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

Javier Olias, Rubén Martín-Clemente, Mª Auxiliadora Sarmiento-Vega and Sergio Cruces

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

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

I. Introduction

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

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

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

The main contributions of the article are the following:

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

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

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

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

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

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

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

II. Basic model of the EEG observations

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A. Classical estimation of the class covariance matrices

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

V. Normalization of the power of the sources

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

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

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

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2905894, IEEE Transactions on Neural Systems and Rehabilitation Engineering

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

PREPROCESSING FOR TRAINING & TESTING
Freq. filtering & centering of the data $\forall\tau$.
Compute $\mathbf{C}_{\mathbf{X}_\tau}^{(0)}$ and $\mathbf{C}_{\mathbf{X}_\tau}^{(1)}$, $\forall\tau$, using (7)-(8).
Determine $\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_1}}^{(1)}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_2}}^{(1)}$ with (9).
METHOD FOR DIM. REDUCTION (STANDARD-CSP)
% Obtain the spatial filters solving (4)
$[\mathbf{V},\mathbf{D}] = eig(\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_1}}^{(1)}, \hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_2}}^{(1)})$
% Sort the N_x solutions
$[\sim, ind] = sort(diag(\mathbf{D}))$, $\mathbf{V} = \mathbf{V}(:, ind)$,
% Select the extreme p eigenvectors
$\mathbf{W} = [\mathbf{V}(:, 1:p/2), \mathbf{V}(:, N_x - p/2) + 1 : N_x]$
% Spatial filtering
$\mathbf{C}_{\mathbf{Y}_\tau} = \mathbf{W}^T \mathbf{C}_{\mathbf{X}_\tau}^{(0)} \mathbf{W}, \quad \tau = 1, \ldots, N_\tau.$
% Transf. for obtaining normal-like features
$\mathbf{f}_\tau = \log\left(\mathrm{diag}(\mathbf{C}_{\mathbf{Y}_\tau}) / \mathrm{sum}(\mathrm{diag}(\mathbf{C}_{\mathbf{Y}_\tau}))\right)$ (F1)
LEARNING THE LDA (BINARY) CLASSIFIER
% Using the training pairs $(c(\tau), \mathbf{f}_\tau)$
$\boldsymbol{\mu}_k = \frac{1}{N_{c_k}} \sum_{\tau: c(\tau) = c_k} \mathbf{f}_\tau \quad$ for $\quad k = 1, 2.$
$\hat{\mathbf{\Sigma}}_{\mathbf{f}|\mathbf{c_k}} = \frac{1}{N_{c_k}} \sum_{\tau: c(\tau) = c_k} (\mathbf{f}_\tau - \boldsymbol{\mu}_k)(\mathbf{f}_\tau - \boldsymbol{\mu}_k)^T, \quad k = 1, 2.$
$p(c_k) = N_{c_k} / (N_{c_1} + N_{c_2}), \quad k = 1, 2.$
$\hat{\mathbf{\Sigma}}_{\mathbf{f}} = p(c_1) \hat{\mathbf{\Sigma}}_{\mathbf{f}|\mathbf{c_1}} + p(c_2) \hat{\mathbf{\Sigma}}_{\mathbf{f}|\mathbf{c_2}}$
$\boldsymbol{\alpha} = \hat{\mathbf{\Sigma}}_{\mathbf{f}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$
$\boldsymbol{\beta} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{\log p(c_1) - \log p(c_2)}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \hat{\mathbf{\Sigma}}_{\mathbf{f}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$
CLASSIFICATION OF TEST DATA
% Implementation using LDA with equal cov.
$\mathbf{C}_{\mathbf{Y}_\tau} = \mathbf{W}^T \mathbf{C}_{\mathbf{X}_\tau}^{(0)} \mathbf{W}, \quad \tau \in Set\ of\ test\ trials.$
Evaluate $\mathbf{f}_\tau$ using the same formula as in (F1)
$\hat{c}(\tau) = c_k \ \ where \ \ k = 1.5 + 0.5\,\mathrm{sign}(\boldsymbol{\alpha}^T (\mathbf{f}_\tau - \boldsymbol{\beta}))$.

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

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

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

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

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

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

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

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

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

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

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

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

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

and to an improved estimate of the global covariance matrix

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Therefore, the covariance matrices estimates of each trial

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

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

TABLE II
PSEUDOCODE OF THE INSTANTANEOUS POWER-NORMALIZATION, WHICH PARTICULARIZES TO A VERSION OF TYLER'S METHOD IN STATISTICS FOR OBTAINING A ROBUST ESTIMATOR OF SCATTER.

PREPROCESSING FOR TRAINING TRIALS
Freq. filtering & centering of the data $\forall\tau$.
Set i=0, $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(0)} =< \mathbf{C}_{\mathbf{X}_\tau}^{(0)} >_\tau$
Repeat
i=i+1
$(\hat{\sigma}_{\tau,t}^{(i-1)})^2 = \frac{1}{N_x}\mathrm{Tr}\{(\boldsymbol{x}_\tau(t))^T(\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)})^{-1}\boldsymbol{x}_\tau(t)\}\ \forall\tau,t$
$\mathbf{C}_{\mathbf{X}_\tau}^{(i)} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\sigma}_{\tau,t}^{(i-1)})^{-2}\ \boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T\ \forall\tau$
$\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)} = \frac{1}{N_\tau}\sum_{\tau=1}^{N_\tau}\mathbf{C}_{\mathbf{X}_\tau}^{(i)}$
Until $\|\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)} - \hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)}\|_F / \|\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i)}\|_F < \epsilon$
Return $\mathbf{C}_{\mathbf{X}_\tau}^{(i)}\ \forall\tau$ and
$\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}^{(i)} = \frac{1}{N_{c_k}}\sum_{\tau:c(\tau)=c_k}\mathbf{C}_{\mathbf{X}_\tau}^{(i)} \quad k=1,\ldots,K.$
PREPROCESSING FOR A TESTING TRIAL τ
Freq. filtering & centering of the trial.
Given the last used estimate $\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)}$ in training...
Evaluate $(\hat{\sigma}_{\tau,t}^{(i-1)})^2 = \frac{1}{N_x}\mathrm{Tr}\{(\boldsymbol{x}_\tau(t))^T(\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{(i-1)})^{-1}\boldsymbol{x}_\tau(t)\}\ \forall t$
Return $\mathbf{C}_{\mathbf{X}_\tau}^{(i)} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\sigma}_{\tau,t}^{(i-1)})^{-2}\ \boldsymbol{x}_\tau(t)(\boldsymbol{x}_\tau(t))^T\ \forall\tau$

*Note: after this preprocessing the evaluation of the features no longer needs normalization, i.e.,

$\mathbf{f}_\tau = \log\left(\mathrm{diag}(\mathbf{C}_{\mathbf{Y}_\tau})\right)$ (F2)

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

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

VII. Gaussian Shrinkage LDA

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

overfitting. Moreover, the implementation of the classifier uses the within-class precision matrix of the features $\hat{\mathbf{\Sigma}}_{\mathbf{f}}^{-1}$, which in this situation may be poorly conditioned.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

VIII. Experimental results

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

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

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

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

$$D_R(\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}\ ,\ \mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}) = \min_{\alpha} \delta_R(\alpha\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}\ ,\ \mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}) \tag{38}$$

$$= \min_{\alpha}\left(\sum_{i=1}^{N_x}\log^2\frac{\lambda_i}{\alpha}\right)^{\frac{1}{2}} \tag{39}$$

$$= \left(\sum_{i=1}^{N_x}\log^2\frac{\lambda_i}{e^{\frac{1}{N_x}\sum_{j=1}^{N_x}\log\lambda_j}}\right)^{\frac{1}{2}} \tag{40}$$

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2905894, IEEE Transactions on Neural Systems and Rehabilitation Engineering

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

TABLE III

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

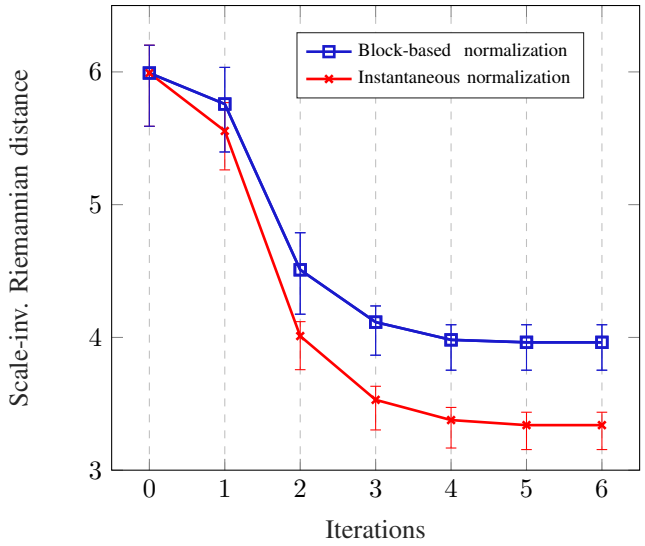


Fig. 1. Variations of the scale-invariant Riemannian distance (between the reference $\mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}$ and estimated $\hat{\mathbf{\Sigma}}_{\mathbf{x}|\mathbf{c_k}}$ covariance matrices) with respect to the number of iterations of the proposed power-normalization procedures.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

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

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

B. Experiments using the BCI competitions datasets

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

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

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

of training and testing trials is set to 40 and the number of spatial filters is set 8, except for those cases where a range of these values is specified.

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

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

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

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

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

IX. Conclusion

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

Appendix

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

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

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

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

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

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

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

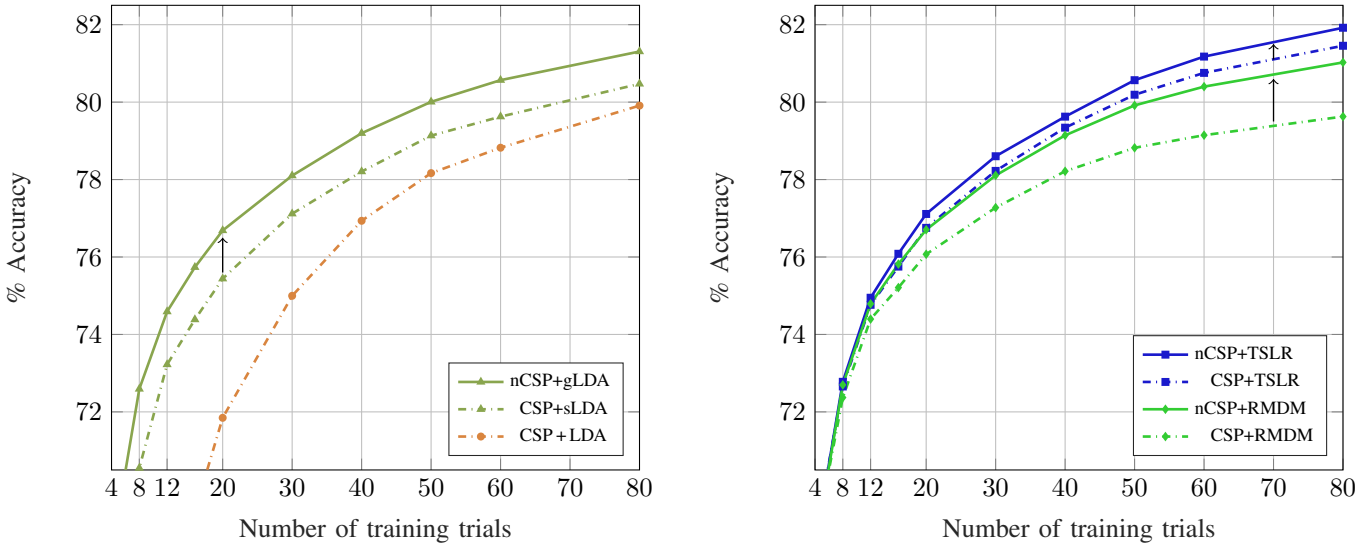
B. Equivalence with Tyler's method for estimation of scatter

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

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

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2905894, IEEE Transactions on Neural Systems and Rehabilitation Engineering



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

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

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

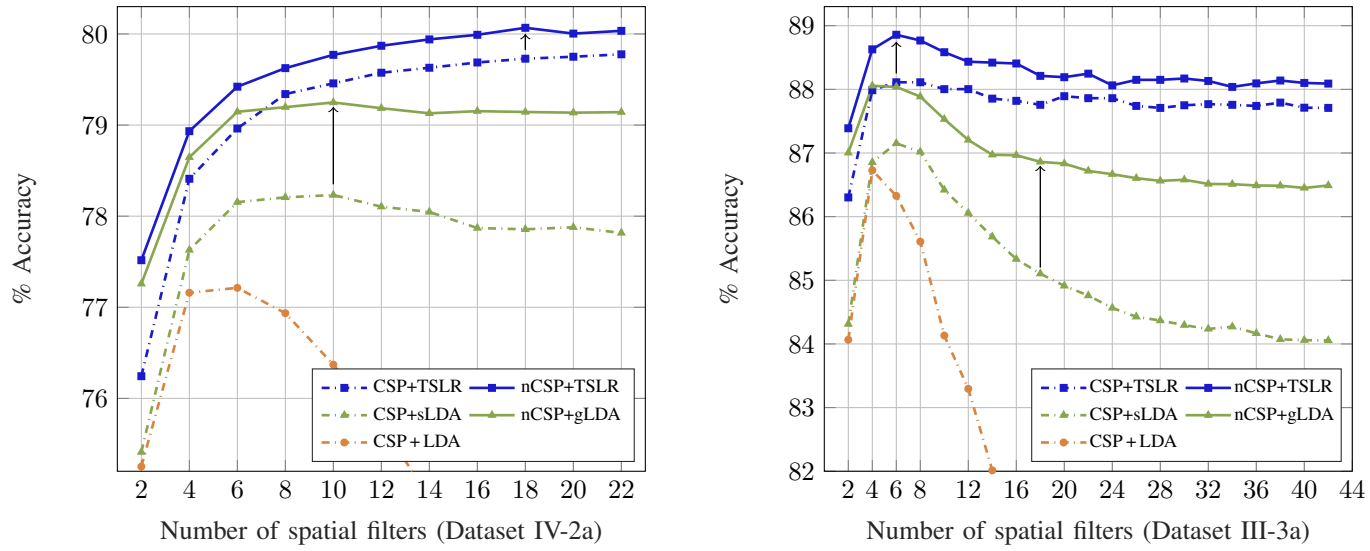


Fig. 3. Variations in performance of the MI-BCI binary classification methods with respect to p, the number of spatial filters. The results confirm the advantages of using nCSP in combination with the state-of-the-art classifiers to improve the expected user accuracy.

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

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

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

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

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

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

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

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

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

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

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

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

C. Procedure for generating random and locally perturbed covariance matrices

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

$$\mathbf{G} = \text{randn}(N_x);\ \mathbf{H}_0 = (\mathbf{G}+\mathbf{G}^T)/2 \tag{48}$$

$$\mathbf{H} = \text{rand}(1)\ (2.5\sqrt{N_x})\ \mathbf{H}_0/\|\mathbf{H}_0\|_{\mathbf{\Sigma}_{\mathbf{x}|\mathbf{c_k}}} \tag{49}$$

where the natural norm in the tangent space is given by

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

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

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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