entropy

MDPI

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, 41092 Seville, Spain; folias@us.es (J.O.); thiyamdeepa@gmail.com (D.B.T.)
2 Department of Sensor and Biomedical Technology, School of Electronics Engineering, VIT University, Vellore, Tamil Nadu 632014, India
3 Skolkovo Institute of Science and Technology (Skoltech), Moscow 143026, Russia; A.Cichocki@skoltech.ru or a.cichocki@riken.jp
4 Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
5 Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, 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: 2 January 2018

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.

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 [3–6]. 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 [1,4,7]. Among them, motor imagery (MI)-based BCI systems seem to be the most promising option [6,8–10]. 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 [11,12], 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 [13] 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 [14]. 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 [11,14]. Finally, imagery of tongue movements can be detected on the primary motor cortex and the premotor cortex [15]. 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 [16,17]. Overall, current BCI approaches achieve success rates of over 90%, although much depends on the person from whom the EEG data are recorded [14].

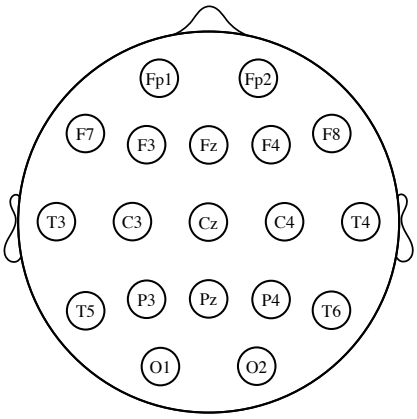


Figure 1. Electrode locations of the international 10–20 system for EEG recording. The letters "F", "T", "C", "P" and "O" stand for frontal, temporal, central, parietal and occipital lobes, respectively. Even numbers correspond to electrodes placed on the right hemisphere, whereas odd numbers refer to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

The common spatial patterns (CSP) method [4,18–22] 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 (AB-LD) 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 [4,18–22,32,33]. 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:

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

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 [12].

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:

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

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

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

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 $\boldsymbol{w}^{\top}\mathbf{\Sigma}_i\boldsymbol{w}$ 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; $\mathbf{Y}_i(k) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_i)$ is an s-dimensional column vector of latent variables (s has to be estimated from the data) and $\mathbf{N}_i(k) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Delta}_i)$ is a D-dimensional vector of noise, independent of the data. Here, the covariance matrices $\mathbf{\Gamma}_i$ and $\mathbf{\Delta}_i$ 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., $\mathbf{X}|c_k\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_k)$, the spatially-filtered data are also from a normal distribution, i.e., $\mathbf{Y}|c_k\sim\mathcal{N}(\mathbf{0},\bar{\boldsymbol{\Sigma}}_k)$, where:

$$\bar{\boldsymbol{\Sigma}}_k=\mathbf{W}^\top\boldsymbol{\Sigma}_k\mathbf{W}\in\mathbb{R}^{d\times d},$$

$k=1,2$. The KL divergence between two d-dimensional multivariate Gaussian densities $f_1=\mathcal{N}_1(\mathbf{0},\bar{\boldsymbol{\Sigma}}_1)$ and $f_2=\mathcal{N}_2(\mathbf{0},\bar{\boldsymbol{\Sigma}}_2)$, that is,

$$Div_{KL}\left(f_1\|f_2\right)=\int\mathcal{N}_1(\mathbf{0},\bar{\boldsymbol{\Sigma}}_1)\log\frac{\mathcal{N}_1(\mathbf{0},\bar{\boldsymbol{\Sigma}}_1)}{\mathcal{N}_2(\mathbf{0},\bar{\boldsymbol{\Sigma}}_2)}d\boldsymbol{y},$$

can be shown to be (after some algebra):

$$Div_{KL}\left(f_1\|f_2\right)=\frac{1}{2}\left[\log\frac{|\bar{\boldsymbol{\Sigma}}_2|}{|\bar{\boldsymbol{\Sigma}}_1|}-d+\operatorname{trace}(\bar{\boldsymbol{\Sigma}}_2)^{-1}(\bar{\boldsymbol{\Sigma}}_1)\right], \tag{10}$$

where $|\cdot|$ stands for "determinant". The symmetrized Kullback–Leibler (sKL) divergence between the probability distributions of the two classes is now defined as:

$$\begin{aligned} Div_{sKL}\left(f_1 \| f_2\right) &= Div_{KL}\left(f_1 \| f_2\right) + Div_{KL}\left(f_2 \| f_1\right) \\ &= \frac{1}{2}\,\mathrm{trace}\left((\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W})\right. \\ &\quad \left. +(\mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W})\right) - d, \end{aligned}$$

where we show again the explicit dependency on $\mathbf{W}$.

We can naturally extend this formula to define the equivalent sKL matrix divergence:

$$\begin{aligned} D_{sKL}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W} \| \mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}) = \; & \frac{1}{2}\,\mathrm{trace}\left((\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W})\right. \\ & \left. +(\mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W})\right) - d. \end{aligned} \tag{11}$$

It has been shown in [23] that the subspace of the filters that maximize the sKL matrix divergence,

$$\mathbf{W}_{sKL} = \arg\max_{\mathbf{W}} D_{sKL}(\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W} \| \mathbf{W}^\top \boldsymbol{\Sigma}_2 \mathbf{W}), \tag{12}$$

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:

$$\mathrm{span}(\mathbf{W}_{sKL}) = \mathrm{span}(\mathbf{W}_{CSP}), \tag{13}$$

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:

$$Div_{KL}\left(\mathcal{N}(0, \mathbf{W}^\top \boldsymbol{\Sigma}_{i,k} \mathbf{W}) \| \mathcal{N}(0, \mathbf{W}^\top \boldsymbol{\Sigma}_i \mathbf{W})\right) = \tfrac{1}{2}\left[\log \frac{|\mathbf{W}^\top \boldsymbol{\Sigma}_i \mathbf{W}|}{|\mathbf{W}^\top \boldsymbol{\Sigma}_{i,k} \mathbf{W}|} - d + \mathrm{trace}(\mathbf{W}^\top \boldsymbol{\Sigma}_i \mathbf{W})^{-1}(\mathbf{W}^\top \boldsymbol{\Sigma}_{i,k} \mathbf{W})\right]. \tag{16}$$

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 (AB-LD) divergence, 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].

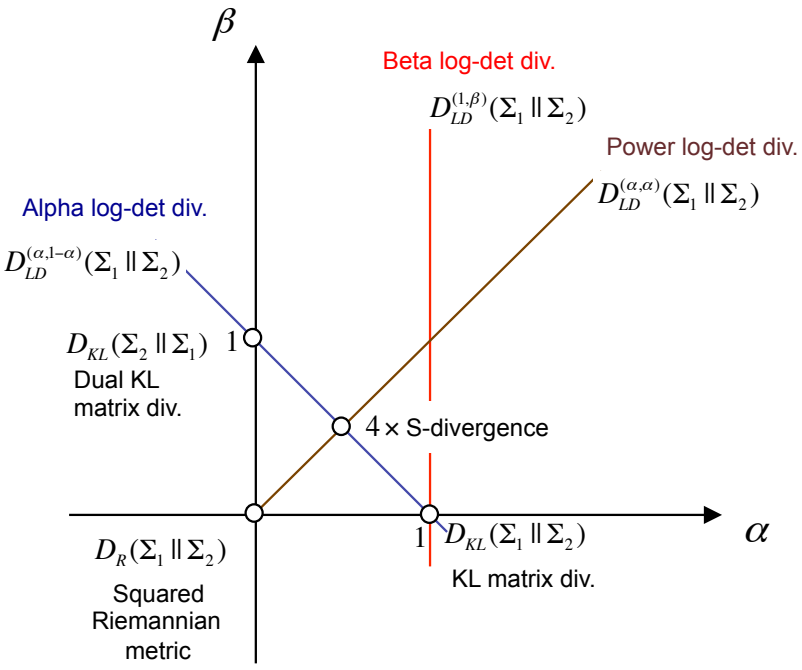


Figure 2. Illustration of the Alpha-Beta log-det divergence (AB-LD) divergence $D_{LD}^{(\alpha,\beta)}(\boldsymbol{\Sigma}_1 \| \boldsymbol{\Sigma}_2)$ in the (α, β)-plane. Note that the position of each divergence is specified by the value of the hyperparameters (α, β). This parameterization smoothly 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-LD 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 3.1–3.3 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.

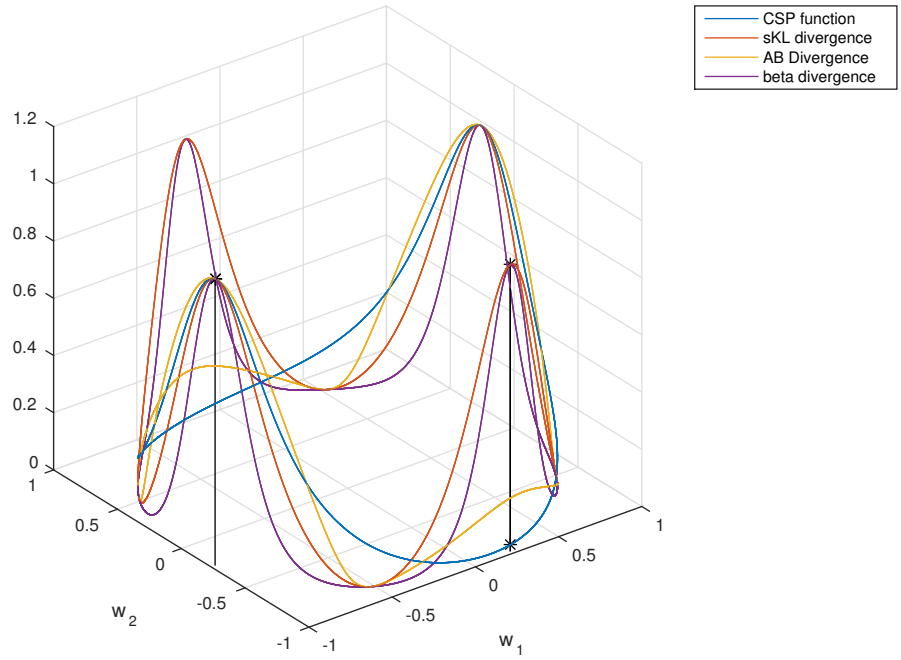


Figure 3. This figure shows the evolution of the common spatial patterns (CSP) criterion function (in blue line), the symmetrized Kullback–Leibler divergence (sKL) (in red line), the symmetrized beta divergence (in purple line) and the AB-LD divergence (in yellow line), all of them as a function of the 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:

$$\begin{aligned} D(\mathbf{W}^\top\boldsymbol{\Sigma}_1\mathbf{W}\|\mathbf{W}^\top\boldsymbol{\Sigma}_2\mathbf{W}) &= D(\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{R}}^\top\|\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{R}}^\top) \\ &= D(\mathbf{I}_d\mathbf{R}\tilde{\boldsymbol{\Sigma}}_1\mathbf{R}^\top\mathbf{I}_d\|\mathbf{I}_d\mathbf{R}\tilde{\boldsymbol{\Sigma}}_2\mathbf{R}^\top\mathbf{I}_d) \\ &\equiv J(\mathbf{R}) \end{aligned} \tag{25}$$

under the constraint that $\mathbf{R}$ is an orthogonal matrix, where $\tilde{\boldsymbol{\Sigma}}_i = \mathbf{P}\boldsymbol{\Sigma}_i\mathbf{P}^\top$.

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 $\bar{\mathbf{R}}\bar{\mathbf{R}}^\top = \mathbf{I} + o(\boldsymbol{\mu}^2)$. 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:

$$< \partial J(\mathbf{R})|\nabla J(\mathbf{R}) > \quad = \quad \frac{1}{2} < \nabla J(\mathbf{R})|\nabla J(\mathbf{R}) > \tag{31}$$

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 $\mathbf{R}$ 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, $\mathbf{R}$ 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 $\mathbf{R}$ given in (34) may be also considered as an update for $\mathbf{M}$ from the zero matrix to its actual value $\mathbf{M}_t$, the algorithm is as follows:

1. Start at the zero matrix $\mathbf{0}$.
2. Move from $\mathbf{0}$ 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 $\mathbf{M}$ 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:

$$\begin{aligned} \tilde{J}_{ITFE}(\boldsymbol{w}) \equiv & -\tfrac{1}{2}\sum_{k=1}^{n_c} P(c_k)\log_2\left(\boldsymbol{w}^\top \boldsymbol{\Sigma}_k \boldsymbol{w}\right) - \tfrac{3}{16}\left(\sum_{k=1}^{n_c} P(c_k)\left((\boldsymbol{w}^\top \boldsymbol{\Sigma}_k \boldsymbol{w})^2 - 1\right)\right)^2 \\ & \approx J_{ITFE}(\boldsymbol{w}), \end{aligned} \tag{42}$$

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 [7] 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 [10,65,66]. 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].

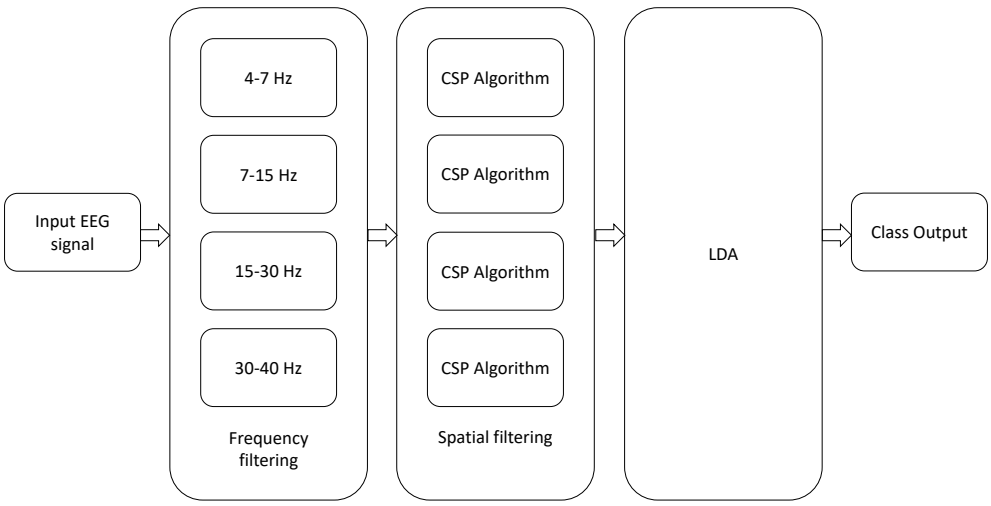


Figure 4. Architecture of filter bank CSP. LDA is shorthand for Linear Discriminant Analysis.

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 [72]. 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:

$$s^2 = \frac{s^2_{\Delta f}}{M}, \tag{50}$$

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:

$$P - VAL = Prob(t > T - STAT|H0). \tag{52}$$

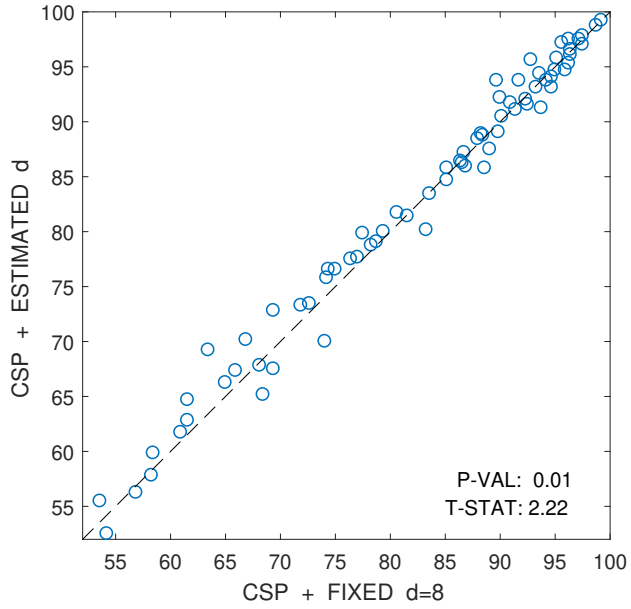
The more positive is $T - STAT$, the smaller is the $P - VAL$, and the probability of observing a t-statistics larger than $T - STAT$ decreases under the null hypothesis. When the p-value falls below the 0.05 threshold of significance, the hypothesis of not having a performance improvement when using 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.

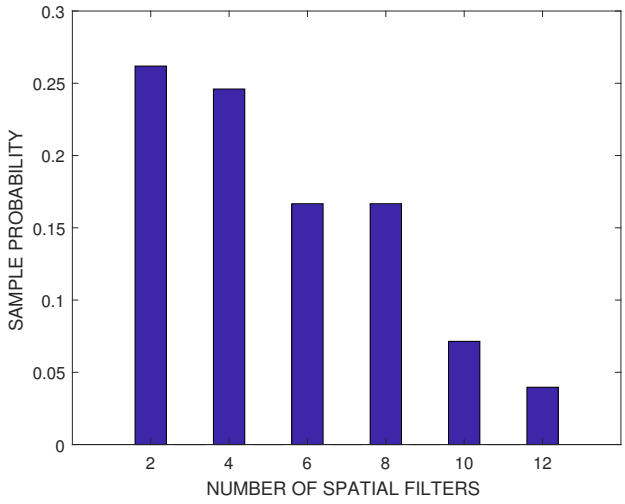
We briefly name and describe below some of the implementations that optimize the already mentioned criteria for dimensionality reduction in MI-BCIs. Because of the substantially higher computational complexity of most of the alternatives to CSP (see Table 1), it is not practical to develop a specific 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 [18,49].
- 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 [73].
- 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 [72].
- 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.



(**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 [72], 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 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.

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.

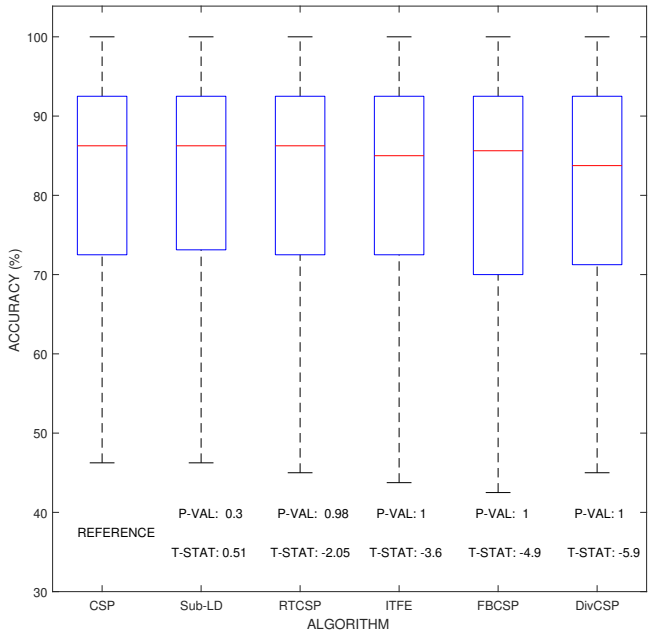


Figure 6. Comparison of the expected accuracy percentages obtained by each of the considered algorithms. The figure shows box-plot illustrations where the median is shown in red line, while the 25% and 75% percentiles are respectively at the bottom and top of each box. Larger positive values $T - STAT \gg 0$ 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.

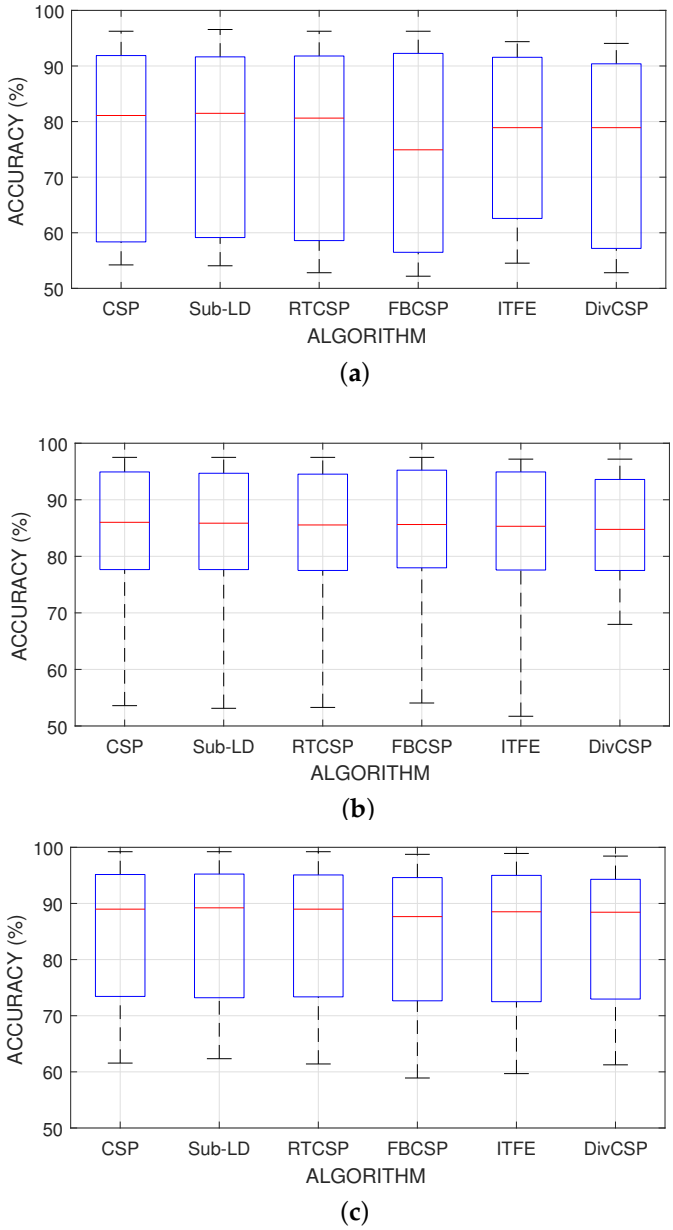
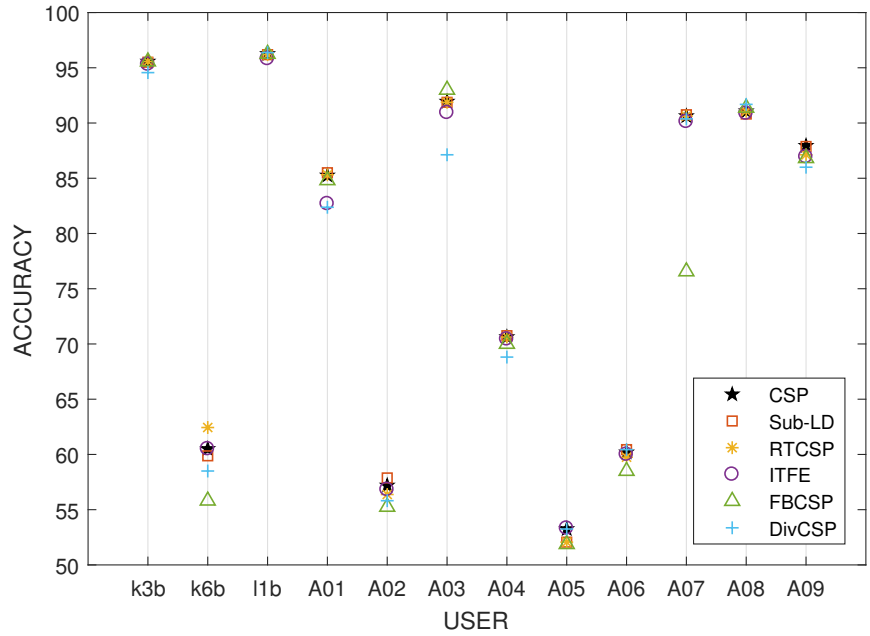
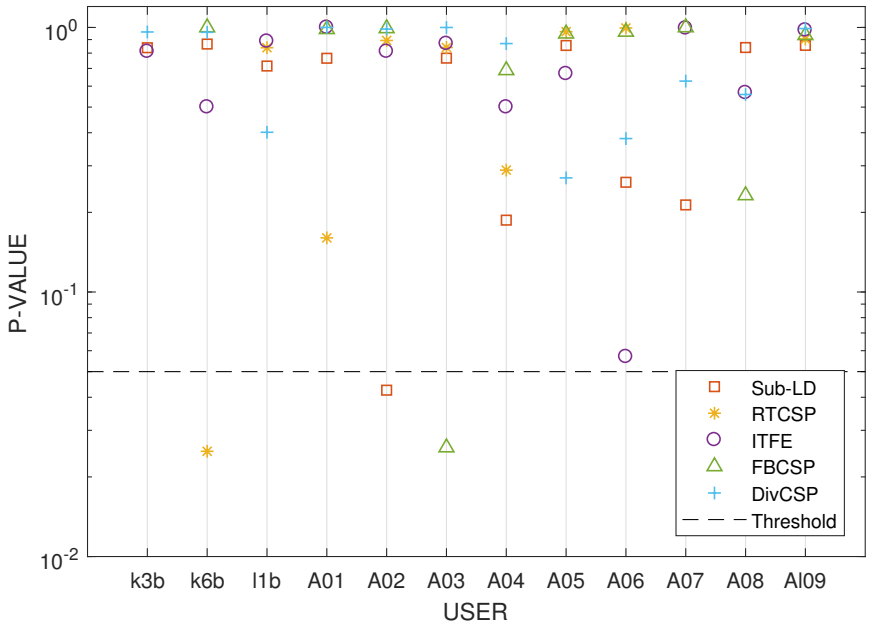


Figure 7. Performance of the algorithms for different motor imagery combinations involving the right hand. (**a**) Right-hand versus left-hand motor imagery classification; (**b**) right-hand versus feet motor imagery classification; (**c**) right-hand versus tongue motor imagery classification.

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.



(**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%.

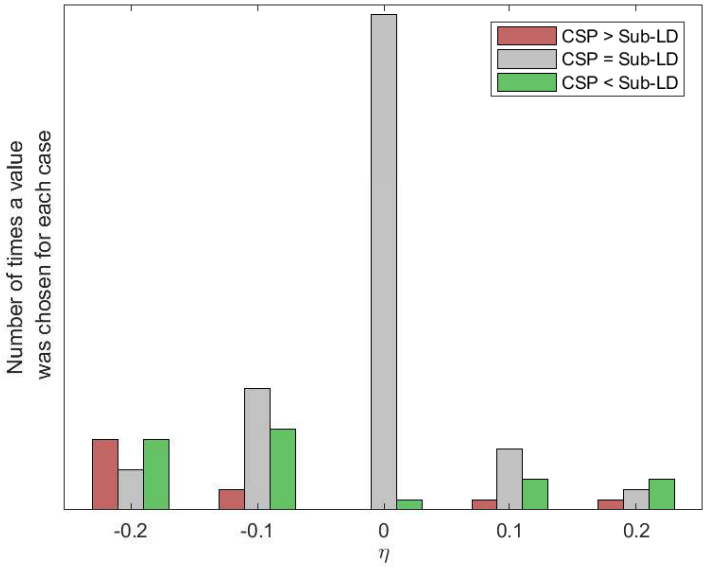


Figure 9. Histogram of the values of the regularization parameter in the Sub-LD algorithm that have been chosen by cross-validation.

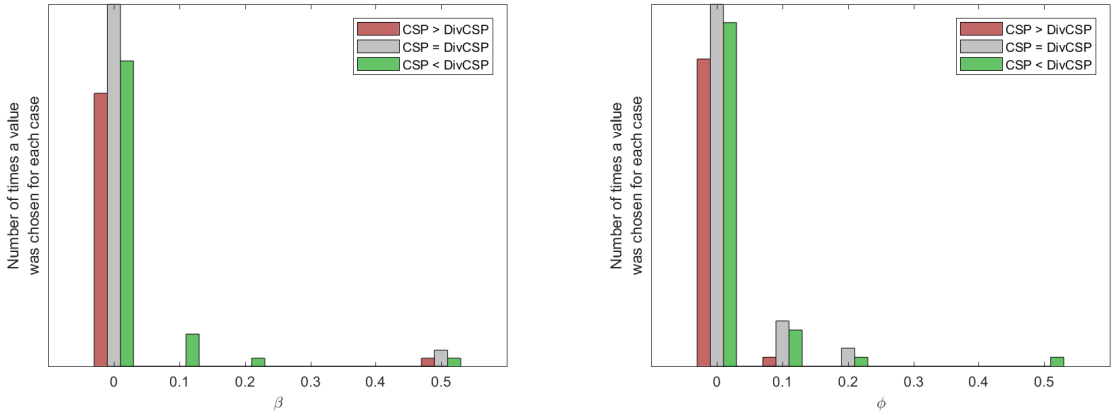


Figure 10. Histogram of the hyper-parameters of the DivCSP algorithm selected by cross-validation. (**a**) Case with $\beta \in [0, 0.5]$ and $\phi = 0$; (**b**) case with $\beta = 0.5$ and $\phi \in [0, 0.5]$.

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.

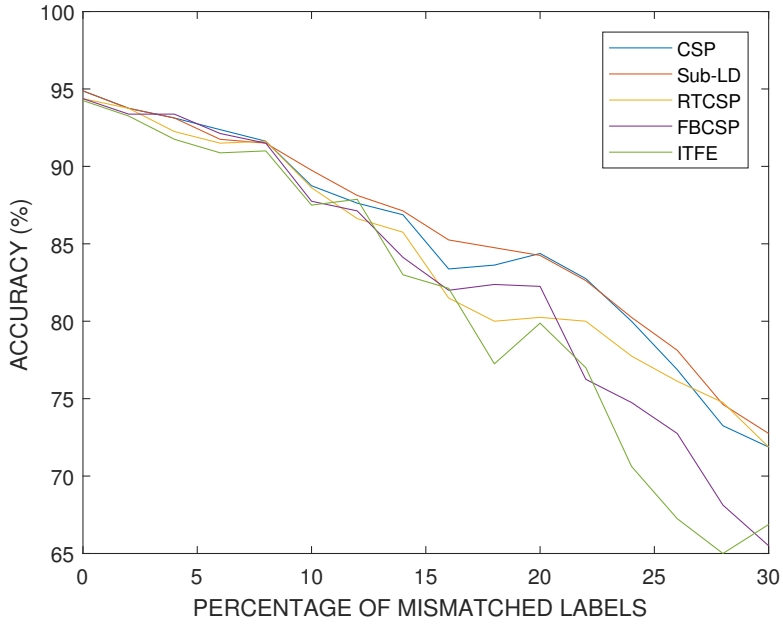


Figure 11. Comparison of the accuracy percentages obtained by each of the considered algorithms with respect to the percentage of mismatched labels in the training set. This experiment illustrates deterioration of the performance of the algorithms with respect to the increase of the percentage of randomly 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.

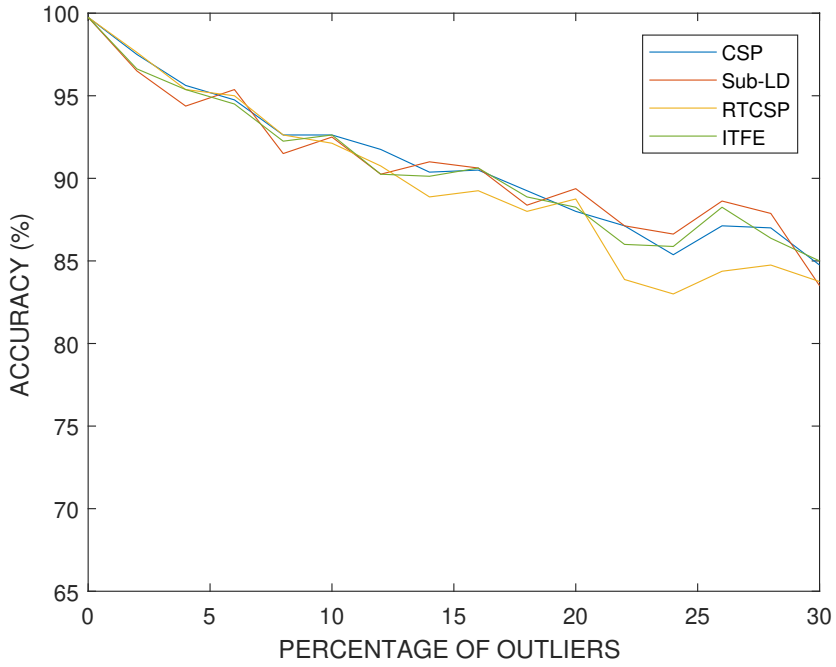


Figure 12. Accuracy percentages versus the percentage of training trials with outliers in a synthetic classification experiment.

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. 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,
5. Samek, W.; Meinecke, F.C.; Müller, K.-R. Transferring subspaces between subjects in brain–computer interfacing. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2289–2298.
6. 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.
7. Grosse-Wentrup, M.; Liefhold, C.; Gramann, K.; Buss, M. Beamforming in noninvasive brain-computer interfaces. *IEEE Trans. Biomed. Eng.* **2009**, *56*, 1209–1219.
8. 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.
9. 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.
10. 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.
11. Graimann, B.; Allison, B.; Pfurtscheller, G. Brain-computer interfaces: A gentle introduction. In *Brain-Computer Interfaces*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–27.
12. Pfurtscheller, G.; Lopes Da Silva, F.H. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clin. Neurophysiol.* **1999**, *110*, 1842–1857.
13. 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).
14. 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.
15. 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.
16. 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.
17. Perdikis, S.; Leeb, R.; Millán, J.D. Context-aware adaptive spelling in motor imagery BCI. *J. Neural Eng.* **2016**, *13*, 036018, doi:10.1088/1741-2560/13/3/036018.
18. Ramoser, H.; Müller-Gerking, J.; Pfurtscheller, G. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehabil. Eng.* **2000**, *8*, 441–446.
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, doi:10.1088/1741-2560/12/3/031001.
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. Available online: https://www.researchgate.net/profile/Motoaki_Kawanabe/publication/216887788_Group-wise_Stationary_Subspace_Analysis_-_A_Novel_Method_for_Studying_Non-Stationarities/links/02e7e51d7fec25159b000000.pdf (accessed on 19 December 2017)
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 5 June 2017).
72. 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.
73. Fabien Lotte. Matlab Codes and Software. Available online: https://sites.google.com/site/fabienlotte/code-and-softwares (accessed on 12 November 2017).
74. Wojciech Samek. The Divergence Methods Web Site . Available online: http://divergence-methods.org (accessed on 12 January 2017).