5% threshold of significance, revealing that, in this experiment, one cannot reject the null hypotheses. It follows that the expected accuracies of the alternative algorithms are not significantly higher than the expected accuracies obtained with CSP. Supporting this conclusion, Figure 7 represents the specific boxplots that corresponds to MI movements involving the right hand. Additionally, we have tested, in the case "left hand versus right hand", whether the improvement obtained by using the alternative algorithms is significant or not. The accuracy in the classification and the corresponding p-values of the tests are shown in Figure 8. The results reveal that, in general and except in a few isolated cases, the null hypothesis that the other methods do not significantly improve performance over CSP cannot be discarded.

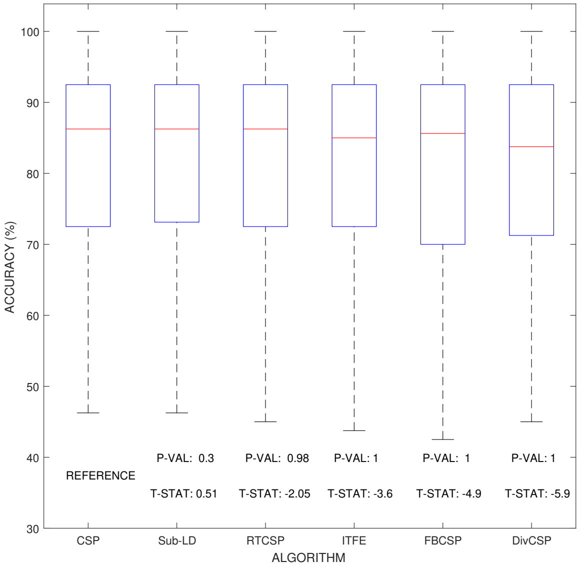


Figure 6. Comparison of the expected accuracy percentages obtained by each of the considered algorithms. The figure shows box-plot illustrations where the median is shown in red line, while the 25% and 75% percentiles are respectively at the bottom and top of each box. Larger positive values $T - STAT \gg 0$ and smaller $P - VAL \ll 1/2$ would correspond with greater expected improvements over CSP. However, none of the p-values, which are shown below their respective box-plots, is able to attain the 5% threshold level of significance ($P - VAL < 0.05$), so the possible improvements cannot be claimed to be statistically significant with respect to those obtained by CSP.

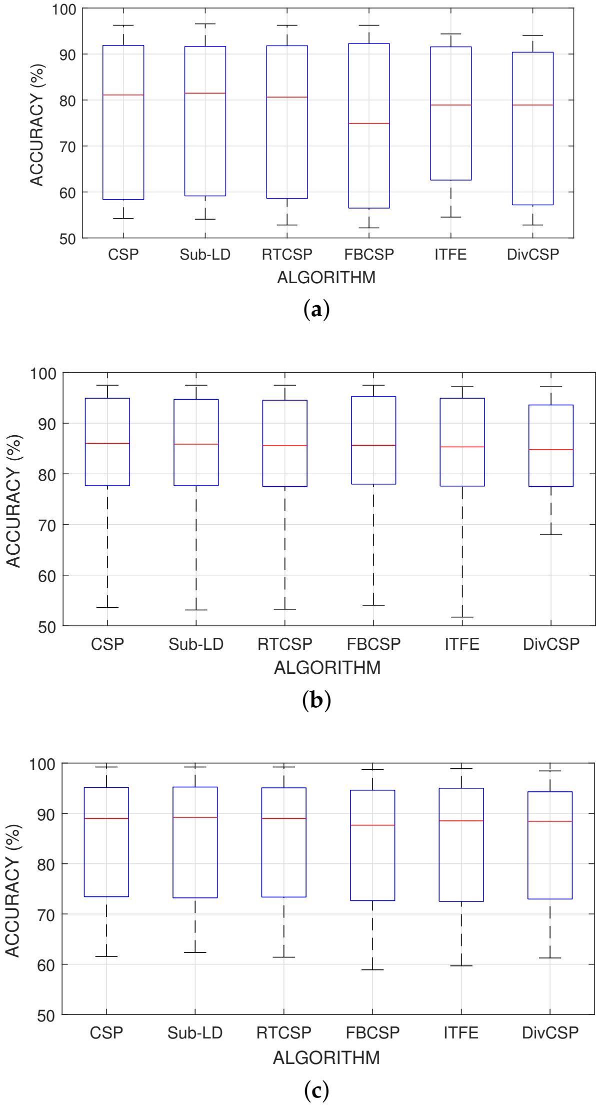
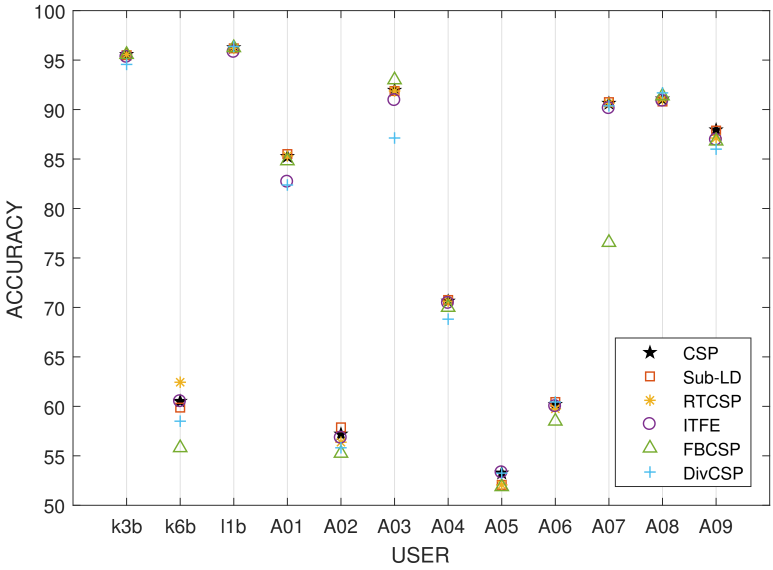
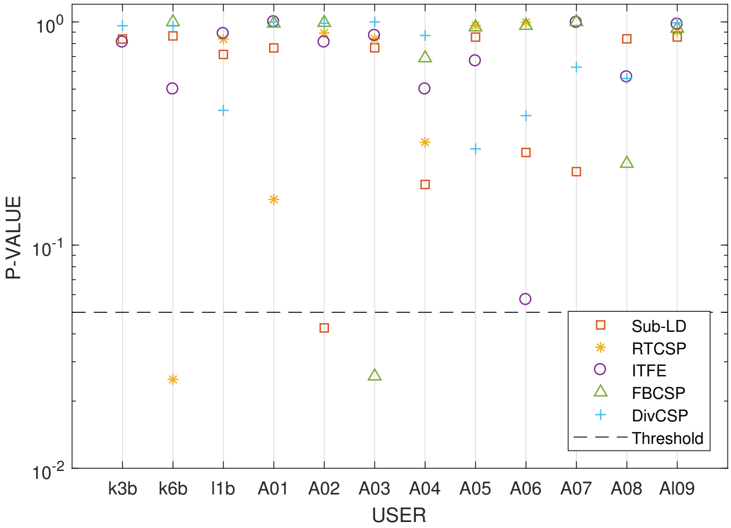


Figure 7. Performance of the algorithms for different motor imagery combinations involving the right hand. (**a**) Right-hand versus left-hand motor imagery classification; (**b**) right-hand versus feet motor imagery classification; (**c**) right-hand versus tongue motor imagery classification.



(**a**)



(**b**)

Figure 8. Accuracy percentages and p-values for the testing of an improvement in performance over CSP when the right hand versus left hand movement imagination are discriminated. The results reveal that, in general and except in a few isolated cases, the null hypothesis that the other methods do not significantly improve the performance over CSP cannot be discarded. (**a**) Average accuracy obtained by the algorithms for each subject; (**b**) p-values of the t-tests that compare whether the performance of the alternative algorithms is significantly better than the one obtained by CSP. The horizontal dashed line represents the threshold level of significance of 5%.

The results of Figure 6 were obtained by choosing through cross-validation the best possible values for the different parameters of the algorithms. Figures 9 and 10 show how many times each value of the parameters has been selected after cross-validation. They also show the number of times that CSP outperformed the corresponding algorithm, the number of times that the algorithm outperformed CSP or the cases in which both of them were equivalent. Without limiting the foregoing, it must be also remarked that the alternative algorithms perform better than CSP for some subjects and MI movements.

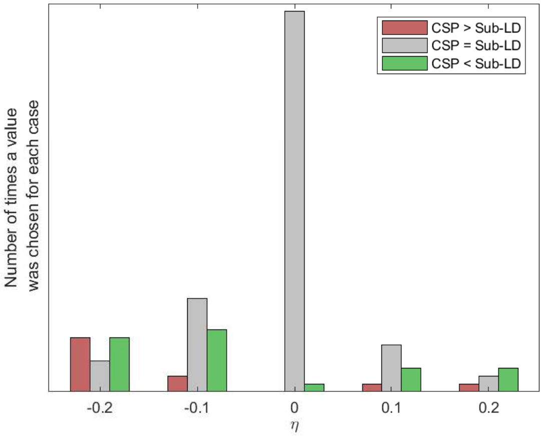


Figure 9. Histogram of the values of the regularization parameter in the Sub-LD algorithm that have been chosen by cross-validation.

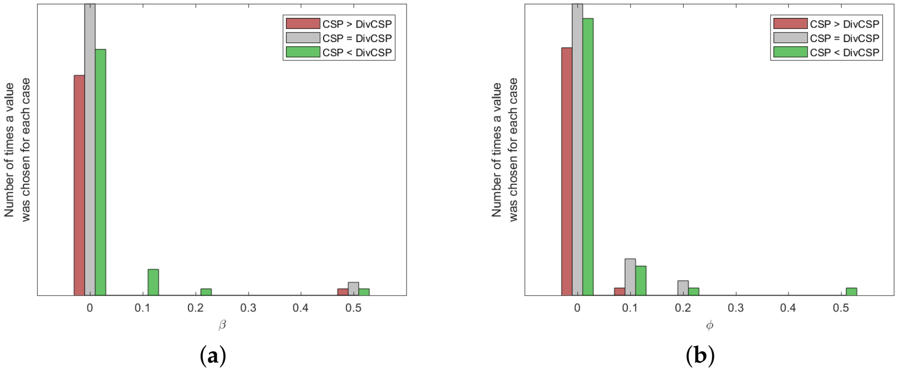


Figure 10. Histogram of the hyper-parameters of the DivCSP algorithm selected by cross-validation. (**a**) Case with $\beta \in [0, 0.5]$ and $\phi = 0$; (**b**) case with $\beta = 0.5$ and $\phi \in [0, 0.5]$.

6.1. Results on Artificially Perturbed Data

In order to study the performance of the algorithms under artificial perturbations of the datasets we have conducted two experiments. The first one consists of introducing random label changes in the real datasets, while the second one defines sample EEG covariance matrices for each condition and artificially introduces outlier covariance matrices in the training procedure to quantify the resulting deterioration in performance.

Exchanging labels of the training set at random is one of the most harmful perturbations that one can consider in a real experiment. It models the failure of the subjects to imagine the correct target MI movements due to fatigue or lack of concentration. For this experiment, we selected a subject who has a relatively good performance in absence of perturbations. Figure 11 presents the progressive degradation of the accuracy of the algorithms as the percentage of mismatched labels increases.

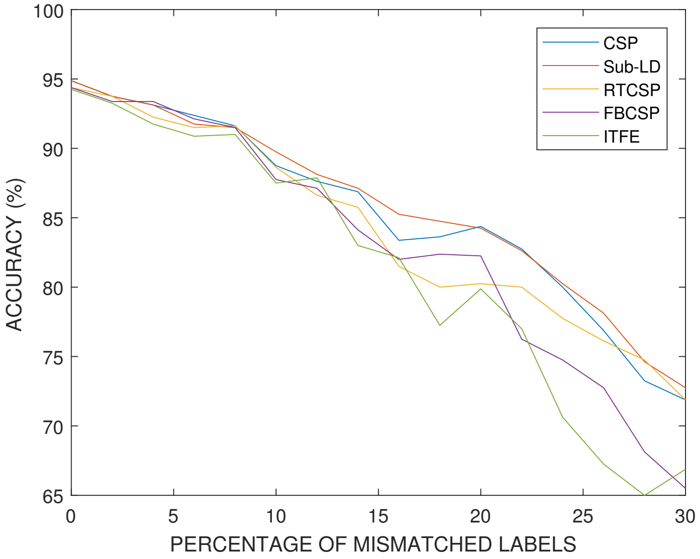


Figure 11. Comparison of the accuracy percentages obtained by each of the considered algorithms with respect to the percentage of mismatched labels in the training set. This experiment illustrates deterioration of the performance of the algorithms with respect to the increase of the percentage of randomly switched labels of the motor imagery movements.

In the second experiment, we have created artificial EEG data and consider the effect of adding random outliers. The artificial data were generated starting from two auxiliary covariance matrices $\mathbf{C}_k$, $k = 1, 2$ for the construction of the conditional covariance matrices of each class. These covariances were generated randomly by drawing two random Gaussian matrices $\mathbf{A}^{(k)}$ with i.i.d. elements $a_{ij}^{(i)} \sim \mathcal{N}(0, 1)$ and forming the covariance matrices with $\mathbf{C}_k = \mathbf{A}^{(k)}(\mathbf{A}^{(k)})^\top$, $k = 1, 2$. In order to control the difficulty of the classification problem, we introduce a dissimilitude parameter $\delta \in [0, 1]$ that interpolates between the two auxiliary covariance matrices as follows:

$$\mathbf{\Sigma}_1 = \mathbf{C}_1^{1/2}(\mathbf{C}_1^{-1/2}\mathbf{C}_2\mathbf{C}_1^{-1/2})^{\frac{(1-\delta)}{2}}\mathbf{C}_1^{1/2} \tag{53}$$

$$\mathbf{\Sigma}_2 = \mathbf{C}_2^{1/2}(\mathbf{C}_2^{-1/2}\mathbf{C}_1\mathbf{C}_2^{-1/2})^{\frac{(1-\delta)}{2}}\mathbf{C}_2^{1/2} \tag{54}$$

In this way, when $\delta = 0$, the two interpolated covariance matrices coincide $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, and it is impossible to distinguish between them. On the contrary, when $\delta = 1$, we obtain the original randomly generated matrices $\mathbf{\Sigma}_1 = \mathbf{C}_1$ and $\mathbf{\Sigma}_2 = \mathbf{C}_2$. The matrices $\mathbf{\Sigma}_k$ are used as the expected covariance matrix of the observations for class k, while the sample covariance matrices for each trial are generated from a Wishart distribution with scale matrix $\frac{1}{T}\mathbf{\Sigma}_k$ and T degrees of freedom (where T denotes the trial length). The outlier matrices have been generated following a similar scheme, though interpolation is not used and the resulting covariances are scaled by a factor of five.

In our simulations with artificial data, we have set the dissimilitude parameter to $\delta = 0.1$. The results obtained for artificial data and with different percentages of outlier covariance matrices in the training set are shown in Figure 12. One can observe how the performance progressively deteriorates with the number of outliers, similarly for all the methods, although at a smaller rate than in the case having the same percentage of mismatched labels. The parameters of the algorithms have been selected by cross-validation.

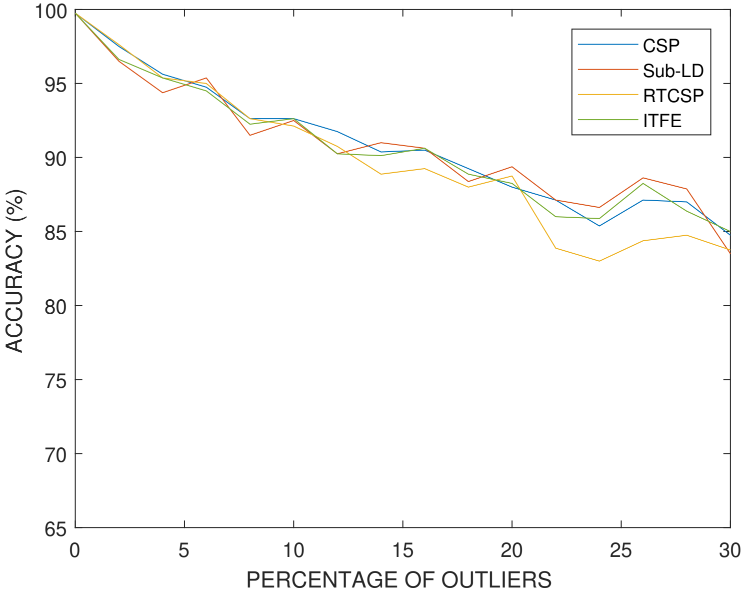


Figure 12. Accuracy percentages versus the percentage of training trials with outliers in a synthetic classification experiment.

7. Conclusions

In this paper, we have reviewed several information theoretic approaches for motor-imagery BCI systems. In particular, we have focused on those based on the Kullback–Leibler divergence, Beta divergence, Alpha-Beta Log-Det divergence and information theoretic feature extraction, exploring the existing links with common spatial patterns, which is a widely-used technique for spatial filtering in BCI applications. The performance of all these methods has been evaluated through experimental simulations using real and synthetic data. In general, the results obtained for real data from BCI competitions reveal a similar performance for all the considered criteria in terms of their percentages of accuracy. However, CSP clearly outperforms the other methods when comparing the required computational burdens. In the case of synthetic data with outliers, a comparison of the divergence-based methods with small regularization parameters reveals that they can slightly increase the frequency of obtaining a better performance, although the average accuracy results are still similar to those obtained with CSP. Therefore, although these divergence-based methods are not yet a practical alternative to CSP, this line of research is in its infancy, and divergence-based methods can have an underlying potential for improvements in performance that remains to be explored.

Acknowledgments: Part of this work was supported by the Spanish Government under MICINN project TEC2014-53103-P. Andrzej Cichocki was partially supported by the MES Russian Federation grant 14.756.31.0001. We also thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

Author Contributions: Rubén Martín-Clemente and Sergio Cruces collaborated in writing the paper and coordinating the study. Andrzej Cichocki critically revised the manuscript by providing inspiring comments. Javier Olias and Deepa Beeta Thiyam conducted the experimental work. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Saeid, S.; Chambers, J.A. *EEG Signal Processing*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
2. Sörnmo, L.; Laguna, P. *Bioelectrical Signal Processing in Cardiac and Neurological Applications*; Academic Press: Cambridge, MA, USA, 2005; Volume 8.
3. Devlaminck, D.; Wyns, B.; Grosse-Wentrup, M.; Otte, G.; Santens, P. Multisubject learning for common spatial patterns in motor-imagery BCI. *Comput. Intell. Neurosci.* **2011**, 217987, doi:10.1155/2011/217987.
4. Ramoser, H.; Müller-Gerking, J.; Pfurtscheller, G. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehabil. Eng.* **2000**, 8, pp. 441–446.
5. Lotte, F. A tutorial on EEG signal-processing techniques for mental-state recognition in brain-computer interfaces. In *Guide to Brain-Computer Music Interfacing*; Springer: London, UK, 2014; pp. 133–161,
6. Samek, W.; Meinecke, F.C.; Müller, K.-R. Transferring subspaces between subjects in brain–computer interfacing. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2289–2298.
7. Wu, W.; Gao, X.; Hong, B.; Gao, S. Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL). *IEEE Trans. Biomed. Eng.* **2008**, *55*, 1733–1743.
8. Grosse-Wentrup, M.; Liefhold, C.; Gramann, K.; Buss, M. Beamforming in noninvasive brain-computer interfaces. *IEEE Trans. Biomed. Eng.* **2009**, *56*, 1209–1219.
9. Gouy-Pailler, C.; Congedo, M.; Brunner, C.; Jutten, C.; Pfurtscheller, G. Nonstationary brain source separation for multiclass motor imagery. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 469–478.
10. Sun, G.; Hu, J.; Wu, G. A novel frequency band selection method for common spatial pattern in motor imagery based brain computer interface. In Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 18–23 July 2010; pp. 1–6.
11. Thomas, K.P.; Guan, C.; Lau, C.T.; Vinod, A.P.; Ang, K.K. A new discriminative common spatial pattern method for motor imagery brain-computer interfaces. *IEEE Trans. Biomed. Eng.* **2009**, *56*, 2730–2733.
12. Graimann, B.; Allison, B.; Pfurtscheller, G. Brain-computer interfaces: A gentle introduction. In *Brain-Computer Interfaces*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–27.
13. Pfurtscheller, G.; Lopes Da Silva, F.H. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clin. Neurophysiol.* **1999**, *110*, 1842–1857.
14. Lotte, F.; Bougrain, L.; Cichocki, A.; Clerc, M.; Congedo, M.; Rakotomamonjy, A.; Yger, F. A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update. *J. Neural Eng.* **2018**, (in print).
15. Schlögl, A.; Lee, F.; Bischof, H.; Pfurtscheller, G. Characterization of four-class motor imagery EEG data for the BCI-competition 2005. *J. Neural Eng.* **2005**, *2*, L14–L22.
16. Ehrsson, H.; Geyer, S.; Naito, E. Imagery of Voluntary Movement of Fingers, Toes, and Tongue Activates Corresponding Body-Part-Specific Motor Representations. *J. Neurophysiol.* **2003**, *90*, 3304–3316.
17. Dagaev, N.; Volkova, K.; Ossadtchi, A. Latent variable method for automatic adaptation to background states in motor imagery BCI. *J. Neural Eng.* **2017**, doi: 10.1088/1741-2552/aa8065.
18. Perdikis, S.; Leeb, R.; Millán, J.D. Context-aware adaptive spelling in motor imagery BCI. *J. Neural Eng.* **2016**, *13*, 036018.
19. Brandl, S.; Müller, K.-R.; Samek, W. Robust common spatial patterns based on Bhattacharyya distance and Gamma divergence. In Proceedings of the 2015 3rd International Winter Conference on Brain-Computer Interface (BCI), Sabuk, Korea, 12–14 January 2015; pp. 1–4.
20. Lotte, F.; Guan, C. Spatially regularized common spatial patterns for EEG classification. In Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 23–26 August 2010; pp. 3712–3715.
21. Lu, H.; Plataniotis, K.N.; Venetsanopoulos, A.N. Regularized common spatial patterns with generic learning for EEG signal classification. In Proceedings of the 2009 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 3–6 September 2009; pp. 6599–6602.
22. Samek, W.; Vidaurre, C.; Müller, K.-R.; Kawanabe, M. Stationary common spatial patterns for brain-computer interfacing. *J. Neural Eng.* **2012**, *9*, 026013.
23. Samek, W.; Kawanabe, M.; Muller, K.-R. Divergence-based framework for common spatial patterns algorithms. *IEEE Rev. Biomed. Eng.* **2014**, *7*, 50–72.

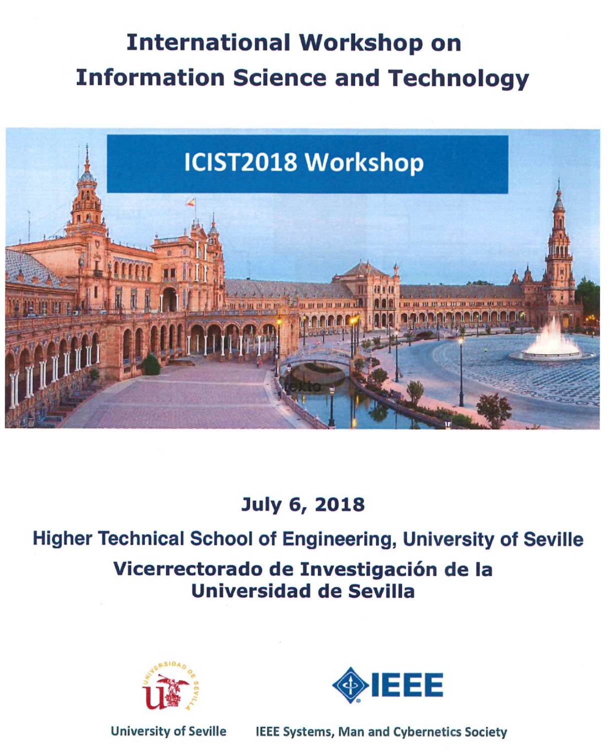
24. Wang, H. Harmonic mean of Kullback–Leibler divergences for optimizing multiclass EEG spatio-temporal filters. *Neural Process. Lett.* **2012**, *36*, 161–171.
25. Samek, W.; Müller, K.-R. Tackling noise, artifacts and nonstationarity in BCI with robust divergences. In Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 2741–2745.
26. Lawhern, V.; David Hairston, W.; McDowell, K.; Westerfield, M.; Robbins, K. Detection and classification of subject-generated artifacts in EEG signals using autoregressive models. *J. Neurosci. Methods* **2012**, *208*, 181–189.
27. Delorme, A.; Sejnowski, T.; Makeig, S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage* **2007**, *34*, 1443–1449,
28. Uusitalo, M.; Ilmoniemi, R.J. Signal-space projection method for separating MEG or EEG into components. *Med. Biol. Eng. Comput.* **1997**, *35*, 135–140,
29. Urigüen, J.A.; García-Zapirain, B. EEG artifact removal-state-of-the-art and guidelines. *J. Neural Eng.* **2015**, *12*, 031001.
30. Ang, K.K.; Chin, Z.Y.; Zhang, H.; Guan, C. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In Proceedings of the IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 2390–2397.
31. Dornhege, G.; Blankertz, B.; Krauledat, M.; Losch, F.; Curio, G.; Muller, K.-R. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 2274–2281.
32. Kang, H.; Nam, Y.; Choi, S. Composite common spatial pattern for subject-to-subject transfer. *IEEE Signal Process. Lett.* **2009**, *16*, 683–686.
33. Ang, K.; Chin, Z.Y.; Zang, H.; Guan, C. Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs. *Pattern Recognit.* **2012**, *45*, 2137–2144.
34. Koles, Z.; Lind, J.; Flor-Henry, P. Spatial patterns in the background EEG underlying mental disease in man. *Electroencephalogr. Clin. Neurophysiol.* **1994**, *91*, 319–328.
35. Wu, W.; Chen, Z.; Gao, S.; Brown, E. A probabilistic framework for robust common spatial patterns. In Proceedings of the Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), Minneapolis, MN, USA, 3–6 September 2009; pp. 4658–4661.
36. Kang, H.; Choi, S. Probabilistic models for common spatial patterns: Parameter extended EM and variational bayes. In Proceedings of the XXVI AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012; pp. 970–976.
37. Kawanabe, M.; Vidaurre, C. Improving BCI performance by modified common spatial patterns with robustly averaged covariance matrices. In *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*; Springer: Munich, Germany, 7–12 September, 2009, pp, 279–282.
38. Yong, X.; Ward, R.K.; Birch, G.E. Robust common spatial patterns for EEG signal preprocessing. In Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada, 20–25 August 2008; pp. 2087–2090.
39. Samek, W.; Kawanabe, M.; Vidaurre, C. Group-wise stationary subspace analysis—A novel method for studying non-stationarities. *Proc. Int. Brain Comput. Interfaces Conf.* 2011, pp. 16–20.
40. Arvaneh, M.; Guan, C.; Ang, K.K.; Quek, C. Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *24*, 610–619.
41. Samek, W.; Blythe, D.; Müller, K.-R.; Kawanabe, M. Robust spatial filtering with beta divergence. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2013; pp. 1007–1015.
42. Beeta Thyam, D.; Cruces, S.; Olías, J.; Chichocki, A. Optimization of Alpha-Beta Log-Det divergences and their application in the spatial filtering of two class motor imagery movements. *Entropy* **2017**, *19*, 89.
43. Cichocki, A.; Cruces, S.; Amari, S.-I. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
44. Plumbley, M.D. Geometrical methods for non-negative ICA: Manifolds, Lie groups and toral subalgebras. *Neurocomputing* **2005**, *67*, 161–197.
45. Edelman, A.; Arias, T.A.; Smith, S.T. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **1998**, *20*, 303–353,

46. Moler, C.; Van Loan, C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.* **2003**, *45*, 3–49.
47. Huang, W.; Absil, P.-A.; Gallivan, K.A. A Riemannian BFGS Method for Nonconvex Optimization Problems. In *Numerical Mathematics and Advanced Applications ENUMATH 2015*; Springer: Cham, Switzerland, 2016; pp. 627–634.
48. Boumal, N.; Mishra, B.; Absil, P.-A.; Sepulchre, R. Manopt, a Matlab Toolbox for Optimization on Manifolds. *J. Mach. Learn. Res.* **2014**, *15*, 1455–1459.
49. Grosse-Wentrup, M.; Buss, M. Multiclass common spatial patterns and information theoretic feature extraction. *IEEE Trans. Biomed. Eng.* **2008**, *55*, 1991–2000.
50. Feder, M.; Merhav, N. Relations between entropy and error probability. *IEEE Trans. Inf. Theory* **1994**, *40*, 259–266.
51. Jones, M.C.; Sibson, R. What is projection pursuit? (with discussion). *J. R. Stat. Soc. Ser. A* **1987**, *150*, 1–36.
52. Wang, H.; Tang, Q.; Zheng, W. L1-norm-based common spatial patterns. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 653–662.
53. Daly, I.; Nicolaou, N.; Nasuto, S.; Warwick, K. Automated artifact removal from the electroencephalogram: A comparative study. *Clin. EEG Neurosci.* **2013**, *44*, 291–306.
54. Fatourechi, M.; Bashashati, A.; Ward, R.; Birch, G. EMG and EOG artifacts in brain-computer interface systems: A survey. *Clin. Neurophysiol.* **2007**, *118*, 480–494.
55. Wang, H.; Li, X. Regularized filters for L1-norm-based common spatial patterns. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2016**, *24*, 201–211.
56. Arvaneh, M.; Guan, C.; Ang, K.K.; Quek, C. Optimizing the channel selection and classification accuracy in EEG-based BCI. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 1865–1873.
57. Park, J.; Chung, W. Common spatial patterns based on generalized norms. In Proceedings of the 2013 International Winter Workshop on Brain-Computer Interface (BCI), Jeongseon, Korea, 18–20 February 2013; pp. 39–42.
58. Lotte, F.; Guan, C. Learning from other subjects helps reducing brain-computer interface calibration time. In Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, TX, USA, 14–19 March 2010; pp. 614–617.
59. Blankertz, B.; Kawanabe, M.; Tomioka, R.; Hohlefeld, F.U.; Nikulin, V.V.; Müller, K.-R. Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In Proceedings of the Advances in Neural Information Processing Systems 20 (NIPS 2007), Vancouver, BC, Canada, 3–5 December 2007; pp. 113–120.
60. Wojcikiewicz, W.; Vidaurre, C.; Kawanabe, M. Stationary common spatial patterns: Towards robust classification of non-stationary EEG signals. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech, 22–27 May 2011; pp. 577–580.
61. Wojcikiewicz, W.; Vidaurre, C.; Kawanabe, M. Improving classification performance of BCIs by using stationary common spatial patterns and unsupervised bias adaptation. In Proceedings of the International Conference on Hybrid Artificial Intelligence Systems, Wroclaw, Poland, 23–25 May 2011; pp. 34–41.
62. Kawanabe, M.; Vidaurre, C.; Scholler, S.; Muuller, K.-R. Robust common spatial filters with a maxmin approach. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 3–6 September 2009; pp. 2470–2473.
63. Kawanabe, M.; Samek, W.; Müller, K.-R.; Vidaurre, C. Robust common spatial filters with a maxmin approach. *Neural Comput.* **2014**, *26*, 349–376.
64. Lotte, F.; Guan, C. Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 355–362.
65. Suk, H.-I.; Lee, S.-W. A novel bayesian framework for discriminative feature extraction in brain-computer interfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 286–299.
66. Wang, H.; Zheng, W. Local temporal common spatial patterns for robust single-trial EEG classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2008**, *16*, 131–139.
67. Dornhege, G.; Blankertz, B.; Curio, G.; Müller, K.-R. Increase Information Transfer Rates in BCI by CSP Extension to Multi-class. In Proceedings of the Advances in Neural Information Processing Systems 16, Vancouver and Whistler, BC, Canada, 8–13 December 2003.

68. Yang, Y.; Chevallier, S.; Wiart, J.; Bloch, I. Time-frequency optimization for discrimination between imagination of right and left hand movements based on two bipolar electroencephalography channels. *EURASIP J. Adv. Signal Process.* **2014**, 38, doi:10.1186/1687-6180-2014-38
69. Yang, Y.; Chevallier, S.; Wiart, J.; Bloch, I. Subject-specific time-frequency selection for multi-class motor imagery-based BCIs using few Laplacian EEG channels. *Biomed. Signal Process. Control* **2017**, *38*, 302–311.
70. Yang, Y.; Chevallier, S.; Wiart, J.; Bloch, I. Subject-Specific Channel Selection Using Time Information for Motor Imagery Brain-Computer Interfaces. *Cogn. Comput.* **2016**, *8*, 505–518.
71. BCI Competitions. Available online: http://www.bbci.de/competition/ (accessed on 05/06/2017).
72. Fabien Lotte. Matlab codes and software. Available online: https://sites.google.com/site/fabienlotte/code-and-softwares (accessed on 12/11/2017).
73. Yang, Y.; Chevallier, S.; Wiart, J.; Bloch, I. Automatic selection of the number of spatial filters for motor-imagery BCI. In Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, 25–27 April 2012; pp. 109–114.
74. Wojciech Samek. The Divergence Methods Web Site . Available online: http://divergence-methods.org (accessed on 12/01/2017).

Oral exposition at international workshop: Review and experimental comparison in motor imagery BCI systems

The author also contributed with an oral exposition at the international workshop on Information Science and Technology that took place in the Escuela Técnica Superior de Ingeniería in Seville. This oral exposition consisted in a introduction, review and experimental comparison of the main techniques used in the MI-BCI systems. Since it does not have an abstract we simply attach the program.



International Workshop on Information Science and Technology

ICIST2018 Workshop

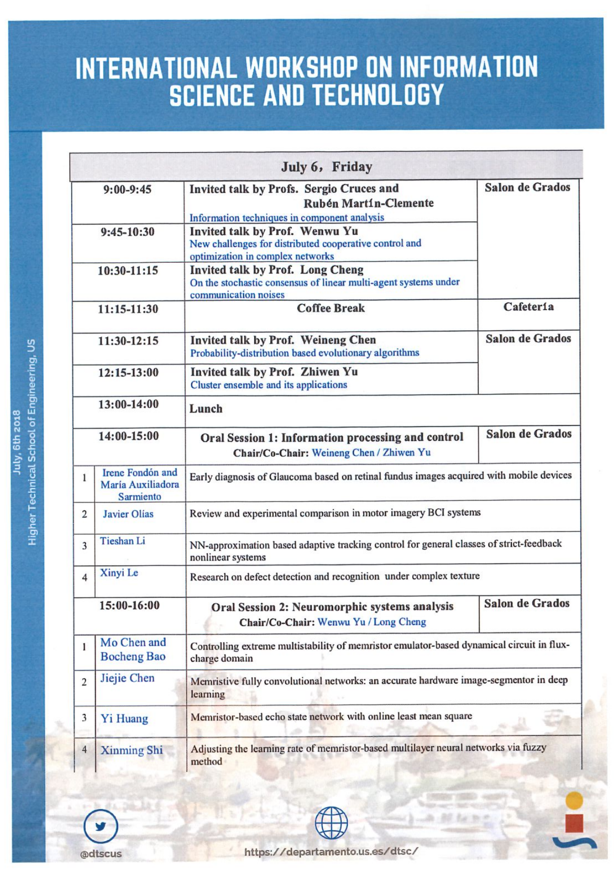
July 6, 2018

Higher Technical School of Engineering, University of Seville

Vicerrectorado de Investigación de la Universidad de Sevilla

University of Seville

IEEE Systems, Man and Cybernetics Society



INTERNATIONAL WORKSHOP ON INFORMATION SCIENCE AND TECHNOLOGY

July, 6th 2018
Higher Technical School of Engineering, US

July 6，Friday			
9:00-9:45		Invited talk by Profs. Sergio Cruces and Rubén Martín-Clemente Information techniques in component analysis	Salon de Grados
9:45-10:30		Invited talk by Prof. Wenwu Yu New challenges for distributed cooperative control and optimization in complex networks	
10:30-11:15		Invited talk by Prof. Long Cheng On the stochastic consensus of linear multi-agent systems under communication noises	
11:15-11:30		Coffee Break	Cafetería
11:30-12:15		Invited talk by Prof. Weineng Chen Probability-distribution based evolutionary algorithms	Salon de Grados
12:15-13:00		Invited talk by Prof. Zhiwen Yu Cluster ensemble and its applications	
13:00-14:00		Lunch	
14:00-15:00		Oral Session 1: Information processing and control Chair/Co-Chair: Weineng Chen / Zhiwen Yu	Salon de Grados
1	Irene Fondón and María Auxiliadora Sarmiento	Early diagnosis of Glaucoma based on retinal fundus images acquired with mobile devices	
2	Javier Olías	Review and experimental comparison in motor imagery BCI systems	
3	Tieshan Li	NN-approximation based adaptive tracking control for general classes of strict-feedback nonlinear systems	
4	Xinyi Le	Research on defect detection and recognition under complex texture	
15:00-16:00		Oral Session 2: Neuromorphic systems analysis Chair/Co-Chair: Wenwu Yu / Long Cheng	Salon de Grados
1	Mo Chen and Bocheng Bao	Controlling extreme multistability of memristor emulator-based dynamical circuit in flux-charge domain	
2	Jiejie Chen	Memristive fully convolutional networks: an accurate hardware image-segmentor in deep learning	
3	Yi Huang	Memristor-based echo state network with online least mean square	
4	Xinming Shi	Adjusting the learning rate of memristor-based multilayer neural networks via fuzzy method	

@dtscus

https://departamento.us.es/dtsc/

Publication D

Unsupervised Common Spatial Patterns

- **Title:** Unsupervised Common Spatial Patterns.
- **Authors:** Rubén Martín-Clemente; Javier Olias; Sergio Cruces; Vicente Zarzoso.
- **DOI:** 10.1109/TNSRE.2019.2936411.
- **Published in:** IEEE Transactions on Neural Systems and Rehabilitation Engineering (Volume: 27 , Issue: 10).
- **Impact factor:** 3972;
- **Quartile:** Q1 (Engineering – Biomedical, Rehabilitation)
- **Date:** May 2019.
- **Pages:** 2135 - 2144.
- **Publisher:** IEEE.

- **Abstract:** The common spatial pattern (CSP) method is a dimensionality reduction technique widely used in brain-computer interface (BCI) systems. In the two-class CSP problem, training data are linearly projected onto directions maximizing or minimizing the variance ratio between the two classes. The present contribution proves that kurtosis maximization performs CSP in an unsupervised manner, i.e., with no need for labeled data, when the classes follow Gaussian or elliptically symmetric distributions. Numerical analyses on synthetic and real data validate these findings in various experimental conditions, and demonstrate the interest of the proposed unsupervised approach.

Summary of Publication D

As we have seen in Chapter 4 the CSP algorithm reduces the dimensionality of the covariances in a supervised manner. Therefore, it relays in the availability and correct labeling of the trials, this is one of the reasons why most MI-BCI systems need a training phase.

The CSP algorithm provides us with a set of orthogonal filters or linear combination of them ($\mathbf{w}_i$) that maximizes the variance of a class while minimizing the variance of the other. The filters are sorted and we can choose the ones that best separate the classes, since each filter there is a maximum or a minimum of the following ratio:

$$\frac{var(\mathbf{w}_i{}^\top \mathbf{X}_{C_1})}{var(\mathbf{w}_i{}^\top \mathbf{X}_{C_2})} = \frac{\sigma^2_{(\mathbf{w}_i,1)}}{\sigma^2_{(\mathbf{w}_i,2)}}, \tag{D.33}$$

where we have use the notation $\sigma^2_{(k,\mathbf{w}_i)}$ to reference the variance of the EEG signal related to the class k when it is filtered with the filter $\mathbf{w}_i$.

Contribution 1: *In this publication we show that CSP can be performed in a blind context.*

We present a method to compute the CSP filters in a blind manner without relaying in the labels of the trials and using the kurtosis. The kurtosis is a statistic measure that can be seen as the normalized fourth order moment of a random distribution and is defined as:

$$\kappa(x) := \frac{\mathbb{E}\left[(x-\mu_x)^4\right]}{\mathbb{E}\left[(x-\mu_x)^2\right]^2}, \tag{D.34}$$

for any random variable x with mean equal to μ_x. The kurtosis is interpreted as a measure of the heaviness of the tails of a distribution. The samples that are farther away from the mean value are those that contribute the most to the kurtosis value. Therefore, it is considered to be a good measure to check whether there is or not outliers in a distribution.

To establish the link between the kurtosis and the CSP algorithm we start by defining the EEG data as the combination of two independent classes samples, which are represented as two Gaussian distribution with zero mean and different covariance matrices:

$$\mathbf{X} \sim \pi_1 \mathscr{N}(\mathbf{0}, \mathbf{\Sigma}_1) + \pi_2 \mathscr{N}(\mathbf{0}, \mathbf{\Sigma}_2), \tag{D.35}$$

where the variables π_1,π_2 represent the probabilities of occurrence of class one and two respectively. With this model, we define an arbitrary lineal combination of the channels weighted by the vector $\mathbf{a}$ with variance:

$$var\left(\mathbf{a}^\top \mathbf{X}\right) = \mathbb{E}\left[(\mathbf{a}^\top \mathbf{X})^2\right] = \pi_1 \sigma^2_{(1,\mathbf{a})} + \pi_2 \sigma^2_{(2,\mathbf{a})}. \tag{D.36}$$

Now we compute the fourth other moment, which in the case of Gaussian distribution is given by $\mathbb{E}\left[(x-\mu_x)^4\right] = 3\sigma_x^4$ and as we impose that the distributions from the two classes are independent, the fourth order momentum of the lineal combination results in:

$$\mathbb{E}\left[(\mathbf{a}^\top \mathbf{X})^4\right] = 3\pi_1 \sigma^4_{(1,\mathbf{a})} + 3\pi_2 \sigma^4_{(2,\mathbf{a})}. \tag{D.37}$$

Now, we can plug (D.36) and (D.37) in (D.34) to obtain the following equation:

$$\kappa(\mathbf{a}^\top \mathbf{X}) = \frac{\mathbb{E}\left[(\mathbf{a}^\top \mathbf{X})^4\right]}{\mathbb{E}\left[(\mathbf{a}^\top \mathbf{X})^2\right]^2} = \frac{3\pi_1 \sigma^4_{(1,\mathbf{a})} + 3\pi_2 \sigma^4_{(2,\mathbf{a})}}{\left(\pi_1 \sigma^2_{(1,\mathbf{a})} + \pi_2 \sigma^2_{(2,\mathbf{a})}\right)^2}. \tag{D.38}$$

By maximizing the previous equation, we arrive to the CSP solution. In the appendix of the paper we can find a full explanation, which is based on Lagrange multipliers. But in an intuitive manner, because of the parabola in the denominator, we can see that the equation (D.38) has its maximum values when either $\sigma_{(1,\mathbf{a})}$ increases with respect to $\sigma_{(2,\mathbf{a})}$ or vice-versa, or in other words, when the ratio in (D.33) is either maximum or minimum.

Contribution 2: *The publication extends the unsupervised CSP algorithm to the case of elliptical distributions.*

We extend the previous result to the case of elliptically distributed data. This allows us to address much more general problems related to actual practice.

Contribution 3: *The solution of the unsupervised CSP can be found through existing optimization algorithms.*

The proposed algorithm allows us to use algorithms that already exists for the optimization of the kurtosis. These algorithms are computationally efficient and also establish a link with other problems, such as Independent Component Analysis (ICA) which also relies on the kurtosis.

Contribution 4: *The publication provides with experiments over synthetic and real data, performing unsupervised classification over the BCI competitions.*

In *section V* we separate two subsection. In the first one we center the experiment on simulated data that show us the similitude between the CSP algorithm and the Kurtosis maximization technique.

Later in *section V.B* we experiment over real data and perform classification in an unsupervised fashion using the GMM algorithm, obtaining fair accuracy results.

Contribution 5: *This publication may open a new research line.*

We show that unsupervised training of BCI is possible, which opens the door to develop a new paradigm with innovative applications and it is more robust against the appearance of unlabeled data.

Unsupervised common spatial patterns

Rubén Martín-Clemente, *Member, IEEE*, Javier Olias, Sergio Cruces, *Senior Member, IEEE* and Vicente Zarzoso, *Senior Member, IEEE*

***Abstract*—The common spatial pattern (CSP) method is a dimensionality reduction technique widely used in brain-computer interface (BCI) systems. In the two-class CSP problem, training data are linearly projected onto directions maximizing or minimizing the variance ratio between the two classes. The present contribution proves that kurtosis maximization performs CSP in an unsupervised manner, i.e., with no need for labelled data, when the classes follow Gaussian or elliptically symmetric distributions. Numerical analyses on synthetic and real data validate these findings in various experimental conditions, and demonstrate the interest of the proposed unsupervised approach.**

***Index Terms*—Common spatial patterns, Brain Computer Interfaces, Kurtosis**

I. Introduction

Common spatial patterns (CSP) is a dimension reduction technique widely used in brain-computer interface (BCI) systems [1], [2], [3], [4]. Typically, electroencephalogram (EEG) samples acquired under two different experimental conditions provide a multivariate data set with two classes. CSP linearly projects the data onto directions where the variance of the projected data points is significantly higher for one class than for the other [5], [6], [7]. The projected data variances can then be used as features for classification. CSP is a supervised technique, whose performance relies heavily on the availability of correctly labelled data.

The present contribution proves that CSP can be also performed in an *unsupervised* fashion by maximizing the kurtosis (normalized fourth-order moment) of the projected data. Unsupervised operation spares the need for training labels and is thus immune to erroneous labelling. Apart from its theoretical interest, this result is useful, for instance, in applications where the training labels are not available or may be uncertain. A mathematical proof is derived for data drawn from a mixture of Gaussian densities, and then generalized to elliptically symmetric distributions. Our experimental evaluation on synthetic and real data corroborates the theoretical findings. Unsupervised techniques are not unknown in EEG processing: e.g., [8] shows that it is possible to perform unsupervised workload classification using EEG spectral features. We can expect that they will become increasingly common in the near future.

The paper is organized as follows: Section II reviews the mathematical formulation of the common spatial patterns method. Section III, the core of our contribution, establishes the link between the kurtosis and the CSP criterion under the assumption of a Gaussian mixture model for the data. Section IV extends this result to elliptically distributed classes. Illustrative examples supporting the theoretical derivations are presented and discussed in Section V. The concluding remarks of Section VI bring the paper to an end. For the sake of clarity, proofs of the theoretical results have been deferred to the Appendices.

II. Common Spatial Patterns

Consider that we are given a set of observations of a random variable X in $\mathbb{R}^p$, in which each observation belongs to one of two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. CSP is usually applied to problems where the class means are null, and this assumption is made in the sequel. A one-dimensional projection of the point cloud can be represented by $Y = \boldsymbol{a}^\intercal X$, where $\boldsymbol{a} \in \mathbb{R}^p$. Denoting by $\boldsymbol{\Sigma}_k = \mathrm{Cov}(X \mid \mathcal{C}_k)$ the covariance matrix of the data in class $\mathcal{C}_k$, with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, it holds that the variance of class $\mathcal{C}_k$, after the projection, equals

$$\sigma_k^2 = \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_k \boldsymbol{a}, \qquad k = 1, 2. \tag{1}$$

The idea behind CSP is to maximize σ_1^2 while minimizing σ_2^2 or *vice versa* [9]. To this end, the objective function is defined as the power ratio

$$R(\boldsymbol{a}) = \frac{\sigma_1^2}{\sigma_2^2}. \tag{2}$$

Note that σ_k^2, $k = 1, 2$, depend on $\boldsymbol{a}$ through relation (1). Also, ratio (2) is scale invariant, i.e., $R(\boldsymbol{a}) = R(c\boldsymbol{a})$, for all $c \in \mathbb{R} \setminus \{0\}$, and therefore only the direction of the projection is significant but not the overall scaling. To find the optimal $\boldsymbol{a}^*$ corresponding to the extremum (maximum or minimum) of CSP criterion (2) we set its gradient $\nabla R(\boldsymbol{a})$ to zero, readily yielding

$$\boldsymbol{\Sigma}_1 \boldsymbol{a}^* = R(\boldsymbol{a}^*) \boldsymbol{\Sigma}_2 \boldsymbol{a}^*. \tag{3}$$

It follows that $\boldsymbol{a}^*$ is a generalized eigenvector of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ [10]. Solving this problem we get the eigenvector corresponding to the maximum (respectively, minimum) eigenvalue, which maximizes (resp. minimizes) the CSP objective function. When more features are needed for the posterior classification stage, a common practice is to project the data onto other eigenvectors from both ends of the eigenvalue spectrum [11]. In the generalized eigenvalue (GEVD) problem (3), $\boldsymbol{a}_i$ denote the eigenvectors and their corresponding eigenvalues are assumed to be sorted in decreasing order: $R(\boldsymbol{a}_i) > R(\boldsymbol{a}_j)$, for $i < j$, $i = 1, 2, \ldots, p$.

Manuscript submitted December 19, 2019. This work was supported in part by the Spanish Ministry of Economy and Competitiveness under project no. TEC2017-82807-P. R. Martín-Clemente held an invited professorship at the I3S Laboratory in July 2018 funded by an IFA scholarship from Université Côte d'Azur. Most of this work was carried while V. Zarzoso was a member of the *Institut Universitaire de France*.

R. Martín-Clemente, J. Olias and S. Cruces are with the Department of Signal Theory and Communications, University of Seville, Avda. Descubrimientos s/n, 41092 Seville, Spain. E-mails: {ruben, folias, sergio}@us.es. V. Zarzoso is with Université Côte d'Azur, CNRS, I3S Laboratory, 06903 Sophia Antipolis Cedex, France. E-mail: vicente.zarzoso@univ-cotedazur.fr.

It is important to remark that a training set of correctly classified observations is required to estimate matrices $\mathbf{\Sigma}_k$, $k = 1, 2$. It is in this sense that CSP can be considered a *supervised* technique. The remaining of this paper presents a fully data-driven procedure that does not require labelled samples, thus performing CSP in an unsupervised manner.

III. Kurtosis as a Blind CSP Criterion

We first focus on the important case where X is distributed as a mixture of two Gaussian densities [11]:

$$X \sim \pi_1 \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2) \tag{4}$$

where π_k stands for the prior probability of class $\mathcal{C}_k$, $k = 1, 2$, with $\pi_1 + \pi_2 = 1$, and $\mathbf{0}$ is a p-dimensional vector of zeros. The distribution of the one-dimensional projection $Y = \boldsymbol{a}^\intercal X$ is also a mixture of Gaussians, that is, $Y \sim \pi_1 \mathcal{N}(0, \sigma_1^2) + \pi_2 \mathcal{N}(0, \sigma_2^2)$, where σ_1^2 and σ_2^2 are defined as in eqn. (1). From the properties of the Gaussian distribution, it follows that

$$\mathrm{E}\{Y^2\} = \pi_1\sigma_1^2 + \pi_2\sigma_2^2, \qquad \mathrm{E}\{Y^4\} = 3\pi_1\sigma_1^4 + 3\pi_2\sigma_2^4$$

where $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation operator. Now, the kurtosis is a statistic defined as the normalized fourth-order moment [12]

$$\kappa_Y(\boldsymbol{a}) \coloneqq \mathrm{E}\{Y^4\}/\mathrm{E}\{Y^2\}^2. \tag{5a}$$

Under data model (4), the kurtosis can be expressed as:

$$\kappa_Y(\boldsymbol{a}) = \frac{3\pi_1\sigma_1^4 + 3\pi_2\sigma_2^4}{(\pi_1\sigma_1^2 + \pi_2\sigma_2^2)^2}. \tag{5b}$$

Some preliminary manipulations show that $\nabla\kappa(\boldsymbol{a}) = \mathbf{0}$ if and only if we have $(R(\boldsymbol{a}) - 1)\nabla R(\boldsymbol{a}) = \mathbf{0}$, meaning that the critical points of $R(\boldsymbol{a})$ are also critical points of $\kappa(\boldsymbol{a})$. Indeed, a thorough analysis detailed in Appendix A leads to the following result:

Theorem 1: Under the working assumptions of CSP recalled in Sec. II, the local maximizers of the kurtosis (5b) maximize or minimize the CSP criterion (2). In particular, the maximizers of kurtosis are $\boldsymbol{a}_1$ if $R(\boldsymbol{a}_1) > 1$ and $\boldsymbol{a}_p$ if $R(\boldsymbol{a}_p) < 1$.

In other words, CSP can be performed blindy, that is, without the need for labelled samples, by projecting the observed data points onto the direction that maximizes the kurtosis of the projections. Consequently, the new method will be referred to as kurtosis-based unsupervised CSP (k-uCSP). Under different working assumptions, namely that the data are a combination of statistically independent variables, the optimization of the kurtosis gives rise to independent component analysis (ICA) [13], [14]. Furthermore, in clustering problems, maximizing the kurtosis can produce the same results as Fisher linear discriminant analysis [15]. We see that the kurtosis is a widely-used tool in signal processing. It is the true model the data follow that determines the outcome achieved.

It should be remarked that the expression of kurtosis given in eqn. (5b) is only exploited to prove the equivalence between kurtosis optimization and common spatial pattern analysis in our theoretical derivations, but we use eqn. (5a) in the actual algorithm to find the patterns. Therefore, the proposed technique is fully unsupervised.

Finally, Theorem 1 also admits an intuitive interpretation. Dividing both the numerator and denominator of (5b) by σ_2^4, the right-hand part of this formula can be expressed in terms of $R(\boldsymbol{a})$,

$$\kappa_Y(\boldsymbol{a}) = \frac{3\pi_1 R^2(\boldsymbol{a}) + 3\pi_2}{(\pi_1 R(\boldsymbol{a}) + \pi_2)^2}, \tag{6}$$

thus revealing that there exists a relationship between the supervised and unsupervised criteria. Fig. 1 shows an example plot of κ_Y against R when $\pi_1 = \pi_2$ and R is in the range $[0.5, 4]$. We make the observation that the maximizers of the kurtosis can lie only in the strictly increasing/decreasing range of (6). Consequently, (6) is also increasing/decreasing in some neighborhood of them and, therefore, invertible. Then, a decrease in the value of the kurtosis in that neighborhood results in a decrease/increase in the value of R. As it is intuitive, this implies that the local maximizers of the kurtosis maximize or minimize the CSP criterion. In Figure 1, additionally, it is not hard to show that as (6) is decreasing for all vectors sufficiently near $\boldsymbol{a}_p$, then (6) has a maximum at $\boldsymbol{a}_p$. Virtually the same argument shows that there is another maximum at $\boldsymbol{a}_1$.

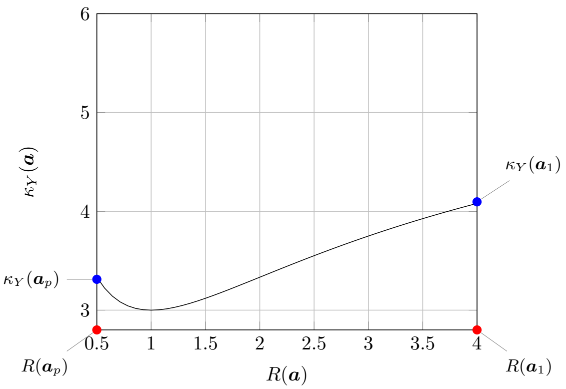


Fig. 1: The kurtosis versus the CSP objective function for $\pi_1 = \pi_2 = 1/2$ with $0.5 = R(\boldsymbol{a}_p) \leq R(\boldsymbol{a}) \leq R(\boldsymbol{a}_1) = 4$.

IV. Extension to Elliptic Distributions

The above result can be extended to the case where the data follow a mixture of zero-mean elliptically symmetric distributions [16]. These distributions generalize the class of multivariate Gaussians by allowing for both heavier-than-Gaussian and lighter-than-Gaussian distribution tails. Examples include the *t* and Laplace distributions. Before continuing, a zero-mean p-dimensional random variable Z is *elliptically distributed* if its characteristic function $\varphi_Z(\boldsymbol{a}) \coloneqq \mathrm{E}\{e^{j\boldsymbol{a}^\intercal Z}\}$, $j = \sqrt{-1}$, $\boldsymbol{a} \in \mathbb{R}^p$, is of the form $\varphi_Z(\boldsymbol{a}) = \phi\left(-\frac{1}{2}\boldsymbol{a}^\intercal\mathbf{\Sigma}\boldsymbol{a}\right)$ for some nonnegative-definite matrix $\mathbf{\Sigma}$. Function $\phi(\cdot)$ is called *characteristic generator* of the distribution. For example, for Gaussian variables $\phi(\alpha) = \exp(\alpha)$ and for multivariate Laplacian variables $\phi(\alpha) = 1/(1 - \alpha)$. The characteristic function of the univariate random variable $W = \boldsymbol{a}^\intercal Z$ is given by $\varphi_W(t) := \mathrm{E}\{e^{jtW}\} = \mathrm{E}\{e^{jt\boldsymbol{a}^\intercal Z}\} = \varphi_Z(t\boldsymbol{a}) = \phi\left(-\frac{1}{2}t^2\boldsymbol{a}^\intercal\boldsymbol{\Sigma}\boldsymbol{a}\right)$, $t \in \mathbb{R}$. It follows that, if the variance of W exists, then

$$\sigma^2 := E\{W^2\} = -\left.\frac{\partial^2}{\partial t^2}\varphi_W(t)\right|_{t=0} = \phi'(0)\boldsymbol{a}^\intercal\boldsymbol{\Sigma}\boldsymbol{a} \quad (7)$$

$$E\{W^4\} = \left.\frac{\partial^4}{\partial t^4}\varphi_W(t)\right|_{t=0} = 3\gamma\sigma^4 \quad (8)$$

where we have defined $\gamma := \frac{\phi''(0)}{\phi'(0)^2}$, which is a strictly positive number. Notations ϕ' and ϕ'' represent, respectively, the first- and second-order derivative of ϕ.

Now, let us replace the Gaussian distributions in (4) with zero-mean elliptically symmetric distributions having characteristic generators of the same type $\phi(\cdot)$ but defined by matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, i.e., $\phi(-\boldsymbol{a}^\intercal\boldsymbol{\Sigma}_1\boldsymbol{a}/2)$ and $\phi(-\boldsymbol{a}^\intercal\boldsymbol{\Sigma}_2\boldsymbol{a}/2)$, respectively. Then, the moments of Y become

$$\mathrm{E}\{Y^2\} = \pi_1\sigma_1^2 + \pi_2\sigma_2^2, \qquad \mathrm{E}\{Y^4\} = 3\pi_1\gamma\sigma_1^4 + 3\pi_2\gamma\sigma_2^4$$

with σ_k^2, $k = 1, 2$, given by (7), and their ratio turns out to be a positively scaled version of (5b):

$$\kappa_Y(\boldsymbol{a}) = \gamma\frac{3\pi_1\sigma_1^4 + 3\pi_2\sigma_2^4}{(\pi_1\sigma_1^2 + \pi_2\sigma_2^2)^2}. \quad (9)$$

It readily follows that Theorem 1 also holds for data classes with elliptic distributions defined by the same type of characteristic generator $\phi(\cdot)$.

V. Experimental Assessment

A number of experiments are performed to validate the theoretical study of the unsupervised CSP criterion developed in this paper and to test its performance in a variety of experimental conditions. These include synthetically generated as well as real EEG data. To perform the numerical optimization of the kurtosis statistic (5a), we employ the algorithm presented in [24]. We point out that this algorithm does not require any class labels as inputs. A free MATLAB implementation of the algorithm is provided in [25].

A. Simulated Data

To illustrate Theorem 1, let us first consider a mixture of equiprobable classes in a two-dimensional space, i.e., $p = 2$. We consider three cases that only differ in the choice of matrix $\boldsymbol{\Sigma}_2$. Case 1 assumes that $\boldsymbol{\Sigma}_1 = \mathrm{diag}(2, 1)$ and $\boldsymbol{\Sigma}_2 = \mathrm{diag}(0.5, 2)$; in case 2, $\boldsymbol{\Sigma}_2$ is replaced by $\mathrm{diag}(0.2, 0.8)$; in case 3, finally, we set $\boldsymbol{\Sigma}_2 = \mathrm{diag}(2.5, 10)$. In all cases, the generalized eigenvector of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ that maximizes the CSP target function (2) is $\boldsymbol{a}_1 = \pm[1, 0]^\intercal$, while the corresponding minimizer is $\boldsymbol{a}_2 = \pm[0, 1]^\intercal$.

Figure 2 plots the CSP criterion R [eqn. (2)] and the kurtosis criterion κ_Y [computed from the ratio of expectations in eqn. (5b)] in dotted and solid lines, respectively, for $\boldsymbol{a} = [\cos(\theta), \sin(\theta)]^\intercal$. Red and black solid lines correspond to Gaussian and Laplacian data, respectively, where the latter were generated as explained in [17]. The results are just as predicted by Theorem 1: the maxima of κ_Y are always maxima or minima of the CSP criterion R, with different patterns of correspondence depending on the generalized eigenvalues of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. In case 1 (Figure 2a), the maxima and minima of R transform into maxima of κ_Y, as $R(\boldsymbol{a}_1) = 4 > 1 > R(\boldsymbol{a}_2) = 0.5$. In case 2 (Figure 2b), κ_Y has the same maxima and minima as R, because both eigenvalues are greater than one: $R(\boldsymbol{a}_1) = 10$, $R(\boldsymbol{a}_2) = 1.25$. Finally, in case 3 (Figure 2c), the minima of R are transformed into maxima of κ_Y and *vice versa*, since both eigenvalues are lower than one: $R(\boldsymbol{a}_1) = 0.8$, $R(\boldsymbol{a}_2) = 0.1$. Additionally, the point $R = 1$, reached in case 1, defines a global minimum of the kurtosis.

As an additional validation experiment, let us also test the case where samples from the same class are correlated, resulting in non diagonal covariance matrices. The following matrices are selected at random:

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 3.8152 & -3.4131 \\ -3.4131 & 3.3104 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2.8465 & 0.5267 \\ 0.5267 & 1.2446 \end{bmatrix}.$$

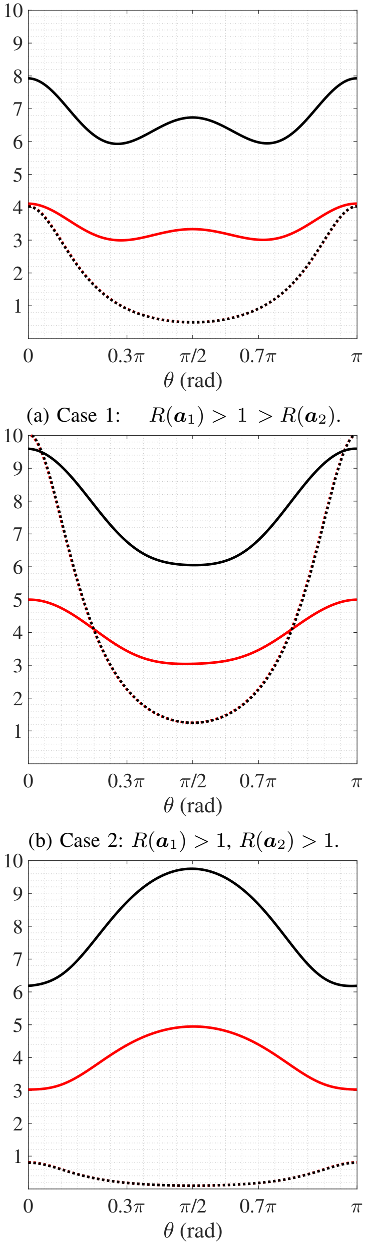
For these covariances, the maximizer $\boldsymbol{a}_1$ of R is the unit vector that makes an angle of $\theta_1 \approx 2\pi/3$ radians with the positive x-axis. Similarly, the minimizer $\boldsymbol{a}_2$ is at an angle $\theta_2 \approx \pi/4$ radians. Figure 3 (top) plots R and κ_Y in a format similar to that of Figure 2. An alternative representation is given in Figure 3 (bottom). The color-coded circles indicate the kurtosis of the entire projected data (outer circle) and the variance ratio of the projected classes (inner circle) as a function of angle θ. For the reader's convenience, we complete the figure by drawing a pair of dashed lines in the direction of vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. Additionally, we draw in different colours the scatterplots of the classes and the probability density functions of the data points after being projected onto $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, which illustrates well the disparity between the variances resulting from the projection.

Furthermore, it is known that whitening the data reduces the initial generalized eigenvalue problem (3) to a standard eigenvalue problem for Hermitian matrices (see Appendix B), for which the eigenvectors are orthogonal. To show this property, we whiten the previous data before calculating R and κ_Y. As a result, Fig. 3, obtained before whitening, transforms into Fig. 4 after whitening. As expected, Fig. 4 (bottom) shows that the directions that maximize the kurtosis become perpendicular. Yet the equivalence between the CSP and kurtosis criteria established by Theorem 1 still holds in this case.

B. Real Data

1) Supervised vs. unsupervised CSP of brain data: In typical BCI implementations, subjects are instructed to imagine movements of, e.g., their right or left hand, one after the other, while their p-channel EEG is being recorded. Then, CSP is applied to the data, previously bandpass-filtered in the band of interest, and the power evolution of the resulting time series allows the BCI system to discriminate between the two classes of motor imagery [18], [19], [20].

The purpose of this experiment is to illustrate the performance of the k-uCSP approach on real EEG data. The proposed kurtosis-based approach is tested using datasets from the BCI competition IV (dataset 2a) [21]. These contain EEG acquired on two different daysÂ from nine subjects with a



(a) Case 1: $R(\boldsymbol{a}_1) > 1 > R(\boldsymbol{a}_2)$.

(b) Case 2: $R(\boldsymbol{a}_1) > 1$, $R(\boldsymbol{a}_2) > 1$.

(c) Case 3: $R(\boldsymbol{a}_1) < 1$, $R(\boldsymbol{a}_2) < 1$.

Fig. 2: Criteria $R(\boldsymbol{a})$ (dotted) and $\kappa_Y(\boldsymbol{a})$ (solid) as a function of angle θ defining the projection direction $\boldsymbol{a} = [\cos(\theta), \sin(\theta)]^\intercal$. Red curves are calculated from Gaussian data, while black curves correspond to Laplacian classes. Dotted curves (R) overlap as the pair $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ is the same for both distributions. As predicted by Theorem 1, the local maximizers of kurtosis (solid lines) either maximize or minimize the CSP criterion (dotted lines) depending on the generalized eigenvalues $R(\boldsymbol{a}_1)$ and $R(\boldsymbol{a}_2)$ of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Statistics are estimated from $T = 1000$ random samples, with 500 i.i.d. samples per class, and each curve is the average of 100 independent experiments.

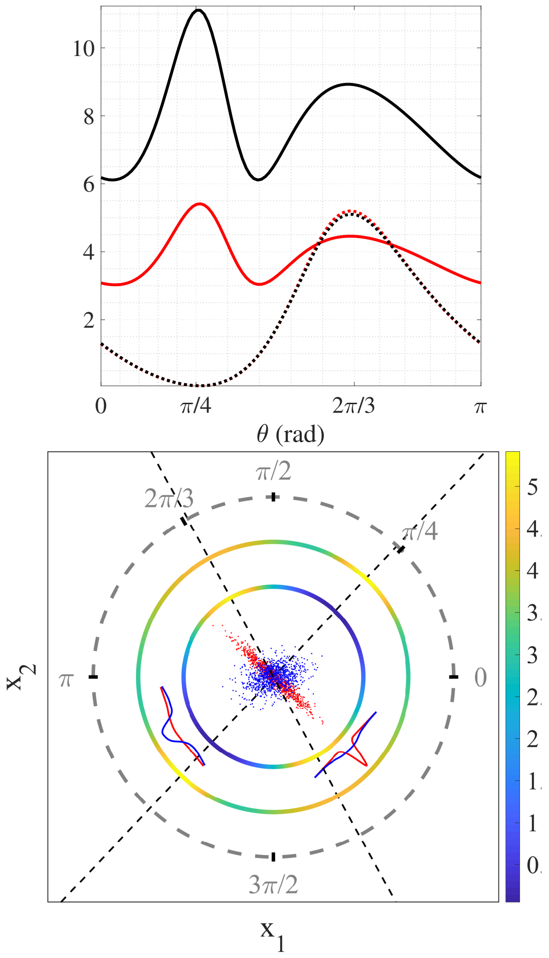


Fig. 3: Supervised vs. unsupervised CSP when samples from the same class are correlated, with $R(\boldsymbol{a}_1) \approx 5.4 > 1 > R(\boldsymbol{a}_2) \approx 0.06$. (Top) Criteria $R(\boldsymbol{a})$ (dotted) and $\kappa_Y(\boldsymbol{a})$ (solid) as a function of angle θ defining the projection direction $\boldsymbol{a} = [\cos(\theta), \sin(\theta)]^\intercal$. As in Fig. 2, red curves are calculated from random Gaussian samples, and black curves from Laplacian data. (Bottom) Alternative representation as a scatterplot of the two correlated Gaussian distributions. Samples from one class are shown as red dots, and from the other as blue dots. The two color-coded circles around the scatterplots indicate the value of $R(\boldsymbol{a})$ (inner circle) and $\kappa_Y(\boldsymbol{a})$ (outer circle). Straight dashed lines mark the directions representing the CSP and kurtosis projections, which coincide in this experiment. The probability density functions of the projected classes are also shown in red and blue lines.

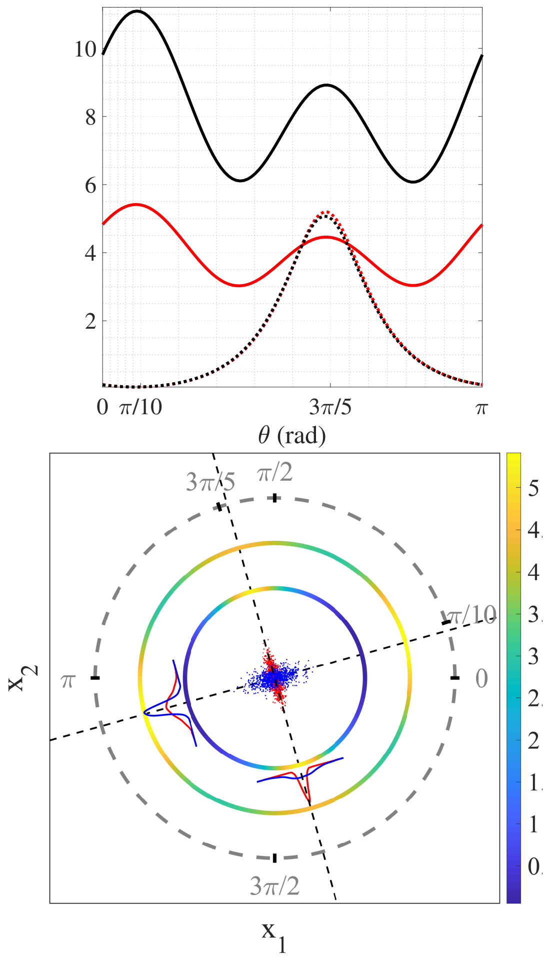


Fig. 4: Results for the data in Fig. 3 after whitening. Apart from the whitening operation, details about the plots are as in the caption of Fig. 3.

p = 22-channel EEG system at a sampling rate of 250 Hz. As pre-processing, we bandpass filter the EEG to $8 - 30$ Hz. This is usual in BCI and ensures that the data are zero-mean. Electrooculogram (EOG) channels are available and ocular artifacts are removed using Signal Space Projections (SSP) [22], [23].

In each trial, an arrow pointing either to left, right, down or up is shown on a display for a short time, and the subject is required to respectively imagine left hand, right hand, both feet or tongue movements in response. The imagined action lasts about three seconds but only the final two are kept to avoid initial transient effects. Thus, each trial data matrix consists of 22 rows of channels and 2 (seconds) × 250 (samples/seconds) = 500 columns of time samples. Data matrices are also normalized as suggested in [26], which generalizes the procedure in [27]. The test statistic proposed in [28] shows that about $2/3$ of the data follow an elliptical distribution with a standard level of significance of $\alpha = 0.01$. Finally, 72 trials of each movement are performed per subject and day.

Experiments are conducted on groups of trials. Each group comprises 144 trials, half from one of the four classes (e.g., left hand) and half from another (e.g., right hand), all from the same subject and recorded the same day. The corresponding 144 trial data matrices are concatenated along the temporal dimension, producing one large observation matrix with dimension 22 × 72000. In total, there are 9 (subjects) × 6 (possible combinations of four classes) × 2 (days) = 108 groups. The projections of the p-channel EEG onto the p-directions maximizing the kurtosis are computed for each group of trials by the algorithm in [24], [25]. Specifically, after applying a whitening transformation, we maximize p-times the kurtosis, one after the other, under the constraint that the direction obtained in the nth maximization is orthogonal to the previously calculated directions (see Appendix B for details).

For illustration purposes, Figure 5 shows several projected trials with maximum kurtosis, calculated from 'left-hand' and 'right-hand' EEG data recorded from subject 1 during the second session. In this particular experiment, the power of the left-hand motor imagery projections (denoted '1') is clearly higher than that of the other hand (denoted '2'), which facilitates class discrimination. A scatter plot of log-variances for all projected trials is also shown in Figure 6, using the supervised projection on the x-axis and the unsupervised kurtosis-based projection on the y-axis.

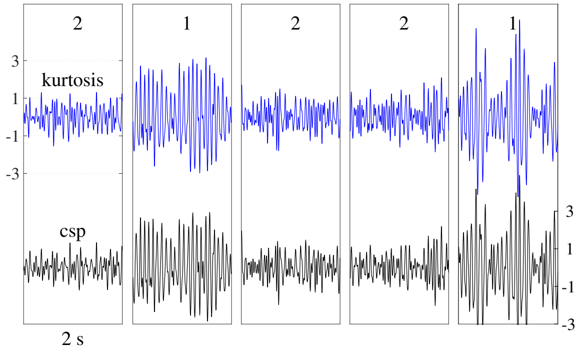


Fig. 5: One-dimensional projection of EEG data with maximal kurtosis (blue line). Several projected trials from the first user are shown. Labels '1' and '2' correspond to data classes ('1' = left hand, '2'= right hand). Visible to the naked eye, the power of one of the projected classes (namely, class '1') is significantly higher, as would be expected from the application of CSP. For comparison, the projection of the same trials onto the corresponding supervised CSP filter are also shown (bottom black curves).

As a more formal test of the relationship between both criteria (supervised and unsupervised), we also project the EEG data onto the p = 22 generalized eigenvectors of the matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. For each projection direction, we calculate the ratio of variances of the projected classes (σ_1^2/σ_2^2). These ratios are then arranged in decreasing order along a

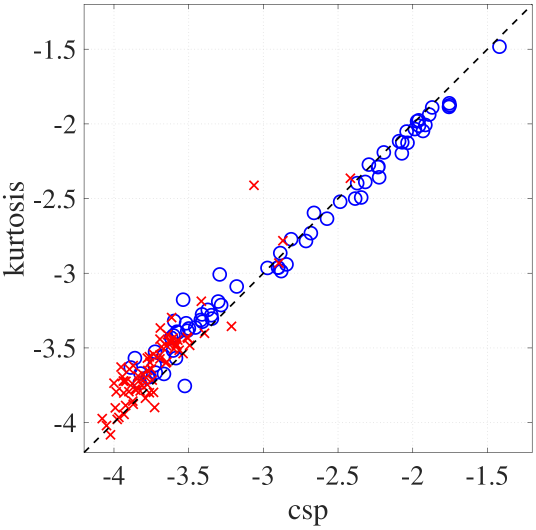


Fig. 6: Scatter plot of log-variances for all ‘left-hand’ vs ‘right-hand’ trials recorded the second day from subject 1, using the supervised projection which maximizes R on the horizontal axis and the corresponding unsupervised projection with maximal kurtosis on the vertical axis. Legend: blue circles = left-hand, red crosses = right-hand.

p-dimensional vector $\boldsymbol{b}_{CSP}$. A distinct vector is generated for each group of trials, as the projection directions differ from one another. Next, we repeat the experiment with the difference that this second time the data are projected onto the p directions maximizing the kurtosis. The new variance ratios are stored in a vector $\boldsymbol{b}_{Kurt}$. A strong similarity is found between the vectors $\boldsymbol{b}_{CSP}$ and $\boldsymbol{b}_{Kurt}$ calculated from the same data. The Pearson correlation coefficient between the components of these two vectors, averaged for all movements, is 0.9692 for subject 6 (best), and 0.9484 for subject 9 (worst). The correlation coefficient averaged over all groups of trials is 0.9586 with standard deviation 0.0257. Both approaches therefore provide very similar variance ratios, though one method is supervised while the other is not.

2) Unsupervised discrimination: The k-uCSP approach may be combined with some unsupervised classification technique in order to design a full data-driven BCI system. Some preliminary illustrative experiments are presented in the remaining of this Section. Let $Y_1, \ldots, Y_p$ be the unsupervised projections that result from projecting a single trial data matrix. These matrices are 22×500, so that each Y_i consists of 500 values. As in [27], their variances are used to calculate the following $p = 22$ features

$$F_i = \log\left(\frac{\mathrm{var}(Y_i)}{\sum_{n=1}^{p} \mathrm{var}(Y_n)}\right), \quad i = 1, \ldots, p, \tag{10}$$

which are arranged in a p-vector.

For illustrative purposes only, Figure 7 draws the scatterplot of 80 log-variances F_i computed from ‘left vs right-hand’ trials from subject 1. Superimposed, we draw the contour plot of the best fitting two-component Gaussian mixture model (GMM) of the point cloud [29]. By assigning a point to the Gaussian distribution it most probably belongs to, we partition the space in two regions with the hope that these regions accurately reflect the original classes. The performance of the classification can be evaluated by using the true labels as ground truth. For example, in Figure 7, when the subject thinks of a ‘left-hand’ movement, 27 times out of 40 (about 2 times out of 3) it is assigned to Region 1. Similarly, ‘right-hand’ movements are 36 times of 40 (9 times out of 10) assigned to Region 2. As both kind of movements are equiprobable, the average probability of success or accurate classification can be calculated as

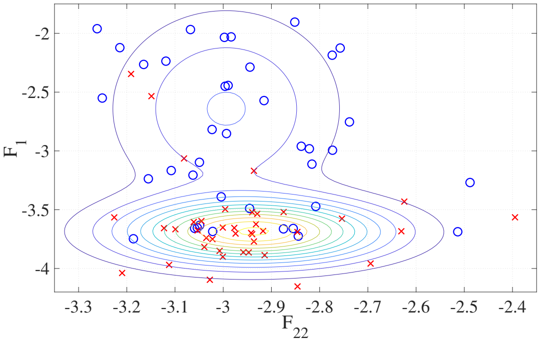


Fig. 7: Scatter plot of the k-uCSP features F_1 and F_{22}, calculated by (10) from ‘left vs right-hand’ trials from subject 1 (circles = left-hand, crosses = right-hand). The features are sorted according a variance criterion: $\mathrm{var}(F_1) > \ldots > \mathrm{var}(F_p)$, where the variance is calculated across all the trials within a group. Superimposed we show the contour plot of the two-component best fitting Gaussian mixture model of the points.

$$P = \frac{27}{40} \times \frac{1}{2} + \frac{36}{40} \times \frac{1}{2} = 0.7875.$$

Our last experiments are conducted on each group of 144 trials. We run a 10 fold cross-validation. To this end, the trials are divided into training (130 trials) and testing (14 trials). In the training phase, we calculate the unsupervised projection directions from the 130 trials in the training set. Then, we fit a two-component p-dimensional GMM to the resulting 130 log-variance feature vectors. Next, in the testing phase, the 14 trials of the testing set are projected onto the previously calculated directions, and the GMM is used for classifying the corresponding 14 p-vectors of log variance features. This is repeated 10 times, with different training and testing sets, and the results are averaged out.

We observe that the performance largely depends on the type of motor imagery. For example, for subject 1, best results are obtained for the classification of ‘right-hand’ vs ‘tongue’ movements: we get 96.48% of accuracy in session 1 and 97.19% in session 2, 96.83% in average. However, ‘feet’ and ‘tongue’ movements are hardly distinguishable one from another: only the 56.19% (session 1), 61.81% (session 2) and 59.0% (average) of the corresponding trials are correctly classified. The following averaged classification accuracies are obtained for the combination of imagined movements with the best performance:

- Subject 1: 96.83% ('right-hand' vs 'tongue'),
- Subject 2: 65.57% ('feet' vs 'tongue'),
- Subject 3: 80.45% ('right-hand' vs 'tongue'),
- Subject 4: 72.52% ('left-hand' vs 'feet'),
- Subject 5: 62.88% ('left-hand' vs 'tongue'),
- Subject 6: 64.33% ('right-hand' vs 'feet'),
- Subject 7: 79.57% ('left-hand' vs 'tongue'),
- Subject 8: 93.07% ('left-hand' vs 'tongue'),
- Subject 9: 91.59% ('left-hand' vs 'tongue').

The average of all these numbers equals 78.31%, not far from the mean classification accuracy (82.3%) obtained by the unsupervised workload classification approach in [8]. It is a good performance considering its unsupervised nature. For comparison, Table I also gives the results obtained by the supervised CSP approach, where supervised classification is carried out with Fisher linear discriminant analysis (LDA) [30].

The complete results are shown in Figure 8, grouped by subject. For example, the accuracy of subject 1 ranges from 56.19% to 97.19%, depending on the motor imagery under consideration. Table II presents the mean value for each subject. For comparison, Table II also shows the accuracy obtained by using the supervised CSP approach and Fisher linear discriminant.

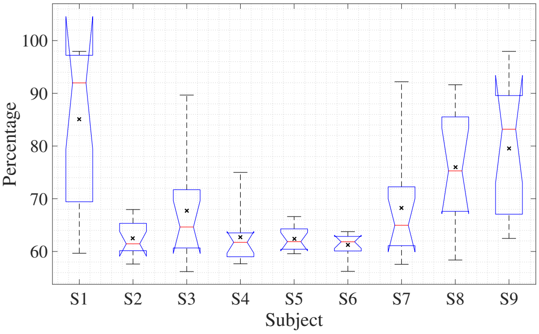


Fig. 8: Box plots of the unsupervised classification accuracy for each user. The boxes extend from the 25th to the 75th percentiles. Lines at either end of the boxes cover the maximum and minimum values. The red line in the middle is the median (50th percentile). Notches represent a confidence interval around the median. The black crosses are the mean values (see also the first column of Table II).

Complementarily, Fig. 9 and Table III show the performance of the criterion for each imagined movement, averaged across the subjects.

A natural question to ask is why there is a marked difference in some data sets between the performance of the traditional supervised approach and the proposed unsupervised one. To address this question, we have investigated the separation between the classes in the feature space using t-Distributed Stochastic Neighbor Embedding (t-SNE) [31], [32]. This is a popular technique for the visualisation of high-dimensional data. It maps each data point to a location in a low (2 or 3) dimensional space. Such a mapping preserves the local structure of the data in the sense that if the data points form well-separated clusters in the original high-dimensional space, they will too in the dimension reduced space.

Specifically, t-SNE is used to map the vectors $(F_1, \ldots, F_p)$ that contain the log-variance features to a space of two

TABLE I: Classification accuracy (in percentage) associated with the combination of imagined movements with the best performance for the unsupervised (k-uCSP with GMM classifier) and supervised (CSP and Fisher LDA) approaches.

Subject	Unsupervised	Supervised
S1	96.83	98.59
S2	65.57	86.69
S3	80.45	94.78
S4	72.52	77.9
S5	62.88	70.85
S6	64.33	69.11
S7	79.57	97.19
S8	93.07	93.40
S9	91.59	93.04

TABLE II: Mean classification accuracy (in percentage), averaged for each subject, comparing the unsupervised (k-uCSP and GMM classifier) and supervised (CSP and Fisher LDA) approaches.

Subject	Unsupervised	Supervised
S1	85.07	89.07
S2	62.52	78.33
S3	67.73	90.44
S4	62.67	74.14
S5	62.35	64.39
S6	61.28	67.26
S7	68.21	91.25
S8	76.02	88.83
S9	79.56	85.06

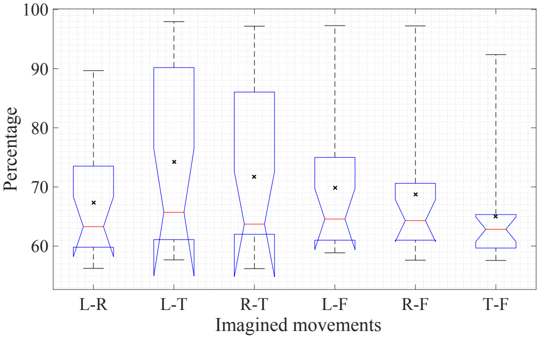


Fig. 9: Box plots of the unsupervised classification accuracy for each imaginary movement ('L-R': left-hand vs right-hand, 'L-T': left-hand vs tongue, 'R-T': right-hand vs tongue, 'L-F': left-hand vs feet, 'R-F': right-hand vs feet, 'T-F': tongue vs feet).

TABLE III: Mean classification accuracy (in percentage) for imaginary movements ('L-R': left-hand vs right-hand, 'L-T': left-hand vs tongue, 'R-T': right-hand vs tongue, 'L-F': left-hand vs feet, 'R-F': right-hand vs feet, 'T-F': tongue vs feet), averaged across subjects, comparing unsupervised and supervised approaches.

Imagined Mov.	Unsupervised	Supervised
L-R	67.3677	76.2513
L-T	74.2751	83.6349
R-T	71.7037	82.2540
L-F	69.8704	82.4471
R-F	68.6958	83.1058
T-F	65.0238	78.1587

dimensions. Let us see an illustrative example: Figure 10a visualizes the mapping of the 'tongue' and 'left-hand' k-uCSP log-variance features calculated from user 7 during Session 1. Each point represents one of the feature vectors. For comparison, 10b depicts the corresponding CSP-based log variance features. We see that the classes are separated, but not always well separated. Consequently, the unsupervised classification algorithm may not be able to differentiate the two groups. In practice, this will lead users to determine on the fly, probably by trial and error, which imagined movements are optimal for themselves (in view of Table III, it may be the combination 'hand' vs 'tongue' for most people).

Before closing the section, it should be remarked that the choice of the GMM classifier based on the features of eqn. (10) is made for illustration purposes only, but is most probably suboptimal. More optimal choices of features and unsupervised classifiers, which are beyond the scope of the present work, are expected to improve unsupervised BCI results.

VI. Conclusions

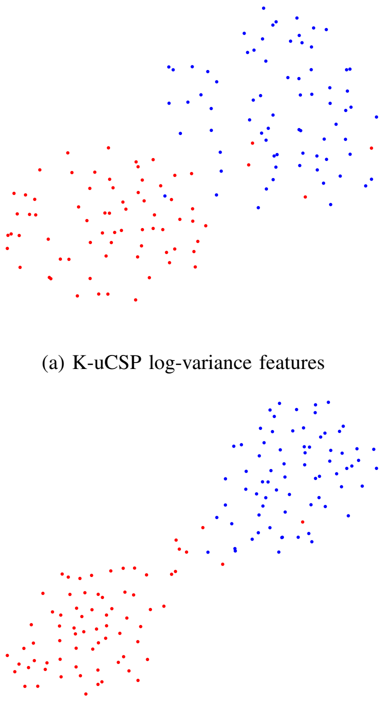
Kurtosis maximization performs CSP in an unsupervised fashion, sparing the need for training data and correct labelling. Further work would also aim at more elaborate BCI systems based on the proposed blind CSP approach.

Acknowledgment

We would like to thank the anonymous Reviewers for their thoughtful comments and efforts towards improving our manuscript.

Appendix A
Proof of Theorem 1

Maximizing (5b) is equivalent to maximizing $f(\boldsymbol{a}) := \pi_1\sigma_1^4 + \pi_2\sigma_2^4$ subject to the constraint $h(\boldsymbol{a}) := \pi_1\sigma_1^2 + \pi_2\sigma_2^2 - 1 = 0$. Because of that constraint, the objective function $f(\boldsymbol{a})$ is defined on a closed interval. This, together with the fact that $f(\boldsymbol{a})$ is continuous, ensures that it has both a maximum and a minimum value that it can attain. Let $\mathcal{L}(\boldsymbol{a}, \lambda) = f(\boldsymbol{a}) + \lambda\, h(\boldsymbol{a})$ be the Lagrangian function and let $\boldsymbol{L}(\boldsymbol{a}, \lambda)$ be the Hessian matrix of $\mathcal{L}(\boldsymbol{a}, \lambda)$ with respect to $\boldsymbol{a}$, having as (i, j)-entry $[\boldsymbol{L}]_{ij} = \frac{\partial^2 \mathcal{L}}{\partial a_i \partial a_j}(\boldsymbol{a}, \lambda)$. Additionally, the tangent plane of $h(\boldsymbol{a})$ at $\boldsymbol{a}^*$ is defined as the set $T(\boldsymbol{a}^*) = \{\boldsymbol{v} : \boldsymbol{v}^\intercal \nabla h(\boldsymbol{a}^*) = 0\}$, where ∇ represents the gradient with respect to $\boldsymbol{a}$. We recall the following result [33, Chap. 20]:



(a) K-uCSP log-variance features

(b) CSP log-variance features

Fig. 10: Visualisation of the 'tongue vs left-hand' log-variance features for user 7. Each class has a different colour set of spots.

Theorem 2: Let $\boldsymbol{a}^*$ be a local maximizer of $f(\boldsymbol{a})$ subject to $h(\boldsymbol{a}) = 0$. Then, there exits $\lambda^* \in \mathbb{R}$ such that

C1) $\nabla f(\boldsymbol{a}^*) + \lambda^* \nabla h(\boldsymbol{a}^*) = 0$, and

C2) for all $\boldsymbol{v} \in T(\boldsymbol{a}^*)$, we have $\boldsymbol{v}^\intercal \boldsymbol{L}(\boldsymbol{a}^*, \lambda^*)\boldsymbol{v} < 0$.

If $\boldsymbol{a}^*$ is a local minimizer, condition C2 becomes $\boldsymbol{v}^\intercal \boldsymbol{L}(\boldsymbol{a}^*, \lambda^*)\boldsymbol{v} > 0$.

Next we check conditions C1 and C2 for our particular problem.

Condition C1: We begin by computing the gradients

$$\nabla f(\boldsymbol{a}) = 4\left(\pi_1\sigma_1^2\boldsymbol{\Sigma}_1\boldsymbol{a} + \pi_2\sigma_2^2\boldsymbol{\Sigma}_2\boldsymbol{a}\right)$$
$$\nabla h(\boldsymbol{a}) = 2\left(\pi_1\boldsymbol{\Sigma}_1\boldsymbol{a} + \pi_2\boldsymbol{\Sigma}_2\boldsymbol{a}\right).$$

Setting the gradient of the Lagrange function to zero, we obtain the equation

$$2\left(\pi_1\sigma_1^2\boldsymbol{\Sigma}_1\boldsymbol{a} + \pi_2\sigma_2^2\boldsymbol{\Sigma}_2\boldsymbol{a}\right) + \lambda\left(\pi_1\boldsymbol{\Sigma}_1\boldsymbol{a} + \pi_2\boldsymbol{\Sigma}_2\boldsymbol{a}\right) = 0. \quad (11)$$

Premultiplying by $\boldsymbol{a}^\intercal$ we easily find that

$$\lambda^* = -2\left(\pi_1\sigma_1^4 + \pi_2\sigma_2^4\right) / \left(\pi_1\sigma_1^2 + \ \pi_2\sigma_2^2\right).$$

Substituting into (11), we get $\sigma_1^2\left(\sigma_2^2-\sigma_1^2\right)\mathbf{\Sigma}_2\,\boldsymbol{a} = \sigma_2^2\left(\sigma_2^2-\sigma_1^2\right)\mathbf{\Sigma}_1\,\boldsymbol{a}$. This equation has two solutions:

$$\text{S1)}\quad \sigma_2^2\mathbf{\Sigma}_1\boldsymbol{a}^* = \sigma_1^2\mathbf{\Sigma}_2\boldsymbol{a}^* \tag{12}$$

$$\text{S2)}\quad \sigma_1^2 = \sigma_2^2. \tag{13}$$

Observe that S1 corresponds to the CSP solution (3).

Condition C2: To ascertain whether these solutions are maximizers or minimizers of the criterion, we need to consider the second-order condition C2. The Hessian matrix of the Lagrangian can also be decomposed as $\boldsymbol{L}(\boldsymbol{a},\lambda) = \boldsymbol{F}(\boldsymbol{a}) + \lambda\boldsymbol{H}(\boldsymbol{a})$, where $\boldsymbol{F}$ and $\boldsymbol{H}$ are the Hessian matrices of $f(\boldsymbol{a})$ and $h(\boldsymbol{a})$, respectively. We first focus on solution (12), which, after some tedious algebraic manipulations, leads to

$$\begin{aligned}\boldsymbol{H}(\boldsymbol{a}^*) &= 2\left(\pi_1\mathbf{\Sigma}_1+\pi_2\mathbf{\Sigma}_2\right)\\ \boldsymbol{F}(\boldsymbol{a}^*) &= 4\sum_{i=1}^{2}\left(2\pi_i\mathbf{\Sigma}_i\boldsymbol{a}^*\boldsymbol{a}^{*\mathsf{T}}\mathbf{\Sigma}_i+\pi_i\sigma_i^2\mathbf{\Sigma}_i\right).\end{aligned}$$

In addition, the tangent plane $T(\boldsymbol{a}^*)$ is the set of vectors $\boldsymbol{v}\in\mathbb{R}^p$ such that $\boldsymbol{v}^\mathsf{T}\nabla h(\boldsymbol{a}^*)=0$, implying that

$$\boldsymbol{v}^\mathsf{T}(\pi_1\mathbf{\Sigma}_1+\pi_2\mathbf{\Sigma}_2)\boldsymbol{a}^*=0 \;\Rightarrow\; \boldsymbol{v}^\mathsf{T}\mathbf{\Sigma}_1\boldsymbol{a}^*=\boldsymbol{v}^\mathsf{T}\mathbf{\Sigma}_2\boldsymbol{a}^*=0 \tag{14}$$

where we have exploited in the last implication the fact that $\mathbf{\Sigma}_1\boldsymbol{a}^*=\frac{\sigma_1^2}{\sigma_2^2}\mathbf{\Sigma}_2\boldsymbol{a}^*$, as follows from (12). Let us denote $s_i^2 := \boldsymbol{v}^\mathsf{T}\mathbf{\Sigma}_i\boldsymbol{v}$ the variance of class i after projection onto $\boldsymbol{v}$, $i=1,2$. It follows that

$$\boldsymbol{v}^\mathsf{T}\boldsymbol{L}(\boldsymbol{a}^*,\lambda^*)\boldsymbol{v}=\eta\,\sigma_2^2 s_2^2\left(\sigma_1^2-\sigma_2^2\right)\left(R(\boldsymbol{v})-R(\boldsymbol{a}^*)\right) \tag{15}$$

where $\eta\overset{\text{def}}{=}4\pi_1\pi_2/(\pi_1 s_1^2+\pi_2 s_2^2)>0$ and we have recognized that s_1^2/s_2^2 is, by definition, equal to $R(\boldsymbol{v})$. To check C2, we distinguish the following cases:

Case 1: $\boldsymbol{a}^*=\boldsymbol{a}_1$, the dominant eigenvector of the GEVD problem defining CSP solution (3). This vector is also the maximizer of the CSP criterion $R(\cdot)$ defined in eqn. (2), as recalled in Sec. II, and therefore $R(\boldsymbol{v})<R(\boldsymbol{a}_1)$, $\forall\boldsymbol{v}\in T(\boldsymbol{a}_1)$. Now, if $R(\boldsymbol{a}_1)>1$, then $\sigma_1^2>\sigma_2^2$, and expression (15) is negative. In other words, $\boldsymbol{a}_1$ is a maximizer of the kurtosis (5b). Otherwise, if $R(\boldsymbol{a}_1)<1$, then $\sigma_1^2<\sigma_2^2$ and expression (15) is positive, so that $\boldsymbol{a}_1$ defines a minimum of the kurtosis.

Case 2: $\boldsymbol{a}^*=\boldsymbol{a}_p$, the least significant eigenvector of GEVD problem (3). As seen in Sec. II, this vector is also the minimizer of $R(\cdot)$, so that $R(\boldsymbol{v})>R(\boldsymbol{a}_p)$, $\forall\boldsymbol{v}\in T(\boldsymbol{a}_p)$. Using the same reasoning as in the above case, $\boldsymbol{a}_p$ is a minimizer of the kurtosis criterion if $R(\boldsymbol{a}_p)>1$, whereas it defines a maximum if $R(\boldsymbol{a}_p)<1$.

Case 3: $\boldsymbol{a}^*=\boldsymbol{a}_i$, $1<i<p$, any of the remaining eigenvectors of GEVD problem (3). Here we exploit the fact that the generalized eigenvectors enjoy an orthogonality property with respect to covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, i.e., $\boldsymbol{a}_i^\mathsf{T}\mathbf{\Sigma}_k\boldsymbol{a}_j=0$, for all $1\le i\ne j\le p$, $k=1,2$. From (14), it follows that both $\boldsymbol{a}_1$ and $\boldsymbol{a}_p$ belong to $T(\boldsymbol{a}_i)$. Hence, the sign of expression (15) when $\boldsymbol{v}=\boldsymbol{a}_1$ is different from that when $\boldsymbol{v}=\boldsymbol{a}_p$, thus defining a saddle point of the kurtosis criterion (5b).

In the light of the above cases, we realize that only the maximizer $\boldsymbol{a}_1$ and/or the minimizer $\boldsymbol{a}_p$ of CSP criterion (2) maximize the kurtosis. More precisely, the maximizer of kurtosis is $\boldsymbol{a}_1$ if $R(\boldsymbol{a}_1)>1$ and $\boldsymbol{a}_p$ if $R(\boldsymbol{a}_p)<1$. Because, by definition, $R(\boldsymbol{a}_1)>R(\boldsymbol{a}_p)$ (Sec. II), we can never have $R(\boldsymbol{a}_1)<1$ and $R(\boldsymbol{a}_p)>1$ simultaneously, the only possibility for neither point to maximize kurtosis. To conclude the proof of Theorem 1, it remains to study solution S2 given in eqn. (13). This solution corresponds to the case where the mixture of projected classes simplifies into a single Gaussian distribution, with $\kappa_Y=3$, and turns out to be a global minimizer of kurtosis, as follows from next lemma:

Lemma 1: $\kappa_Y(\boldsymbol{a})\ge 3$, $\forall\boldsymbol{a}\in\mathbb{R}^p$.

Proof: The kurtosis (5b) can be expressed as $\kappa(\boldsymbol{a})=\frac{3\|\boldsymbol{b}\|^2\|\boldsymbol{c}\|^2}{(\boldsymbol{b}\cdot\boldsymbol{c})^2}$, with $\boldsymbol{b}\cdot\boldsymbol{c}\overset{\text{def}}{=}\boldsymbol{b}^\mathsf{T}\boldsymbol{D}\boldsymbol{c}$, $\boldsymbol{b}=[\sigma_1^2,\sigma_2^2]^\mathsf{T}$, $\boldsymbol{c}=[1,1]^\mathsf{T}$, $\boldsymbol{D}=\operatorname{diag}(\pi_1,\pi_2)$. By the Cauchy-Schwarz inequality, $|\boldsymbol{b}\cdot\boldsymbol{c}|<\|\boldsymbol{b}\|\|\boldsymbol{c}\|$, and then $\kappa_Y(\boldsymbol{a})\ge 3$, with equality if and only $\boldsymbol{b}=\ell\boldsymbol{c}$, $\ell\in\mathbb{R}\setminus\{0\}$, implying $\sigma_1^2=\sigma_2^2$.

Appendix B
Computing Several Projection Directions

The whole set of eigenvectors of (12) can be computed as follows. It can be always assumed that $\mathrm{E}\{XX^\mathsf{T}\}=\boldsymbol{I}$, where the $\boldsymbol{I}$ is the identity matrix, under the assumption that the data has been whitened or sphered in a pre-processing step. Since in general the covariance matrix of the data is given by $\mathrm{E}\{XX^\mathsf{T}\}=\pi_1\mathbf{\Sigma}_1+\pi_2\mathbf{\Sigma}_2$, whitening imposes that $\pi_2\mathbf{\Sigma}_2=\boldsymbol{I}-\pi_1\mathbf{\Sigma}_1$. Substituting in (3), we get a usual eigenvalue problem $\mathbf{\Sigma}_1\boldsymbol{a}=\delta\boldsymbol{a}$, where $\delta=R(\boldsymbol{a})/(\pi_1 R(\boldsymbol{a})+\pi_2)$. Finally, as the eigenvectors of a Hermitian matrix are orthogonal, they can be determined by maximizing the kurtosis under the constraint that the direction obtained in the nth step is orthogonal to the previously computed directions. A possible implementation is presented as pseudo-code in Algorithm 1, where μ is the step size and the gradient $\nabla\kappa_Y(\boldsymbol{a})$ can be easily obtained from eqn. (5a), i.e., $\kappa_Y(\boldsymbol{a}):=\mathrm{E}\{Y^4\}/\mathrm{E}\{Y^2\}^2$, where $Y=\boldsymbol{a}^\mathsf{T}X$, as

$$\begin{aligned}\nabla\kappa_Y(\boldsymbol{a}) &= \frac{\partial\kappa_Y(\boldsymbol{a})}{\partial\boldsymbol{a}}\\ &= \frac{4}{\mathrm{E}\{Y^2\}^2}\left(\mathrm{E}\{XY^3\}-\frac{\mathrm{E}\{XY\}\mathrm{E}\{Y^4\}}{\mathrm{E}\{Y^2\}}\right).\end{aligned} \tag{16}$$

Observe that this expression can be always evaluated without any knowledge of the data labels by just replacing the expectations with their sample average estimates. For example, note that [24], [25] implement this pseudo-code in Matlab. Furthermore, [24], [25] seek the optimal value of μ using an algebraic line search approach.

References

[1] Y. Han and H. Bin, “Brain-Computer Interfaces Using Sensorimotor Rhythms: Current State and Future Perspectives”, *IEEE Transactions on Biomedical Engineering*, Vol. 61, No. 5., pp. 1425-1435, 2014.

Algorithm 1 Compute orthogonal directions maximizing the kurtosis

```
Center and whiten the data,
    $X \leftarrow \Sigma^{-1/2}\left(X - \bar{X}\right)$,
    where $\bar{X}$ is the mean and $\Sigma$ the covariance matrix of $X$
Choose randomly the initial projection vectors $\mathbf{a}_1, \dots, \mathbf{a}_p$.
repeat
    for $k = 1$ to $p$ do
        $\boldsymbol{a}_k \leftarrow \boldsymbol{a}_k + \mu \nabla \kappa_Y(\boldsymbol{a}_k)$ {Gradient-ascend update rule, see eqn. (16)}
        $\boldsymbol{a}_k \leftarrow \boldsymbol{a}_k - \sum_{n=1}^{k-1} \left(\boldsymbol{a}_k^{\intercal} \boldsymbol{a}_n\right) \boldsymbol{a}_n$ {Gram-Schmidt orthogonalization}
        $\boldsymbol{a}_k \leftarrow \frac{\boldsymbol{a}_k}{\|\boldsymbol{a}_k\|_2}$ {Normalization}
    end for
until convergence
```

[2] F. Lotte, "A Tutorial on EEG Signal Processing Techniques for Mental State Recognition in Brain-Computer Interfaces", E.R. Miranda (Ed.), *Guide to Brain-Computer Music Interfacing*, Springer, 2014.

[3] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy and F. Yger, "A Review of Classification Algorithms for EEG-Based Brain-Computer Interfaces: a 10 Year Update", *Journal of Neural Engineering*, Vol. 15, No. 3, 031005, 2018.

[4] H. Wang, "Harmonic Mean of Kullback-Leibler Divergences for Optimizing Multi-Class EEG Spatio-Temporal Filters", *Neural Processing Letters*, Vol. 36, No. 2, pp. 161-171, 2012.

[5] A. S. Aghaei, M. S. Mahanta, K. N. Plataniotis, "Separable Common Spatio-Spectral Patterns for Motor Imagery BCI Systems", *IEEE Transactions on Biomedical Engineering*, Vol. 63, No. 1, pp. 15-29, 2016.

[6] T. Jiang et al, "Characterization and Decoding the Spatial Patterns of Hand Extension/Flexion using High-Density ECoG", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 25, No. 4, pp. 370-379, 2017.

[7] R. Martín-Clemente, J. Olias, D. B. Thiyam, A. Cichocki and S. Cruces, "Information Theoretic Approaches for Motor-Imagery BCI Systems: Review and Experimental Comparison" , *Entropy*, Vol. 20, No. 1, 2018, https://doi.org/10.3390/e20010007

[8] M. Schultze-Kraft, S. Dähne, M., Gugler, G. Curio and B. Blankertz, "Unsupervised classification of operator workload from brain signals", *Journal of Neural Engineering*, Vol. 13, No. 2, 2016, https://doi.org/10.1088/1741-2560/13/3/036008

[9] W. Samek, C. Vidaurre, K.R. Müller and M. Kawanabe, "Stationary Common Spatial Patterns for Brain–Computer Interfacing", *Journal of neural engineering*, Vol. 9, No. 2, 2012.

[10] G. W. Steward, *Matrix Algorithms. Vol. II: Eigensystems*, Siam, 2001.

[11] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe and K. R. Müller, "Optimizing Spatial Filters for Robust EEG Single-Trial Analysis", *IEEE Signal Processing Magazine*, Vol. 25, No. 1, pp. 41–56, 2008.

[12] P. Westfall, "Kurtosis as Peakedness, 1905 – 2014. R.I.P", *The American statistician*, Vol. 68, No. 3, pp. 191-195, 2014.

[13] N. Delfosse and P. Loubaton, "Adaptive Blind Separation of Independent Sources: a Deflation Approach", *Signal Processing*, Vol. 45, No. 1, pp. 59-83, 1995.

[14] R. Martín-Clemente, V. Zarzoso, "On the link between L1-PCA and ICA", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 3, pp. 515-528, 2017, https://doi.org/10.1109/TPAMI.2016.2557797

[15] D. Peña and F. Prieto, "Cluster Identification Using Projections", *Journal of the American Statistical Association*, Vol. 96, No. 456, pp. 1433-1445, 2001.

[16] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, 2004.

[17] T. Eltoft, Taesu Kim and Te-Won Lee, "On the Multivariate Laplace Distribution", *IEEE Signal Processing Letters*, Vol. 13, No. 5, pp. 300-303, 2006.

[18] S-H. Park, D. Lee and S-G. Lee, "Filter Bank Regularized Common Spatial Pattern Ensemble for Small Sample Motor Imagery Classification", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 26, No. 2, pp. 498-505, 2018.

[19] H. Higashi and T. Tanaka. "Simultaneous Design of FIR Filter Banks and Spatial Patterns for EEG Signal Classification", *IEEE Transactions on Biomedical Engineering*, Vol. 60, No. 4, pp. 1100-1110, 2013.

[20] I. Daly, R. Scherer, M. Billinger and G. Müller-Putz, "FORCe: Fully Online and Automated Artifact Removal for Brain-Computer Interfacing", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 23, No. 5, pp. 725-736, 2015.

[21] C. Brunner *et al*, Institute for Knowledge Discovery, Graz University of Technology. Dataset available at http://www.bbci.de/competition/iv/.

[22] C. D. Tesche, M. A. Uusitalo, R. J. Ilmoniemi, M. Huotilainen, M. Kajola, and O. Salonen, "Signal-Space Projections of MEG Data Characterize Both Distributed and Well-Localized Neuronal Sources", *Electroencephalography and Clinical Neurophysiology*, Vol. 95, pp. 189-200, 1995.

[23] M. A. Uusitalo and R. J. Ilmoniemi, "Signal-Space Projection Method for Separating MEG or EEG into Components", *Medical & Biological Engineering & Computing*, Vol. 35, pp. 135-40, 1997.

[24] V. Zarzoso and P. Comon, "Robust Independent Component Analysis by Iterative Maximization of the Kurtosis Contrast with Algebraic Optimal Step Size", *IEEE Transactions on Neural Networks* , Vol. 21, No. 2, pp. 248–261, 2010.

[25] http://www.i3s.unice.fr/~zarzoso/robustica.html (free `MATLAB` code for the kurtosis maximization algorithm proposed in [24]).

[26] J. Olias, R. Martín-Clemente, M. A. Sarmiento-Vega, S. Cruces, "EEG Signal Processing in MI-BCI Applications with Improved Covariance Matrix Estimators", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, https://doi.org/10.1109/TNSRE.2019.2905894, 2019.

[27] H. Ramoser, J. Müler-Gerking and G. Pfurtscheller, "Optimal Spatial Filtering of Single Trial EEG During Imagined Hand Movement", *IEEE Transactions on Rehabilitation Engineering*, Vol. 8, No. 4, pp. 441-446, 2000.

[28] J. R. Schott, " Testing for elliptical symmetry in covariance-matrix-based analyses", *Statistics & Probability Letters*, Vol. 60, No.e 4, pp. 395-404, 2002.

[29] T. K. Moon and W. C. Stirling, "Mathematical Methods and Algorithms for Signal Processing", Prentice Hall, 2000.

[30] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning. Data Mining, Inference, and Prediction", Springer, 2009.

[31] L. van der Maaten and G.E. Hinton, Visualizing High-Dimensional Data Using t-SNE, *Journal of Machine Learning Research*, Vol. 9, pp. 2579-2605, 2008.

[32] https://lvdmaaten.github.io/tsne/code/tSNE_matlab.zip (free `MATLAB` code for the algorithm proposed in [31]).

[33] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, 4th Edition, John Wiley & Sons, 2013.

Conclusions

Along this thesis we have studied the application of statistical signal processing and machine learning methods for improving the performance of existing Motor Imagery Brain Computer Interfaces. As we have commented previously, the elucidation of the movements intended by the subject is a process that is typically decomposed in the following main stages: signal preprocessing, feature extraction (usually involving a dimensionality reduction) and classification.

In this work, we have tried to put forward some novel proposals for each of the previous stages. In some cases, these humble proposals seem to lead to certain advances compared to the state-of-the-art methods in this field. In the following, we summarize the main contributions of this thesis:

- The preprocessing of the EEG observations usually relies on some classical techniques that have been revealed to be very effective in practice. These include the use of referential measurements, its bandpass filtering and a suitable normalization of the covariance matrices of the resulting observations. Any advances that could be obtained in this initial stage of the process have a great potential, in the sense that they can be immediately applied to the plethora of existing techniques for solving the MI-BCI problem.

  In this regard, one of our contributions has focused on tackling the non-stationarity of the EEG measurements, which can happen at different levels. The classical inter-subject and inter-session variabilities have been frequently addressed in the literature. On the contrary, in this thesis, we have shifted our attention to the less studied variabilities that happen between trials and also within samples of the same trial. The signals generated by the brain are non-stationary in power at the trial and sample levels. Our hypothesis was that this power variability hinders the correct estimation of the covariance matrices of the trials. Therefore, we proposed an overcomplete model of the EEG observations (that regards the noise as an additional source in the mixture) and developed an iterative amplitude correction method for the measurements that aims to implicitly equalize the power of the effective EEG sources, i.e., of those components of the sources that effectively contributes to the observations. Furthermore, this normalization also attenuates heavily those samples that are less related to the rest of the data, minimizing the negative effects of the artifacts and making the whole system more robust.

This normalization has several advantages: it is free of hyper-parameters to tune and can be used to obtain improved covariance matrix estimates not only in training, but also during the test trials. When the method is applied to a unique training trial, it is reduced to Tyler's procedure to obtain an improved scatter matrix estimator. Still, the performance improves further when considering the whole set of training trials, a scenario of heterogeneous observations that seems to be covered by our proposal while it is beyond the hypotheses required by the derivation of Tyler's method.

Besides, we have also shown that standard power normalization of the observations, which is widely used in the preprocessing of the EEG data for MI-BCI, coincides (for a suitable isotropic initialization) with the result obtained in the initial iteration of the proposed procedure. In general, this leads to a useful correction of the observations, although it is sub-optimal compared to the improvement in the covariance matrix estimates that can be obtained by just a few more iterations (since the convergence of the proposed procedure is quite fast).

The experimental results obtained with synthetic and real data have corroborated our hypothesis that the correction of the power variability of the effective EEG source signals, in general, leads to improved covariance matrix estimates. This, in turn, allows to obtain transversal improvements in accuracy (and in most cases clearly significant) when the proposed normalization is combined with the existing MI-BCI classification algorithms.

- The second stage considered in a MI-BCI system, the feature extraction stage, is one of the most active research areas in the MI-BCI field. We have shown that there are two differentiated lines of investigation: the algorithms based in Riemann's metric and the ones based on CSP. Although both lines can be applied to the same problems, the Riemannian metric does not need to reduce dimensionality, while CSP is a dimensionality reduction technique.

  CSP works with two classes by projecting linearly the data obtained from EEG samples onto directions where the variance of the projected data points is significantly higher for one class than for the other. The projected data variances can then be used as features for classification. Although we have studied many of its variations, we have seen that, in general, the results obtained for real data from BCI competitions reveal a similar performance for many variations of CSP, in terms of accuracy.

  However, CSP and its variants depend on the availability of correctly labeled data, making them fall in the category of supervised algorithms. In this regard, one of our contributions has proven that CSP can also be performed in an unsupervised or blind way, that is, using unlabeled data. For this, we assume that the classes follow an elliptically symmetric distribution (which includes Gaussian distributions as a particular case). The proposed algorithm is based on the maximization of the kurtosis using a gradient descend approach. Among the benefits of using an unsupervised algorithm, we can highlight that it spares the need for training labels and it is not affected by incorrect labeled data. Also, it may be used when training labels are not available or may be uncertain. We have proved that with the maximization of the kurtosis we obtain similar results as the ones obtained with CSP, using both real and simulated data.

- Regarding the last and final step of the system, the classification stage is based on algorithms widely used in the field of Machine Learning, such as SVM or LR. However, one of the most used classifiers in the MI-BCI field is the LDA algorithm. It has been seen that this algorithm was greatly improved by the shrinkage method proposed by Ledoit and Wolf, which is known as sLDA.

  Following this line of research, we have experimentally observed that the shrinkage method developed by Chen et. al. combined with LDA yields better results, specially when we are working with very few training data. The reason why we obtain better results with this combination, which we have denominated gLDA, is due to the fact that both LDA and the shrinking method assume that the features follow Gaussian distributions.

  Our last contribution was to show that it is possible to classify imagery movements in almost a blind context thanks to the unsupervised CSP and the GMM algorithms.

List of Figures

List of Tables

Bibliography

[1] N. Waytowich. (2014) Bci control of a motorized wheelchair for disabled individuals using a calibrationless ssvep system. [Online]. Available: https://www.youtube.com/watch?v=qhK572LJhSc

[2] M. S. Bascil, A. Y. Tesneli, and F. Temurtas, "Spectral feature extraction of EEG signals and pattern recognition during mental tasks of 2-d cursor movements for BCI using SVM and ANN," *Australasian Physical & Engineering Sciences in Medicine*, vol. 39, no. 3, pp. 665–676, jul 2016.

[3] L. van Dokkum, T. Ward, and I. Laffont, "Brain computer interfaces for neurorehabilitation – its current status as a rehabilitation strategy post-stroke," *Annals of Physical and Rehabilitation Medicine*, vol. 58, no. 1, pp. 3–8, feb 2015.

[4] Tech@facebok, "Imagining a new interface: Hands-free communication without saying a word," 2019. [Online]. Available: https://tech.fb.com/imagining-a-new-interface-hands-free-communication-without-saying-a-word/

[5] J. J. Vidal, "Toward direct brain-computer communication," *Annual Review of Biophysics and Bioengineering*, vol. 2, no. 1, pp. 157–180, 1973, pMID: 4583653. [Online]. Available: https://doi.org/10.1146/annurev.bb.02.060173.001105

[6] E. Hortal, E. Iáñez, A. Úbeda, C. Perez-Vidal, and J. M. Azorín, "Combining a brain–machine interface and an electrooculography interface to perform pick and place tasks with a robotic arm," *Robotics and Autonomous Systems*, vol. 72, pp. 181 – 188, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889015001244

[7] Y. He, D. Eguren, J. M. Azorín, R. G. Grossman, T. P. Luu, and J. L. Contreras-Vidal, "Brain–machine interfaces for controlling lower-limb powered robotic systems," *Journal of Neural Engineering*, vol. 15, no. 2, p. 021004, feb 2018. [Online]. Available: https://doi.org/10.1088%2F1741-2552%2Faaa8c0

[8] V. Guy, M.-H. Soriani, M. Bruno, T. Papadopoulo, C. Desnuelle, and M. Clerc, “Brain computer interface with the p300 speller: Usability for disabled people with amyotrophic lateral sclerosis,” *Annals of Physical and Rehabilitation Medicine*, vol. 61, no. 1, pp. 5 – 11, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877065717304104

[9] M. Ahn, M. Lee, J. Choi, and S. C. Jun, “A review of brain-computer interface games and an opinion survey from researchers, developers and users,” *Sensors (Basel, Switzerland)*, vol. 14, no. 8, pp. 14 601–14 633, Aug 2014, 25116904[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25116904

[10] C. Guger, A. Brendan, and E. C. Leuthardt, *Brain-Computer Interface Research. A State-of-the-Art Summary -2 Interface Researchs*, 2014.

[11] S. Nieuwenhuis, G. Aston-Jones, and J. D. Cohen, “Decision making, the p3, and the locus coeruleus–norepinephrine system.” *Psychological Bulletin*, vol. 131, no. 4, pp. 510–532, 2005.

[12] J. Fell, T. Dietl, T. Grunwald, M. Kurthen, P. Klaver, P. Trautner, C. Schaller, C. E. Elger, and G. Fernández, “Neural bases of cognitive ERPs: More than phase reset,” *Journal of Cognitive Neuroscience*, vol. 16, no. 9, pp. 1595–1604, nov 2004.

[13] L. A. Farwell and E. Donchin, “Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials,” *Electroencephalography and Clinical Neurophysiology*, vol. 70, pp. 510–523, 1988.

[14] R. Fazel-Rezai and W. Ahmad, “P300-based brain-computer interface paradigm design,” in *Recent Advances in Brain-Computer Interface Systems*, R. Fazel-Rezai, Ed. Rijeka: IntechOpen, 2011, ch. 4. [Online]. Available: https://doi.org/10.5772/14858

[15] K. Nakayama and M. Mackeben, “Steady state visual evoked potentials in the alert primate,” *Vision Research*, vol. 22, no. 10, p. 1261–1271, 1982.

[16] C. Herrmann, “Human eeg responses to 1-100 hz flicker: Resonance phenomena in visual cortex and their potential correlation to cognitive phenomena,” *Experimental Brain Research*, vol. 137, no. 3-4, pp. 346–353, 2001, cited By 496. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035086024&doi=10.1007%2fs002210100682&partnerID=40&md5=e8e1549f048ab276483ff72311a130ec

[17] M. Cheng, X. Gao, S. Gao, and D. Xu, “Design and implementation of a brain-computer interface with high transfer rates,” *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 10, pp. 1181–1186, 2002, cited By 485. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0036785277&doi=10.1109%2fTBME.2002.803536&partnerID=40&md5=f80f7389adaf00c681955796b63aa3e4

[18] A. González-Mendoza, J. L. Pérez-Benítez, J. A. Pérez-Benítez, and J. H. Espina-Hernández, “Brain computer interface based on ssvep for controlling a remote control car,” in *2015 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Feb 2015, pp. 93–97.

[19] M. Wang, R. Li, R. Zhang, G. Li, and D. Zhang, “A wearable ssvep-based bci system for quadcopter control using head-mounted device,” *IEEE Access*, vol. 6, pp. 26 789–26 798, 2018.

[20] P. Stawicki, F. Gembler, C. Y. Chan, M. Benda, A. Rezeika, A. Saboor, R. Grichnik, and I. Volosyak, “Ssvep-based bci in virtual reality - control of a vacuum cleaner robot,” in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, oct 2019, pp. 534–537, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062234537&doi=10.1109%2fSMC.2018.00749&partnerID=40&md5=887c6cdf79c1a8d4eee8b247aa83b939

[21] C. Bledowski, D. Prvulovic, K. Hoechstetter, M. Scherg, M. Wibral, R. Goebel, and D. E. J. Linden, “Localizing p300 generators in visual target and distractor processing: A combined event-related potential and functional magnetic resonance imaging study,” *Journal of Neuroscience*, vol. 24, no. 42, pp. 9353–9360, 2004. [Online]. Available: http://www.jneurosci.org/content/24/42/9353

[22] D. B. Egolf, *uman Communication and the Brain : Building the Foundation for the Field of Neurocommunication*, L. Books, Ed., 2012.

[23] S. Sanei and J. Chambers, *EEG Signal Processing*. John Wiley & Sons Ltd,, sep 2007.

[24] S. Herculano-Houzel, “The human brain in numbers: A linearly scaled-up primate brain,” *Frontiers in human neuroscience*, vol. 3, 2009.

[25] S. Kumar, F. Yger, and F. Lotte, “Towards adaptive classification using riemannian geometry approaches in brain-computer interfaces,” in *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, feb 2019.

[26] J. Solé-Casals and F.-B. Vialatte, “Towards semi-automatic artifact rejection for the improvement of alzheimer’s disease screening from EEG signals,” *Sensors*, vol. 15, no. 8, pp. 17 963–17 976, jul 2015.

[27] A. Jafarifarmand and M. Badamchizadeh, “Eeg artifacts handling in a real practical brain-computer interface controlled vehicle,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1200–1208, 2019, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067233577&doi=10.1109%2fTNSRE.2019.2915801&partnerID=40&md5=e92e8926ef52521713e53bf35d875488

[28] J. W. Kam, S. Griffin, A. Shen, S. Patel, H. Hinrichs, H.-J. Heinze, L. Y. Deouell, and R. T. Knight, “Systematic comparison between a wireless eeg

system with dry electrodes and a wired eeg system with wet electrodes," *NeuroImage*, vol. 184, pp. 119 – 129, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811918307961

[29] C. Tabernig, L. Carrere, C. Lopez, and C. Ballario, "Eeg event-related desynchronization of patients with stroke during motor imagery of hand movement," in *Journal of Physics: Conference Series (2016) 705(1)*, vol. 705, no. 1. IOP Publishing, apr 2016, p. 012059, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84978127759&doi=10.1088%2f1742-6596%2f705%2f1%2f012059&partnerID=40&md5=f7487cae5092aa208b63136c9ef23bc4

[30] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000, cited By 1265. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0034470472&doi=10.1109%2f86.895946&partnerID=40&md5=9f4bc0e02f2a4d0c1b3d6afd14e09fc0

[31] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller. (2008) Bci competition 2008 – graz data set a. [Online]. Available: http://www.bbci.de/competition/iv/desc_2a.pdf

[32] Y. Höller, J. Bergmann, A. Thomschewski, M. Kronbichler, P. Höller, J. S. Crone, E. V. Schmid, K. Butz, R. Nardone, and E. Trinka, "Comparison of EEG-features and classification methods for motor imagery in patients with disorders of consciousness," *PLoS ONE*, vol. 8, no. 11, p. e80479, nov 2013.

[33] F. Lotte, "A tutorial on EEG signal-processing techniques for mental-state recognition in brain–computer interfaces," in *Guide to Brain-Computer Music Interfacing*. Springer London, 2014, pp. 133–161.

[34] G. Pfurtscheller, "Functional brain imaging based on ERD/ERS," *Vision Research*, vol. 41, no. 10-11, pp. 1257–1260, may 2001.

[35] Y. Renard, F. Lotte, G. Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand, and A. Lécuyer, "OpenViBE: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 19, no. 1, pp. 35–53, feb 2010.

[36] C. S. Nam, A. Nijholt, and F. Lotte, *A Step-by-Step Tutorial for a Motor Imagery–Based BCI.* CRC Press, jan 2018, p. 746. [Online]. Available: https://hal.inria.fr/hal-01655819

[37] G. Pfurtscheller and F. L. da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, nov 1999.

[38] W. Samek, M. Kawanabe, and K.-R. Muller, “Divergence-based framework for common spatial patterns algorithms,” *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[39] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, “Filter Bank Common Spatial Pattern (FBCSP) in brain-computer interface,” in *Proceedings of the International Joint Conference on Neural Networks*, 2008, pp. 2390–2397. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-56349090127&doi=10.1109%2fIJCNN.2008.4634130&partnerID=40&md5=b6a886b233ce9ba1468241b210d88fcd

[40] H. Higashi and T. Tanaka, “Simultaneous design of FIR filter banks and spatial patterns for EEG signal classification,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1100–1110, apr 2013.

[41] C. A. Kothe and S. Makeig, “BCILAB: a platform for brain–computer interface development,” *Journal of Neural Engineering*, vol. 10, no. 5, p. 056014, aug 2013.

[42] X. Jiang, G.-B. Bian, and Z. Tian, “Removal of artifacts from EEG signals: A review,” *Sensors*, vol. 19, no. 5, p. 987, feb 2019.

[43] A. Delorme and S. Makeig, “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, mar 2004.

[44] F.-B. Vialatte, J. Solé-Casals, M. Maurice, C. Latchoumane, N. Hudson, S. Wimalaratna, J. Jeong, and A. Cichocki, “Improving the quality of EEG data in patients with alzheimer’s disease using ICA,” in *Advances in Neuro-Information Processing*. Springer Berlin Heidelberg, 2009, pp. 979–986.

[45] L. F. Nicolas-Alonso and J. Gomez-Gil, “Brain computer interfaces, a review,” *Sensors*, vol. 12, no. 2, pp. 1211–1279, jan 2012.

[46] J. Solé-Casals, C. F. Caiafa, Q. Zhao, and A. Cichocki, “Brain-computer interface with corrupted EEG data: a tensor completion approach,” *Cognitive Computation*, vol. 10, no. 6, pp. 1062–1074, jul 2018.

[47] A. Barachant and G. Titericz. (2016) Kaggle melbourne university aes/mathworks/nih seizure prediction challenge - winning solution. [Online]. Available: https://github.com/alexandrebarachant/kaggle-seizure-prediction-challenge-2016

[48] Q. Barthelemy, L. Mayaud, D. Ojeda, and M. Congedo, “The riemannian potato field: A tool for online signal quality index of EEG,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 2, pp. 244–255, feb 2019.

[49] I. Rejer and P. Gorski, “Independent component analysis in a motor imagery brain computer interface,” in *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*. IEEE, jul 2017.

[50] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*, I. S. D. Academic Press Professional and C. A. USA ©1990, Eds. Elsevier Science Publishing Co Inc, 1990. [Online]. Available: https://www.ebook.de/de/product/3598641/keinosuke_fukunaga_introduction_to_statistical_pattern_recognition.html

[51] Z. J. K. S. L. Z. Zhou, "Spatial patterns underlying population differences in the background eeg," *Brain Topography*, vol. 2, no. 4, pp. 275–284, 1990.

[52] R. Bhatia, *Matrix Analysis*. Springer New York, 1996. [Online]. Available: https://www.ebook.de/de/product/3713864/rajendra_bhatia_matrix_analysis.html

[53] D. Thiyam, S. Cruces, J. Olias, and A. Cichocki, "Optimization of alpha-beta log-det divergences and their application in the spatial filtering of two class motor imagery movements," *Entropy*, vol. 19, no. 3, p. 89, feb 2017.

[54] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 8, pp. 1991–2000, aug 2008.

[55] A. Jafarifarmand and M. A. Badamchizadeh, "Real-time multiclass motor imagery brain-computer interface by modified common spatial patterns and adaptive neuro-fuzzy classifier," *Biomedical Signal Processing and Control*, vol. 57, p. 101749, mar 2020.

[56] X. Yong, R. K. Ward, and G. E. Birch, "Robust common spatial patterns for EEG signal preprocessing," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, aug 2008.

[57] M. Kawanabe and C. Vidaurre, "Improving BCI performance by modified common spatial patterns with robustly averaged covariance matrices," in *IFMBE Proceedings*. Springer Berlin Heidelberg, 2009, pp. 279–282.

[58] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 610–619, apr 2013.

[59] W. Samek, F. C. Meinecke, and K.-R. Muller, "Transferring subspaces between subjects in brain–computer interfacing," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, aug 2013.

[60] W. Samek, D. Blythe, K.-R. Müller, and M. Kawanabe, "Robust spatial filtering with beta divergence," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 1007–1015. [Online]. Available: http://papers.nips.cc/paper/4922-robust-spatial-filtering-with-beta-divergence.pdf

[61] A. Cichocki, S. Cruces, and S. ichi Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, jan 2011.

[62] W. Förstner and B. Moonen, "A metric for covariance matrices," in *Geodesy-The Challenge of the 3rd Millennium*. Springer Berlin Heidelberg, 2003, pp. 299–309.

[63] A. Barachant, S. Bon, M. Congedo, and C. Jutten, "Common spatial pattern revisited by riemannian geometry," in *2010 IEEE International Workshop on Multimedia Signal Processing*. IEEE, oct 2010.

[64] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, apr 2012.

[65] P. T. Fletcher and S. Joshi, "Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004, pp. 87–98.

[66] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: A review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1753–1762, oct 2017.

[67] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, feb 2011.

[68] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain–computer interfacing," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026013, feb 2012.

[69] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2009.

[70] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 7, pp. 179–188, 1936.

[71] C. Cortes and V. Vapnik, *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[72] C.-C. Chang and C.-J. Lin, "LIBSVM," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, apr 2011.

[73] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, apr 2018.

[74] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a riemannian-based kernel for BCI applications," *Neurocomputing*, vol. 112, pp. 172–178, jul 2013.

[75] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt,

and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[76] E. M. Bruce Bowerman, *Regression Analysis: Unified Concepts, Practical Applications, Computer Implementation*. LIGHTNING SOURCE INC, 2014. [Online]. Available: https://www.ebook.de/de/product/24905349/bruce_bowerman_emily_murphree_regression_analysis_unified_concepts_practical_applications_computer_implementation.html

[77] A. Sarmiento, I. Fondón, I. Durán-Díaz, and S. Cruces, "Centroid-based clustering with $\alpha\beta$-divergences," *Entropy*, vol. 21, no. 2, p. 196, feb 2019.

[78] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, mar 1982.

[79] O. C. Carrasco, "Gaussian mixture models explained," Towards Data Science, Tech. Rep., 2019. [Online]. Available: https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

[80] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, aug 2011.

[81] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, feb 2004.

[82] B. Rajaratnam and D. Vincenzi, "A theoretical study of stein's covariance estimator," *Biometrika*, vol. 103, no. 3, pp. 653–666, sep 2016.

[83] L. R. Haff, "Empirical bayes estimation of the multivariate normal covariance matrix," *The Annals of Statistics*, vol. 8, no. 3, pp. 586–597, may 1980.

[84] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387–399, may 2011.

[85] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components — a tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, may 2011.

[86] O. Ledoit and M. N. Wolf, "Honey, i shrunk the sample covariance matrix," *SSRN Electronic Journal*, 2003.

[87] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5016–5029, oct 2010.

[88] Y. Chen, A. Wiesel, and A. O. Hero, “Robust shrinkage estimation of high-dimensional covariance matrices,” *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4097–4107, sep 2011.

[89] R. Martín-Clemente, J. Olias, D. Beeta, A. Cichocki, and S. Cruces, “Information Theoretic Approaches for Motor-Imagery BCI Systems: Review and Experimental Comparison,” *Entropy*, vol. 20, no. 7, 2018.

[90] R. A. Maronna, “Robust m-estimators of multivariate location and scatter,” *The Annals of Statistics*, vol. 4, no. 1, pp. 51–67, 1976. [Online]. Available: http://www.jstor.org/stable/2957994

[91] D. E. Tyler, “A distribution-free m-estimator of multivariate scatter,” *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987. [Online]. Available: http://www.jstor.org/stable/2241079

[92] M. Rangaswamy, “Statistical analysis of the nonhomogeneity detector for non-gaussian interference backgrounds,” *IEEE Transactions on Signal Processing*, vol. 53, no. 6, pp. 2101–2111, jun 2005.

[93] S. Chevallier, E. Kalunga, Q. Barthélemy, and F. Yger, *Brain–Computer Interfaces Handbook:Technological and Theoretical Advances*, 2018. [Online]. Available: https://hal.uvsq.fr/hal-01710089

[94] F. Yger, F. Lotte, and M. Sugiyama, “Averaging covariance matrices for EEG signal classification based on the CSP: An empirical study,” in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, aug 2015.

[95] W. Samek, S. Nakajima, M. Kawanabe, and K.-R. Müller, “On robust parameter estimation in brain–computer interfacing,” *Journal of Neural Engineering*, vol. 14, no. 6, p. 061001, nov 2017. [Online]. Available: https://doi.org/10.1088%2F1741-2552%2Faa8232

[96] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, oct 1998.

[97] R. Bhavsar, Y. Sun, N. Helian, N. Davey, D. Mayor, and T. Steffert, “The correlation between EEG signals as measured in different positions on scalp varying with distance,” *Procedia Computer Science*, vol. 123, pp. 92–97, 2018.

[98] L.-W. Ko, C.-T. Lin, Y.-C. Lu, H. Bustince, Y.-C. Chang, Y. Chang, J. Ferandez, Y.-K. Wang, J. A. Sanz, and G. P. Dimuro, “Multimodal fuzzy fusion for enhancing the motor-imagery-based brain computer interface,” *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 96–106, feb 2019.

[99] Y. Yang, S. Chevallier, J. Wiart, and I. Bloch, “Automatic selection of the number of spatial filters for motor-imagery bci,” in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, April 2012, pp. 109–114.

Glossary

AB-LD Alpha-Beta Log-Det divergence. 28

BCI Brain Computer Interface. 3

CSP Common Spatial Patterns. 22

CV Cross Validation. 11

ECoG Electrocorticography. 8

EEG Electroencephalogram. 7

ERD Event Related with the Desynchronization. 10

ERS Event Related with the Synchronization. 10

FBCSP Filter Bank Common Spatial Patterns. 18

GMM Gaussian Mixture Model. 40

ICA Independent Component Analysis. 19

ITFE Information Theory Feature Extraction. 25

JADE Joint Approximation Diagonalization of Eigen-matrices. 25

LDA Linear Discriminant Analysis. 33

LR Logistic Regression. 38

MI-BCI Motor Imagery Brain Computer Interface. 9

ML Maximum Likelihood. 45

MSE Minimum Square Error. 38

PCA Principal Component Analysis. 19

sLDA Shrinkage Linear Discriminant Analysis. 34

SNR Signal to Noise Ratio. 8

SSVEP Steady State Visual Evoked Potential. 5

SVD Singular Value Decomposition. 45

SVM Support Vector Machine. 33