entropy

MDPI

Article

Optimization of Alpha-Beta Log-Det Divergences and their Application in the Spatial Filtering of Two Class Motor Imagery Movements

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

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

Academic Editor: Kevin H. Knuth
Received: 13 December 2016; Accepted: 22 February 2017; Published: 25 February 2017

Abstract: The Alpha-Beta Log-Det divergences for positive definite matrices are flexible divergences that are parameterized by two real constants and are able to specialize several relevant classical cases like the squared Riemannian metric, the Steins loss, the S-divergence, etc. A novel classification criterion based on these divergences is optimized to address the problem of classification of the motor imagery movements. This research paper is divided into three main sections in order to address the above mentioned problem: (1) Firstly, it is proven that a suitable scaling of the class conditional covariance matrices can be used to link the Common Spatial Pattern (CSP) solution with a predefined number of spatial filters for each class and its representation as a divergence optimization problem by making their different filter selection policies compatible; (2) A closed form formula for the gradient of the Alpha-Beta Log-Det divergences is derived that allows to perform optimization as well as easily use it in many practical applications; (3) Finally, in similarity with the work of Samek et al. 2014, which proposed the robust spatial filtering of the motor imagery movements based on the beta-divergence, the optimization of the Alpha-Beta Log-Det divergences is applied to this problem. The resulting subspace algorithm provides a unified framework for testing the performance and robustness of the several divergences in different scenarios.

Keywords: similarity measures; Common Spatial Pattern (CSP); generalized divergences for symmetric positive definite matrices; Brain Computer Interface (BCI); Alpha-Beta Log-Det divergence

1. Introduction

Over the last few years, the use of specialized metrics and divergences measures in the successful design of dimensionality reduction techniques has been progressively acquiring much recognition [1–5]. There are numerous real scenarios and applications for which the parameters of interest belong to non-flat manifolds, and where the Euclidean geometry results are unsuitable to evaluate the similarities. Indeed, this is usual case in the comparison of probability density functions and also of their associated covariance matrices. The present contribution may be seen as a continuation of the work in [6], where we defined the Alpha-Beta Log-Det family of divergences between Symmetric and Positive Definite (SPD) matrices and studied its properties. The Alpha-Beta Log-Det family unifies under the same framework many existing Log-Det divergences and connects them smoothly, through intermediate

versions, with the help of two real hyperparameters: α and β. In [7] a recent extension of the the Alpha-Beta Log-Determinant divergences was also proposed for the infinite-dimensional setting.

The evaluation of the Alpha-Beta Log-Det divergences depends on the generalized eigenvalues of the compared SPD matrices, and this makes its optimization non-trivial. In this paper, we motivate the use of these divergences with the illustrative application of dimensionality reduction in Brain-Computer Interface (BCI) and explain how to perform their optimization. The electroencephalogram (EEG) data has a typical high-dimensionality, a low signal to noise ratio and may have artifacts/outliers. The dimensionality reduction is then a necessary processing of the EEG signals for extracting those subspaces where the features have highest discriminative power.

Brain-Computer Interface has gained lots of interest in neuroscience and rehabilitation engineering. BCI [8,9] systems enable a person to operate external devices by using brain signals. The motor imagery (MI) based BCI systems are the most preferable BCI systems among others. It uses the brain signals of the MI movements as control commands for external devices without using the peripheral nervous system. During the imagination process, an alteration in the rhythmic activity of the brain can be observed in the *mu* and β rhythms at the corresponding area of the sensory-motor cortex. This phenomenon is known as event-related synchronization (ERS) or event-related desynchronization (ERD) [10]. The MI-based BCI systems use theses activities as control commands. Such a system can potentially serves as a communication aid for the people suffering from amyotrophic lateral sclerosis, multiple sclerosis and completely locked-in.

One of the most popular and efficient algorithms used for MI-based BCI applications is the common spatial pattern (CSP) algorithm [11]. It was first used to detect the abnormalities present in EEG signals [12] and later, was introduced in BCI applications [13]. The main objective of the CSP is to obtain the spatial filters by maximizing the variance of one class, at the same time minimizing that of the other class variance. It has been reported that this algorithm provides excellent classification accuracy for MI-based BCI systems. Besides, being the most popular method, its performance is easily affected by the presence of artifacts and nonstationarities. Since the computation of the spatial filters mainly depends on the covariance matrix, the presence of artifacts such as blinking of the eyes, eye movements and improper placement of the electrodes contribute to the poor computation of the covariance matrix which leads to the poor classification performance.

The main contributions of this work are the following:

- The existing link between the CSP method and the symmetric KL divergence (see [1]), is extended to the case of the minimax optimization of the AB log-det divergences. In absence of regularization, their solutions are shown to be equivalent whenever these methods apply the same divergence-based criterion for choosing the spatial filters. Although, in general, this is not the case when the CSP method adopts the popular practical criterion of a priori fixing the number of spatial filters for each class, we show that the equivalence with the solution of the optimization of AB log-det divergences can be still preserved if a suitable scaling factor κ is used in one of the arguments of the divergence.
- The details on how to perform the optimization of the AB log-det divergence are presented. The explicit expression of the gradient of this divergence with respect to the spatial filters is obtained. Expression which generalizes and extends the gradient of several more established well-known divergences, for instance, the gradient of the Alpha–Gamma divergence and the gradient of the Kullback–Leibler divergence between SPD matrices.
- The robustness property of the AB log-det divergence with respect to outliers has been analyzed. The study reveals that the hyperparameters of the divergence can be chosen to underweight or overweight, at convenience, the influence of the larger and smaller generalized eigenvalues in the estimating equations.
- Motivated by the success of criteria based on the Beta divergence [1] in the robust spatial filtering of the motor imagery movements, in this work, we consider the use of a criterion based on AB log-det divergences for the same purpose. A subspace optimization algorithm based on

regularized AB log-det divergences is proposed for obtaining the relevant subset of spatial filters. Some exemplary simulations illustrate its robustness over synthetic and real datasets.

This article is organized as follows: Section 2 presents the fundamental model of the observations and paper notation. Section 3 reviews the CSP algorithm while Section 4 discusses CSP via the divergence optimization. In Section 5, we present the family of AB log-det divergences and provide a new upper-bounds and conditions for the equivalence between this divergence optimization and the robust CSP solution. Section 6 explains how to obtain closed-form formulas for computing the gradient of the AB log-det divergence, which is useful for its optimization. The analysis of the robustness of the divergence in terms of its hyperparameters is the objective of Section 7. Section 8 briefly reviews several related techniques, while Section 9 presents the regularized version of the criterion based on AB log-det divergences, as well as, the subspace algorithm that optimizes it. Section 10 presents the experimental datasets, the steps involved in the preprocessing, feature extraction and classification. The results of the simulations are presented and discussed in Section 11. Finally, the paper summarizes main results in Section 12.

2. Notation and Model of the Measurements

Throughout this paper, the following notations are adopted. Vectors are typically denoted by bold letters, the capital bold letters are reserved for the matrices, while the random variables appear in italic capital letters. The operators $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ round the value of their argument to the nearest lower and higher integers respectively. All the covariance matrices, which are denoted by $Cov(\cdot)$, are assumed to be positive definite and hence invertible.

Let us now describe the statistical model of the observations. As usual, the raw EEG observations are initially preprocessed by a bandpass filter that retains the activity in the bands of the *mu* and β rhythms and are later normalized for each trial so as to keep their total spatial power constant. One can define a statistical model of these "normalized" observations as $\boldsymbol{x}(t) = [x_1(t), \ldots, x_n(t)]^T$ conditioned on the true imagery movement, which here will be represented by a member of the class $c \in \{c_1, c_2\}$. In general, the EEG observations are noisy and high-dimensional, while the number of recorded trials is quite limited. Therefore, the learning of the discriminative features is quite sensitive to overfitting, a situation that would severely degrade the prediction accuracy over test samples. In this case, it is worth sacrificing the bias by choosing a simpler (less complex) model in which parameters can be estimated with a smaller variance. For this reason, we adopt usual convention [14] of considering the observations from each class as drawn from the independent and identically distributed (i.i.d.) Gaussian random vectors as represented as $\boldsymbol{X}|c$ of zero mean and with covariance matrix as $Cov(\boldsymbol{X}|c)$, which in turn is set equal to the sample covariance matrix of the class, i.e.,

$$Cov(\boldsymbol{X}|c) = Cov(\boldsymbol{x}|c) \quad \text{for} \quad c \in \{c_1, c_2\}. \tag{1}$$

The observations are then modeled by the mixture distribution

$$p(\boldsymbol{x}) \quad = \quad p(c_1)p(\boldsymbol{x}|c_1) + p(c_2)p(\boldsymbol{x}|c_2), \tag{2}$$

where $p(c)$ refers to the sample probabilities of each class in the training data. When $\bar{\boldsymbol{x}}$ denotes the sample mean of the observations, their sample covariance matrix is obtained by

$$Cov(\boldsymbol{x}) \quad = \quad \frac{1}{T}\sum_{t=1}^{T} \left(\boldsymbol{x}(t) - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}(t) - \bar{\boldsymbol{x}}\right)^T \tag{3}$$

$$= \quad p(c_1)Cov(\boldsymbol{x}|c_1) + p(c_2)Cov(\boldsymbol{x}|c_2). \tag{4}$$

and its eigenvalue decomposition is

$$Cov(\boldsymbol{x}) \quad = \quad \mathbf{U}_1 \boldsymbol{\Delta} \mathbf{U}_1^T \tag{5}$$

where $\mathbf{\Delta}$ and $\mathbf{U}_1$, respectively denote the matrix of eigenvalues and eigenvectors of $Cov(\boldsymbol{x})$.

We define $\boldsymbol{w}_i = [w_{1i}, w_{2i}, ..., w_{pi}]^T$ as the vector with the coefficients of the i-th-esime spatial filter for $i = 1, \ldots, p$. The collection of p spatial filters forms the overall filter matrix $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2 ..... \boldsymbol{w}_p]$, which is used to reduce the dimensionality of the observations by projecting them onto the p-dimensional subspace spanned by the filter outputs

$$\boldsymbol{y} = \mathbf{W}^T \boldsymbol{x} \in \mathbb{R}^p, \tag{6}$$

where $p \ll n$. The model for the estimated conditional distribution $p(\boldsymbol{y}|c)$ is a multidimensional Gaussian of zero mean and covariance matrix $Cov(\boldsymbol{Y}|c) = \mathbf{W}^T Cov(\boldsymbol{x}|c)\mathbf{W}$, i.e., for each class

$$\boldsymbol{Y}|c \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^T Cov(\boldsymbol{x}|c)\mathbf{W}) \,. \tag{7}$$

3. The Common Spatial Patterns Algorithm

The development of the CSP algorithm as a technique for feature selection in classification problems can be traced back to the work of [11], while later, [12,13] considered its practical application for the study of EEG signals. This technique exploits the event-related desynchronization during the limbs movement imagination process that alters the rhythmic activity in a class dependent area of the motor cortex. The objective of the algorithm is to obtain a set of most discriminative spatial filters, i.e., those that hierarchically maximize the output activity of one class, while at the same time; they minimize the activity of the other class. Since only the direction of the spatial filters (i.e., not the scale) are of interest, the technique starts with a linear transformation $\boldsymbol{y} = \mathbf{W}^T \boldsymbol{x}$ that whitens the sample covariance of the outputs

$$\begin{aligned} Cov(\boldsymbol{y}) &= p(c_1) Cov(\boldsymbol{y}|c_1) + p(c_2) Cov(\boldsymbol{y}|c_2) && (8)\\ &= \mathbf{W}^T \, Cov(\boldsymbol{x}) \, \mathbf{W} && (9)\\ &= \mathbf{I}_p \,. && (10) \end{aligned}$$

With the help of the eigenvalue decomposition of $Cov(\boldsymbol{x})$, the general expression of the spatial filter matrix that preserves the whitening constraint can be found as

$$\mathbf{W}^T = \mathbf{\Omega}^T \mathbf{\Delta}^{-\frac{1}{2}} \mathbf{U}_1^T \,. \tag{11}$$

Note that $\mathbf{W} \in \mathbb{R}^{n \times p}$ is specified up to the ambiguity in the choice of the semi-orthogonal matrix $\mathbf{\Omega} \in \mathbb{R}^{n \times p}$ (i.e., $\mathbf{\Omega}^T \mathbf{\Omega} = \mathbf{I}_p$) which parameterizes the relevant degrees of freedom for finding the most discriminative directions. Then, the objective of the CSP criterion [11] is implemented by first choosing one part of the spatial filters from the constrained maximization of the conditional covariances of the outputs of the first class

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} \boldsymbol{w}^T Cov(\boldsymbol{x}|c_1) \boldsymbol{w} \qquad i = 1, \ldots, k, \tag{12}$$

and later choosing the other part of the filters to hierarchically maximize the conditional covariances of the outputs of the second class

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} \boldsymbol{w}^T Cov(\boldsymbol{x}|c_2) \boldsymbol{w} \qquad i = k+1, \ldots, p, \tag{13}$$

where, in both cases, the maximization with respect to the spatial filters takes place under the whitening or ($Cov(\boldsymbol{x})$-orthonormality) constraints

$$\boldsymbol{w}_i^T Cov(\boldsymbol{x}) \boldsymbol{w}_j = \delta_{ij} \qquad \forall j \leq i \,. \tag{14}$$

The number of spatial filters k that hierarchically maximize (12) can be determined by a chosen filter selection policy. For simplicity, in most cases it is usually set k close to $\frac{p}{2}$ with the aim to balance the number of spatial filters devoted to each of the classes.

The maximization in (12) can be alternatively posed as the constrained optimization of the quotient

$$\boldsymbol{w}_i \quad = \quad \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^T Cov(\boldsymbol{x}|c)\boldsymbol{w}}{\boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w}} \quad \text{subject to} \quad \boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w} = \delta_{ij} \quad \forall j \leq i \tag{15}$$

which, in terms of the transformed and normalized spatial vectors

$$\boldsymbol{r}_i = \frac{(Cov(\boldsymbol{x}))^{\frac{1}{2}}\boldsymbol{w}_i}{\|(Cov(\boldsymbol{x}))^{\frac{1}{2}}\boldsymbol{w}_i\|_2}, \tag{16}$$

is rewritten as a quadratic optimization under orthogonality constraints

$$\boldsymbol{w}_i = (Cov(\boldsymbol{x}))^{-\frac{1}{2}} \times \arg\max_{\boldsymbol{r}} \left\{ \boldsymbol{r}^T \, (Cov(\boldsymbol{x}))^{-\frac{1}{2}} Cov(\boldsymbol{x}|c)(Cov(\boldsymbol{x}))^{-\frac{1}{2}} \, \boldsymbol{r} \right\} \quad \text{s.t.} \quad \boldsymbol{r}^T \boldsymbol{r}_j = \delta_{ij} \quad \forall j \leq i \tag{17}$$

At this point, the straightforward application of the Courant–Fisher–Weyl minimax principle ([15], p. 58) yields the variational description of the desired spatial filters as the minimax solution of the Rayleigh quotients for each class

$$\boldsymbol{w}_i \quad = \quad (Cov(\boldsymbol{x}))^{-\frac{1}{2}} \times \arg \min_{\dim\{\mathcal{R}\}=n-i+1} \max_{\substack{\boldsymbol{r}\in\mathcal{R} \\ \|\boldsymbol{r}\|=1}} \left\{ \boldsymbol{r}^T \, (Cov(\boldsymbol{x}))^{-\frac{1}{2}} Cov(\boldsymbol{x}|c)(Cov(\boldsymbol{x}))^{-\frac{1}{2}} \, \boldsymbol{r} \right\} \tag{18}$$

$$= \quad \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} \frac{\boldsymbol{w}^T Cov(\boldsymbol{x}|c)\boldsymbol{w}}{\boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w}} \tag{19}$$

$$= \quad \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} \frac{Cov(y_i|c)}{Cov(y_i)}. \tag{20}$$

By the same principle, the generalized eigenvectors $\boldsymbol{v}_i^{(c)}$ of the matrix pencil $(p(c)\ Cov(\boldsymbol{y}|c)$, $Cov(\boldsymbol{y}))$, are the minimax solutions of the Rayleigh quotient, while the values that takes the criterion at these solutions are the generalized eigenvalues

$$\lambda_i^{(c)} = p(c)\frac{\boldsymbol{v}_i^{(c)T} Cov(\boldsymbol{x}|c)\boldsymbol{v}_i^{(c)}}{\boldsymbol{v}_i^{(c)T} Cov(\boldsymbol{x})\boldsymbol{v}_i^{(c)}} = p(c) \min_{\dim\{\mathcal{W}\}=i} \max_{\boldsymbol{w}\in\mathcal{W}} \frac{\boldsymbol{w}^T Cov(\boldsymbol{x}|c)\boldsymbol{w}}{\boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w}}, \tag{21}$$

which are sorted according to the descent in their magnitude, $\lambda_1^{(c)} \geq \lambda_2^{(c)} \geq \ldots \geq \lambda_n^{(c)}$.

The generalized eigenvectors of the two quotients (one for each class) coincide, except for their ordering which are reversed [11], i.e., $\boldsymbol{v}_i^{(c_1)} = \boldsymbol{v}_{n-i+1}^{(c_2)}$, while the weighted sum of generalized eigenvalues is bounded by

$$\lambda_i^{(c_1)} + \lambda_{n-i+1}^{(c_2)} = \frac{\boldsymbol{v}_i^{(c_1)T} \left(p(c_1)Cov(\boldsymbol{x}|c_1) + p(c_2)Cov(\boldsymbol{x}|c_2)\right) \boldsymbol{v}_i^{(c_1)}}{\boldsymbol{v}_i^{(c_1)T} Cov(\boldsymbol{x})\boldsymbol{v}_i^{(c_1)}} = 1. \tag{22}$$

Therefore, a direction of maximum variance for one class will simultaneously minimize the variance of the other class, and vice versa. Hence, the standard CSP solution is obtained when the spatial filters match with the principal and minor eigenvectors of the generalized eigendecomposition problem [11–13]

$$Cov(\boldsymbol{x}|c_1)\boldsymbol{v}_i^{(c_1)} = \lambda_i^{(c_1)} Cov(\boldsymbol{x})\boldsymbol{v}_i^{(c_1)} \quad i = 1, \ldots, n. \tag{23}$$

After sorting the eigenvalues according to its magnitude, CSP explicitly selects k spatial filters $\boldsymbol{v}_i^{(c_1)}$ from the principal eigenvectors and $p-k$ spatial filters from the minor eigenvectors, to form the spatial filter matrix

$$\begin{aligned} \mathbf{W}_{CSP} &\equiv [\boldsymbol{w}_1, \boldsymbol{w}_2 ..... \boldsymbol{w}_p] && (24) \\ &= [\boldsymbol{v}_1^{(c_1)}, \ldots, \boldsymbol{v}_k^{(c_1)}, \boldsymbol{v}_{n-(p-k)+1}^{(c_1)}, \ldots, \boldsymbol{v}_n^{(c_1)}]. && (25) \end{aligned}$$

4. The Divergence Optimization Interpretation of CSP

Under the appropriate selection policy for the number of spatial filters for each class, the solution obtained by the CSP algorithm admits an interpretation in terms of the optimization divergence measures (here denoted by $Div(\cdot\|\cdot)$) between the Gaussian pdfs outputs for each class

$$\boldsymbol{w}_i = \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} Div(p(y_i|c_1)\|p(y_i|c_2)), \quad (26)$$

except for a probable permutation in the ordering of some of the spatial filters.

The problem can be formulated using the following optimization problem

$$\boldsymbol{w}_i = \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D(Cov(y_i|c_1)\|Cov(y_i|c_2)) \quad (27)$$

where $D(\cdot\|\cdot)$ refers to a divergence between the covariances of the conditional densities of the outputs. As a consequence of the assumption of zero mean Gaussian densities, the covariances are the only necessary statistics that summarize all the relevant information of the conditional data.

In particular, the solution of the CSP algorithm was linked in [1,11,16,17] with the optimization of the symmetric Kullback–Leibler divergence (sKL)

$$\begin{aligned} Div_{sKL}(p(y_i|c_1)\|p(y_i|c_2)) &= \int p(y_i|c_1)\log\frac{p(y_i|c_1)}{p(y_i|c_2)}dy_i + \int p(y_i|c_2)\log\frac{p(y_i|c_2)}{p(y_i|c_1)}dy_i, && (28) \\ &= \int (p(y_i|c_1)-p(y_i|c_2))\log\frac{p(y_i|c_1)}{p(y_i|c_2)}dy_i. && (29) \end{aligned}$$

This divergence measures can be simplified to the symmetric Kullback–Leibler (sKL) divergence between the class conditional covariances

$$\begin{aligned} Div_{sKL}(p(y_i|c_1)\|p(y_i|c_2)) &= \frac{1}{2}\frac{Cov(y_i|c_1)}{Cov(y_i|c_2)} + \frac{1}{2}\frac{Cov(y_i|c_2)}{Cov(y_i|c_1)} - 1 && (30) \\ &\equiv D_{sKL}(Cov(y_i|c_1)\|Cov(y_i|c_2)). && (31) \end{aligned}$$

In this paper, we propose an extension of the existing KL to the criterion of the Alpha-Beta log-det divergence (AB log-det) between the class-conditional covariances defined as [6]

$$D_{AB}^{(\alpha,\beta)}(Cov(y_i|c_1)\|Cov(y_i|c_2)) = \frac{1}{\alpha\beta}\log\left|\frac{\alpha\left(\frac{Cov(y_i|c_1)}{Cov(y_i|c_2)}\right)^{\beta}+\beta\left(\frac{Cov(y_i|c_2)}{Cov(y_i|c_1)}\right)^{\alpha}}{\alpha+\beta}\right|_+ \quad (32)$$

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

where

$$|x|_+ = \begin{cases} x & x\geq 0, \\ 0, & x<0, \end{cases} \quad (33)$$

denotes the non-negative truncation operator. When the arguments covariances are scalars and $\alpha, \beta > 0$, the AB log-det divergence can also be rewritten as the logarithmic ratio between the weigthed arithmetic mean of the scaled covariances $(Cov^{\alpha+\beta}(y_i|c_1), Cov^{\alpha+\beta}(y_i|c_2))$ and their weighted geometric mean, i.e.,

$$D_{AB}^{(\alpha,\beta)}(Cov(y_i|c_1)\|Cov(y_i|c_2)) = \frac{1}{\alpha\beta}\log\frac{\left(\frac{\alpha}{\alpha+\beta}Cov^{\alpha+\beta}(y_i|c_2)+\frac{\beta}{\alpha+\beta}Cov^{\alpha+\beta}(y_i|c_1)\right)}{(Cov^{\alpha+\beta}(y_i|c_2))^{\frac{\alpha}{\alpha+\beta}}(Cov^{\alpha+\beta}(y_i|c_1))^{\frac{\beta}{\alpha+\beta}}}. \quad (34)$$

Additionally if $\alpha+\beta=1$, the AB log-det divergence between covariances is proportional to the Alpha–Gamma divergence [18] between the conditional densities

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(Cov(y_i|c_1)\|Cov(y_i|c_2)) &\equiv 2\,Div_{AG}^{(\beta,\alpha)}\left(p(y_i|c_1)\|p(y_i|c_2)\right) \\ &= \frac{2}{\alpha\beta}\log\frac{\left(\int_\Omega p(y_i|c_1)\,dy_i\right)^{\beta}\left(\int p(y_i|c_2)\,dy_i\right)^{\alpha}}{\int p^{\beta}(y_i|c_1)\,p^{\alpha}(y_i|c_2)\,dy_i} \\ &\quad \text{for} \quad \alpha>0,\ \beta>0,\ \alpha+\beta=1. \end{aligned} \quad (35)$$

In Section 5.3, it is proven that, under certain conditions, the simple optimization of an AB log-det divergence also leads to the solution of the CSP algorithm. Although, the potential of these divergences does not rely on their plain optimization but instead rely on their optimization in the presence of some regularization terms that help to specify the desired solutions.

Recently, several divergence criteria have been proposed for the extraction of the spatial dimensions with maximum discriminative power. Among these, the multiclass approach based on the maximization of the harmonic mean of Kullback–Leibler divergences [16] and the regularization framework based on the beta divergences [1,17] are the most noteworthy methods. Another approach based on Bhattacharyya distance and Gamma divergence has also been proposed for classification of motor imagery movements [19]. Our proposal is motivated by the success of these methods in improving the classification accuracy and the robustness against the outliers. The distinctive property of the AB log-det divergence is that it smoothly connects (through its hyperparameters) a quite broad family of log-det divergences for SPD matrices, covering several relevant classical cases like: the KL divergence, the dual KL divergence, the Beta log-det family, the Alpha log-det family, the Power log-det family, as well as the Affine Invariant Riemannian divergence.

5. The Definition of the AB Log-Det Divergence

Henceforth, we will work on the multidimensional observation vectors $\boldsymbol{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$. In order to simplify the notation, the covariance matrices of the two classes are renamed as follows

$$\mathbf{P} \equiv Cov(\boldsymbol{x}|c_1), \quad (36)$$

$$\mathbf{Q} \equiv Cov(\boldsymbol{x}|c_2). \quad (37)$$

The AB log-det divergence is a directed divergence that evaluates the dissimilarity between two multidimensional covariance matrices. It was defined in [6] as

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{\alpha\beta}\log\left|\frac{\alpha(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{\beta}+\beta(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{-\alpha}}{\alpha+\beta}\right|_+ \quad (38)$$

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

while, for the singular cases, its definition is given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \frac{1}{\alpha^2}\left[\mathrm{tr}\left((\mathbf{Q}^{\frac{1}{2}}\mathbf{P}^{-1}\mathbf{Q}^{\frac{1}{2}})^{\alpha}-\mathbf{I}\right)-\alpha\log|\mathbf{Q}^{\frac{1}{2}}\mathbf{P}^{-1}\mathbf{Q}^{\frac{1}{2}}|\right] & \text{for } \alpha\neq 0,\ \beta=0, \\ \frac{1}{\beta^2}\left[\mathrm{tr}\left((\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{\beta}-\mathbf{I}\right)-\beta\log|\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}}|\right] & \text{for } \alpha=0,\ \beta\neq 0, \\ \frac{1}{\alpha^2}\log\left|(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{\alpha}(\mathbf{I}+\log(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{-\alpha})\right|_{+} & \text{for } \alpha=-\beta, \\ \frac{1}{2}||\log(\mathbf{Q}^{\frac{1}{2}}\mathbf{P}^{-1}\mathbf{Q}^{\frac{1}{2}})||_F^2 & \text{for } \alpha,\ \beta=0. \end{cases} \tag{39}$$

The divergence depends only on the eigenvalues $\mathbf{\Lambda}=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$ of the Symmetric Positive Definite (SPD) matrix $\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2}$, which also coincide with the eigenvalues of the matrix $\mathbf{Q}^{-1}\mathbf{P}$, although their eigenspaces differ. Given the eigenvalue decomposition

$$\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}}=\mathbf{V}_1\mathbf{\Lambda}\mathbf{V}_1^T, \tag{40}$$

where $\mathbf{V}_1$ is the orthogonal matrix of eigenvectors, and $\mathbf{\Lambda}=\mathrm{diag}\{\lambda_1,\lambda_2,\ldots,\lambda_n\}$ is the diagonal matrix with positive eigenvalues $\lambda_i>0,\ i=1,2,\ldots,n$. One of the properties of the AB log-det divergence is that it is invariant under a common change of basis on its matrix arguments, i.e., an invertible congruence transformation. Since, with the help of this specific transformation, we have

$$\mathbf{P} \ \rightarrow\ (\mathbf{V}_1^T\mathbf{Q}^{-\frac{1}{2}})\mathbf{P}(\mathbf{V}_1^T\mathbf{Q}^{-\frac{1}{2}})^T=\mathbf{\Lambda}, \tag{41}$$

$$\mathbf{Q} \ \rightarrow\ (\mathbf{V}_1^T\mathbf{Q}^{-\frac{1}{2}})\mathbf{Q}(\mathbf{V}_1^T\mathbf{Q}^{-\frac{1}{2}})^T=\mathbf{I}, \tag{42}$$

it can be inferred that the divergence is separable (over the generalized eigenvalues of the matrix pencil $(\mathbf{P},\mathbf{Q})$) in a sum of marginal divergences that measure how far are each of the generalized eigenvalues from the unity, i.e.,

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\,\|\mathbf{Q})=D_{AB}^{(\alpha,\beta)}\,(\mathbf{\Lambda}\|\,\mathbf{I}_n)=\sum_{i=1}^{n}D_{AB}^{(\alpha,\beta)}(\lambda_i\|\,1). \tag{43}$$

Hence,

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \ = \ \frac{1}{\alpha\beta}\sum_{i=1}^{n}\log\left|\frac{\alpha\lambda_i^{\beta}+\beta\lambda_i^{-\alpha}}{\alpha+\beta}\right|_{+},\ \ \alpha,\ \beta,\ \alpha+\beta\neq 0. \tag{44}$$

Similarly, for the singular cases, the divergence is

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \ = \ \begin{cases} \frac{1}{\alpha^2}\left[\sum_{i=1}^{n}\left(\lambda_i^{-\alpha}-\log(\lambda_i^{-\alpha})\right)-n\right] & \text{for } \alpha\neq 0,\ \beta=0 \\ \frac{1}{\beta^2}\left[\sum_{i=1}^{n}\left(\lambda_i^{\beta}-\log(\lambda_i^{\beta})\right)-n\right] & \text{for } \alpha=0,\ \beta\neq 0 \\ \frac{1}{\alpha^2}\left[\sum_{i=1}^{n}\log\left|\frac{\lambda_i^{\alpha}}{1+\log\lambda_i^{\alpha}}\right|_{+}\right] & \text{for } \alpha=-\beta\neq 0 \\ \frac{1}{2}\sum_{i=1}^{n}\log^2(\lambda_i) & \text{for } \alpha,\ \beta=0. \end{cases} \tag{45}$$

This divergence compares two symmetric positive definite matrices and returns its dissimilarity, i.e., a positive value when they are non-coincident and $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = 0$ iff $\mathbf{P} = \mathbf{Q}$. As it can be observed in Figure 1 the AB log-det divergence generalizes several existing log-det matrix divergences, like: the Stein's loss, 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 (see Table 1 in [6] for a comprehensive list).

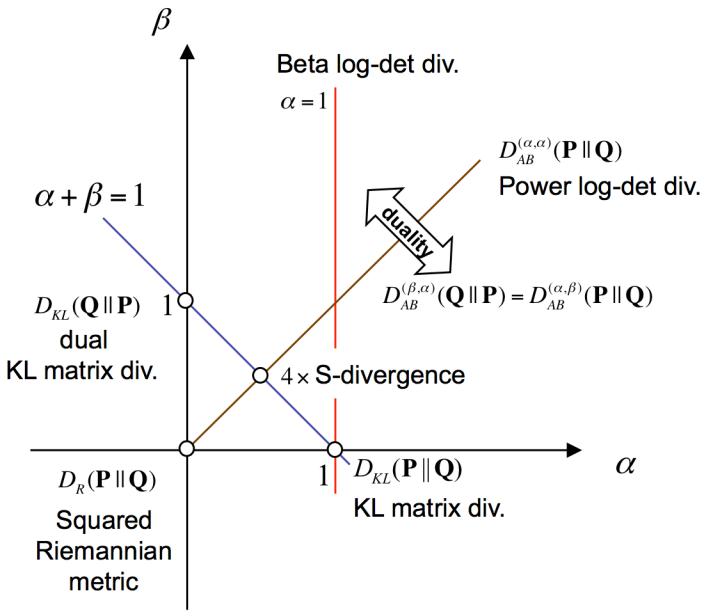


Figure 1. This illustration shows the AB log-det divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ positioned in a plane as a function of their real pair of hyperparameters (α, β). It is clear from the figure, that the parameterization smoothly connects several relevant positive definite matrix divergences, like: 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.

5.1. A Tight Upper-Bound for the AB Log-Det Divergences

The divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ depends on the generalized eigenvalues $\lambda_1, \cdots, \lambda_n$ of the matrix pencil $(\mathbf{P}, \mathbf{Q})$ which, for convenience, are assumed to have a simple spectrum (the eigenvalues are unique or non-coincident) and can be sorted in descending order

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0. \tag{46}$$

In practice, the assumption is plausible because the real symmetric matrices with unique eigenvalues are known to form an open dense set in the space of all the real symmetric matrices [20].

Although the space of the observations is high-dimensional, most of the discriminative information between the two conditions is confined into a low-dimensional subspace. Thus, the spatial filter matrix $\mathbf{W} \in \mathbb{R}^{n\times p}$ is used to reduce the dimensionality of the samples from n to p with the linear compression transformation $\boldsymbol{y} = \mathbf{W}^T\boldsymbol{x} \in \mathbb{R}^p$. It is shown in [6] that, after applying this compression to the arguments of the divergence, the resulting output covariance matrices $\mathbf{W}^T\mathbf{P}\mathbf{W}$ and $\mathbf{W}^T\mathbf{Q}\mathbf{W}$ are more similar than in the original space, as shown in the below equation

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\,\|\mathbf{W}^T\mathbf{Q}\mathbf{W}) = \sum_{i=1}^{p} D_{AB}^{(\alpha,\beta)}(\mu_i\|1) \leq \sum_{i=1}^{n} D_{AB}^{(\alpha,\beta)}(\lambda_i\|1) = D_{AB}^{(\alpha,\beta)}(\mathbf{P}\,\|\mathbf{Q}), \tag{47}$$

where $\mu_1 \geq \cdots \geq \mu_p > 0$ are the generalized eigenvalues of the matrix pencil $(\mathbf{W}^T\mathbf{P}\mathbf{W}, \mathbf{W}^T\mathbf{Q}\mathbf{W})$. However, this upper bound is loose for the case of interest (dimensionality reduction), i.e., when $p < n$. In Appendix A.1, the possible way to tighten the previous upper-bound with the following new proposal is shown

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W} \,\|\mathbf{W}^T\mathbf{Q}\mathbf{W}) \leq \sum_{i=1}^{p} D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_i}\| \, 1), \tag{48}$$

where π defines the permutation of the indices $1, \ldots, n$ that sorts the divergence of the eigenvalues from the unity in descending order

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_1}\| \, 1) \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_2}\| \, 1) \geq \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_n}\| \, 1). \tag{49}$$

Moreover, the equality with the upper-bound is only obtained for those extraction matrices $\mathbf{W}$ that lie within the span of the p generalized eigenvectors of the matrix pencil $(\mathbf{P}, \mathbf{Q})$ which are associated with the eigenvalues $\lambda_{\pi_1}, \ldots, \lambda_{\pi_p}$ that maximize the divergence from unity in (49).

5.2. Relationship between the Generalized Eigenvalues and Eigenvectors of the Matrix Pencils $(\mathbf{P}, \mathbf{Q})$ *and* $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$

We have seen in the previous section that the tight upper-bound of the divergence is attained by a subset of the generalized eigenvectors of the matrix pencil $(\mathbf{P}, \mathbf{Q})$, whereas, the CSP solution in (24) depends on a subset of the generalized eigenvectors of another matrix pencil $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$. In this section we addresses the close relationship between both eigendecompositions. For this purpose, we recognize $\mathbf{\Lambda}$ as the matrix of eigenvalues of $\mathbf{Q}^{-1}\mathbf{P}$ and $\mathbf{\Lambda}^{(c_1)}$ as the matrix of eigenvalues of $(Cov(\boldsymbol{x}))^{-1} \, p(c_1)\mathbf{P}$. Then, we write

$$(p(c_2)\mathbf{Q})^{-1} \, (p(c_1)\mathbf{P}) \quad = \quad [(Cov(\boldsymbol{x}))^{-1}(p(c_2)\mathbf{Q})]^{-1} \, [(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})], \tag{50}$$

and use the decomposition of $Cov(\boldsymbol{x})$ in (4) to substitute $p(c_2)\mathbf{Q} = Cov(\boldsymbol{x}) - p(c_1)\mathbf{P}$ in the previous equation. In this way, we obtain

$$(p(c_2)\mathbf{Q})^{-1}(p(c_1)\mathbf{P}) \quad = \quad [\mathbf{I}_n - (Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]^{-1} \, [(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]. \tag{51}$$

The matrix of eigenvectors $\mathbf{V}$ of $\mathbf{Q}^{-1}\mathbf{P}$ diagonalizes both sides of the previous equation

$$\begin{aligned}
\mathbf{\Lambda}^{\frac{p(c_1)}{p(c_2)}} &= \mathbf{V}^{-1}\left[\tfrac{p(c_1)}{p(c_2)}\mathbf{Q}^{-1}\mathbf{P}\right]\mathbf{V} && (52)\\
&= (\mathbf{V}^{-1}[\mathbf{I}_n - (Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]^{-1}\mathbf{V}) \, (\mathbf{V}^{-1}[(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]\mathbf{V}) && (53)\\
&= [\mathbf{I}_n - \mathbf{V}^{-1}(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P}\mathbf{V})]^{-1} \, (\mathbf{V}^{-1}[(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]\mathbf{V}) && (54)\\
&= (\mathbf{I}_n - \mathbf{\Lambda}^{(c_1)})^{-1}\mathbf{\Lambda}^{(c_1)}. && (55)
\end{aligned}$$

Hence, we have the explicit relationship between the two sets of eigenvalues

$$\lambda_i^{\frac{p(c_1)}{p(c_2)}} = \frac{\lambda_i^{(c_1)}}{1 - \lambda_i^{(c_1)}} \equiv g(\lambda_i^{(c_1)}), \qquad i = 1, \ldots, n, \tag{56}$$

where $g(\lambda_i^{(c_1)})$, as can be seen in Figure 2, is a strictly monotonous ascending function over the domain of $\lambda_i^{(c_1)} \in (0, 1)$. Moreover, the Equations (52)–(55) imply that the matrix $\mathbf{V}$ of generalized eigenvectors of the matrix pencil $(\mathbf{P}, \mathbf{Q})$ exactly coincides with the matrix $\mathbf{V}^{(c_1)}$ of generalized eigenvectors of the other matrix pencil $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$.

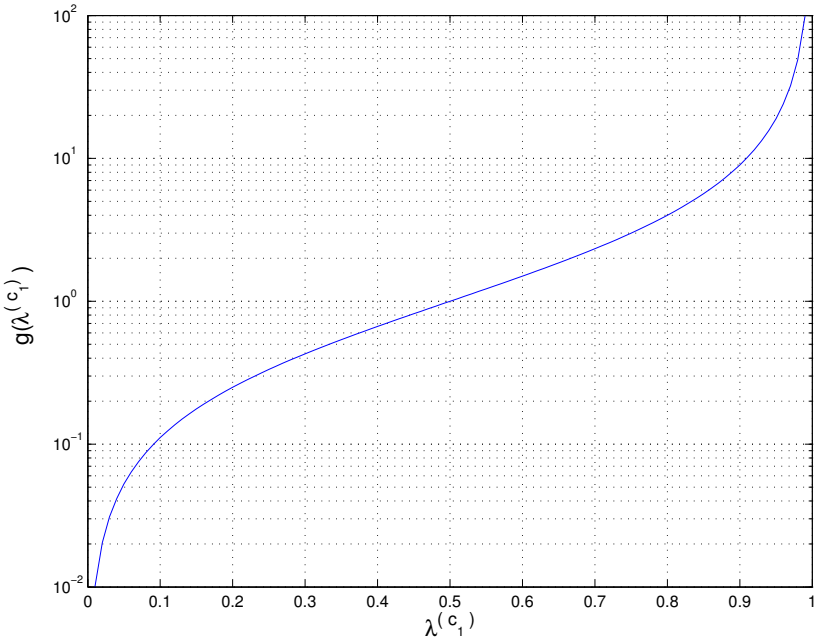


Figure 2. Illustration of the strictly monotonous ascending transformation $g(\cdot)$ that, through Equation (56), maps eigenvalues of the matrix pencil $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$ into the eigenvalues of the matrix pencil $(\mathbf{P}, \mathbf{Q})$, in a case where the sample probabilities of the classes are uniform $p(c_1) = p(c_2)$. Note that the eigenvalues of the first pencil are bounded in the interval $(0, 1)$, while the domain of the eigenvalues of the second pencil is $(0, \infty)$.

5.3. Linking the Optimization of the Divergence and the CSP Solution

There is a link between the solutions of the CSP method and the solutions obtained with the optimization of the symmetric KL divergence between the class conditional covariances, which was studied in previous works [1,11,16]. This subsection shows that under the appropriate filter selection criteria the link also extends to the optimization of other divergences, like the AB log-det family of divergences.

We have previously assumed that generalized eigenvalues are ordered and can be regarded as non-equal. Therefore, we can cluster them in the following three sets of principal, inner and minor eigenvalues of the matrix pencil $(\mathbf{P}, \mathbf{Q})$:

$$\underbrace{\lambda_1 > \cdots > \lambda_k}_{k \text{ principal eigenvalues}} > \underbrace{\lambda_{k+1} > \cdots > \lambda_{n-(p-k)}}_{\text{inner eigenvalues}} > \underbrace{\lambda_{n-(p-k)+1} > \cdots > \lambda_n}_{(p-k) \text{ minor eigenvalues}}. \tag{57}$$

The following sequence of optimizations induces an alternative ordering of the generalized eigenvalues

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_i} \| \, 1) = \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T \mathbf{P} \boldsymbol{w}_i \parallel \boldsymbol{w}_i^T \mathbf{Q} \boldsymbol{w}_i), \qquad i = 1, \ldots, n, \tag{58}$$

according to a permutation π that sorts their marginal divergences from 1 in descending order

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_1} \| \, 1) \geq \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_p} \| \, 1) \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_{p+1}} \| \, 1) \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_n} \| \, 1). \tag{59}$$

For building the matrix of spatial filters $\mathbf{W}_{Div} \equiv [\boldsymbol{w}_1, \boldsymbol{w}_2 ..... \boldsymbol{w}_p]$, one possible selection policy is to retain only the p most discriminative spatial filters for the considered divergence optimization problem, i.e., those that solve (58) for $i = 1, \ldots, p$. The filters consist in p eigenvectors ($\boldsymbol{v}_{\pi_i}$ with $i = 1, \ldots, p$)

of the matrix pencil $(\mathbf{P}, \mathbf{Q})$ that are arranged according to the permutation π. From the one-to-one relationship that exists between the generalized eigenvalues and eigenvectors of the matrix pencils $(\mathbf{P}, \mathbf{Q})$ and $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$ (see the previous subsection) the solution takes the following form

$$\mathbf{W}_{Div} = [\boldsymbol{v}_{\pi_1}, \ldots, \boldsymbol{v}_{\pi_p}] \quad (60)$$

$$= [\boldsymbol{v}_{\pi_1}^{(c_1)}, \ldots, \boldsymbol{v}_{\pi_p}^{(c_1)}]. \quad (61)$$

This result tells us that the optimization of different divergences (in absence of other regularizing terms) only differs in the selection criteria for the spatial filters, which eventually determine the chosen subindices $\pi_1, \ldots, \pi_p$.

Now, the question of whether these spatial filters that solve the sequence of minimax divergence optimization problems

$$\min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T\mathbf{P}\boldsymbol{w}_i \parallel \boldsymbol{w}_i^T\mathbf{Q}\boldsymbol{w}_i), \qquad i = 1, \ldots, p, \quad (62)$$

essentially coincide (up to a possible permutation in the order of the spatial filters) with the spatial filters of the CSP solution in (63)

$$\mathbf{W}_{CSP} = [\underbrace{\boldsymbol{v}_1^{(c_1)}, \ldots, \boldsymbol{v}_k^{(c_1)}}_{k \text{ principal eigenvectors}}, \underbrace{\boldsymbol{v}_{n-(p-k)+1}^{(c_1)}, \ldots, \boldsymbol{v}_n^{(c_1)}}_{p-k \text{ minor eigenvectors}}], \quad (63)$$

has a simple answer. The straightforward comparison between (61) and (63) reveals that both solutions should essentially coincide when the subindices $\pi_1, \ldots, \pi_p$ are a permutation of the integers $1, \ldots, k, n-(p-k)+1, \ldots, n$. Thus, the link between both techniques happens whenever CSP method adopts the filter selection policy of the divergence criterion in (59).

However, many of the CSP implementations find satisfactory to choose the number of spatial filters for each class a priori, respectively as k and $p-k$ (we will refer to this case as the original CSP filter selection policy), where k is close to $p/2$ in order to approximately balance the number of spatial filters for each class [13,21].

In general, the use of a divergence based selection policy does not ensure a balanced representation of the spatial filters for each of the classes. For instance, consider the synthetic but illustrative situation for $n = 100$, where we wish to select $p = 8$ spatial filters. If the generalized eigenvalues of the matrix pencil $(\mathbf{P}, \mathbf{Q})$ are shifted towards to zero, for instance, equal to $\{10, 0.99, 0.98, \ldots, 0.03, 0.02, 0.01\}$. In most cases, the solution $\mathbf{W}_{Div}$ will select as its columns: only $k = 1$ principal eigenvectors and $p - k = 7$ minor eigenvectors, an unbalanced choice.

In view of this potential limitation, an interesting question is whether it would be possible to modify the AB log-det divergence criterion so as to enforce that its solution essentially coincides with the one obtained by the CSP method with its original filter selection policy. We will show in the following that this requires only a suitable scaling $\kappa \in \mathbb{R}^+$ in one of the arguments of the divergence. Without loss of generality, we assume scaling in the second argument of the divergence. As it is shown in the Appendix A.2, there is a permutation π' of the indices of the spatial filters $1, \ldots, p$ that links the CSP solution in (24) with the optimization of the divergence

$$\boldsymbol{w}_{\pi_i'} = \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T\mathbf{P}\boldsymbol{w}_i \parallel \kappa\, \boldsymbol{w}_i^T\mathbf{Q}\boldsymbol{w}_i), \qquad i = 1, \ldots, p, \quad (64)$$

for any given

$$\kappa \in (\kappa_{\inf}, \kappa_{\sup}) \quad (65)$$

with

$$\kappa_{\text{inf}} \equiv \mathcal{K}(\lambda_{k+1}\,,\,\lambda_{n-(p-k)+1}) \tag{66}$$

$$\kappa_{\text{sup}} \equiv \mathcal{K}(\lambda_{k}\,,\,\lambda_{n-(p-k)}) \tag{67}$$

where the function

$$\mathcal{K}(a,b) = \begin{cases} \left(\dfrac{(a^{\beta}-b^{\beta})/\beta}{(a^{-\alpha}-b^{-\alpha})/(-\alpha)}\right)^{\frac{1}{\alpha+\beta}} & \text{for } \alpha,\beta,\alpha+\beta\neq 0 \\ \left(\dfrac{\log(a/b))}{(a^{-\alpha}-b^{-\alpha})/(-\alpha)}\right)^{\frac{1}{\alpha}} & \text{for } \alpha\neq 0,\ \beta=0 \\ \left(\dfrac{(a^{\beta}-b^{\beta})/\beta}{\log(a/b))}\right)^{\frac{1}{\beta}} & \text{for } \alpha=0,\ \beta\neq 0 \\ \exp\left(\dfrac{a^{\alpha}\log(eb^{\alpha})-b^{\alpha}\log(ea^{\alpha})}{\alpha(a^{\alpha}-b^{\alpha})}\right) & \text{for } \alpha=-\beta\neq 0 \\ \sqrt{a\,b} & \text{for } \alpha=\beta=0 \end{cases} \tag{68}$$

determines the value of the constant $\kappa = \mathcal{K}(a,b) \in \mathbb{R}$ that equalizes the value of the AB log-det divergences between any arbitrary $a, b \in \mathbb{R}$ constants (in the first argument) and κ (in the second argument), i.e.,

$$D_{AB}^{(\alpha,\beta)}(a\|\,\kappa) = D_{AB}^{(\alpha,\beta)}(b\|\,\kappa). \tag{69}$$

Note that the only role of the scaling factor κ is to adjust the reference value in one of the arguments of the divergence to ensure the exact balance in the number of spatial filters that are specialized in each class. As it is shown in the Appendix A.2, this scaling factor prevents that the minimax solution for $i = 1,\ldots,p$, could be attained by some eigenvectors associated with elements of the inner set of eigenvalues in (57), so the chosen subset of eigenvectors have to essentially coincide with the principal and minor eigenvectors that form the CSP solution in (63). In practice, a value of κ which is closer to unity and meets the required bounds can be obtained from the truncated choice

$$\kappa_{\star} = \begin{cases} \kappa_{\text{inf}}+\varepsilon & \text{for } \kappa_{\text{inf}} \geq 1 \\ 1 & \text{for } 1 \in (\kappa_{\text{inf}},\kappa_{sup}) \\ \kappa_{\text{sup}}-\varepsilon & \text{for } \kappa_{\text{sup}} \leq 1 \end{cases} \tag{70}$$

for an arbitrary small value of the constant $\epsilon \ll \kappa_{sup} - \kappa_{inf}$.

6. The Gradient of the AB Log-Det Divergence

The AB log-det divergence between the conditional covariance of the outputs $\mathbf{Y} = \mathbf{W}^T\boldsymbol{x}$ for each of the classes

$$f(\mathbf{W}) = D_{AB}^{(\alpha,\beta)}(Cov(\mathbf{Y}|c_1)\,\|\,Cov(\mathbf{Y}|c_2)) \tag{71}$$

$$= D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\|\mathbf{W}^T\mathbf{Q}\mathbf{W})\,, \tag{72}$$

is a function of the matrix $\mathbf{W} \in \mathbb{R}^{n\times p}$.

The optimization of this function with respect to $\mathbf{W}$ is non-trivial, so in this section, we show how the gradient of the AB log-det divergences can be derived. One may note that this is not only naturally interesting for the optimization that we would like to perform in this work, but it also

contributes to pave the way for the potential practical use of the AB log-det divergence in other scenarios and applications.

As we have shown previously, the divergence is separable

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\|\mathbf{W}^T\mathbf{Q}\mathbf{W}) \quad = \quad D_{AB}^{(\alpha,\beta)}(\mathbf{M}\|\mathbf{I}_p) = \sum_{i=1}^{p} D_{AB}^{(\alpha,\beta)}(\mu_i(\mathbf{W})\|\,1) \tag{73}$$

over the eigenvalues of the matrix

$$\mathbf{M} \quad = \quad (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{W}^T\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \tag{74}$$

$$\quad = \quad \mathbf{U}\,\mathrm{diag}\{\mu_1,\cdots,\mu_n\}\mathbf{U}^T \tag{75}$$

where $\mathrm{diag}\{\mu_1,\cdots,\mu_p\}$ and $\mathbf{U} = [\boldsymbol{u}_1,\cdots,\boldsymbol{u}_p]$, respectively denote the matrices of eigenvalues and eigenvectors of $\mathbf{M}$, which are functions of the matrix $\mathbf{W}$.

The differential of $f(\mathbf{W})$ can be expressed as

$$df(\mathbf{W}) \quad = \quad \mathrm{tr}\left\{d\mathbf{W}^T\,\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}}\right\} \tag{76}$$

where

$$\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} = \left[\frac{\partial f(\mathbf{W})}{\partial W_{ij}}\right]_{ij} \in \mathbb{R}^{n\times p} \tag{77}$$

denotes the gradient of the function. The divergence directly depends on the generalized eigenvalues, which in turn depend on the matrix $\mathbf{W}$. The suitable tool to obtain the gradient of this composition of functions is the chain rule, which can be written as

$$\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} \quad = \quad \sum_{i=1}^{p}\frac{\partial \mu_i}{\partial \mathbf{W}}\frac{\partial f(\mathbf{W})}{\partial \mu_i}. \tag{78}$$

So, the gradient can be evaluated after finding $\frac{\partial f(\mathbf{W})}{\partial \mu_i}$ and $\frac{\partial \mu_i}{\partial \mathbf{W}}$.

Since the divergence is a separable function of the generalized eigenvalues, the first term is easier to obtain,

$$\frac{\partial f(\mathbf{W})}{\partial \mu_i} = \frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_i\|\,1)}{\partial \mu_i} = \begin{cases} \dfrac{\mu_i^{\beta-1}-\mu_i^{-\alpha-1}}{\alpha\mu_i^{\beta}+\beta\mu_i^{-\alpha}} = \dfrac{\mu_i^{\alpha+\beta}-1}{\mu_i(\alpha\mu_i^{\alpha+\beta}+\beta)} & \text{for } \alpha+\beta\neq 0 \\ \dfrac{\log\mu_i}{\mu_i(1+\alpha\log\mu_i)} & \text{for } \alpha+\beta=0\,. \end{cases} \tag{79}$$

Obtaining the second term $\frac{\partial \mu_i}{\partial \mathbf{W}}$ is not so easy and requires to employ our previous plausible assumption that the generalized eigenvalues have a simple spectrum. Under this condition, the *Hadamard first variation formula* can be used to write the differential of the eigenvalues as

$$d\mu_i = \boldsymbol{u}_i^T d\mathbf{M}\boldsymbol{u}_i, \tag{80}$$

where $\boldsymbol{u}_i$ denotes the normalized eigenvector ($\|\boldsymbol{u}_i\|_2 = 1$) corresponding to each eigenvalue μ_i.

With the help of the product rule for differentials, we obtain

$$\begin{aligned} d\mathbf{M} &= d(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M} + \\ &\quad \mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\ d(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} + \\ &\quad (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ (d\mathbf{W}^T\mathbf{P}\mathbf{W} + \mathbf{W}^T\mathbf{P}d\mathbf{W})\ (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}. \end{aligned} \tag{81}$$

As we show in the Appendix A.3, it can be simplified as follows

$$\begin{aligned} d(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}} &= -\frac{1}{2}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ d(\mathbf{W}^T\mathbf{Q}\mathbf{W})\ (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} & (82)\\ &= -\frac{1}{2}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ (d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W})\ (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} & (83) \end{aligned}$$

hence

$$\begin{aligned} d\mathbf{M} &= -\frac{1}{2}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}(d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W})(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ \mathbf{M} \\ &\quad -\frac{1}{2}\left[(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}(d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W})(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ \mathbf{M}\right]^T \\ &\quad +(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ (d\mathbf{W}^T\mathbf{P}\mathbf{W} + \mathbf{W}^T\mathbf{P}d\mathbf{W})\ (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}. \end{aligned} \tag{84}$$

Thus, after substituting (84) in (80) and using the invariance of the trace under transpositions ($\mathrm{tr}\{\mathbf{A}\} = \mathrm{tr}\{\mathbf{A}^T\}$) and the cyclic shifts ($\mathrm{tr}\{\mathbf{AB}\} = \mathrm{tr}\{\mathbf{BA}\}$), the following values are obtained

$$\begin{aligned} d\mu_i &= \boldsymbol{u}_i^T d\mathbf{M}\boldsymbol{u}_i & (85)\\ &= \mathrm{tr}\left\{\boldsymbol{u}_i^T d\mathbf{M}\ \boldsymbol{u}_i\right\} & (86)\\ &= -\frac{1}{2}\,\mathrm{tr}\left\{\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}(d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W})(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ \mathbf{M}\boldsymbol{u}_i\right\} \\ &\quad -\frac{1}{2}\,\mathrm{tr}\left\{\boldsymbol{u}_i^T\left[(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}(d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W})(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ \mathbf{M}\right]^T\boldsymbol{u}_i\right\} \\ &\quad +\mathrm{tr}\left\{\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ (d\mathbf{W}^T\mathbf{P}\mathbf{W} + \mathbf{W}^T\mathbf{P}d\mathbf{W})\ (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\boldsymbol{u}_i\right\} & (87)\\ &= -2\,\mathrm{tr}\left\{d\mathbf{W}^T\ \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{M}\ \boldsymbol{u}_i\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\right\} \\ &\quad +2\,\mathrm{tr}\left\{d\mathbf{W}^T\ \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\boldsymbol{u}_i\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\right\}. & (88) \end{aligned}$$

At this point, we can use the identity for the differential

$$d\mu_i = \mathrm{tr}\left\{d\mathbf{W}^T\,\frac{\partial\mu_i}{\partial\mathbf{W}}\right\} \tag{89}$$

in (88) to identify the second desired term

$$\frac{\partial\mu_i}{\partial\mathbf{W}} = -2\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{M}\ \boldsymbol{u}_i\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} + 2\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\boldsymbol{u}_i\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}. \tag{90}$$

Substituting the expressions (79) and (90) in (78), we obtain

$$\begin{aligned} \frac{\partial f(\mathbf{W})}{\partial\mathbf{W}} &= \sum_{i=1}^{p}\frac{\partial\mu_i}{\partial\mathbf{W}}\frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_i\|\ 1)}{\partial\mu_i} \\ &= -2\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{M}\ \mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} + 2\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \end{aligned} \tag{91}$$

where, for convenience, the matrix is defined as following

$$\mathbf{Z} = \sum_{i=1}^{p} \boldsymbol{u}_i \frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_i \| \, 1)}{\partial \mu_i} \boldsymbol{u}_i^T \tag{92}$$

$$= \mathbf{U} \, \mathrm{diag} \left\{ \frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_1 \| \, 1)}{\partial \mu_1}, \ldots, \frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_p \| \, 1)}{\partial \mu_p} \right\} \mathbf{U}^T. \tag{93}$$

The matrix $\mathbf{Z}$ can also be represented directly in terms of the matrix $\mathbf{M}$ (which we have defined previously in Equation (74)) as

$$\mathbf{Z} = \begin{cases} \mathbf{M}^{-1}(\alpha \mathbf{M}^{\alpha+\beta} + \beta \mathbf{I})^{-1}(\mathbf{M}^{\alpha+\beta} - \mathbf{I}) & \text{for } \alpha + \beta \neq 0 \\ \mathbf{M}^{-1}((\log \mathbf{M})^{-1} + \alpha \mathbf{I})^{-1} & \text{for } \alpha + \beta = 0 \end{cases} \tag{94}$$

where $\log(\cdot)$ for matrix arguments denotes the matrix logarithm functional. After the grouping of common terms in (91) we obtain the final gradient expression, which is given by

$$\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} = 2[\mathbf{P}\mathbf{W} - \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{P}\mathbf{W})](\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \mathbf{Z} (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}. \tag{95}$$

6.1. Validation of Equation (95) with the Gradient of the KL Divergence

The Kullback–Leibler (KL) divergence between the Gaussian densities $p(\boldsymbol{x}|c_2)$ and the $p(\boldsymbol{x}|c_1)$, of zero mean and the respective covariance matrices $Cov(\mathbf{Y}|c_1)$ and $Cov(\mathbf{Y}|c_2)$, is given by

$$Div_{KL}(p(\boldsymbol{x}|c_2) \parallel p(\boldsymbol{x}|c_1)) = \int p(\boldsymbol{x}|c_2) \log \frac{p(\boldsymbol{x}|c_2)}{p(\boldsymbol{x}|c_1)} d\boldsymbol{x} \tag{96}$$

$$= \tfrac{1}{2} \log |Cov(\mathbf{Y}|c_1)| - \tfrac{1}{2} \log |Cov(\mathbf{Y}|c_2)| + \tfrac{1}{2} \, \mathrm{tr}\{Cov^{-1}(\mathbf{Y}|c_1) Cov(\mathbf{Y}|c_2) - \mathbf{I}_p\}.$$

Since this divergence only involves trace and log-det operators, as it is shown in the Appendix A.4, its gradient with respect to $\mathbf{W}$, i.e.,

$$\frac{\partial}{\partial \mathbf{W}} Div_{KL}(p(\boldsymbol{x}|c_2) \parallel p(\boldsymbol{x}|c_1)) = -\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} + \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} + \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}$$
$$+ \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}, \tag{97}$$

is relatively easy to obtain. Then, we can use the fact that the KL divergence is proportional to the AB log-det divergence between the class conditional covariance matrices, as long as the conditional covariance matrices appear in the AB log-det divergence interchanged in position with respect to class conditional density arguments of the KL divergence. So for the specific case of $\alpha = 1$ and $\beta = 0$, i.e.,

$$D_{AB}^{(1,0)}(Cov(\mathbf{Y}|c_1) \parallel Cov(\mathbf{Y}|c_2)) = 2 \, Div_{KL}(p(\boldsymbol{x}|c_2) \parallel p(\boldsymbol{x}|c_1)), \tag{98}$$

to test whether there is coherence between the obtained gradient formula in (95) and twice the gradient of the KL divergence that was independently obtained in the Appendix A.4. For this purpose, in the specific case of $\alpha = 1$ and $\beta = 0$, from (94) the following auxiliary matrices are evaluated

$$\mathbf{Z} = \mathbf{M}^{-1}(\mathbf{I}_p - \mathbf{M}^{-1}) \tag{99}$$

$$(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \mathbf{Z} (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} = (\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} - (\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \tag{100}$$

and are substituted in the expression of the gradient of the AB log-det divergence (95). After the following straightforward simplifications,

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{W}} D_{AB}^{(1,0)}(Cov(\mathbf{Y}|c_1) \parallel Cov(\mathbf{Y}|c_2)) &= 2[\mathbf{PW} - \mathbf{QW}(\mathbf{W}^T\mathbf{QW})^{-1}(\mathbf{W}^T\mathbf{PW})] \\
&\quad \times[(\mathbf{W}^T\mathbf{QW})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{QW})^{-\frac{1}{2}}] \qquad (101)\\
&= 2[-\mathbf{QW}(\mathbf{W}^T\mathbf{QW})^{-1}(\mathbf{W}^T\mathbf{PW}) + \mathbf{PW}] \\
&\quad \times[(\mathbf{W}^T\mathbf{PW})^{-1} - (\mathbf{W}^T\mathbf{PW})^{-1}(\mathbf{W}^T\mathbf{QW})(\mathbf{W}^T\mathbf{PW})^{-1}] \qquad (102)\\
&= 2\Big[-\mathbf{QW}(\mathbf{W}^T\mathbf{QW})^{-1} + \mathbf{PW}(\mathbf{W}^T\mathbf{PW})^{-1} + \mathbf{QW}(\mathbf{W}^T\mathbf{PW})^{-1} \\
&\quad + \mathbf{PW}(\mathbf{W}^T\mathbf{PW})^{-1}(\mathbf{W}^T\mathbf{QW})^{-1}(\mathbf{W}^T\mathbf{PW})^{-1}\Big] \qquad (103)\\
&= 2\,\frac{\partial}{\partial \mathbf{W}} Div_{KL}(p(\boldsymbol{x}|c_2) \parallel p(\boldsymbol{x}|c_1))\,. \qquad (104)
\end{aligned}$$

the proportionality between the gradient of $D_{AB}^{(1,0)}(Cov(\mathbf{Y}|c_1)\|Cov(\mathbf{Y}|c_2))$ and the gradient of the KL divergence in (97) is confirmed.

6.2. Validation of Equation (95) with the Gradient of the AG Divergence

The Alpha–Gamma divergence between the Gaussian densities $p(\boldsymbol{x}|c_2)$ and $p(\boldsymbol{x}|c_1)$, of zero mean and with respective covariance matrices $Cov(\mathbf{Y}|c_1) = \mathbf{W}^T\mathbf{PW}$ and $Cov(\mathbf{Y}|c_2) = \mathbf{W}^T\mathbf{QW}$, is equal to

$$\begin{aligned}
Div_{AG}^{(\alpha,\beta)}\left(p(y_i|c_2)\|p(y_i|c_1)\right) &\equiv \frac{1}{\alpha\beta}\log\frac{\left(\int_\Omega p(y_i|c_1)\,dy_i\right)^\beta\left(\int p(y_i|c_2)\,dy_i\right)^\alpha}{\int p^\beta(y_i|c_1)\,p^\alpha(y_i|c_2)\,dy_i} \qquad (105)\\
&= \tfrac{1}{2\alpha\beta}\log|\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W}| - \tfrac{1}{2\beta}\log|\mathbf{W}^T\mathbf{PW}| - \tfrac{1}{2\alpha}\log|\mathbf{W}^T\mathbf{QW}| \\
&\quad \text{for} \quad \alpha>0,\ \beta>0,\ \alpha+\beta=1. \qquad (106)
\end{aligned}$$

Due to the constraint $\alpha+\beta=1$, we assume that β is determined by α, i.e., $\beta = 1-\alpha$ along this subsection. Since

$$\nabla_{\mathbf{W}}\log|(\mathbf{W}^T\mathbf{PW})| = 2\mathbf{PW}(\mathbf{W}^T\mathbf{PW})^{-1}, \qquad (107)$$

the gradient of the AG divergence with respect to $\mathbf{W}$ is given by

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{W}} Div_{AG}^{(\alpha,\beta)}(p(\boldsymbol{x}|c_2) \parallel p(\boldsymbol{x}|c_1)) &= \tfrac{2}{2\alpha\beta}(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W}\left(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W}\right)^{-1} \qquad (108)\\
&\quad -\tfrac{2}{2\alpha}\mathbf{QW}(\mathbf{W}^T\mathbf{QW})^{-1} - \tfrac{2}{2\beta}\mathbf{PW}(\mathbf{W}^T\mathbf{PW})^{-1} \\
&= -\tfrac{1}{\beta}\mathbf{PW}[(\mathbf{W}^T\mathbf{PW})^{-1} - \left(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W}\right)^{-1}] \\
&\quad -\tfrac{1}{\alpha}\mathbf{QW}[(\mathbf{W}^T\mathbf{QW})^{-1} - \left(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W}\right)^{-1}].
\end{aligned}$$

Then, we can use the equivalence between the AG divergence and the AB log-det divergence between the class conditional covariance matrices

$$D_{AB}^{(\alpha,\beta)}(Cov(\mathbf{Y}|c_1) \parallel Cov(\mathbf{Y}|c_2)) = 2\,Div_{AG}^{(\alpha,\beta)}\left(p(y_i|c_2)\|p(y_i|c_1)\right), \qquad (109)$$

which is valid for the specific case of $\alpha+\beta=1$ and $\alpha,\beta>0$, to also test the coherence between the obtained gradient formula in (95) and twice the gradient of the AG divergence. For $\alpha+\beta=1$, the auxiliary matrices in the definition of the gradient are

$$\mathbf{Z} = (\alpha\mathbf{M}+\beta\mathbf{I})^{-1}[\mathbf{M}^{-1}(\mathbf{M}-\mathbf{I})] = (\alpha\mathbf{M}+\beta\mathbf{I})^{-1} - (\alpha\mathbf{M}^2+\beta\mathbf{M})^{-1} \qquad (110)$$

and

$$
\begin{aligned}
(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} &= (\alpha(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}+\beta\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} \\
&\quad -(\alpha(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}} \\
&\quad +\beta(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}})^{-1} \qquad (111)\\
&= [\mathbf{I}-(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})](\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1}\,. \qquad (112)
\end{aligned}
$$

After substituting this last expression in the gradient of the AB log-det divergence (95), we obtain

$$
\begin{aligned}
&\frac{\partial}{\partial\mathbf{W}}D_{AB}^{(\alpha,\beta)}(Cov(\mathbf{Y}|c_1)\,\|\,Cov(\mathbf{Y}|c_2)) \\
&\qquad = 2[\mathbf{P}\mathbf{W}-\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{P}\mathbf{W})](\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \\
&\qquad = +2(\mathbf{P}\mathbf{W}+\mathbf{Q}\mathbf{W})(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1} \\
&\qquad\quad -\frac{2}{\beta}\alpha\mathbf{P}\mathbf{W}[(\mathbf{W}^T\alpha\mathbf{P}\mathbf{W})+(\mathbf{W}^T\alpha\mathbf{P}\mathbf{W})(\mathbf{W}^T\beta\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\alpha\mathbf{P}\mathbf{W})]^{-1} \\
&\qquad\quad -\frac{2}{\alpha}\beta\mathbf{Q}\mathbf{W}[(\mathbf{W}^T\beta\mathbf{Q}\mathbf{W})+(\mathbf{W}^T\beta\mathbf{Q}\mathbf{W})(\mathbf{W}^T\alpha\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\beta\mathbf{Q}\mathbf{W})]^{-1}\,.
\end{aligned} \qquad (113)
$$

With the help of the particular form of the Woodbury identity for the matrix inverse

$$
[\mathbf{A}+\mathbf{A}\mathbf{B}^{-1}\mathbf{A}]^{-1}=\mathbf{A}^{-1}-(\mathbf{A}+\mathbf{B})^{-1} \qquad (114)
$$

we simplify the terms within the brackets. Finally, we use the fact that $\alpha+\beta=1$ to confirm the proportionality with the gradient of the AG divergence given in (108),

$$
\begin{aligned}
\frac{\partial}{\partial\mathbf{W}}D_{AB}^{(\alpha,\beta)}(Cov(\mathbf{Y}|c_1)\,\|\,Cov(\mathbf{Y}|c_2)) &= +2(\mathbf{P}\mathbf{W}+\mathbf{Q}\mathbf{W})(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1} \\
&\quad -\frac{2}{\beta}\alpha\mathbf{P}\mathbf{W}[(\mathbf{W}^T\alpha\mathbf{P}\mathbf{W})^{-1}-(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1}] \\
&\quad -\frac{2}{\alpha}\beta\mathbf{Q}\mathbf{W}[(\mathbf{W}^T\beta\mathbf{Q}\mathbf{W})^{-1}-(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1}] \qquad (115)\\
&= +2(\mathbf{P}\mathbf{W}+\mathbf{Q}\mathbf{W})(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1} \\
&\quad -\frac{2}{\beta}\mathbf{P}\mathbf{W}[(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}-\alpha(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1}] \\
&\quad -\frac{2}{\alpha}\mathbf{Q}\mathbf{W}[(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}-\beta(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1}] \qquad (116)\\
&= +2((1+\tfrac{\alpha}{\beta})\mathbf{P}\mathbf{W}+(1+\tfrac{\beta}{\alpha})\mathbf{Q}\mathbf{W})(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1} \\
&\quad -\frac{2}{\beta}\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}-\frac{2}{\alpha}\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} \qquad (117)\\
&= +2(\tfrac{1}{\beta}\mathbf{P}\mathbf{W}+\tfrac{1}{\alpha}\mathbf{Q}\mathbf{W})(\mathbf{W}^T(\alpha\mathbf{P}+\beta\mathbf{Q})\mathbf{W})^{-1} \\
&\quad -\frac{2}{\beta}\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}-\frac{2}{\alpha}\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} \qquad (118)\\
&= 2\frac{\partial}{\partial\mathbf{W}}Div_{AG}^{(\alpha,\beta)}(p(\boldsymbol{x}|c_2)\,\|\,p(\boldsymbol{x}|c_1))\,. \qquad (119)
\end{aligned}
$$

7. Robustness of the AB Log-Det Divergence in Terms of α and β

The squared Riemann metric is known to be the natural distance in the manifold of SPD matrices, as it measures the squared length of the geodesic path between the arguments of the divergence [3]. However, in the real data there are usually several model contaminations (mismatches), including outliers or artifacts, that could make other robust divergences preferable. In this section, we study how the hyperparameters α and β can influence robustness of the AB log-det divergence with respect to the behavior of the squared Riemann metric, which is used as a reference.

For convenience, we denote the AB log-det divergence as a function of the spatial filter matrix $\mathbf{W}$ by

$$f_{(\alpha,\beta)}(\mathbf{W}) \equiv D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\|\mathbf{W}^T\mathbf{Q}\mathbf{W}), \tag{120}$$

and we consider its gradient expression given by Equation (78). The spatial filters that maximize this divergence should satisfy the following estimating equations

$$\frac{\partial f_{(\alpha,\beta)}(\mathbf{W})}{\partial \mathbf{W}} = \sum_{i=1}^{p} \frac{\partial \mu_i}{\partial \mathbf{W}} \psi_{(\alpha,\beta)}(\mu_i) = 0, \tag{121}$$

where μ_i, $i = 1, \ldots, p$, are the eigenvalues of matrix $\mathbf{M}$, which was defined in Equation (74), and

$$\psi_{(\alpha,\beta)}(\mu_i) = \frac{\partial f_{(\alpha,\beta)}(\mathbf{W})}{\partial \mu_i}, \quad i = 1, \ldots, p, \tag{122}$$

may be regarded as influence functions for each pair (α, β) that account for the penalty variation in the divergence with respect to μ_i. The complementary term to $\psi_{(\alpha,\beta)}(\mu_i)$ in (121), i.e., $\frac{\partial \mu_i}{\partial \mathbf{W}}$, is a matrix of partial derivatives of the generalized eigenvalues μ_i with respect to the elements of the spatial filters $\mathbf{W}$ and, therefore, it is independent of the considered divergence. It is easy to observe that, in the particular case of $\alpha = \beta = 0$, the expression in (121) represents the estimating equation for the squared Riemann metric

$$\frac{\partial f_{(0,0)}(\mathbf{W})}{\partial \mathbf{W}} = \sum_{i=1}^{p} \frac{\partial \mu_i}{\partial \mathbf{W}} \psi_{(0,0)}(\mu_i) = 0. \tag{123}$$

In order to study the relative robusness to outliers, one can rewrite the estimating equation for a chosen pair of hyperparameters (α, β) in terms of the influence function for the squared Riemannian metric as

$$\frac{\partial f_{(\alpha,\beta)}(\mathbf{W})}{\partial \mathbf{W}} = \sum_{i=1}^{p} \left(\frac{\partial \mu_i}{\partial \mathbf{W}} \psi_{(0,0)}(\mu_i) \right) w_{(\alpha,\beta)}(\mu_i) = 0, \tag{124}$$

where the scalar term

$$w_{(\alpha,\beta)}(\mu) = \frac{\psi_{(\alpha,\beta)}(\mu)}{\psi_{(0,0)}(\mu)} \tag{125}$$

acts as a weight function that controls, for a given pair (α, β), the magnitude of the effect in the estimation equation of departures of μ_i from unity.

The presence of outliers in the real data, typically results in eigenvalues μ_i that are too far from unity. However, depending on the problem, the higher prevalence of outliers may be stronger only for the greatest eigenvalues, or for the smallest eigenvalues, or simultaneously for the greatest and smaller eigenvalues. Those hyperparameters (α, β) that are able to down-weight the contribution of the outliers, are considered more robust. Therefore, the shape of the weight functions $w_{(\alpha,\beta)}(\mu_i)$ is useful to study the relative immunity of the AB log-det divergence to outliers.

Figure 3a shows the squared Riemannian metric ($\alpha = \beta = 0$) and its weight function, which is flat since this divergence is taken as reference. Figure 3b presents a similar plot for the Power Log-det divergence with $\alpha = \beta = 1$. In this case, the bell shape of the weight function is an indicator of the robustness with respect to the presence of outliers in the greatest and smallest eigenvalues, since they will be down-weighted in the estimating Equation (124). Similar plots can be done by increasing the magnitude of $\alpha = \beta$, which progressively enhances the robustness. When $\alpha \neq \beta$ the divergence is asymmetric. The Figure 4a,b respectively present the Kullback–Leibler divergence for SPD matrices ($\alpha = 1, \beta = 0$) and its dual version ($\alpha = 0, \beta = 1$), together with their associated weight functions. These plots illustrate the asymmetric cases in situations where $\alpha + \beta > 0$ and reveal that, when $\alpha \gg \beta$,

the AB log-det divergences tend to be more robust against outliers in the large eigenvalues while, for $\alpha \ll \beta$, the robustness tends to be with respect to the outliers in small eigenvalues.

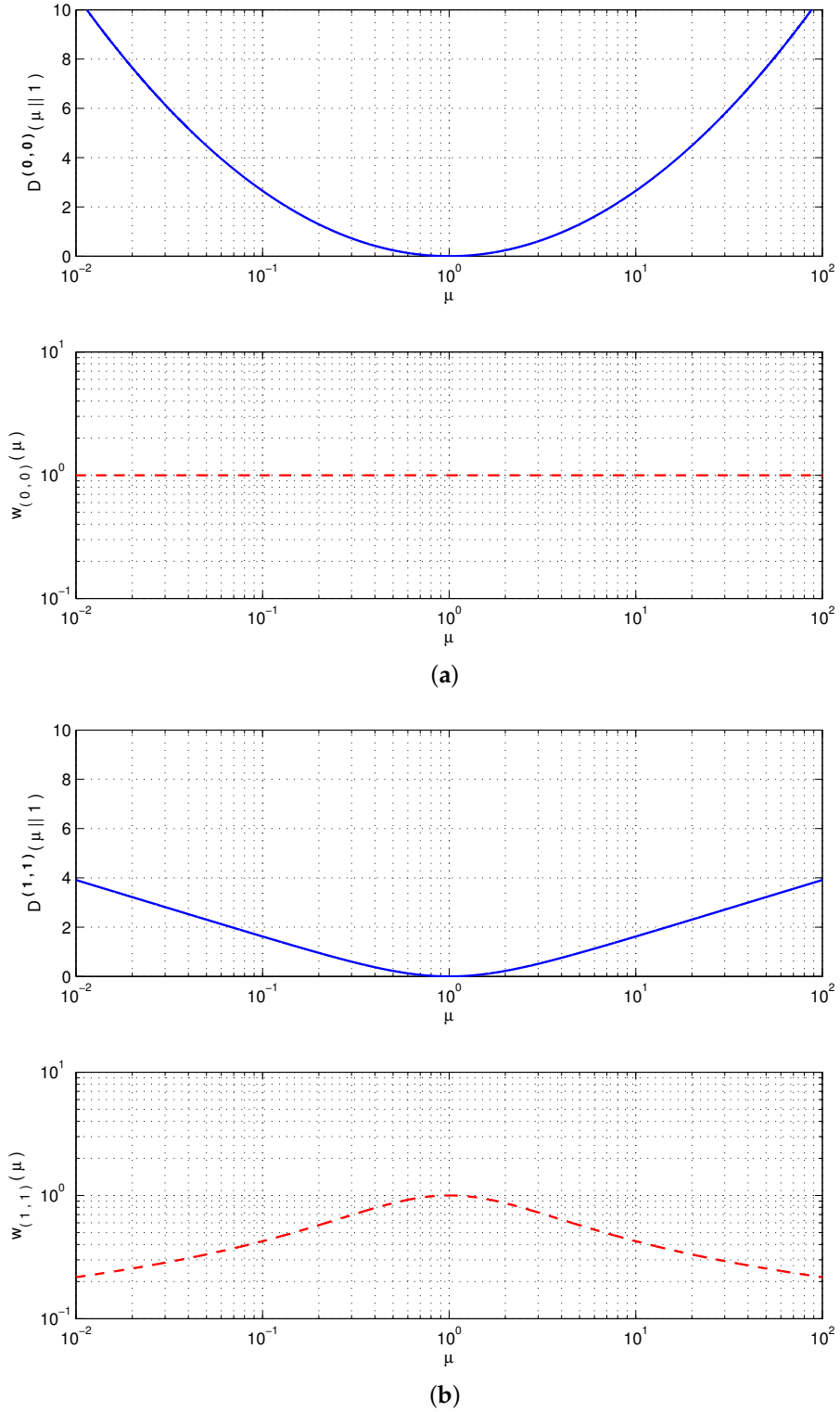
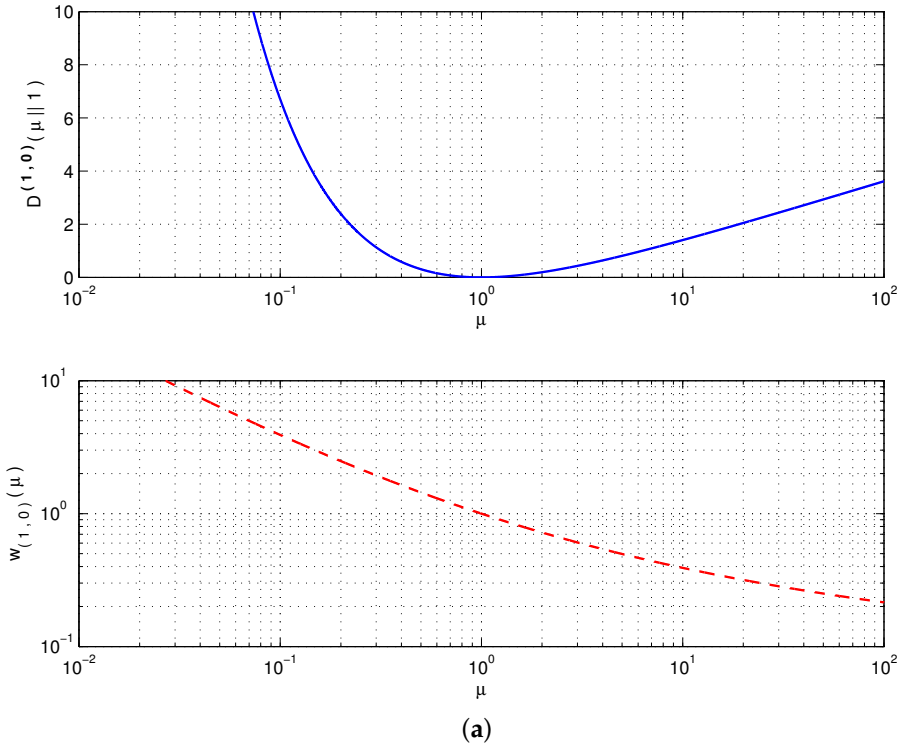
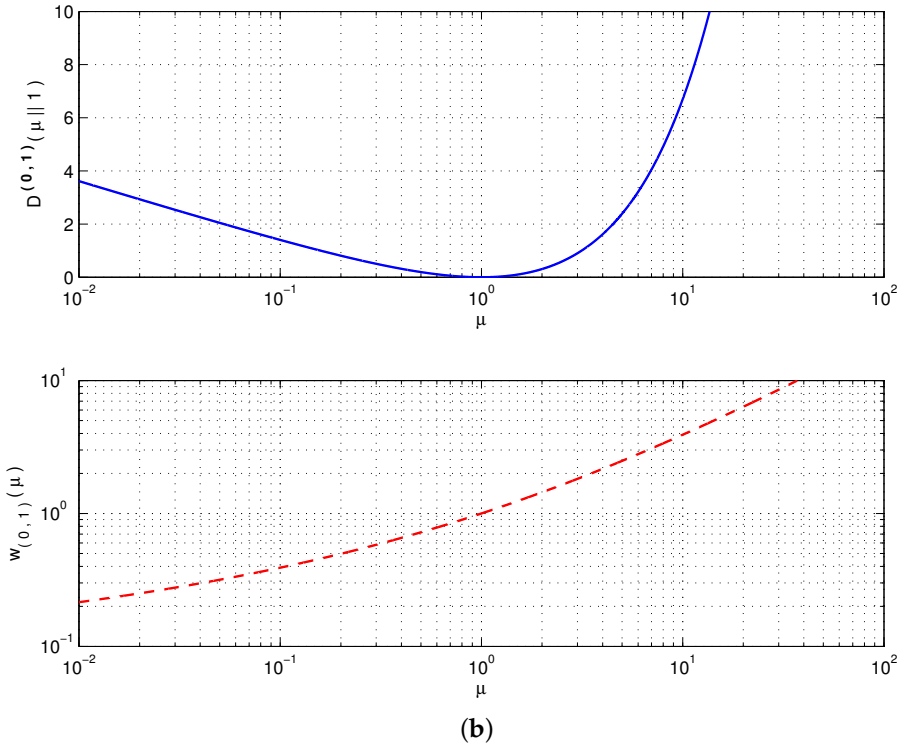


Figure 3. Illustration of the behavior of the AB log-det divergence $D_{AB}^{(\alpha,\beta)}(\mu, 1)$, and of its associated weight function $w_{\alpha,\beta}(\mu)$, versus μ for different values of $\alpha = \beta$. Note that μ is shown in log-scale. (**a**) Squared Riemannian metric for $\alpha = \beta = 0$ (upper plot) and its weight function (lower plot); (**b**) Power Log-det divergence for $\alpha = \beta = 1$ (upper plot) and its weight function (lower plot).



(**a**)



(**b**)

Figure 4. Illustration of the behavior of the AB log-det divergence $D_{AB}^{(\alpha,\beta)}(\mu,1)$, and of its associated weight function $w_{\alpha,\beta}(\mu)$, versus μ for different values of $\alpha \neq \beta$. Note that μ is shown in log-scale. (**a**) Kullback–Leibler (KL) positive definite matrix divergence for $\alpha = 1, \beta = 0$, and its weight function (lower plot); (**b**) Dual KL positive definite matrix div. for $\alpha = 0, \beta = 1$, and its weight function (lower plot).

8. Review of Some Related Techniques for the Spatial Filtering of Motor Imagery Movements

In this section, we will review the regularized variants of CSP that have been proposed to improve the classification performance. The regularization approaces of CSP are mainly done either in the estimation of the covariance matrices or by modifying the CSP objective function.

Most of them combine the estimation of the covariance matrices for each class with the regularization of the CSP objective function using penalty terms. Some of the approaches include the previous information [22], other subject data [23,24] and previous session data [25] for estimating the class covariance matrix. Another approach used M-estimators to compute the robust class covariance matrices [26] and yet another approach obtained the covariance matrices by finding the minimum squared error [27]. The authors of [28] applied Multiple Kernel Learning (MKL) to combine the information from different subjects.

It has been shown in [29] that the regularization of the objective function is more useful than regularizing the estimated covariance matrix. Several approaches have been proposed by regularizing the objective function. The authors of [21] have additionally incorporated the electrooculogram (EOG) signals for reducing the ocular artifacts. Other authors have tried to ensure robustness by selecting only the important channels and produce sparse spatial filters [30–32]. Another approach is to robustify the system by obtaining only the stationary features. A robustify maximin CSP method was proposed that used a set of covariance matrices instead of an individual covariance matrix without using any other user data or data from the previous sessions [33,34]. In order to avoid the presence of the outlier, the CSP objective function has been formulated using l_p-norm in [35,36]. The Stationary Subspace Analysis (SSA) algorithm was proposed to obtain the stationary subspaces of the time series EEG signals by considering only the stationary components of the signals. The limitation of this method is the detection of dissimilarity of the different class as a non-stationary feature [37]. The group wise SSA (gwSSA) algorithm aims at obtaining the non-stationarities by dividing the dataset into different groups and calculating the minimum KL divergence between estimated source distribution of each trial in a group and the average distribution of the corresponding group. This algorithm not only allows the combining of the multisubject data but also the multiclass data [38]. But, the gwSSA algorithm cannot find the discriminative information between the classes. The same group proposed a new approach for extracting the discriminative information, by subtracting the inter class divergences from the gwSSA objective function [39]. To overcome the limitation of the SSA algorithm, two-step approaches have been proposed where the initial extraction of the stationary sources was done using the SSA method and later, the CSP was used for the computation of the spatial filters[40]. Another approach to extract the stationary features is to reduce the nonstationarities between the two sessions. The supervised and unsupervised methods for adaptation of the data space have been proposed using KL divergence between the intersession data [41]. Recently, the authors of [14] presented maximum a posteriori-CSP (MAP-CSP) algorithm by deriving the probabilistic model of CSP to resolve the issue of overfitting of the baseline CSP algorithm.

One of the limitations of the CSP algorithm is that it is mainly suitable only for the discrimination of two classes, while, in general, for an efficient BCI system more than two motor imagery movements are required. In order to formulate it for the multiclass system, the authors of [42,43] have reduced the multiclass problem to a binary problem. The authors of [44] proposed two approaches for the multiclass problem; firstly to find the spatial filters for one class with respect to all the other classes and secondly, by simultaneous diagonalization methods. Other approaches, like [45], proposed to solve the multiclass problems by combining information theoretic criteria with joint diagonalization methods. Several other methods have been proposed for the multiclass paradigm using independent component analysis [46] and Riemannian geometry to obtain the spatial filters [47]. The authors of [48] derived a relation between Bayes classification error and Rayleigh quotient and used this approach to solve the multiclass problem. In spite of all these different approaches, the performance of MI based BCI systems is degraded due to the presence of non-stationarities and outliers, which is a challenge

for the BCI systems in a real application. Hence, a robust feature extraction algorithm is needed to increase the overall performance of the system.

9. Proposed Criterion and Algorithm for Spatial Filtering

For the presentation of the proposed criterion some additional notation needs to be defined. Let $\tilde{\boldsymbol{x}}^{(j)}(t)|c$ denote the output of the passband filtering of the raw observations at time t and for the jth trial of class $c \in \{c_1, c_2\}$. The power of the trials of a given class c is normalized by the operation

$$\boldsymbol{x}^{(j)}(t) = \frac{\tilde{\boldsymbol{x}}^{(j)}(t)}{\sqrt{\mathrm{tr}\{Cov(\tilde{\boldsymbol{x}}^{(j)}|c)\}}}, \tag{126}$$

where

$$Cov(\boldsymbol{x}^{(j)}|c) = \frac{1}{L}\sum_{t=1}^{L}\left(\boldsymbol{x}^{(j)}(t) - \bar{\boldsymbol{x}}^{(j)}\right)\left(\boldsymbol{x}^{(j)}(t) - \bar{\boldsymbol{x}}^{(j)}\right)^T \quad \text{with} \quad \bar{\boldsymbol{x}}^{(j)} = \frac{1}{L}\sum_{t=1}^{L}\boldsymbol{x}^{(j)}(t) \tag{127}$$

denotes the sample covariance matrix the j^{th} trial $\boldsymbol{x}^{(j)}$ of class c, and L is the size in samples of each trial. In order to simplify the notation, the covariance matrices of the two classes are renamed as

$$\mathbf{P}_j \equiv Cov(\boldsymbol{x}^{(j)}|c_1) \quad \text{and} \quad \mathbf{Q}_j \equiv Cov(\boldsymbol{x}^{(j)}|c_2), \tag{128}$$

and their averaged versions (the centroids of each class) are denoted as

$$\mathbf{P} \equiv <\mathbf{P}_j> = \frac{1}{N_1}\sum_{j=1}^{N_1}\mathbf{P}_j \quad \text{and} \quad \mathbf{Q} \equiv <\mathbf{Q}_j> = \frac{1}{N_2}\sum_{j=1}^{N_2}\mathbf{Q}_j. \tag{129}$$

The classification of imagery movements involves extracting the relevant features of the observations and the classification of the observed patterns in the feature space. In the considered application, the data is high-dimensional but only a few features are sufficient to capture the discriminative information about the intended movements. Thus, the extraction of the relevant features involves a dimensionality reduction step for the observations from $\mathbb{R}^n$ to $\mathbb{R}^p$ where $p \ll n$. This step is implemented through the spatial filtering, i.e., by projecting the n-dimensional observations onto a p-dimensional subspace which should allow a good discrimination of the cluster centroids and, at the same time, guarantee a compact representation of the clusters.

As mentioned earlier, the CSP solution will be obtained by a minimax optimization of the divergence between the projected and scaled centroids of the classes, i.e., $D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T\mathbf{P}\boldsymbol{w}_i \parallel \kappa\, \boldsymbol{w}_i^T\mathbf{Q}\boldsymbol{w}_i)$. However, since this solution completely ignores the within-class dispersion of the samples, it is quite sensitive to artifact and outlier in the training dataset. In similarity with the divergence framework presented in [1] and with some variants of Fisher LDA, p. 366 in [49], one can regularize the previous problem by controlling the dispersion of the trials of each class around their centroids and also by exploiting the degrees of freedom in the selection of the hyperparameters of the divergences. Then, a robust criterion based on the AB log-det divergence takes the following form

$$F(\mathbf{W}) \quad = \quad D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W} \| \,\kappa\, \mathbf{W}^T\mathbf{Q}\mathbf{W}) - \eta\,(p(c_1)\mathbf{R}_1 + p(c_2)\mathbf{R}_2)\ , \tag{130}$$

where the penalties associated to the within-class dispersion involve the averaged divergences

$$\mathbf{R}_1 = \frac{1}{N_1}\sum_{j=1}^{N_1} D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}_j\mathbf{W}\|\mathbf{W}^T\mathbf{P}\mathbf{W}), \tag{131}$$

$$\mathbf{R}_2 = \frac{1}{N_2}\sum_{j=1}^{N_2} D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{Q}_j\mathbf{W}\|\mathbf{W}^T\mathbf{Q}\mathbf{W}), \tag{132}$$

and the parameter $\eta \in \mathbb{R}^+$ controls the balance between the maximization of the between-class scatter and the minimization of the within-class scatter. Note that in (132) we have used the fact that the AB log-det divergence is invariant under the common scaling of its arguments, to simplify $D_{AB}^{(\alpha,\beta)}(\kappa\mathbf{W}^T\mathbf{Q}_j\mathbf{W}\|\kappa\mathbf{W}^T\mathbf{Q}\mathbf{W}) = D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{Q}_j\mathbf{W}\|\mathbf{W}^T\mathbf{Q}\mathbf{W})$.

The optimization of the criterion in (130) can be performed simultaneously, for all the spatial filters, with the use of subspace techniques [1]. In the next section, we present a subspace optimization algorithm based on AB log-det divergences.

The Subspace Optimization Algorithm (Sub-ABLD)

The subspace method aims to extract the desired set of p spatial filters in two steps. The idea is to first use a robust method to determine the discriminative subspace of the spatial filters, for instance, considering the optimization of a robust criterion like (130). Later, another criterion is used to identify the individual spatial filters within the subspace. Since the influence of outliers on the solution is significantly reduced after the discriminative subspace is determined. In the second step, the standard CSP criterion can be safely used to determine the final spatial directions within the chosen subspace.

The input parameters of the subspace optimization algorithm based on AB log-det divergences (Sub-ABLD) are the set of covariance matrices for each class ($\mathbf{P}_j$, $\mathbf{Q}_j$), the dimension of subspace to be extracted p, and the hyperparameters α, β and η. The method starts with the computation of the sample prior probabilities as well as the average covariance matrices for each class, i.e., $p(c_1), p(c_2)$ and ($\mathbf{P}$, $\mathbf{Q}$). The spatial filter matrix decomposes as $\mathbf{W}^T = \mathbf{\Omega}^T\mathbf{T}$ into the product of a whitening transformation matrix $\mathbf{T}$ of the observations and a semi-orthogonal matrix $\mathbf{\Omega}^T$, which satisfies $\mathbf{\Omega}^T\mathbf{\Omega} = \mathbf{I}_p$. The whitening transformation is obtained from eigenvalue decomposition of $Cov(\boldsymbol{x}) = p(c_1)\mathbf{P} + p(c_2)\mathbf{Q} = \mathbf{U}_1\mathbf{\Delta}\mathbf{U}_1^T$ as follows

$$\mathbf{T} = \mathbf{\Delta}^{-\frac{1}{2}}\mathbf{U}_1^T, \tag{133}$$

where $\mathbf{\Delta}$ and $\mathbf{U}_1$ represent the matrices of eigenvalues and eigenvectors. This transformation is applied to both sides of the covariance matrices to obtain the whitened trial covariances

$$\check{\mathbf{P}}_j = \mathbf{T}\,\mathbf{P}_j\mathbf{T}^T, \qquad \check{\mathbf{Q}}_j = \mathbf{T}\,\mathbf{Q}_j\mathbf{T}^T, \tag{134}$$

and their averaged versions

$$\check{\mathbf{P}} = \mathbf{T}\,\mathbf{P}\,\mathbf{T}^T, \qquad \check{\mathbf{Q}} = \mathbf{T}\,\mathbf{Q}\,\mathbf{T}^T. \tag{135}$$

The scaling parameter κ, which pursues the balance of the number of features for each class in absence of regularizers, is determined with the truncation procedure proposed in (70). The semiorthogonal matrix $\mathbf{\Omega}^T$ that projects the whitened observations onto a p-dimensional subspace is initialized from the identity matrix of dimension $n \times p$. This is equivalent to start the optimization projecting onto the principal p-dimensional subspace of the observations, which ensures a good initial

signal to noise ratio. Once the whitening transformation is fixed, the criterion to optimize $F(\mathbf{W})$ can be rewritten, in terms of $\mathbf{\Omega}$, as the following function

$$\begin{aligned} f(\mathbf{\Omega}) &= D_{AB}^{(\alpha,\beta)}(\mathbf{\Omega}^T\check{\mathbf{P}}\mathbf{\Omega}\|\kappa\mathbf{\Omega}^T\check{\mathbf{Q}}\mathbf{\Omega}) \\ &\quad -\eta\left(p(c_1)\frac{1}{N_1}\sum_{j=1}^{N_1}D_{AB}^{(\alpha,\beta)}(\mathbf{\Omega}^T\check{\mathbf{P}}_j\mathbf{\Omega}\|\mathbf{\Omega}^T\check{\mathbf{P}}\mathbf{\Omega})\right. \\ &\quad \left.+p(c_2)\frac{1}{N_2}\sum_{j=1}^{N_2}D_{AB}^{(\alpha,\beta)}(\mathbf{\Omega}^T\check{\mathbf{Q}}_j\mathbf{\Omega}\|\mathbf{\Omega}^T\check{\mathbf{Q}}\mathbf{\Omega})\right), \end{aligned} \tag{136}$$

which ordinary gradient can be determined from (95), to obtain

$$\begin{aligned} \frac{\partial f(\mathbf{\Omega})}{\partial\mathbf{\Omega}} &= 2[\check{\mathbf{P}}\mathbf{\Omega}-\kappa\check{\mathbf{Q}}\mathbf{\Omega}(\kappa\mathbf{\Omega}^T\check{\mathbf{Q}}\mathbf{\Omega})^{-1}(\mathbf{\Omega}^T\check{\mathbf{P}}\mathbf{\Omega})](\kappa\mathbf{\Omega}^T\check{\mathbf{Q}}\mathbf{\Omega})^{-\frac{1}{2}}\mathbf{Z}_1(\kappa\mathbf{\Omega}^T\check{\mathbf{Q}}\mathbf{\Omega})^{-\frac{1}{2}} \\ &\quad -\eta\left(p(c_1)\frac{2}{N_1}\sum_{j=1}^{N_1}[\check{\mathbf{P}}_j\mathbf{\Omega}-\check{\mathbf{P}}\mathbf{\Omega}(\mathbf{\Omega}^T\check{\mathbf{P}}\mathbf{\Omega})^{-1}(\mathbf{\Omega}^T\check{\mathbf{P}}_j\mathbf{\Omega})](\mathbf{\Omega}^T\check{\mathbf{P}}\mathbf{\Omega})^{-\frac{1}{2}}\mathbf{Z}_2(\mathbf{\Omega}^T\check{\mathbf{P}}\mathbf{\Omega})^{-\frac{1}{2}}\right. \\ &\quad \left.+p(c_2)\frac{2}{N_2}\sum_{j=1}^{N_2}[\check{\mathbf{Q}}_j\mathbf{\Omega}-\check{\mathbf{Q}}\mathbf{\Omega}(\mathbf{\Omega}^T\check{\mathbf{Q}}\mathbf{\Omega})^{-1}(\mathbf{\Omega}^T\check{\mathbf{Q}}_j\mathbf{\Omega})](\mathbf{\Omega}^T\check{\mathbf{Q}}\mathbf{\Omega})^{-\frac{1}{2}}\mathbf{Z}_3(\mathbf{\Omega}^T\check{\mathbf{Q}}\mathbf{\Omega})^{-\frac{1}{2}}\right) \end{aligned} \tag{137}$$

where the matrices $\mathbf{Z}_i$ should be defined for each case ($i=1,\ldots,3$) as in (94). However, this gradient is not the fastest ascent direction in the structured manifold of semi-orthogonal matrices (the Stiefel manifold). Instead, the fastest ascent direction is given by the "natural" gradient in this manifold [50,51], which is given by

$$\nabla_\Omega f(\mathbf{\Omega}) = \frac{\partial f(\mathbf{\Omega})}{\partial\mathbf{\Omega}} - \mathbf{\Omega}\left(\frac{\partial f(\mathbf{\Omega})}{\partial\mathbf{\Omega}}\right)^T\mathbf{\Omega}. \tag{138}$$

Let $\mathbf{\Omega}^{(i)}$ denote the semi-orthogonal matrix at iteration i and let $\mu^{(i)}$ denotes the step-size, the gradient ascent update is then performed with

$$\mathbf{\Omega}_{tg}^{(i+1)} = \mathbf{\Omega}^{(i)} + \mu^{(i)}\,\nabla_\Omega f(\mathbf{\Omega}^{(i)}). \tag{139}$$

The resulting matrix $\mathbf{\Omega}_{tg}^{(i+1)}$ belongs to the tangent space of the manifold at $\mathbf{\Omega}^{(i)}$ and asymptotically follows the geodesic path of maximum ascent for a sufficient small stepsize $\mu\to 0$. However, for practical stepsizes, like the one that we consider next

$$\mu^{(i)} = \frac{0.02}{\|\nabla_\Omega f(\mathbf{\Omega}^{(i)})\|_F}, \tag{140}$$

the resulting updates $\mathbf{\Omega}_{tg}^{(i+1)}$ are not exactly semi-orthogonal and, in order to restore this property, a retraction procedure onto the manifold is necessary after each iteration. The retraction can be implemented with the help of the MatLab command for a "thin" singular value decomposition as

$$[\mathbf{Q}_L,\mathbf{D},\mathbf{Q}_R] = \mathrm{svd}(\mathbf{\Omega}_{tg}^{(i+1)},0), \tag{141}$$

$$\mathbf{\Omega}^{(i+1)} = \mathbf{Q}_L\mathbf{Q}_R^T. \tag{142}$$

The procedure is then repeated until convergence to a maxima of the criterion at a given iteration i_{max}. After that, the solution $(\mathbf{\Omega}^{(i_{max})})^T\mathbf{T}$ identifies the subspace of the spatial filters, but not each of their individual directions. In order to determine them, one can solve a CSP problem within the previously identified subspace. We compute the generalized eigenvalues of the matrix pencil $((\mathbf{\Omega}^{(i_{max})})^T\check{\mathbf{P}}\mathbf{\Omega}^{(i_{max})}, (\mathbf{\Omega}^{(i_{max})})^T\check{\mathbf{Q}}\mathbf{\Omega}^{(i_{max})})$ and use the resulting principal and minor eigenvectors $\check{\boldsymbol{v}}_j$ to form the spatial filter matrix

$$\check{\mathbf{V}} \quad = \quad [\check{\boldsymbol{v}}_1, \ldots, \check{\boldsymbol{v}}_{\lfloor \frac{p}{2} \rfloor}, \check{\boldsymbol{v}}_{n-p+1+\lfloor \frac{p}{2} \rfloor}, \ldots, \check{\boldsymbol{v}}_n]. \tag{143}$$

The final matrix of spatial filters that solves the problem, is the product of the whitening matrix $\mathbf{T}$, the projection matrix $(\mathbf{\Omega}^{(i_{max})})^T$ and a CSP rotation matrix $\check{\mathbf{V}}^T$ which operates within the subspace, i.e.,

$$\mathbf{W}^T = \check{\mathbf{V}}^T(\mathbf{\Omega}^{(i_{max})})^T\mathbf{T}. \tag{144}$$

The main steps of the Sub-ABLD iteration are summarized in Algorithm 1.

Algorithm 1 Sub-ABLD algorithm

function SUB-ABLD($\{\mathbf{P}_j\}$,$\{\mathbf{Q}_j\}$,p,α,β,η)
  Compute the average covariance matrices $\mathbf{P}$ and $\mathbf{Q}$.
  Compute the total covariance matrix $Cov(\boldsymbol{x}) = p(c_1)\mathbf{P} + p(c_2)\mathbf{Q}$.
  Compute the whitening transform matrix $\mathbf{T}$ using (133).
  Whiten the trial and average covariance matrices to respectively obtain $\{\check{\mathbf{P}}_j\}$,$\{\check{\mathbf{Q}}_j\}$ and $\check{\mathbf{P}}$,$\check{\mathbf{Q}}$.
  Compute the scaling parameter, κ using (70) and initialize the iteration counter: $i = 0$.
  Initialize the semi-orthogonal matrix $\mathbf{\Omega}^{(i)} = \mathbf{I}_{n\times p}$.
  repeat
    Compute the robust criterion $f(\mathbf{\Omega}^{(i)})$ using (136).
    Compute the ordinary gradient $\frac{\partial f(\mathbf{\Omega}^{(i)})}{\partial \mathbf{\Omega}}$ using (137).
    Compute the natural gradient on the Stiefel manifold $\nabla_\Omega f(\mathbf{\Omega}^{(i)})$ using (138).
    Obtain the tangent matrix $\mathbf{\Omega}_{tg}^{(i+1)}$ using (139).
    Obtain the projection matrix $\mathbf{\Omega}^{(i+1)}$ using (141) and (142) (the retraction onto the manifold).
    Increase the iteration counter: $i = i + 1$.
  until convergence at iteration i_{max}.
  Collect in $\check{\mathbf{V}}$ the princip./minor eigenvect. of the pencil $((\mathbf{\Omega}^{(i_{max})})^T\check{\mathbf{P}}\mathbf{\Omega}^{(i_{max})}, (\mathbf{\Omega}^{(i_{max})})^T\check{\mathbf{Q}}\mathbf{\Omega}^{(i_{max})})$.
return $\mathbf{W}^T = \check{\mathbf{V}}^T(\mathbf{\Omega}^{(i_{max})})^T\mathbf{T}$.
end function

The proposed subspace algorithm (Sub-ABLD) is similar in structure to the one presented in [1] for Beta divergences. In spite of the fact that they optimize different criteria, the main difference between both subspace algorithms is in the specific way that the updates of the estimates are implemented. In [1] the authors opted for applying multiplicative updates that require the determination of the gradient of the criterion in the space of skew-symmetric matrices, whereas our proposal performs tangent updates to the manifold of the semi-orthogonal matrices that are followed by a projection or retraction onto the manifold. These updates are quite common in the research field of Independent Component Analysis [50–53].

10. Experimental Study

The discrimination of two class MI movements consists of the following steps. The MI EEG signals are acquired, preprocessed and spatially filtered. The filtered signals are then used for extracting the required features, which are classified using a linear classifier. In the following section, we explain the experimental steps used in the testing and comparison of the proposed algorithm.

10.1. Simulations Data and Preprocessing

Initially, we have explored the robustness of the proposed algorithm in a controlled situation with synthetic data. Two sets of symmetric positive definite matrices (SPD) that represent the trial covariance matrices of the two classes were randomly generated. Each set consists of 200 trials. For further preprocessing, both the sets of matrices were concatenated. The concatenated data are cross-validated using k-fold cross-validation $(k = 10)$. This divides the data into 10 equal subsets in which a single set was used as a testing data and the remaining 9 subsets were used for training the classifier. The performance of the proposed algorithm was studied in the presence of the outliers. The outliers consist of matrices with abnormal higher variances that were inserted in the training set of both the classes. The proposed Sub-ABLD algorithm was tested in the following figure by progressively varying the percentage of outliers in the trials from 0% until 30%. The robustness of Sub-ABLD and its comparison with respect to the other algorithms mentioned in the figure will be addressed in Section 11.

10.2. EEG Dataset and Preprocessing

To evaluate the proposed Sub-ABLD algorithm with BCI competition datasets, we utilized two datasets from competition III: data set 3a, data set 4a (which can be downloaded from [54]) and one dataset from competition IV data set 2a (which can be downloaded from [55]). The data were acquired during the MI movements. The first dataset 3a [56] from BCI competition III [57], were acquired from 3 healthy subjects namely K3, K6 and L1 using 60 channels EEG acquisition system. The signals were recorded while executing the MI movements of the left hand, right hand, foot and tongue. The signals were sampled at a frequency of 250 Hz. The sampled signals were bandpass filtered at the frequency range between 1 to 50 Hz. The data set consists of two sessions i.e., training and testing sessions. For subject K3, both the sessions consist of 45 trials for each class whereas the other two subjects i.e., K6 and L1 performed 30 trials per class in both the sessions. For the second dataset, data set 4a [44] of BCI competition III [57], the signals were acquired from five subjects namely AA, AL, AV, AW and AY using 118 channels EEG system. The acquisition was done during the imagery movements of the left hand, right hand and right foot. Down-sampling of the recorded signals was done at 100 Hz. The band-pass filter between 0.05 to 200 Hz frequency band was applied to the signals. The data set of each subject consists of 280 total trials. The size of the training sessions is different from testing sessions. The training sessions consist of 168, 224, 84, 56, 28 trails for subjects AA, AL, AV, AW, AY and the remaining denotes the testing trails for the corresponding subjects. The last dataset, data set 2a [46] BCI competition IV [58] were acquired from nine subjects (A1 to A9) while performing the left hand, right hand, foot and tongue MI movements using 22 electrodes. The sampling frequency of the signals was 250 Hz. The band-pass filtering of the acquired signals were performed between 0.5 and 100 Hz. For each subject, the data were acquired on different days and each set consists of 72 trials for each class.

In this approach, the performances were obtained using only two MI movements considering all the channels from each dataset. The preprocessing step was implemented similarly for all the algorithms. First, a fifth-order band-pass filter with a cut-off frequency between 8 to 30 Hz was applied to the raw EEG signals. A time window of 2s during the imagination of movements was extracted for each trial. The extracted trials were concatenated for each class and applied a k-fold cross-validation $(k = 10)$ to the concatenated data. The CV process divides the data into 10 equal sets where one set

of data was used as testing data and the remaining 9 sets were used for training. Finally, the optimal spatial filters were obtained using the training dataset. The number of filters selected for each class is $k = 3$, so the total number $p = 6$.

10.3. Feature Extraction and Feature Classification

For both-the synthetic and the BCI datasets, the obtained spatial filters were used for filtering the training and testing data. The training and testing features were obtained by taking the log-variance of the filtered data in order that their distribution be closer to Gaussianity. The Linear Discriminant Analysis (LDA) [59] classifier was used for discriminating the features of the two classes. The classifier was trained using the training features and its performance was obtained using the testing features. The preprocessing, feature extraction and classification steps were repeated 10 times and finally the average performance was obtained.

10.4. Selection of α, β and η Values

The selection of α and β is one of the crucial steps for the proposed algorithm. Depending on the α and β values, the AB Log-Det divergence can be derived into different divergence techniques [6]. The proposed algorithm performed better when $\alpha = \beta$, situation where the AB Log-Det divergence is symmetric or invariant under the permutation of its arguments. In this experiment, we have observed the performance for various values of $\alpha = \beta$ and η, and a suitable configuration of parameters for each dataset was selected.

11. Results and Discussion

The performance of the proposed Sub-ABLD algorithm is compared with the performance of the existing algorithms: CSP, JADE, MAPCSP and divCSP-WS for both the synthetic and the real BCI competition datasets. JADE algorithm performs a joint approximate diagonalization of the trial covariance matrices of the classes [45]. MAPCSP is a Bayesian algorithm that tries to find the maximum a posteriori estimates of the patterns and sources in a generative model with additive Gaussian isotropic noise [14]. The subspace implementation of divCSP-WS finds a balance between the maximization of Beta divergence between the conditional covariances of the filtered outputs for each class and the minimization of the variability within each class [1]. This algorithm contains two hyperparameters, the regularization factor λ and the real scalar β' that specifies the chosen Beta divergence. The factor λ admits an equivalence in terms of the reguralization parameter η in Sub-ABLD which link them through the mapping $\lambda \equiv \eta/(1+\eta)$, while the parameter of the Beta divergence β'_* was chosen in the simulations to maximize the performance .

In order to carry out a fair performance comparison, a total of six features (i.e., $p = 6$) have been selected for all the algorithms. The implementation of the JADE and divCSP-WS algorithms were taken from the webpages of the authors. The baseline divCSP-WS algorithm has been downloaded from [60], while the implementation of JADE algorithm can be found at [61]. The performance comparison between all the algorithms is presented in the following subsections.

11.1. Observations for Simulated Data

To study the performance of the proposed algorithm in the presence of outliers, the experiment was done by increasing the percentage of outlier trials in the training set for both the classes. The performance of the proposed Sub-ABLD algorithm for $(\alpha, \beta) = (1.5, 1.5)$ and $\eta = 1$ was obtained. The performances of the above algorithms with the increasing percentage of outliers in the training set are presented in Figure 5. It can be observed that CSP and MAPCSP perform worse in the presence of the outliers. The performances of JADE and divCSP-WS are much more robust than those of CSP and MAPCSP, but in overall the proposed Sub-ABLD algorithm seems to outperform the compared algorithms in the presence of the outliers.

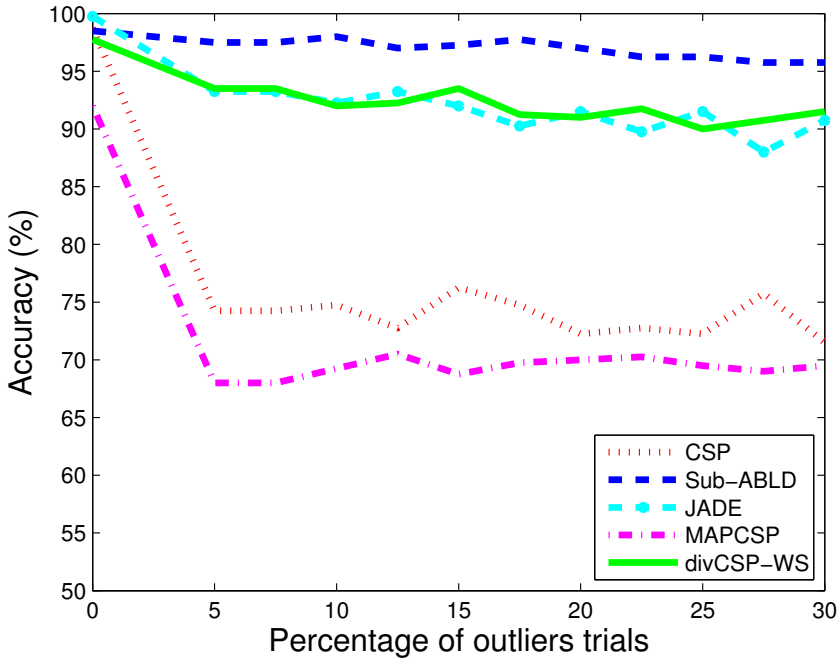


Figure 5. Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 1,\ \alpha = \beta = 1.5$) with CSP, JADE, MAPCSP and divCSP-WS ($\lambda = 0.5,\ \beta'_* = 0.25$), versus the percentage of outlier trials.

11.2. Observations for BCI Competition Datasets

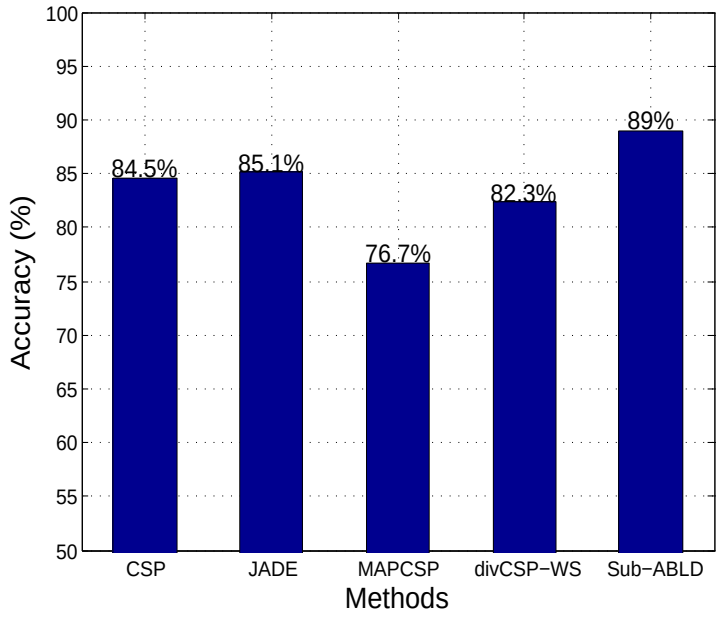
In this section, the proposed algorithm is tested using three BCI competition datasets. For each dataset, the performances of the proposed algorithm for the different values of (α, β) and η were observed. From the observation, the maximum performance of the Sub-ABLD algorithm for the particular (α, β) and η values was selected. The selected performance is compared with the performances of other existing algorithms. Further analysis is done by using a box plot comparison for all the algorithms. The box plot analysis shows the distribution of the performances. In a box plot representation, the line inside the box represents the median performance. The upper and lower hinge of the box denote the 75-th and 25-th percentile of the overall performance distributions. The whiskers are symbolized by the two lines outside the box. The upper and lower whisker represent the maximum and minimum performance observed.

For BCI competition III dataset 3a, the Figure 6a shows the comparison of the highest average performance of the Sub-ABLD algorithm with the average performances of other existing algorithms. From the figure, it is observed that the Sub-ABLD algorithm outperforms the other existing algorithms with an average performance accuracy of 89% for this dataset. The box plot comparison is shown in Figure 6b. Although, the median performance is slightly higher for CSP, JADE and divCSP, their the 25th percentile performance is much smaller than the one of the Sub-ABLD algorithm. As we will see later, is a consequence that with the Sub-ABLD algorithm the most difficult subjects have attained a significant improvement in their classification performance.

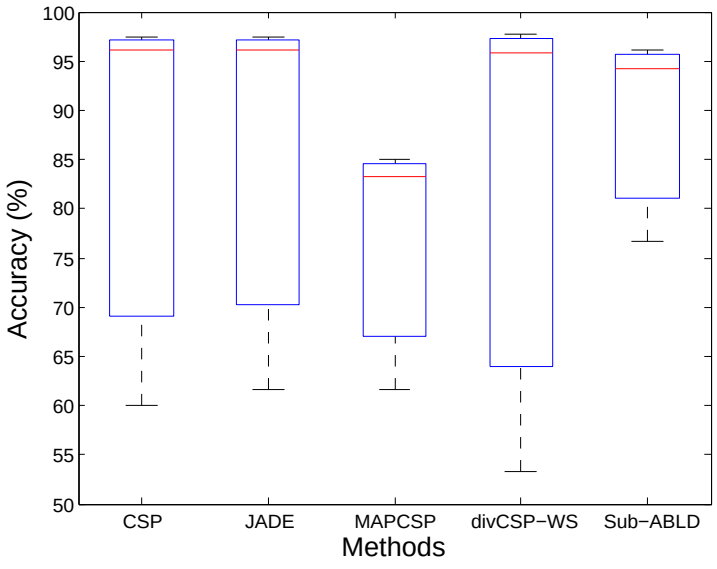
Figure 7 shows the observed average performances using BCI competition III dataset 4a. For this dataset, the algorithms JADE, Sub-ABLD and divCSP-WS perform essentially similar and slightly above than the average performance of CSP, which is 88.1%. From the box plot of the results we can observe that the 25-th percentiles for these four algorithms are also quite close.

Similar results have been obtained for the BCI competition IV dataset 2a, which are shown in Figure 8. Again the algorithms JADE, Sub-ABLD and divCSP-WS perform essentially the same as

CSP, which average performance is 81%. In the box plot we can observe that the quartiles of these algorithms are approximately coincident.

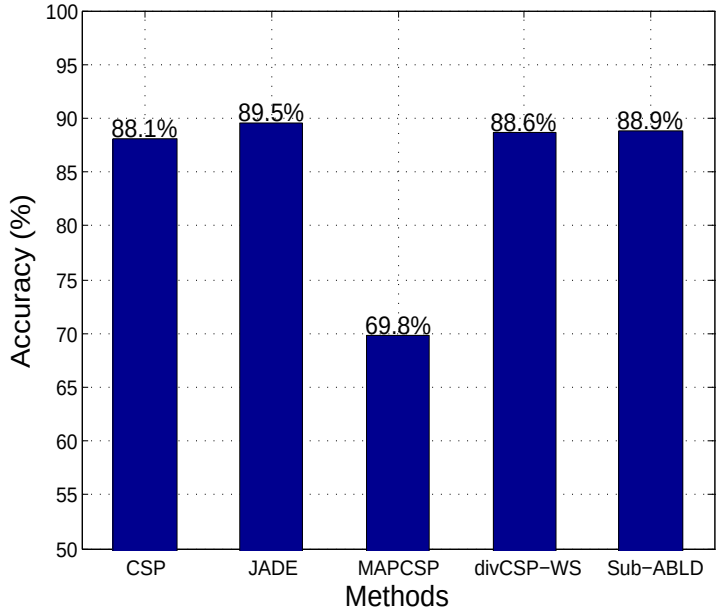


(**a**)

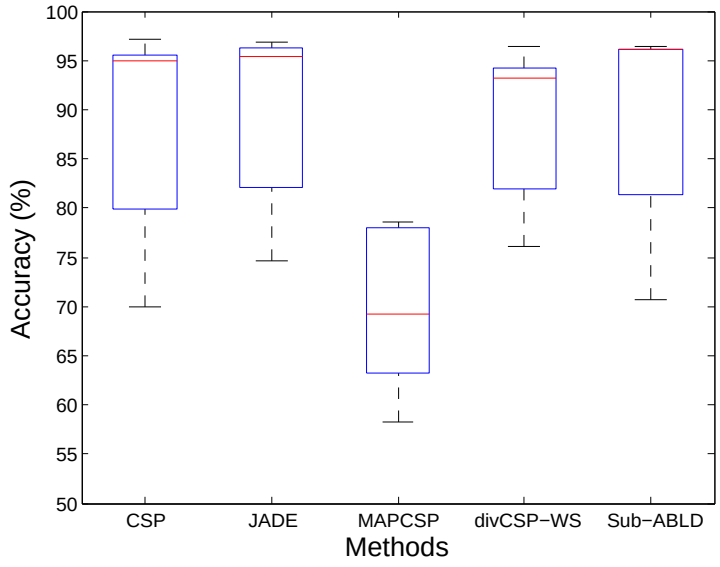


(**b**)

Figure 6. (**a**) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 2$, $\alpha = \beta = 1.5$) with CSP, JADE, MAPCSP and divCSP-WS ($\lambda = 0.66$, $\beta'_* = 1$) using BCI competition III dataset 3a and (**b**) its corresponding boxplot.

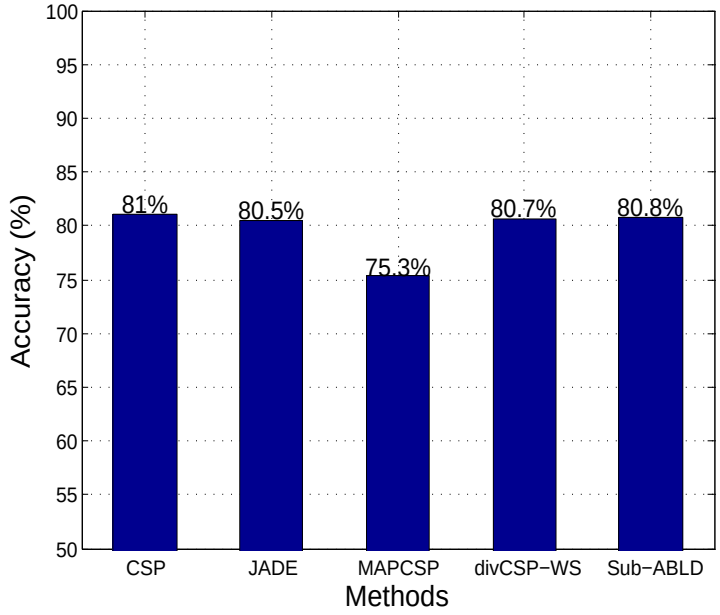


(**a**)

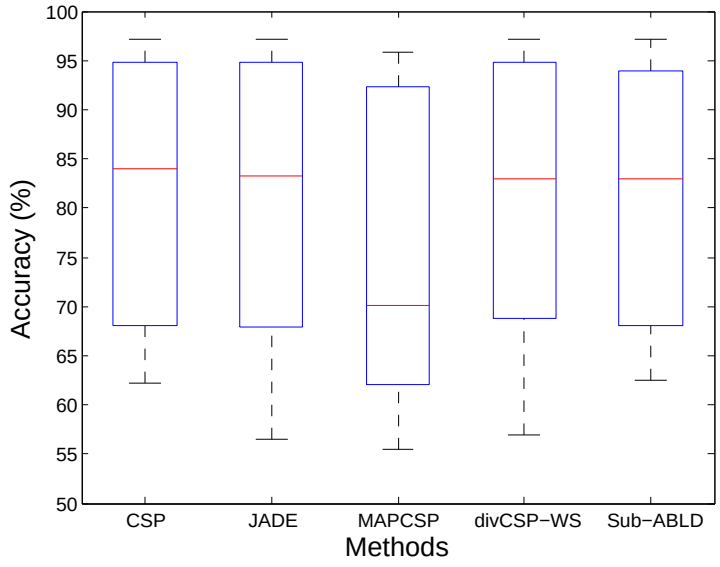


(**b**)

Figure 7. (**a**) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 0.5$, $\alpha = \beta = 2$) with CSP, JADE, MAPCSP and divCSP-WS ($\lambda = 0.33$, $\beta'_* = 0.5$) using BCI competition datasets III dataset 4a and (**b**) its corresponding boxplot.



(**a**)



(**b**)

Figure 8. (**a**) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 0.25$, $\alpha = \beta = 1.25$) with CSP, JADE, MAPCSP and divCSP-WS ($\lambda = 0.2$, $\beta'_* = 0$) using BCI competition datasets IV dataset 2a and (**b**) its corresponding boxplot.

To analyze the effect of performance for different divergences, we varied the parameters (α, β) for a single subject (Subject k6 from BCI competition III dataset 3a, which is one of the subjects with worst performance for the experiment) and obtained the corresponding performance. The values of (α, β) are varied to cover the interval $[0, 2] \times [0, 2]$ with a mesh of 0.1 spacings. The observed performance is shown in Figure 9. This figure reveals a tendency to improve the classification accuracy of the worst user for values of α and β that are close to the diagonal and large enough so they can

effectively down-weight the contribution in the estimating equations coming from the largest and smallest generalized eigenvalues.

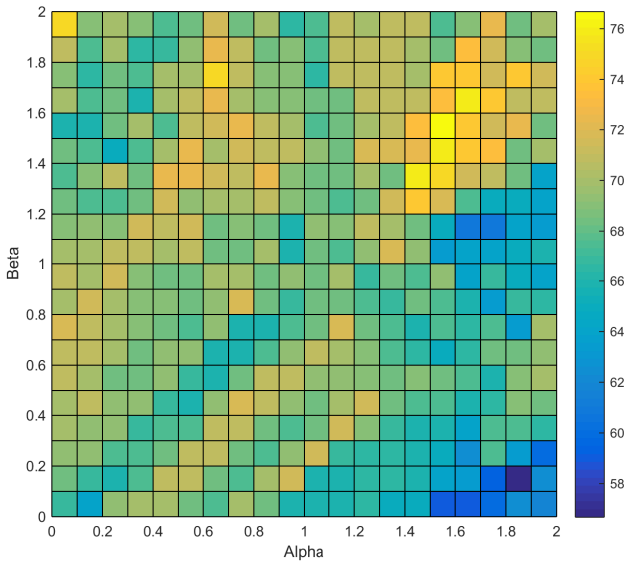


Figure 9. Results of the Sub-ABLD algorithm for the subject k6 from BCI competition III dataset 3a. This figure illustrates the changes in the average classification performance with respect to the variation of the parameters α and β. Relatively good performance results are obtained close to the diagonal and for moderately large values of the parameters.

The proposed Sub-ABLD algorithm has been tested on both simulated and real EEG signals. On one hand, the results with synthetic data indicate that the proposed Sub-ABLD exhibits a certain robustness to the presence of outliers trials in the dataset. On the other hand, the analysis of real EEG signals is also challenging because of the possible presence of artifacts and non-stationarities. We have presented the performance of the Sub-ABLD algorithm using several real BCI datasets. For BCI competition III dataset 3a, we can observe that the proposed Sub-ABLD algorithm also outperforms the other algorithms. Whereas, the performance of the proposed algorithm is almost similar to the one obtained by JADE, divCSP-WS and CSP in the other two datasets, i.e., for the BCI competition III dataset 4a and BCI competition IV dataset 2a. Additionally, the analysis of the box-plots reveals that the proposed Sub-ABLD algorithm increased the classification performance of the subjects that do not perform well for the other methods. At the same time, it retained an almost similar performance for the remaining subjects. These observations meet our initial goal of developing a robust algorithm. The classification performance is also affected by the regularization parameter η that controls the penalty term. In general, the data with outliers give the best performance for the higher values of η and, otherwise, smaller values are preferable. In this study, the value of η has been kept constant across subjects in each dataset.

12. Conclusions

In this paper, we have explained how one can be able to use and optimize the recently proposed family of Alpha-Beta Log-Det divergences. For this purpose, we have summarized the key properties of these divergences and derived an original explicit formula for their gradient. In this work, we have adopted as an illustrative example of application the problem of spatial filter selection for the classification of two class imagery movements. We have reexamined the relation between the Common Spatial Pattern criterion with a predefined number of spatial filters for each class and its interpretation as an Alpha-Beta Log-Det divergence optimization problem, to show that a scaling

factor in one of the arguments of the divergence is necessary for the equivalence of the solutions. We have proposed a subspace algorithm (Sub-ABLD) for obtaining the spatial filters that retain the discriminative information of two class MI movements. This algorithm was tested with synthetic and real datasets and compared with the other existing algorithms. The simulations have confirmed the possibility to tune up the hyperparameters of the divergence so as to improve the robustness of the obtained solutions without deteriorating the expected accuracy.

Acknowledgments: This work was supported by the Spanish Government under the MICINN project TEC2014-53103-P.

Author Contributions: Deepa Beeta Thiyam and Sergio Cruces have collaborated in the tast of writing the manuscript. Sergio Cruces was in charge of developing the theoretical content and has coordinated the proposal, while Deepa Beeta Thiyam was in charge of the design of the proposed algorithm and simulations. Javier Olias and Andrzej Cichocki have respectively collaborated in the experimental part and in the study of the AB log-det divergences. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1 Obtaining the Upper-Bound of the AB Log-Det Divergence

The divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ depends on the generalized eigenvalues of the matrix pencil $(\mathbf{P},\mathbf{Q})$, which have been denoted by λ_i for $i = 1,\ldots,n$. Similarly, the divergence of the compressed arguments $D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\|\ \mathbf{W}^T\mathbf{Q}\mathbf{W})$ depends on μ_i for $i = 1,\ldots,p$, the eigenvalues of the matrix pencil $(\mathbf{W}^T\mathbf{P}\mathbf{W},\mathbf{W}^T\mathbf{Q}\mathbf{W})$. The Cauchy interlacing inequalities [62]

$$\lambda_j \leq \mu_j \leq \lambda_{n-p+j}. \tag{A1}$$

provide upper and lower-bounds for μ_j in terms of the eigenvalues of the uncompressed matrix pencil. This property implies that the eigenvalues μ_j, for each $j = 1,\ldots,p$, should lie in a sequence of possibly partially overlapping intervals given by $[\lambda_j,\lambda_{n-p+j}]$.

The divergence $D_{AB}^{(\alpha,\beta)}(\lambda\|\ 1)$ is minimum (zero) for $\lambda = 1$, strictly monotone descending for $\lambda < 1$ and strictly monotone ascending for $\lambda > 1$. So we can bound the the AB log-det divergence in each interval by

$$D_{AB}^{(\alpha,\beta)}(\mu_j\|\ 1) \quad \leq \quad \max\{D_{AB}^{(\alpha,\beta)}(\lambda_j\|\ 1)\ ,\ D_{AB}^{(\alpha,\beta)}(\lambda_{n-p+j}\|\ 1)\}, \tag{A2}$$

and the maximum value occurs at one of the extreme eigenvalues of the interval. The construction of the interlacing property, prevents that any eigenvalue with a given index could appear more than once as upper extreme of an interval or as a lower extreme of an interval. This fact, combined with the strict monotonicity property of the divergence, implies that the maxima of the divergence for each interval can only be obtained by eigenvalues with different indices. Finally, the result of adding these p maximum values can not exceed the sum of the divergences for those eigenvalues which maximize the divergence from unity,

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\ \|\mathbf{W}^T\mathbf{Q}\mathbf{W}) \quad = \quad \sum_{j=1}^{p} D_{AB}^{(\alpha,\beta)}(\mu_j\|\ 1) \tag{A3}$$

$$\leq \quad \sum_{j=1}^{p} \max\{D_{AB}^{(\alpha,\beta)}(\lambda_j\|\ 1)\ ,\ D_{AB}^{(\alpha,\beta)}(\lambda_{n-p+j}\|\ 1)\} \tag{A4}$$

With the help of the permutation π of the indices $1,\ldots,n$ that sorts the divergence of the eigenvalues from the unity in descending order

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_1}\|\,1) \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_2}\|\,1) \geq \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_n}\|\,1)\,, \tag{A5}$$

we can write

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\,\|\mathbf{W}^T\mathbf{Q}\mathbf{W}) &= \sum_{j=1}^{p} D_{AB}^{(\alpha,\beta)}(\mu_j\|\,1) && \text{(A6)}\\ &\leq \sum_{j=1}^{p} \max\{D_{AB}^{(\alpha,\beta)}(\lambda_j\|\,1)\,,\, D_{AB}^{(\alpha,\beta)}(\lambda_{n-p+j}\|\,1)\} && \text{(A7)}\\ &\leq \sum_{j=1}^{p} D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_j}\|\,1) && \text{(A8)} \end{aligned}$$

which is the desired upper-bound.

Appendix A.2 Proof of the Link between the Optimization of the Divergence and the CSP Solution

The fact that any Rayleigh quotient is bounded by the maximum and minimum eigenvalues of the associated matrix pencil

$$\lambda_1 \leq \frac{\boldsymbol{w}_i^T\mathbf{P}\boldsymbol{w}_i}{\boldsymbol{w}_i^T\mathbf{Q}\boldsymbol{w}_i} \leq \lambda_n \tag{A9}$$

can be used to recursively prove that the minimax value of the divergence is equal to

$$\begin{aligned} \min_{\dim\{\mathcal{W}\}=n-i+1}\ \max_{\boldsymbol{w}\in\mathcal{W}}\ D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T\mathbf{P}\boldsymbol{w}_i\,\|\,\kappa\,\boldsymbol{w}_i^T\mathbf{Q}\boldsymbol{w}_i) &= \min_{\dim\{\mathcal{W}\}=n-i+1}\ \max_{\boldsymbol{w}\in\mathcal{W}}\ D_{AB}^{(\alpha,\beta)}(\tfrac{\boldsymbol{w}_i^T\mathbf{P}\boldsymbol{w}_i}{\boldsymbol{w}_i^T\mathbf{Q}\boldsymbol{w}_i}\,\|\,\kappa)\\ &= D_{AB}^{(\alpha,\beta)}(\lambda_{\pi'_i}\|\,\kappa), \end{aligned} \tag{A10}$$

where permutation π' sorts the divergence of the eigenvalues from κ in descending order

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi'_1}\|\,\kappa) \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi'_2}\|\,\kappa) \geq \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi'_n}\|\,\kappa). \tag{A11}$$

The minimax value is then attained for the eigenvectors

$$\boldsymbol{v}_{\pi'_i} = \arg\ \min_{\dim\{\mathcal{W}\}=n-i+1}\ \max_{\boldsymbol{w}\in\mathcal{W}}\ D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T\mathbf{P}\boldsymbol{w}_i\,\|\,\kappa\,\boldsymbol{w}_i^T\mathbf{Q}\boldsymbol{w}_i), \qquad i=1,\ldots,p. \tag{A12}$$

For the coincidence of the set of solutions $\{\boldsymbol{v}_{\pi'_1},\ldots,\boldsymbol{v}_{\pi'_p}\}$ in (A12) with the set of spatial filters $\{\boldsymbol{v}_1^{(c_1)},\ldots,\boldsymbol{v}_k^{(c_1)},\boldsymbol{v}_{n-(p-k)+1}^{(c_1)},\ldots,\boldsymbol{v}_n^{(c_1)}\}$ that define the $\mathbf{W}_{CSP}$, the eigenvalues $\lambda_{\pi_1},\ldots,\lambda_{\pi_p}$ that maximize their divergence from κ, should all belong to the upper and lower sets of eigenvalues defined in (57). For this to be true, it necessary and sufficient that the divergence of the last selected eigenvalue λ_{π_p} from κ upper-bounds with inequality all the divergences between an inner eigenvalue λ_i and κ, in the sense that

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_p}\|\,\kappa) > \max_{i\in[k+1\,,\,n-(p-k)]} D_{AB}^{(\alpha,\beta)}(\lambda_i\|\,\kappa) \tag{A13}$$

The domain of κ for which this strict inequality holds true is

$$\kappa \in \left(\kappa_{\inf}\,,\,\kappa_{\sup}\right) \tag{A14}$$

where the bounds

$$\kappa_{\text{inf}} \equiv \mathcal{K}(\lambda_{k+1}\,,\,\lambda_{n-(p-k)+1\rfloor}) \tag{A15}$$

$$\kappa_{\text{sup}} \equiv \mathcal{K}(\lambda_{k}\,,\,\lambda_{n-(p-k)}) \tag{A16}$$

respectively equalize the value of the divergences

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_p}\|\,\kappa_{\text{inf}}) = D_{AB}^{(\alpha,\beta)}(\lambda_{k+1}\|\,\kappa_{\text{inf}}) = D_{AB}^{(\alpha,\beta)}(\lambda_{n-(p-k)+1}\|\,\kappa_{\text{inf}}) \tag{A17}$$

and

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_p}\|\,\kappa_{\text{sup}}) = D_{AB}^{(\alpha,\beta)}(\lambda_{k}\|\,\kappa_{\text{sup}}) = D_{AB}^{(\alpha,\beta)}(\lambda_{n-(p-k)}\|\,\kappa_{\text{sup}}). \tag{A18}$$

Appendix A.3 Differential of the Inverse Square Root of a SPD Matrix

Let $\mathbf{X}$ be any symmetric possitive definite matrix (SPD). We would like to obtain the differential of its inverse square root $d\mathbf{X}^{-\frac{1}{2}}$ in terms of the matrix $\mathbf{X}$ and its differential $d\mathbf{X}$, and later use this result to simplify the desired expression $d\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$. For this purpose, we start from the trivial identity $\mathbf{I}_p = \mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$ and take differentials on both sides of this equality, with the help of the product rule for differentials we obtain

$$\mathbf{0} = d\mathbf{I}_p \quad = \quad d(\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}) = d\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} + \mathbf{X}^{-\frac{1}{2}}d\mathbf{X}^{\frac{1}{2}}. \tag{A19}$$

Solving for the differential

$$d\mathbf{X}^{-\frac{1}{2}} \quad = \quad -\mathbf{X}^{-\frac{1}{2}}d\mathbf{X}^{\frac{1}{2}}\mathbf{X}^{-\frac{1}{2}}, \tag{A20}$$

we see it as a function of $\mathbf{X}$ and $d\mathbf{X}^{\frac{1}{2}}$. Then, we simplify $d\mathbf{X}^{\frac{1}{2}}$ with the help of the another trivial identify $\mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}})^T = \mathbf{X}$. We take again differentials on both sides of the equality

$$d(\mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}})^T) = d\mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}})^T + \mathbf{X}^{\frac{1}{2}}(d\mathbf{X}^{\frac{1}{2}})^T \quad = \quad d\mathbf{X} \tag{A21}$$

and obtain the special solution

$$d\mathbf{X}^{\frac{1}{2}} = \tfrac{1}{2}d\mathbf{X}(\mathbf{X}^{-\frac{1}{2}})^T. \tag{A22}$$

The substitution of (A22) in (A22) yields the differential of the inverse symmetric square root of the SPD matrix

$$d\mathbf{X}^{-\frac{1}{2}} \quad = \quad -\mathbf{X}^{-\frac{1}{2}}\left(\tfrac{1}{2}d\mathbf{X}(\mathbf{X}^{-\frac{1}{2}})^T\right)\mathbf{X}^{-\frac{1}{2}}. \tag{A23}$$

Finally, by the symmetry of $\mathbf{X}^{-\frac{1}{2}}$, we prove the desired result

$$d\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} \quad = \quad -\frac{1}{2}\mathbf{X}^{-\frac{1}{2}}d\mathbf{X}\,(\mathbf{X}^{-\frac{1}{2}})^T \tag{A24}$$

$$\quad = \quad -\frac{1}{2}\mathbf{X}^{-\frac{1}{2}}d\mathbf{X}\,\mathbf{X}^{-\frac{1}{2}} \tag{A25}$$

Appendix A.4 The Gradient of the KL Divergence between Gaussian Densities

The Kullback–Leibler (KL) divergence between the Gaussian densities $p(\boldsymbol{x}|c_2)$ and $p(\boldsymbol{x}|c_1)$, of zero mean and with respective covariance matrices $Cov(\mathbf{Y}|c_1) = \mathbf{W}^T\mathbf{P}\mathbf{W}$ and $Cov(\mathbf{Y}|c_2) = \mathbf{W}^T\mathbf{Q}\mathbf{W}$, is equal to

$$\begin{aligned} Div_{KL}(p(\boldsymbol{x}|c_2) \,\|\, p(\boldsymbol{x}|c_1)) &= \tfrac{1}{2}\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| - \tfrac{1}{2}\log|\mathbf{W}^T\mathbf{Q}\mathbf{W}| \\ &\quad + \tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W}) - \mathbf{I}_p\} \end{aligned} \tag{A26}$$

This subsection explains the operations involved in obtaining its gradient. The first differential of the log-determinant terms is

$$d\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| = \mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}d(\mathbf{W}^T\mathbf{P}\mathbf{W})\} \tag{A27}$$

$$= \mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(d\mathbf{W}^T\,\mathbf{P}\mathbf{W} + \mathbf{W}^T\,\mathbf{P}\,d\mathbf{W})\} \tag{A28}$$

$$= 2\,\mathrm{tr}\{[\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}]\,d\mathbf{W}^T\} \tag{A29}$$

By using the relationship between the first differential and the gradient

$$d\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| = \mathrm{tr}\{[\nabla_{\mathbf{W}}\log|\mathbf{W}^T\mathbf{P}\mathbf{W}|]\,d\mathbf{W}^T\} \tag{A30}$$

one can identify from (A29) that

$$\nabla_{\mathbf{W}}\tfrac{1}{2}\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| = \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \tag{A31}$$

and, similarly,

$$\nabla_{\mathbf{W}}[-\tfrac{1}{2}\log|\mathbf{W}^T\mathbf{Q}\mathbf{W}|] = -\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}\,. \tag{A32}$$

On the other hand, the first differential of the trace term simplifies to

$$\begin{aligned} d\left[\tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W}) - \mathbf{I}_p\}\right] &= \tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}d(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &\quad + \tfrac{1}{2}\,\mathrm{tr}\{d(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &= \tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}d(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &\quad + \tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}d(\mathbf{W}^T\mathbf{P}\mathbf{W})(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &= \tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W})\} \\ &\quad + \tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(d\mathbf{W}^T\mathbf{P}\mathbf{W} + \mathbf{W}^T\mathbf{P}d\mathbf{W}) \\ &\qquad \times(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &= \tfrac{1}{2}\,\mathrm{tr}\{2\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}\,d\mathbf{W}^T\} \\ &\quad + \tfrac{1}{2}\,\mathrm{tr}\{2\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}d\mathbf{W}^T\} \end{aligned} \tag{A33}$$

From which one can also identify

$$\begin{aligned} \nabla_{\mathbf{W}}\left[\tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W}) - \mathbf{I}_p\}\right] &= \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\ &\quad + \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \end{aligned} \tag{A34}$$

Once we have obtained in (A31), (A32) and (A34) the gradients of the partial terms that are involved in the definition (A26) of the KL divergence, their simple addition yields the complete gradient of the KL divergence with respect to $\mathbf{W}$, which is given by

$$\begin{aligned}\frac{\partial}{\partial \mathbf{W}} Div_{KL}(p(\boldsymbol{x}|c_2) \,\|\, p(\boldsymbol{x}|c_1)) \;=\; & -\mathbf{QW}(\mathbf{W}^T\mathbf{QW})^{-1} + \mathbf{PW}(\mathbf{W}^T\mathbf{PW})^{-1} + \mathbf{QW}(\mathbf{W}^T\mathbf{PW})^{-1} \\ & + \mathbf{PW}(\mathbf{W}^T\mathbf{PW})^{-1}(\mathbf{W}^T\mathbf{QW})^{-1}(\mathbf{W}^T\mathbf{PW})^{-1}. \end{aligned} \tag{A35}$$

References

1. Samek, W.; Kawanabe, M.; Müller, K.R. Divergence-based framework for common spatial patterns algorithms. *IEEE Rev. Biomed. Eng.* **2014**, *7*, 50–72.
2. Huang, Z.; Wang, R.; Shan, S.; Li, X.; Chen, X. Log-Euclidean Metric Learning on Symmetric Positive Definite Manifold with Application to Image Set Classification. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015.
3. Harandi, M; Salzmann, M.; Hartley, R. Dimensionality Reduction on SPD Manifolds: The Emergence of Geometry-Aware Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, doi:10.1109/ TPAMI.2017.2655048.
4. Sra, S.; Hosseini, R. *Geometric Optimization in Machine Learning Algorithmic Advances in Riemannian Geometry and Applications: For Machine Learning, Computer Vision, Statistics, and Optimization*; Springer: Berlin/Heidelberg, Germany, 2016.
5. Horev, I.; Yger, F.; Sugiyama, M. Geometry-aware principal component analysis for symmetric positive definite matrices. In Proceedings of the 7th Asian Conference on Machine Learning, Hong Kong, China, 20–22 November 2015; pp. 1–16.
6. Cichocki, A.; Cruces, S.; Amari, S.i. Log-Determinant Divergences Revisited: Alpha-Beta and Gamma Log-Det Divergences. *Entropy* **2015**, *17*, 2988–3034.
7. Minh, H.Q. Infinite-dimensional Log-Determinant divergences II: Alpha-Beta divergences. *arXiv* **2017**, arXiv:1610.08087.
8. Dornhege, G. *Toward Brain-Computer Interfacing*; MIT Press: Cambridge, MA, USA, 2007.
9. Wolpaw, J.; Wolpaw, E.W. *Brain-computer Interfaces: Principles and Practice*; Oxford University Press: Oxford, UK, 2012.
10. Pfurtscheller, G.; Da Silva, F.L. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clin. Neurophysiol.* **1999**, *110*, 1842–1857.
11. Fukunaga, K.; Koontz, W.L.G. Application of the Karhunen-Loeve Expansion to Feature Selection and Ordering. *IEEE Trans. Comput.* **1970**, *C-19*, 440–447.
12. Koles, Z.J. The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalogr. Clin. Neurophysiol.* **1991**, *79*, 440–447.
13. 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.
14. Wu, W.; Chen, Z.; Gao, X.; Li, Y.; Brown, E.N.; Gao, S. Probabilistic common spatial patterns for multichannel EEG analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 639–653.
15. Bhatia, R. *Matrix Analysis*; Graduate Texts in Mathematics; Springer: Berlin/Heidelberg, Germany, 1997.
16. Wang, H. Harmonic mean of Kullback–Leibler divergences for optimizing multi-class EEG spatio-temporal filters. *Neural Process. Lett.* **2012**, *36*, 161–171.
17. Samek, W.; Blythe, D.; Müller, K.R.; Kawanabe, M. Robust spatial filtering with beta divergence. In Proceedings of the Advances in Neural Information Processing Systems, Stateline, NV, USA, 5–10 December 2013; pp. 1007–1015.
18. Cichocki, A.; Amari, S. Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities. *Entropy* **2010**, *12*, 1532–1568.
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), Jeongsun-Kun, Korea, 12–14 January 2015; pp. 1–4.

20. Tao, T. *Topics in Random Matrix Theory*; American Mathematical Society: Providence, RI, USA, 2012; Volume 132.
21. Blankertz, B.; Kawanabe, M.; Tomioka, R.; Hohlefeld, F.; Müller, K.R.; Nikulin, V.V. Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007; pp. 113–120.
22. 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.
23. Kang, H.; Nam, Y.; Choi, S. Composite common spatial pattern for subject-to-subject transfer. *IEEE Signal Process. Lett.* **2009**, *16*, 683–686.
24. 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.
25. Lu, H.; Plataniotis, K.N.; Venetsanopoulos, A.N. Regularized common spatial patterns with generic learning for EEG signal classification. 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. 6599–6602.
26. Xinyi Yong, R.K.W.; Birch, G.E. Robust Common Spatial Patterns for EEG Signal Preprocessing. In Proceedings of the IEEE EMBS 30th Annual International Conference, Vancouver, BC, Canada, 20–25 August 2008; pp. 2087– 2090.
27. 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, Munich, Germany, 7–12 September 2009.
28. Samek, W.; Binder, A.; Müller, K.R. Multiple kernel learning for brain-computer interfacing. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 7048–7051.
29. 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.
30. Arvaneh, M.; Guan, C.; Ang, K.K.; Quek, H.C. Spatially sparsed common spatial pattern to improve BCI performance. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 2412–2415.
31. Farquhar, J.; Hill, N.; Lal, T.N.; Schölkopf, B. Regularised CSP for sensor selection in BCI. In Proceedings of the 3rd International BCI workshop, Graz, Austria, 21–24 September 2006; pp. 1–2.
32. Yong, X.; Ward, R.K.; Birch, G.E. Sparse spatial filter optimization for EEG channel reduction in brain-computer interface. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Las Vegas, NV, USA, 31 March–4 April 2008; pp. 417–420.
33. Kawanabe, M.; Vidaurre, C.; Scholler, S.; Müller, 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.
34. Kawanabe, M.; Samek, W.; Müller, K.R.; Vidaurre, C. Robust common spatial filters with a maxmin approach. *Neural Comput.* **2014**, *26*, 349–376.
35. Wang, H.; Tang, Q.; Zheng, W. L1-norm-based common spatial patterns. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 653–662.
36. Park, J.; Chung, W. Common spatial patterns based on generalized norms. In Proceedings of the 2013 International Winter Workshop on Brain-Computer Interface (BCI), Jeongsun-kun, Korea, 18–20 February 2013; pp. 39–42.
37. Von Bünau, P.; Meinecke, F.C.; Király, F.C.; Müller, K.R. Finding stationary subspaces in multivariate time series. *Phys. Rev. Lett.* **2009**, *103*, 214101.
38. Samek, W.; Kawanabe, M.; Vidaurre, C. Group-wise stationary subspace analysis–A novel method for studying non-stationarities. In Proceedings of the International Brain–Computer Interfacing Conference, Graz, Austria, 22–24 September 2011; pp. 16–20.
39. Samek, W.; Müller, K.R.; Kawanabe, M.; Vidaurre, C. Brain-computer interfacing in discriminative and stationary subspaces. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), San Diego, CA, USA, 28 August–1 September 2012 ; pp. 2873–2876.

40. Von Bünau, P.; Meinecke, F.C.; Scholler, S.; Müller, K.R. Finding stationary brain sources in EEG data. In Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Buenos Aires, Argentina, 31 August–4 September 2010; pp. 2810–2813.
41. 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.
42. Müller-Gerking, J.; Pfurtscheller, G.; Flyvbjerg, H. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clin. Neurophysiol.* **1999**, *110*, 787–798.
43. Allwein, E.L.; Schapire, R.E.; Singer, Y. Reducing multiclass to binary: A unifying approach for margin classifiers. *J. Mach. Learn. Res.* **2001**, *1*, 113–141.
44. Dornhege, G.; Blankertz, B.; Curio, G.; Müller, K.R. Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 993–1002.
45. Grosse-Wentrup, M.; Buss, M. Multiclass common spatial patterns and information theoretic feature extraction. *IEEE Trans. Biomed. Eng.* **2008**, *55*, 1991–2000.
46. Naeem, M.; Brunner, C.; Leeb, R.; Graimann, B.; Pfurtscheller, G. Separability of four-class motor imagery data using independent components analysis. *J. Neural Eng.* **2006**, *3*, 208–216.
47. Barachant, A.; Bonnet, S.; Congedo, M.; Jutten, C. Multiclass brain–Computer interface classification by Riemannian geometry. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 920–928.
48. Zhang, H.; Yang, H.; Guan, C. Bayesian learning for spatial filtering in an EEG-based brain–computer interface. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *24*, 1049–1060.
49. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.
50. Edelman, A.; Arias, T.A.; Smith, S.T. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **1998**, *20*, 303–353.
51. Amari, S.I. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276.
52. Cruces, S.; Cichocki, A.; Amari, S. From Blind Signal Extraction to Blind Instantaneous Signal Separation. *IEEE Trans. Neural Netw.* **2004**, *15*, 859–873.
53. Nishimori, Y. Learning algorithm for ICA by geodesic flows on orthogonal group. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Washington, DC, USA, 10–16 July 1999; pp. 933–938.
54. BCI Competition III. Available online: http://www.bbci.de/competition/iii/ (accessed on 26 August 2014).
55. BCI Competition IV. Available online: http://www.bbci.de/competition/iv/ (accessed on 26 August 2014).
56. 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.
57. Blankertz, B.; Müller, K.R.; Krusienski, D.J.; Schalk, G.; Wolpaw, J.R.; Schlögl, A.; Pfurtscheller, G.; Millan, J.R.; Schröder, M.; Birbaumer, N. The BCI competition III: Validating alternative approaches to actual BCI problems. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2006**, *14*, 153–159.
58. Tangermann, M.; Müller, K.R.; Aertsen, A.; Birbaumer, N.; Braun, C.; Brunner, C.; Leeb, R.; Mehring, C.; Miller, K.J.; Mueller-Putz, G.; et al. Review of the BCI competition IV. *Front. Neurosci.* **2012**, *6*, 55.
59. Duda, R.O.; Hart, P.E. *Pattern Classification and Scene Analysis*; Wiley: New York, NY, USA, 1973; Volume 3.
60. The Divergence Methods Web Site. Available online: http://www.divergence-methods.org (accessed on 4 April 2016).
61. Machine Learning in Neural Engineering. Available online: http://brain-computer-interfaces.net/ (accessed on 29 January 2015).
62. Li, R. Rayleigh Quotient Based Optimization Methods For Eigenvalue Problems. In *Summary of Lectures Delivered at Gene Golub SIAM Summer School 2013*; Fudan University: Shanghai, China, 2013; pp. 1–27.