Unsupervised Common Spatial Patterns

Rubén Martín-Clemente, *Member, IEEE*, Javier Olias, Sergio Cruces, *Senior Member, IEEE*, and Vicente Zarzoso, *Senior Member, IEEE*

***Abstract*—The common spatial pattern (CSP) method is a dimensionality reduction technique widely used in brain-computer interface (BCI) systems. In the two-class CSP problem, training data are linearly projected onto directions maximizing or minimizing the variance ratio between the two classes. The present contribution proves that kurtosis maximization performs CSP in an unsupervised manner, i.e., with no need for labeled data, when the classes follow Gaussian or elliptically symmetric distributions. Numerical analyses on synthetic and real data validate these findings in various experimental conditions, and demonstrate the interest of the proposed unsupervised approach.**

***Index Terms*—Common spatial patterns, brain computer interfaces, kurtosis.**

I. Introduction

COMMON spatial patterns (CSP) is a dimension reduction technique widely used in brain-computer interface (BCI) systems [1]–[4]. Typically, electroencephalogram (EEG) samples acquired under two different experimental conditions provide a multivariate data set with two classes. CSP linearly projects the data onto directions where the variance of the projected data points is significantly higher for one class than for the other [5]–[7]. The projected data variances can then be used as features for classification. CSP is a supervised technique, whose performance relies heavily on the availability of correctly labeled data.

The present contribution proves that CSP can be also performed in an *unsupervised* fashion by maximizing the kurtosis (normalized fourth-order moment) of the projected data. Unsupervised operation spares the need for training labels and is thus immune to erroneous labelling. Apart from its theoretical interest, this result is useful, for instance, in applications where the training labels are not available or may be uncertain. A mathematical proof is derived for data drawn from a mixture of Gaussian densities, and then generalized to elliptically symmetric distributions. Our experimental evaluation on synthetic and real data corroborates the theoretical findings. Unsupervised techniques are not unknown in EEG processing: e.g., [8] shows that it is possible to perform unsupervised workload classification using EEG spectral features. We can expect that they will become increasingly common in the near future.

The paper is organized as follows: Section II reviews the mathematical formulation of the common spatial patterns method. Section III, the core of our contribution, establishes the link between the kurtosis and the CSP criterion under the assumption of a Gaussian mixture model for the data. Section IV extends this result to elliptically distributed classes. Illustrative examples supporting the theoretical derivations are presented and discussed in Section V. The concluding remarks of Section VI bring the paper to an end. For the sake of clarity, proofs of the theoretical results have been deferred to the Appendices.

II. Common Spatial Patterns

Consider that we are given a set of observations of a random variable X in $\mathbb{R}^p$, in which each observation belongs to one of two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. CSP is usually applied to problems where the class means are null, and this assumption is made in the sequel. A one-dimensional projection of the point cloud can be represented by $Y = \boldsymbol{a}^\mathsf{T} X$, where $\boldsymbol{a} \in \mathbb{R}^p$. Denoting by $\boldsymbol{\Sigma}_k = \mathrm{Cov}(X|\mathcal{C}_k)$ the covariance matrix of the data in class $\mathcal{C}_k$, with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, it holds that the variance of class $\mathcal{C}_k$, after the projection, equals

$$\sigma_k^2 = \boldsymbol{a}^\mathsf{T} \boldsymbol{\Sigma}_k \boldsymbol{a}, \quad k = 1, 2. \tag{1}$$

The idea behind CSP is to maximize σ_1^2 while minimizing σ_2^2 or *vice versa* [9]. To this end, the objective function is defined as the power ratio

$$R(\boldsymbol{a}) = \frac{\sigma_1^2}{\sigma_2^2}. \tag{2}$$

Note that σ_k^2, $k = 1, 2$, depend on $\boldsymbol{a}$ through relation (1). Also, ratio (2) is scale invariant, i.e., $R(\boldsymbol{a}) = R(c\boldsymbol{a})$, for all $c \in \mathbb{R}\backslash\{0\}$, and therefore only the direction of the projection is significant but not the overall scaling. To find the optimal $\boldsymbol{a}^*$ corresponding to the extremum (maximum or minimum) of CSP criterion (2) we set its gradient $\nabla R(\boldsymbol{a})$ to zero, readily yielding

$$\boldsymbol{\Sigma}_1 \boldsymbol{a}^* = R(\boldsymbol{a}^*) \boldsymbol{\Sigma}_2 \boldsymbol{a}^*. \tag{3}$$

It follows that $\boldsymbol{a}^*$ is a generalized eigenvector of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ [10]. Solving this problem we get the eigenvector

Manuscript received May 8, 2019; revised July 28, 2019; accepted August 5, 2019. Date of publication September 19, 2019; date of current version October 8, 2019. This work was supported in part by the Spanish Ministry of Economy and Competitiveness under Project TEC2017-82807-P and in part by the European Regional Development Fund (ERFD) of the European Commission. The work of R. Martín-Clemente was supported by the I3S Laboratory through the IFA Scholarship from Université Côte d'Azur in July 2018. Most of this work was carried out while V. Zarzoso was a member of the *Institut Universitaire de France. (Corresponding author: Rubén Martín-Clemente.)*

R. Martín-Clemente, J. Olias, and S. Cruces are with the Department of Signal Theory and Communications, University of Seville, 41092 Seville, Spain (e-mail: ruben@us.es; folias@us.es; sergio@us.es).

V. Zarzoso is with Université Côte d'Azur, CNRS, I3S Laboratory, 06903 Sophia Antipolis, France (e-mail: vicente.zarzoso@univ-cotedazur.fr).

Digital Object Identifier 10.1109/TNSRE.2019.2936411

corresponding to the maximum (respectively, minimum) eigenvalue, which maximizes (resp. minimizes) the CSP objective function. When more features are needed for the posterior classification stage, a common practice is to project the data onto other eigenvectors from both ends of the eigenvalue spectrum [11]. In the generalized eigenvalue (GEVD) problem (3), $\boldsymbol{a}_i$ denote the eigenvectors and their corresponding eigenvalues are assumed to be sorted in decreasing order: $R(\boldsymbol{a}_i) > R(\boldsymbol{a}_j)$, for $i < j$, $i = 1, 2, \ldots, p$.

It is important to remark that a training set of correctly classified observations is required to estimate matrices $\boldsymbol{\Sigma}_k$, $k = 1, 2$. It is in this sense that CSP can be considered a *supervised* technique. The remaining of this paper presents a fully data-driven procedure that does not require labeled samples, thus performing CSP in an unsupervised manner.

III. Kurtosis as a Blind CSP Criterion

We first focus on the important case where X is distributed as a mixture of two Gaussian densities [11]:

$$X \sim \pi_1 \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2) \tag{4}$$

where π_k stands for the prior probability of class $\mathcal{C}_k$, $k = 1, 2$, with $\pi_1 + \pi_2 = 1$, and $\mathbf{0}$ is a p-dimensional vector of zeros. The distribution of the one-dimensional projection $Y = \boldsymbol{a}^\top X$ is also a mixture of Gaussians, that is, $Y \sim \pi_1 \mathcal{N}(0, \sigma_1^2) + \pi_2 \mathcal{N}(0, \sigma_2^2)$, where σ_1^2 and σ_2^2 are defined as in eqn. (1). From the properties of the Gaussian distribution, it follows that

$$\mathrm{E}\{Y^2\} = \pi_1\sigma_1^2 + \pi_2\sigma_2^2, \quad \mathrm{E}\{Y^4\} = 3\pi_1\sigma_1^4 + 3\pi_2\sigma_2^4$$

where $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation operator.

Now, the kurtosis is a statistic defined as the normalized fourth-order moment [12]

$$\kappa_Y(\boldsymbol{a}) := \mathrm{E}\{Y^4\}/\mathrm{E}\{Y^2\}^2. \tag{5a}$$

Under data model (4), the kurtosis can be expressed as:

$$\kappa_Y(\boldsymbol{a}) = \frac{3\pi_1\sigma_1^4 + 3\pi_2\sigma_2^4}{(\pi_1\sigma_1^2 + \pi_2\sigma_2^2)^2}. \tag{5b}$$

Some preliminary manipulations show that $\nabla\kappa(\boldsymbol{a}) = \mathbf{0}$ if and only if we have $(R(\boldsymbol{a}) - 1)\nabla R(\boldsymbol{a}) = \mathbf{0}$, meaning that the critical points of $R(\boldsymbol{a})$ are also critical points of $\kappa(\boldsymbol{a})$. Indeed, a thorough analysis detailed in Appendix A leads to the following result:

Theorem 1: Under the working assumptions of CSP recalled in Sec. II, the local maximizers of the kurtosis (5b) maximize or minimize the CSP criterion (2). In particular, the maximizers of kurtosis are $\boldsymbol{a}_1$ if $R(\boldsymbol{a}_1) > 1$ and $\boldsymbol{a}_p$ if $R(\boldsymbol{a}_p) < 1$.

In other words, CSP can be performed blindly, that is, without the need for labeled samples, by projecting the observed data points onto the direction that maximizes the kurtosis of the projections. Consequently, the new method will be referred to as kurtosis-based unsupervised CSP (k-uCSP). Under different working assumptions, namely that the data are a combination of statistically independent variables, the optimization of the kurtosis gives rise to independent component analysis (ICA) [13], [14]. Furthermore, in clustering problems, maximizing the kurtosis can produce the same results as Fisher linear discriminant analysis [15]. We see that the kurtosis is a widely-used tool in signal processing. It is the true model the data follow that determines the outcome achieved.

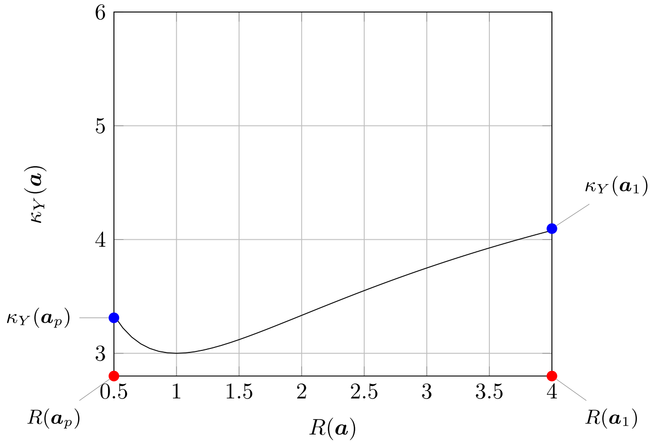


Fig. 1. The kurtosis versus the CSP objective function for $\pi_1 = \pi_2 = 1/2$ with $0.5 = R(\boldsymbol{a}_p) \leqslant R(\boldsymbol{a}) \leqslant R(\boldsymbol{a}_1) = 4$.

It should be remarked that the expression of kurtosis given in eqn. (5b) is only exploited to prove the equivalence between kurtosis optimization and common spatial pattern analysis in our theoretical derivations, but we use eqn. (5a) in the actual algorithm to find the patterns. Therefore, the proposed technique is fully unsupervised.

Finally, Theorem 1 also admits an intuitive interpretation. Dividing both the numerator and denominator of (5b) by σ_2^4, the right-hand part of this formula can be expressed in terms of $R(\boldsymbol{a})$,

$$\kappa_Y(\boldsymbol{a}) = \frac{3\pi_1 R^2(\boldsymbol{a}) + 3\pi_2}{(\pi_1 R(\boldsymbol{a}) + \pi_2)^2}, \tag{6}$$

thus revealing that there exists a relationship between the supervised and unsupervised criteria. Fig. 1 shows an example plot of κ_Y against R when $\pi_1 = \pi_2$ and R is in the range $[0.5, 4]$. We make the observation that the maximizers of the kurtosis can lie only in the strictly increasing/decreasing range of (6). Consequently, (6) is also increasing/decreasing in some neighborhood of them and, therefore, invertible. Then, a decrease in the value of the kurtosis in that neighborhood results in a decrease/increase in the value of R. As it is intuitive, this implies that the local maximizers of the kurtosis maximize or minimize the CSP criterion. In Figure 1, additionally, it is not hard to show that as (6) is decreasing for all vectors sufficiently near $\boldsymbol{a}_p$, then (6) has a maximum at $\boldsymbol{a}_p$. Virtually the same argument shows that there is another maximum at $\boldsymbol{a}_1$.

IV. Extension to Elliptic Distributions

The above result can be extended to the case where the data follow a mixture of zero-mean elliptically symmetric distributions [16]. These distributions generalize the class of multivariate Gaussians by allowing for both heavier-than-Gaussian and lighter-than-Gaussian distribution tails. Examples include the *t* and Laplace distributions. Before continuing, a zero-mean p-dimensional random variable Z is *elliptically distributed* if its characteristic function $\varphi_Z(\boldsymbol{a}) := \mathrm{E}\{e^{j\boldsymbol{a}^\top Z}\}$, $j = \sqrt{-1}$, $\boldsymbol{a} \in \mathbb{R}^p$, is of the form $\varphi_Z(\boldsymbol{a}) = \phi\left(-\frac{1}{2}\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a}\right)$ for some nonnegative-definite matrix $\boldsymbol{\Sigma}$. Function $\phi(\cdot)$ is called *characteristic generator* of the distribution. For example, for Gaussian variables $\phi(\alpha) = \exp(\alpha)$ and for multivariate Laplacian variables $\phi(\alpha) = 1/(1-\alpha)$. The characteristic function of the univariate random variable $W = \boldsymbol{a}^\top Z$ is given by $\varphi_W(t) := \mathrm{E}\{e^{jtW}\} = \mathrm{E}\{e^{jt\boldsymbol{a}^\top Z}\} = \varphi_Z(t\boldsymbol{a}) = \phi\left(-\frac{1}{2}t^2\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a}\right)$, $t \in \mathbb{R}$. It follows that, if the variance of W exists, then

$$\sigma^2 := E\{W^2\} = -\left.\frac{\partial^2}{\partial t^2}\varphi_W(t)\right|_{t=0} = \phi'(0)\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a} \qquad (7)$$

$$E\{W^4\} = \left.\frac{\partial^4}{\partial t^4}\varphi_W(t)\right|_{t=0} = 3\gamma\,\sigma^4 \qquad (8)$$

where we have defined $\gamma := \frac{\phi''(0)}{\phi'(0)^2}$, which is a strictly positive number. Notations ϕ' and ϕ'' represent, respectively, the first- and second-order derivative of ϕ.

Now, let us replace the Gaussian distributions in (4) with zero-mean elliptically symmetric distributions having characteristic generators of the same type $\phi(\cdot)$ but defined by matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, i.e., $\phi(-\boldsymbol{a}^\top\boldsymbol{\Sigma}_1\boldsymbol{a}/2)$ and $\phi(-\boldsymbol{a}^\top\boldsymbol{\Sigma}_2\boldsymbol{a}/2)$, respectively. Then, the moments of Y become

$$\mathrm{E}\{Y^2\} = \pi_1\sigma_1^2 + \pi_2\sigma_2^2, \quad \mathrm{E}\{Y^4\} = 3\pi_1\gamma\,\sigma_1^4 + 3\pi_2\gamma\,\sigma_2^4$$

with σ_k^2, $k = 1, 2$, given by (7), and their ratio turns out to be a positively scaled version of (5b):

$$\kappa_Y(\boldsymbol{a}) = \gamma\,\frac{3\pi_1\sigma_1^4 + 3\pi_2\sigma_2^4}{(\pi_1\sigma_1^2 + \pi_2\sigma_2^2)^2}. \qquad (9)$$

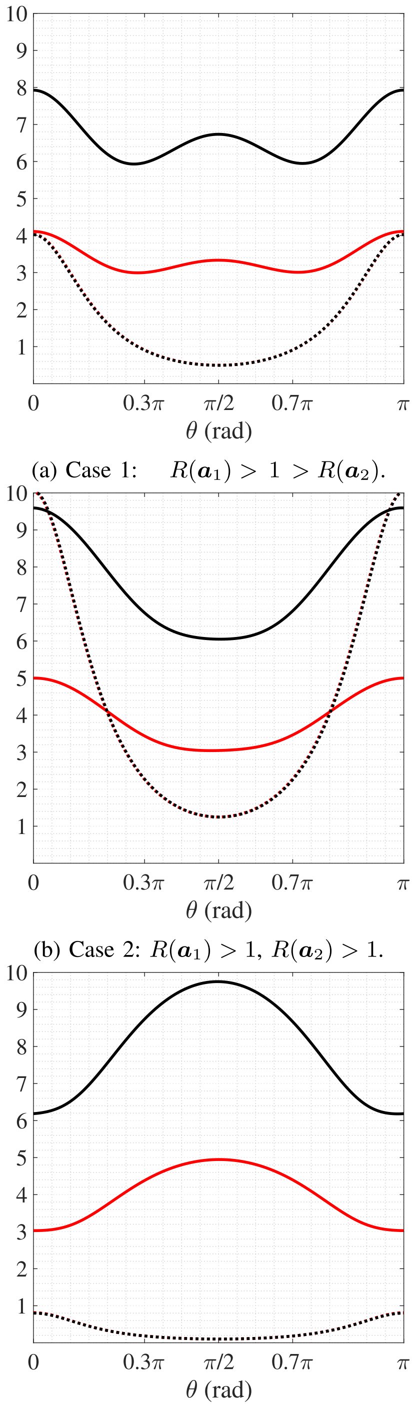
It readily follows that Theorem 1 also holds for data classes with elliptic distributions defined by the same type of characteristic generator $\phi(\cdot)$.

V. Experimental Assessment

A number of experiments are performed to validate the theoretical study of the unsupervised CSP criterion developed in this paper and to test its performance in a variety of experimental conditions. These include synthetically generated as well as real EEG data. To perform the numerical optimization of the kurtosis statistic (5a), we employ the algorithm presented in [24]. We point out that this algorithm does not require any class labels as inputs. A free MATLAB implementation of the algorithm is provided in [25].

A. Simulated Data

To illustrate Theorem 1, let us first consider a mixture of equiprobable classes in a two-dimensional space, i.e., $p = 2$. We consider three cases that only differ in the choice of matrix $\boldsymbol{\Sigma}_2$. Case 1 assumes that $\boldsymbol{\Sigma}_1 = \mathrm{diag}(2, 1)$ and $\boldsymbol{\Sigma}_2 = \mathrm{diag}(0.5, 2)$; in case 2, $\boldsymbol{\Sigma}_2$ is replaced by $\mathrm{diag}(0.2, 0.8)$; in case 3, finally, we set $\boldsymbol{\Sigma}_2 = \mathrm{diag}(2.5, 10)$. In all cases, the generalized eigenvector of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ that maximizes the CSP target function (2) is $\boldsymbol{a}_1 = \pm[1, 0]^\top$, while the corresponding minimizer is $\boldsymbol{a}_2 = \pm[0, 1]^\top$.



(a) Case 1: $R(\boldsymbol{a}_1) > 1 > R(\boldsymbol{a}_2)$.

(b) Case 2: $R(\boldsymbol{a}_1) > 1$, $R(\boldsymbol{a}_2) > 1$.

(c) Case 3: $R(\boldsymbol{a}_1) < 1$, $R(\boldsymbol{a}_2) < 1$.

Fig. 2. Criteria $R(\boldsymbol{a})$ (dotted) and $\kappa_Y(\boldsymbol{a})$ (solid) as a function of angle θ defining the projection direction $\boldsymbol{a} = [\cos(\theta), \sin(\theta)]^\top$. Red curves are calculated from Gaussian data, while black curves correspond to Laplacian classes. Dotted curves (R) overlap as the pair $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ is the same for both distributions. As predicted by Theorem 1, the local maximizers of kurtosis (solid lines) either maximize or minimize the CSP criterion (dotted lines) depending on the generalized eigenvalues $R(\boldsymbol{a}_1)$ and $R(\boldsymbol{a}_2)$ of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Statistics are estimated from $T = 1000$ random samples, with 500 i.i.d. samples per class, and each curve is the average of 100 independent experiments.

Figure 2 plots the CSP criterion R [eqn. (2)] and the kurtosis criterion κ_Y [computed from the ratio of expectations in eqn. (5b)] in dotted and solid lines, respectively, for

$\boldsymbol{a} = [\cos(\theta), \sin(\theta)]^\intercal$. Red and black solid lines correspond to Gaussian and Laplacian data, respectively, where the latter were generated as explained in [17]. The results are just as predicted by Theorem 1: the maxima of κ_Y are always maxima or minima of the CSP criterion R, with different patterns of correspondence depending on the generalized eigenvalues of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. In case 1 (Figure 2a), the maxima and minima of R transform into maxima of κ_Y, as $R(\boldsymbol{a}_1) = 4 > 1 > R(\boldsymbol{a}_2) = 0.5$. In case 2 (Figure 2b), κ_Y has the same maxima and minima as R, because both eigenvalues are greater than one: $R(\boldsymbol{a}_1) = 10$, $R(\boldsymbol{a}_2) = 1.25$. Finally, in case 3 (Figure 2c), the minima of R are transformed into maxima of κ_Y and *vice versa*, since both eigenvalues are lower than one: $R(\boldsymbol{a}_1) = 0.8$, $R(\boldsymbol{a}_2) = 0.1$. Additionally, the point $R = 1$, reached in case 1, defines a global minimum of the kurtosis.

As an additional validation experiment, let us also test the case where samples from the same class are correlated, resulting in non diagonal covariance matrices. The following matrices are selected at random:

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 3.8152 & -3.4131 \\ -3.4131 & 3.3104 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2.8465 & 0.5267 \\ 0.5267 & 1.2446 \end{bmatrix}.$$

For these covariances, the maximizer $\boldsymbol{a}_1$ of R is the unit vector that makes an angle of $\theta_1 \approx 2\pi/3$ radians with the positive x-axis. Similarly, the minimizer $\boldsymbol{a}_2$ is at an angle $\theta_2 \approx \pi/4$ radians. Figure 3 (top) plots R and κ_Y in a format similar to that of Figure 2. An alternative representation is given in Figure 3 (bottom). The color-coded circles indicate the kurtosis of the entire projected data (outer circle) and the variance ratio of the projected classes (inner circle) as a function of angle θ. For the reader's convenience, we complete the figure by drawing a pair of dashed lines in the direction of vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. Additionally, we draw in different colours the scatterplots of the classes and the probability density functions of the data points after being projected onto $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, which illustrates well the disparity between the variances resulting from the projection.

Furthermore, it is known that whitening the data reduces the initial generalized eigenvalue problem (3) to a standard eigenvalue problem for Hermitian matrices (see Appendix B), for which the eigenvectors are orthogonal. To show this property, we whiten the previous data before calculating R and κ_Y. As a result, Fig. 3, obtained before whitening, transforms into Fig. 4 after whitening. As expected, Fig. 4 (bottom) shows that the directions that maximize the kurtosis become perpendicular. Yet the equivalence between the CSP and kurtosis criteria established by Theorem 1 still holds in this case.

B. Real Data

1) Supervised vs. Unsupervised CSP of Brain Data: In typical BCI implementations, subjects are instructed to imagine movements of, e.g., their right or left hand, one after the other, while their p-channel EEG is being recorded. Then, CSP is applied to the data, previously bandpass-filtered in the band of interest, and the power evolution of the resulting time series allows the BCI system to discriminate between the two classes of motor imagery [18]–[20].

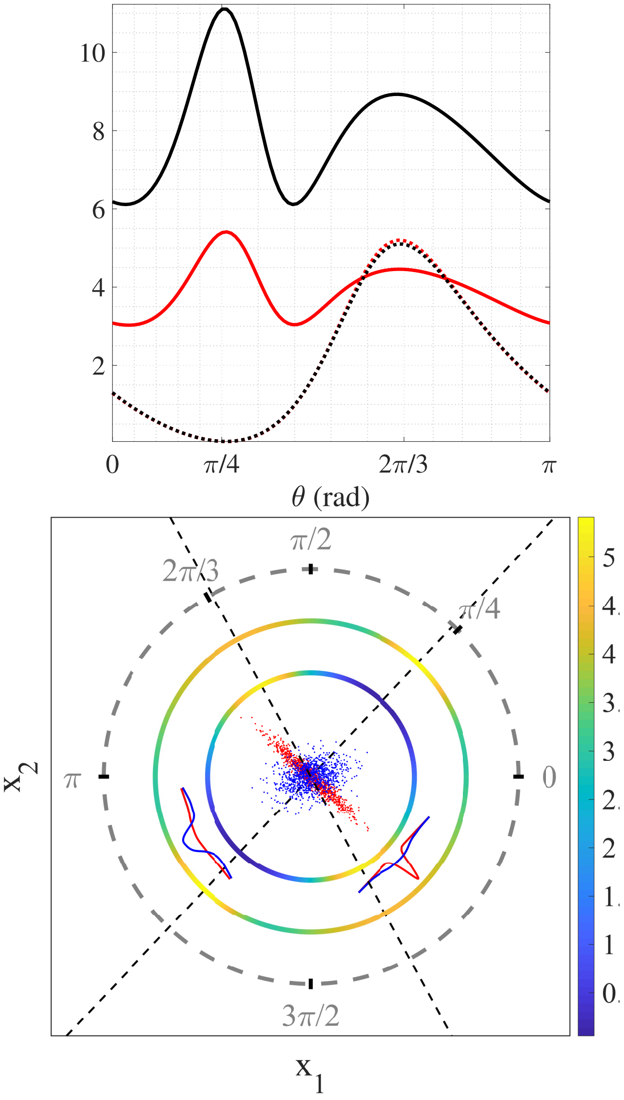


Fig. 3. Supervised vs. unsupervised CSP when samples from the same class are correlated, with $R(\boldsymbol{a}_1) \approx 5.4 > 1 > R(\boldsymbol{a}_2) \approx 0.06$. (Top) Criteria $R(\boldsymbol{a})$ (dotted) and $\kappa_Y(\boldsymbol{a})$ (solid) as a function of angle θ defining the projection direction $\boldsymbol{a} = [\cos(\theta), \sin(\theta)]^\intercal$. As in Fig. 2, red curves are calculated from random Gaussian samples, and black curves from Laplacian data. (Bottom) Alternative representation as a scatterplot of the two correlated Gaussian distributions. Samples from one class are shown as red dots, and from the other as blue dots. The two color-coded circles around the scatterplots indicate the value of $R(\boldsymbol{a})$ (inner circle) and $\kappa_Y(\boldsymbol{a})$ (outer circle). Straight dashed lines mark the directions representing the CSP and kurtosis projections, which coincide in this experiment. The probability density functions of the projected classes are also shown in red and blue lines.

The purpose of this experiment is to illustrate the performance of the k-uCSP approach on real EEG data. The proposed kurtosis-based approach is tested using datasets from the BCI competition IV (dataset 2a) [21]. These contain EEG acquired on two different days from nine subjects with a $p = 22$-channel EEG system at a sampling rate of 250 Hz. As pre-processing, we bandpass filter the EEG to $8 - 30$ Hz. This is usual in BCI and ensures that the data are zero-mean. Electrooculogram (EOG) channels are available and ocular artifacts are removed using Signal Space Projections (SSP) [22], [23].

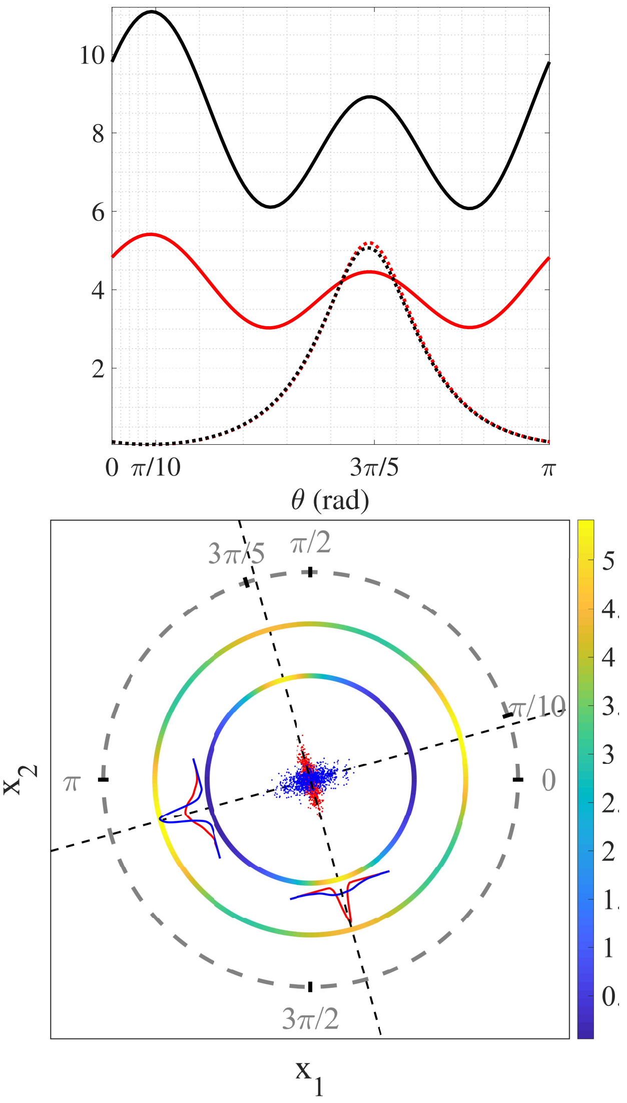


Fig. 4. Results for the data in Fig. 3 after whitening. Apart from the whitening operation, details about the plots are as in the caption of Fig. 3.

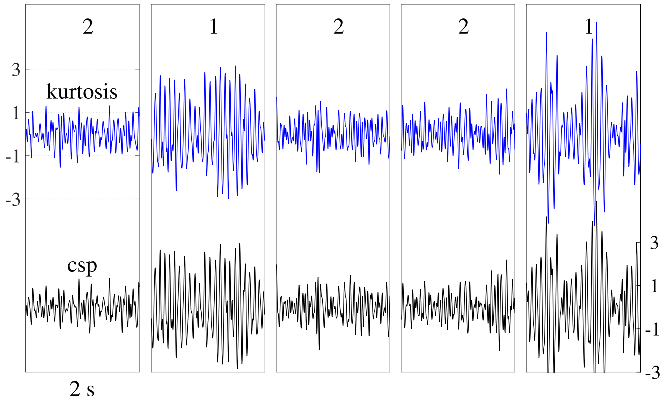


Fig. 5. One-dimensional projection of EEG data with maximal kurtosis (blue line). Several projected trials from the first user are shown. Labels '1' and '2' correspond to data classes ('1' = left hand, '2' = right hand). Visible to the naked eye, the power of one of the projected classes (namely, class '1') is significantly higher, as would be expected from the application of CSP. For comparison, the projection of the same trials onto the corresponding supervised CSP filter are also shown (bottom black curves).

In each trial, an arrow pointing either to left, right, down or up is shown on a display for a short time, and the subject is required to respectively imagine left hand, right hand, both feet or tongue movements in response. The imagined action lasts about three seconds but only the final two are kept to avoid initial transient effects. Thus, each trial data matrix consists of 22 rows of channels and 2 (seconds) × 250 (samples/seconds) = 500 columns of time samples. Data matrices are also normalized as suggested in [26], which generalizes the procedure in [27]. The test statistic proposed in [28] shows that about $2/3$ of the data follow an elliptical distribution with a standard level of significance of $\alpha = 0.01$. Finally, 72 trials of each movement are performed per subject and day.

Experiments are conducted on groups of trials. Each group comprises 144 trials, half from one of the four classes (e.g., left hand) and half from another (e.g., right hand), all from the same subject and recorded the same day. The corresponding 144 trial data matrices are concatenated along the temporal dimension, producing one large observation matrix with dimension 22×72000. In total, there are 9 (subjects) × 6 (possible combinations of four classes) × 2 (days) = 108 groups. The projections of the p-channel EEG onto the p-directions maximizing the kurtosis are computed for each group of trials by the algorithm in [24], [25]. Specifically, after applying a whitening transformation, we maximize p-times the kurtosis, one after the other, under the constraint that the direction obtained in the nth maximization is orthogonal to the previously calculated directions (see Appendix B for details).

For illustration purposes, Figure 5 shows several projected trials with maximum kurtosis, calculated from 'left-hand' and 'right-hand' EEG data recorded from subject 1 during the second session. In this particular experiment, the power of the left-hand motor imagery projections (denoted '1') is clearly higher than that of the other hand (denoted '2'), which facilitates class discrimination. A scatter plot of log-variances for all projected trials is also shown in Figure 6, using the supervised projection on the x-axis and the unsupervised kurtosis-based projection on the y-axis.

As a more formal test of the relationship between both criteria (supervised and unsupervised), we also project the EEG data onto the $p = 22$ generalized eigenvectors of the matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. For each projection direction, we calculate the ratio of variances of the projected classes (σ_1^2/σ_2^2). These ratios are then arranged in decreasing order along a p-dimensional vector $\boldsymbol{b}_{CSP}$. A distinct vector is generated for each group of trials, as the projection directions differ from one another. Next, we repeat the experiment with the difference that this second time the data are projected onto the p directions maximizing the kurtosis. The new variance ratios are stored in a vector $\boldsymbol{b}_{Kurt}$. A strong similarity is found between the vectors $\boldsymbol{b}_{CSP}$ and $\boldsymbol{b}_{Kurt}$ calculated from the same data. The Pearson correlation coefficient between the components of these two vectors, averaged for all movements, is 0.9692 for subject 6 (best), and 0.9484 for subject 9 (worst).

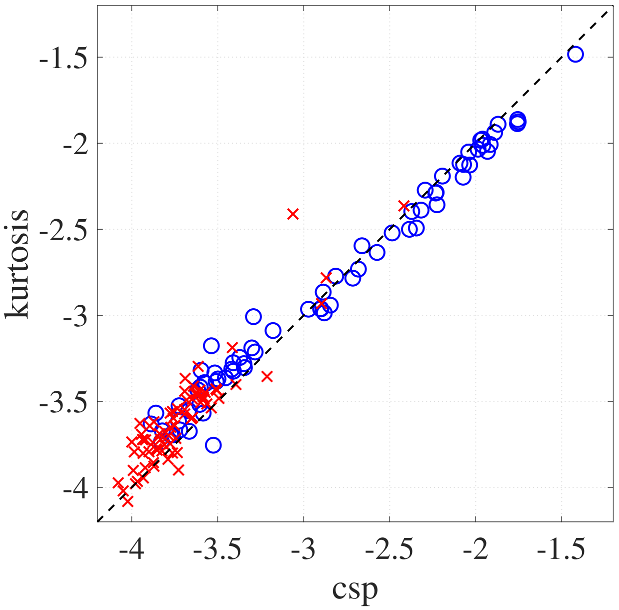


Fig. 6. Scatter plot of log-variances for all ‘left-hand’ vs ‘right-hand’ trials recorded the second day from subject 1, using the supervised projection which maximizes R on the horizontal axis and the corresponding unsupervised projection with maximal kurtosis on the vertical axis. Legend: blue circles = left-hand, red crosses = right-hand.

The correlation coefficient averaged over all groups of trials is 0.9586 with standard deviation 0.0257. Both approaches therefore provide very similar variance ratios, though one method is supervised while the other is not.

2) Unsupervised Discrimination: The k-uCSP approach may be combined with some unsupervised classification technique in order to design a full data-driven BCI system. Some preliminary illustrative experiments are presented in the remaining of this Section. Let $Y_1, \ldots, Y_p$ be the unsupervised projections that result from projecting a single trial data matrix. These matrices are 22×500, so that each Y_i consists of 500 values. As in [27], their variances are used to calculate the following $p = 22$ features

$$F_i = \log\left(\frac{\mathrm{var}(Y_i)}{\sum_{n=1}^{p}\mathrm{var}(Y_n)}\right), \quad i = 1, \ldots, p, \tag{10}$$

which are arranged in a p-vector.

For illustrative purposes only, Figure 7 draws the scatter-plot of 80 log-variances F_i computed from ‘left vs right-hand’ trials from subject 1. Superimposed, we draw the contour plot of the best fitting two-component Gaussian mixture model (GMM) of the point cloud [29]. By assigning a point to the Gaussian distribution it most probably belongs to, we partition the space in two regions with the hope that these regions accurately reflect the original classes. The performance of the classification can be evaluated by using the true labels as ground truth. For example, in Figure 7, when the subject thinks of a ‘left-hand’ movement, 27 times out of 40 (about 2 times out of 3) it is assigned to Region 1. Similarly, ‘right-hand’ movements are 36 times of 40 (9 times out of 10) assigned to Region 2. As both kind of movements are equiprobable, the average probability of success or accurate classification can be calculated as

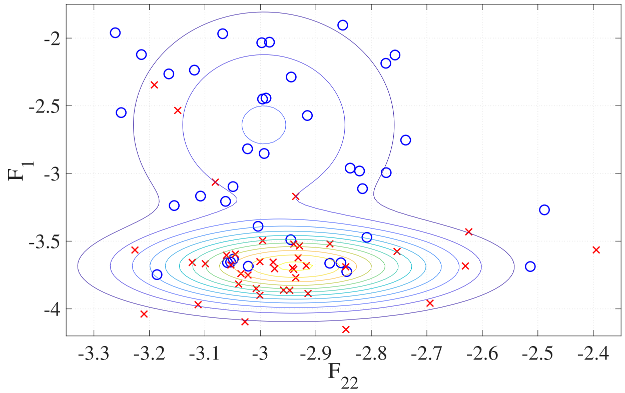


Fig. 7. Scatter plot of the k-uCSP features F_1 and F_{22}, calculated by (10) from ‘left vs right-hand’ trials from subject 1 (circles = left-hand, crosses = right-hand). The features are sorted according a variance criterion: $\mathrm{var}(F_1) > \ldots > \mathrm{var}(F_p)$, where the variance is calculated across all the trials within a group. Superimposed we show the contour plot of the two-component best fitting Gaussian mixture model of the points.

$$P = \frac{27}{40} \times \frac{1}{2} + \frac{36}{40} \times \frac{1}{2} = 0.7875.$$

Our last experiments are conducted on each group of 144 trials. We run a 10 fold cross-validation. To this end, the trials are divided into training (130 trials) and testing (14 trials). In the training phase, we calculate the unsupervised projection directions from the 130 trials in the training set. Then, we fit a two-component p-dimensional GMM to the resulting 130 log-variance feature vectors. Next, in the testing phase, the 14 trials of the testing set are projected onto the previously calculated directions, and the GMM is used for classifying the corresponding 14 p-vectors of log variance features. This is repeated 10 times, with different training and testing sets, and the results are averaged out.

We observe that the performance largely depends on the type of motor imagery. For example, for subject 1, best results are obtained for the classification of ‘right-hand’ vs ‘tongue’ movements: we get 96.48 % of accuracy in session 1 and 97.19 % in session 2, 96.83 % in average. However, ‘feet’ and ‘tongue’ movements are hardly distinguishable one from another: only the 56.19 % (session 1), 61.81 % (session 2) and 59.0 % (average) of the corresponding trials are correctly classified. The following averaged classification accuracies are obtained for the combination of imagined movements with the best performance:

- Subject 1: 96.83 % (‘right-hand’ vs ‘tongue’),
- Subject 2: 65.57 % (‘feet’ vs ‘tongue’),
- Subject 3: 80.45 % (‘right-hand’ vs ‘tongue’),
- Subject 4: 72.52 % (‘left-hand’ vs ‘feet’),
- Subject 5: 62.88 % (‘left-hand’ vs ‘tongue’),
- Subject 6: 64.33 % (‘right-hand’ vs ‘feet’),
- Subject 7: 79.57 % (‘left-hand’ vs ‘tongue’),
- Subject 8: 93.07 % (‘left-hand’ vs ‘tongue’),
- Subject 9: 91.59 % (‘left-hand’ vs ‘tongue’).

The average of all these numbers equals 78.31 %, not far from the mean classification accuracy (82.3 %) obtained

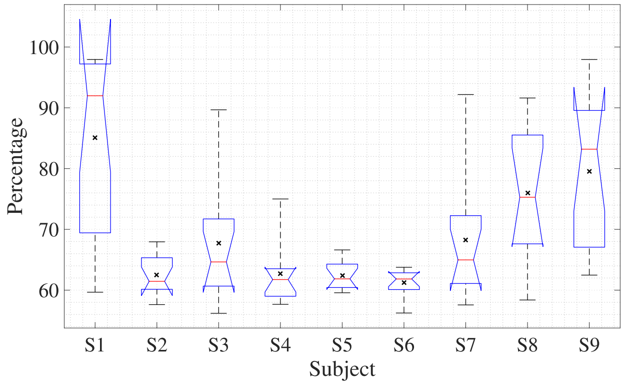


Fig. 8. Box plots of the unsupervised classification accuracy for each user. The boxes extend from the 25th to the 75th percentiles. Lines at either end of the boxes cover the maximum and minimum values. The red line in the middle is the median (50th percentile). Notches represent a confidence interval around the median. The black crosses are the mean values (see also the first column of Table II).

TABLE I

Classification Accuracy (in Percentage) Associated With the Combination of Imagined Movements With the Best Performance for the Unsupervised (k-uCSP With GMM Classifier) and Supervised (CSP and Fisher LDA) Approaches

Subject	Unsupervised	Supervised
S1	96.83	98.59
S2	65.57	86.69
S3	80.45	94.78
S4	72.52	77.9
S5	62.88	70.85
S6	64.33	69.11
S7	79.57	97.19
S8	93.07	93.40
S9	91.59	93.04

by the unsupervised workload classification approach in [8]. It is a good performance considering its unsupervised nature. For comparison, Table I also gives the results obtained by the supervised CSP approach, where supervised classification is carried out with Fisher linear discriminant analysis (LDA) [30].

The complete results are shown in Figure 8, grouped by subject. For example, the accuracy of subject 1 ranges from 59.19 % to 97.19 %, depending on the motor imagery under consideration. Table II presents the mean value for each subject. For comparison, Table II also shows the accuracy obtained by using the supervised CSP approach and Fisher linear discriminant.

Complementarily, Fig. 9 and Table III show the performance of the criterion for each imagined movement, averaged across the subjects.

A natural question to ask is why there is a marked difference in some data sets between the performance of the traditional supervised approach and the proposed unsupervised one. To address this question, we have investigated the separation between the classes in the feature space using t-Distributed Stochastic Neighbor Embedding (t-SNE) [31], [32]. This is a popular technique for the visualisation of high-dimensional data. It maps each data point to a location in a low (2 or 3) dimensional space. Such a mapping preserves the local structure of the data in the sense that if the data points form well-separated clusters in the original high-dimensional space, they will too in the dimension reduced space.

Specifically, t-SNE is used to map the vectors $(F_1, \ldots, F_p)$ that contain the log-variance features to a space of two dimensions. Let us see an illustrative example: Figure 10a visualizes the mapping of the 'tongue' and 'left-hand' k-uCSP log-variance features calculated from user 7 during Session 1. Each point represents one of the feature vectors. For comparison, 10b depicts the corresponding CSP-based log variance features. We see that the classes are separated, but not always well separated. Consequently, the unsupervised classification

TABLE II

Mean Classification Accuracy (in Percentage), Averaged for Each Subject, Comparing the Unsupervised (k-uCSP and GMM Classifier) and Supervised (CSP and Fisher LDA) Approaches

Subject	Unsupervised	Supervised
S1	85.07	89.07
S2	62.52	78.33
S3	67.73	90.44
S4	62.67	74.14
S5	62.35	64.39
S6	61.28	67.26
S7	68.21	91.25
S8	76.02	88.83
S9	79.56	85.06

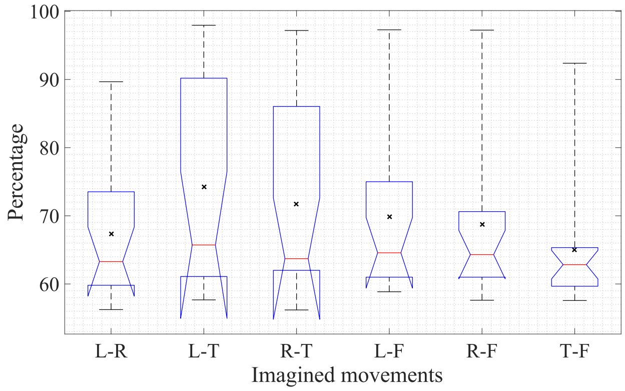
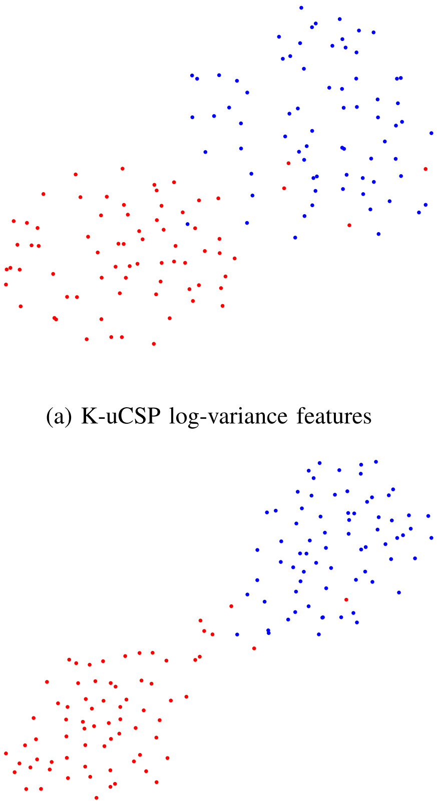


Fig. 9. Box plots of the unsupervised classification accuracy for each imaginary movement ('L-R': left-hand vs right-hand, 'L-T': left-hand vs tongue, 'R-T': right-hand vs tongue, 'L-F': left-hand vs feet, 'R-F': right-hand vs feet, 'T-F': tongue vs feet).

TABLE III

Mean Classification Accuracy (in Percentage) for Imaginary Movements ('L-R': Left-Hand vs Right-Hand, 'L-T': Left-Hand vs Tongue, 'R-T': Right-Hand vs Tongue, 'L-F': Left-Hand vs Feet, 'R-F': Right-Hand vs Feet, 'T-F': Tongue vs Feet), Averaged Across Subjects, Comparing Unsupervised and Supervised Approaches

Imagined Mov.	Unsupervised	Supervised
L-R	67.3677	76.2513
L-T	74.2751	83.6349
R-T	71.7037	82.2540
L-F	69.8704	82.4471
R-F	68.6958	83.1058
T-F	65.0238	78.1587



(a) K-uCSP log-variance features

(b) CSP log-variance features

Fig. 10. Visualisation of the ‘tongue vs left-hand’ log-variance features for user 7. Each class has a different colour set of spots.

algorithm may not be able to differentiate the two groups. In practice, this will lead users to determine on the fly, probably by trial and error, which imagined movements are optimal for themselves (in view of Table III, it may be the combination ‘hand’ vs ‘tongue’ for most people).

Before closing the section, it should be remarked that the choice of the GMM classifier based on the features of eqn. (10) is made for illustration purposes only, but is most probably suboptimal. More optimal choices of features and unsupervised classifiers, which are beyond the scope of the present work, are expected to improve unsupervised BCI results.

VI. Conclusions

Kurtosis maximization performs CSP in an unsupervised fashion, sparing the need for training data and correct labelling. Further work would also aim at more elaborate BCI systems based on the proposed blind CSP approach.

Appendix A
Proof of Theorem 1

Maximizing (5b) is equivalent to maximizing $f(\boldsymbol{a}) := \pi_1\sigma_1^4 + \pi_2\sigma_2^4$ subject to the constraint $h(\boldsymbol{a}) := \pi_1\sigma_1^2 + \pi_2\sigma_2^2 - 1 = 0$. Because of that constraint, the objective function $f(\boldsymbol{a})$ is defined on a closed interval. This, together with the fact that $f(\boldsymbol{a})$ is continuous, ensures that it has both a maximum and a minimum value that it can attain. Let $\mathcal{L}(\boldsymbol{a},\lambda) = f(\boldsymbol{a}) + \lambda\, h(\boldsymbol{a})$ be the Lagrangian function and let $\boldsymbol{L}(\boldsymbol{a},\lambda)$ be the Hessian matrix of $\mathcal{L}(\boldsymbol{a},\lambda)$ with respect to $\boldsymbol{a}$, having as (i,j)-entry $[\boldsymbol{L}]_{ij} = \frac{\partial^2 \mathcal{L}}{\partial a_i \partial a_j}(\boldsymbol{a},\lambda)$. Additionally, the tangent plane of $h(\boldsymbol{a})$ at $\boldsymbol{a}^*$ is defined as the set $T(\boldsymbol{a}^*) = \{\boldsymbol{v} : \boldsymbol{v}^\top \nabla h(\boldsymbol{a}^*) = 0\}$, where ∇ represents the gradient with respect to $\boldsymbol{a}$. We recall the following result [33, Chap. 20]:

Theorem 2: Let $\boldsymbol{a}^*$ be a local maximizer of $f(\boldsymbol{a})$ subject to $h(\boldsymbol{a}) = 0$. Then, there exits $\lambda^* \in \mathbb{R}$ such that

C1) $\nabla f(\boldsymbol{a}^*) + \lambda^* \nabla h(\boldsymbol{a}^*) = 0$, and

C2) for all $\boldsymbol{v} \in T(\boldsymbol{a}^*)$, we have $\boldsymbol{v}^\top \boldsymbol{L}(\boldsymbol{a}^*, \lambda^*)\boldsymbol{v} < 0$.

If $\boldsymbol{a}^*$ is a local minimizer, condition C2 becomes $\boldsymbol{v}^\top \boldsymbol{L}(\boldsymbol{a}^*, \lambda^*)\boldsymbol{v} > 0$.

Next we check conditions C1 and C2 for our particular problem.

Condition C1: We begin by computing the gradients

$$\begin{aligned}\nabla f(\boldsymbol{a}) &= 4\left(\pi_1\sigma_1^2\boldsymbol{\Sigma}_1\boldsymbol{a} + \pi_2\sigma_2^2\boldsymbol{\Sigma}_2\boldsymbol{a}\right)\\ \nabla h(\boldsymbol{a}) &= 2\left(\pi_1\boldsymbol{\Sigma}_1\boldsymbol{a} + \pi_2\boldsymbol{\Sigma}_2\boldsymbol{a}\right).\end{aligned}$$

Setting the gradient of the Lagrange function to zero, we obtain the equation

$$2\left(\pi_1\sigma_1^2\boldsymbol{\Sigma}_1\boldsymbol{a} + \pi_2\sigma_2^2\boldsymbol{\Sigma}_2\boldsymbol{a}\right) + \lambda\left(\pi_1\boldsymbol{\Sigma}_1\boldsymbol{a} + \pi_2\boldsymbol{\Sigma}_2\boldsymbol{a}\right) = 0. \quad (11)$$

Premultiplying by $\boldsymbol{a}^\top$ we easily find that

$$\lambda^* = -2\left(\pi_1\sigma_1^4 + \pi_2\sigma_2^4\right) / \left(\pi_1\sigma_1^2 + \pi_2\sigma_2^2\right).$$

Substituting into (11), we get $\sigma_1^2\left(\sigma_2^2 - \sigma_1^2\right)\boldsymbol{\Sigma}_2\boldsymbol{a} = \sigma_2^2\left(\sigma_2^2 - \sigma_1^2\right)\boldsymbol{\Sigma}_1\boldsymbol{a}$. This equation has two solutions:

$$\text{S1)}\quad \sigma_2^2\boldsymbol{\Sigma}_1\boldsymbol{a}^* = \sigma_1^2\boldsymbol{\Sigma}_2\boldsymbol{a}^* \quad (12)$$

$$\text{S2)}\quad \sigma_1^2 = \sigma_2^2. \quad (13)$$

Observe that S1 corresponds to the CSP solution (3).

Condition C2: To ascertain whether these solutions are maximizers or minimizers of the criterion, we need to consider the second-order condition C2. The Hessian matrix of the Lagrangian can also be decomposed as $\boldsymbol{L}(\boldsymbol{a},\lambda) = \boldsymbol{F}(\boldsymbol{a}) + \lambda\boldsymbol{H}(\boldsymbol{a})$, where $\boldsymbol{F}$ and $\boldsymbol{H}$ are the Hessian matrices of $f(\boldsymbol{a})$ and $h(\boldsymbol{a})$, respectively. We first focus on solution (12), which, after some tedious algebraic manipulations, leads to

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{a}^*) &= 4\sum_{i=1}^{2}\left(2\pi_i\boldsymbol{\Sigma}_i\boldsymbol{a}^*\boldsymbol{a}^{*\top}\boldsymbol{\Sigma}_i + \pi_i\sigma_i^2\boldsymbol{\Sigma}_i\right)\\ \boldsymbol{H}(\boldsymbol{a}^*) &= 2\left(\pi_1\boldsymbol{\Sigma}_1 + \pi_2\boldsymbol{\Sigma}_2\right).\end{aligned}$$

In addition, the tangent plane $T(\boldsymbol{a}^*)$ is the set of vectors $\boldsymbol{v} \in \mathbb{R}^p$ such that $\boldsymbol{v}^\top\nabla h(\boldsymbol{a}^*) = 0$, implying that

$$\boldsymbol{v}^\top(\pi_1\boldsymbol{\Sigma}_1 + \pi_2\boldsymbol{\Sigma}_2)\boldsymbol{a}^* = 0 \Rightarrow \boldsymbol{v}^\top\boldsymbol{\Sigma}_1\boldsymbol{a}^* = \boldsymbol{v}^\top\boldsymbol{\Sigma}_2\boldsymbol{a}^* = 0 \quad (14)$$

where we have exploited in the last implication the fact that $\boldsymbol{\Sigma}_1\boldsymbol{a}^* = \frac{\sigma_1^2}{\sigma_2^2}\boldsymbol{\Sigma}_2\boldsymbol{a}^*$, as follows from (12). Let us denote $s_i^2 := \boldsymbol{v}^\top\boldsymbol{\Sigma}_i\boldsymbol{v}$ the variance of class i after projection onto $\boldsymbol{v}$, $i = 1, 2$. It follows that

$$\boldsymbol{v}^\top\boldsymbol{L}(\boldsymbol{a}^*, \lambda^*)\boldsymbol{v} = \eta\,\sigma_2^2 s_2^2\left(\sigma_1^2 - \sigma_2^2\right)\left(R(\boldsymbol{v}) - R(\boldsymbol{a}^*)\right) \quad (15)$$

where $\eta \overset{\text{def}}{=} 4\pi_1\pi_2/(\pi_1 s_1^2 + \pi_2 s_2^2) > 0$ and we have recognized that s_1^2/s_2^2 is, by definition, equal to $R(\boldsymbol{v})$. To check C2, we distinguish the following cases:

Case 1: $\boldsymbol{a}^* = \boldsymbol{a}_1$, the dominant eigenvector of the GEVD problem defining CSP solution (3). This vector is also the maximizer of the CSP criterion $R(\cdot)$ defined in eqn. (2), as recalled in Sec. II, and therefore $R(\boldsymbol{v}) < R(\boldsymbol{a}_1)$, $\forall \boldsymbol{v} \in T(\boldsymbol{a}_1)$. Now, if $R(\boldsymbol{a}_1) > 1$, then $\sigma_1^2 > \sigma_2^2$, and expression (15) is negative. In other words, $\boldsymbol{a}_1$ is a maximizer of the kurtosis (5b). Otherwise, if $R(\boldsymbol{a}_1) < 1$, then $\sigma_1^2 < \sigma_2^2$ and expression (15) is positive, so that $\boldsymbol{a}_1$ defines a minimum of the kurtosis.

Case 2: $\boldsymbol{a}^* = \boldsymbol{a}_p$, the least significant eigenvector of GEVD problem (3). As seen in Sec. II, this vector is also the minimizer of $R(\cdot)$, so that $R(\boldsymbol{v}) > R(\boldsymbol{a}_p)$, $\forall \boldsymbol{v} \in T(\boldsymbol{a}_p)$. Using the same reasoning as in the above case, $\boldsymbol{a}_p$ is a minimizer of the kurtosis criterion if $R(\boldsymbol{a}_p) > 1$, whereas it defines a maximum if $R(\boldsymbol{a}_p) < 1$.

Case 3: $\boldsymbol{a}^* = \boldsymbol{a}_i$, $1 < i < p$, any of the remaining eigenvectors of GEVD problem (3). Here we exploit the fact that the generalized eigenvectors enjoy an orthogonality property with respect to covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, i.e., $\boldsymbol{a}_i^\mathsf{T}\boldsymbol{\Sigma}_k\boldsymbol{a}_j = 0$, for all $1 \leqslant i \neq j \leqslant p$, $k = 1, 2$. From (14), it follows that both $\boldsymbol{a}_1$ and $\boldsymbol{a}_p$ belong to $T(\boldsymbol{a}_i)$. Hence, the sign of expression (15) when $\boldsymbol{v} = \boldsymbol{a}_1$ is different from that when $\boldsymbol{v} = \boldsymbol{a}_p$, thus defining a saddle point of the kurtosis criterion (5b).

In the light of the above cases, we realize that only the maximizer $\boldsymbol{a}_1$ and/or the minimizer $\boldsymbol{a}_p$ of CSP criterion (2) maximize the kurtosis. More precisely, the maximizer of kurtosis is $\boldsymbol{a}_1$ if $R(\boldsymbol{a}_1) > 1$ and $\boldsymbol{a}_p$ if $R(\boldsymbol{a}_p) < 1$. Because, by definition, $R(\boldsymbol{a}_1) > R(\boldsymbol{a}_p)$ (Sec. II), we can never have $R(\boldsymbol{a}_1) < 1$ and $R(\boldsymbol{a}_p) > 1$ simultaneously, the only possibility for neither point to maximize kurtosis. To conclude the proof of Theorem 1, it remains to study solution S2 given in eqn. (13). This solution corresponds to the case where the mixture of projected classes simplifies into a single Gaussian distribution, with $\kappa_Y = 3$, and turns out to be a global minimizer of kurtosis, as follows from next lemma:

Lemma 1: $\kappa_Y(\boldsymbol{a}) \geqslant 3$, $\forall \boldsymbol{a} \in \mathbb{R}^p$.

Proof: The kurtosis (5b) can be expressed as $\kappa(\boldsymbol{a}) = \frac{3\|\boldsymbol{b}\|^2\|\boldsymbol{c}\|^2}{(\boldsymbol{b}\cdot\boldsymbol{c})^2}$, with $\boldsymbol{b}\cdot\boldsymbol{c} \overset{\text{def}}{=} \boldsymbol{b}^\mathsf{T}\boldsymbol{D}\boldsymbol{c}$, $\boldsymbol{b} = [\sigma_1^2, \sigma_2^2]^\mathsf{T}$, $\boldsymbol{c} = [1, 1]^\mathsf{T}$, $\boldsymbol{D} = \operatorname{diag}(\pi_1, \pi_2)$. By the Cauchy-Schwarz inequality, $|\boldsymbol{b}\cdot\boldsymbol{c}| < \|\boldsymbol{b}\|\|\boldsymbol{c}\|$, and then $\kappa_Y(\boldsymbol{a}) \geqslant 3$, with equality if and only $\boldsymbol{b} = \ell\boldsymbol{c}$, $\ell \in \mathbb{R}\backslash\{0\}$, implying $\sigma_1^2 = \sigma_2^2$.

Appendix B
Computing Several Projection Directions

The whole set of eigenvectors of (12) can be computed as follows. It can be always assumed that $\mathrm{E}\{XX^\mathsf{T}\} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix, under the assumption that the data has been whitened or sphered in a pre-processing step. Since in general the covariance matrix of the data is given by $\mathrm{E}\{XX^\mathsf{T}\} = \pi_1\boldsymbol{\Sigma}_1 + \pi_2\boldsymbol{\Sigma}_2$, whitening imposes that $\pi_2\boldsymbol{\Sigma}_2 = \boldsymbol{I} - \pi_1\boldsymbol{\Sigma}_1$. Substituting in (3), we get a usual eigenvalue problem $\boldsymbol{\Sigma}_1\boldsymbol{a} = \delta\boldsymbol{a}$, where $\delta = R(\boldsymbol{a})/(\pi_1 R(\boldsymbol{a}) + \pi_2)$. Finally, as the eigenvectors of a Hermitian matrix are orthogonal, they can be determined by maximizing the kurtosis under the constraint that the direction obtained in the nth step is orthogonal to the previously computed directions. A possible implementation is presented as pseudo-code in Algorithm 1, where μ is the step size and the gradient $\nabla\kappa_Y(\boldsymbol{a})$ can be easily obtained from eqn. (5a), i.e., $\kappa_Y(\boldsymbol{a}) := \mathrm{E}\{Y^4\}/\mathrm{E}\{Y^2\}^2$, where $Y = \boldsymbol{a}^\mathsf{T}X$, as

$$\nabla\kappa_Y(\boldsymbol{a}) = \frac{\partial\kappa_Y(\boldsymbol{a})}{\partial\boldsymbol{a}} = \frac{4}{\mathrm{E}\{Y^2\}^2}\left(\mathrm{E}\{XY^3\} - \frac{\mathrm{E}\{XY\}\mathrm{E}\{Y^4\}}{\mathrm{E}\{Y^2\}}\right). \tag{16}$$

Observe that this expression can be always evaluated without any knowledge of the data labels by just replacing the expectations with their sample average estimates. For example, note that [24], [25] implement this pseudo-code in Matlab. Furthermore, [24], [25] seek the optimal value of μ using an algebraic line search approach.

Algorithm 1 Compute Orthogonal Directions Maximizing the Kurtosis

```
Center and whiten the data,
    $X \leftarrow \Sigma^{-1/2}\left(X - \bar{X}\right),$
  where $\bar{X}$ is the mean and $\Sigma$ the covariance matrix of $X$
Choose randomly the initial projection vectors $\mathbf{a}_1, \dots, \mathbf{a}_p$.
repeat
  for $k = 1$ to $p$ do
    $\boldsymbol{a}_k \leftarrow \boldsymbol{a}_k + \mu\nabla\kappa_Y(\boldsymbol{a}_k)$ {Gradient-ascend update rule, see eqn. (16)}
    $\boldsymbol{a}_k \leftarrow \boldsymbol{a}_k - \sum_{n=1}^{k-1}\left(\boldsymbol{a}_k^\mathsf{T}\boldsymbol{a}_n\right)\boldsymbol{a}_n$ {Gram-Schmidt orthogonalization}
    $\boldsymbol{a}_k \leftarrow \frac{\boldsymbol{a}_k}{\|\boldsymbol{a}_k\|_2}$ {Normalization}
  end for
until convergence
```

Acknowledgment

The authors would like to thank the anonymous reviewers for their thoughtful comments and efforts towards improving our manuscript.

References

[1] H. Yuan and B. He, "Brain–computer interfaces using sensorimotor rhythms: Current state and future perspectives," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1425–1435, May 2014.

[2] F. Lotte, "A tutorial on EEG signal-processing techniques for mental-state recognition in brain–computer interfaces," in *Guide to Brain-Computer Music Interfacing*, E. R. Miranda, Ed. London, U.K.: Springer, 2014.

[3] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, 2018, Art. no. 031005.

[4] H. Wang, "Harmonic mean of Kullback–Leibler divergences for optimizing multi-class EEG spatio-temporal filters," *Neural Process. Lett.*, vol. 36, no. 2, pp. 161–171, Oct. 2012.

[5] A. S. Aghaei, M. S. Mahanta, and K. N. Plataniotis, "Separable common spatio-spectral patterns for motor imagery BCI systems," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 1, pp. 15–29, Jan. 2016.

[6] T. Jiang *et al.*, "Characterization and decoding the spatial patterns of hand extension/flexion using high-density ECoG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 4, pp. 370–379, Apr. 2017.

[7] R. Martín-Clemente, J. Olias, D. B. Thiyam, A. Cichocki, and S. Cruces, "Information theoretic approaches for motor-imagery BCI systems: Review and experimental comparison," *Entropy*, vol. 20, no. 1, p. 7, 2018. doi: 10.3390/e20010007.

[8] M. Schultze-Kraft, S. Dähne, M. Gugler, G. Curio, and B. Blankertz, "Unsupervised classification of operator workload from brain signals," *J. Neural Eng.*, vol. 13, no. 3, 2016, Art. no. 036008. doi: 10.1088/1741-2560/13/3/036008.

[9] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain–computer interfacing," *J. Neural Eng.*, vol. 9, no. 2, 2012, Art. no. 026013.

[10] G. W. Stewart, *Matrix Algorithms, Eigensystems*, vol. 2. Philadelphia, PA, USA: SIAM, 2001.

[11] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, Dec. 2008.

[12] P. H. Westfall, "Kurtosis as peakedness, 1905–2014. R.I.P," *Amer. Statistician*, vol. 68, no. 3, pp. 191–195, 2014.

[13] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Process.*, vol. 45, no. 1, pp. 59–83, 1995.

[14] R. Martín-Clemente and V. Zarzoso, "On the link between L1-PCA and ICA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 515–528, Mar. 2017. doi: 10.1109/TPAMI.2016.2557797.

[15] D. Peña and F. J. Prieto, "Cluster identification using projections," *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1433–1445, 2001.

[16] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. Hoboken, NJ, USA: Wiley, 2004.

[17] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate Laplace distribution," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 300–303, May 2006.

[18] S.-H. Park, D. Lee, and S.-G. Lee, "Filter bank regularized common spatial pattern ensemble for small sample motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 498–505, Feb. 2018.

[19] H. Higashi and T. Tanaka, "Simultaneous design of FIR filter banks and spatial patterns for EEG signal classification," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 1100–1110, Apr. 2013.

[20] I. Daly, R. Scherer, M. Billinger, and G. Müller-Putz, "FORCe: Fully online and automated artifact removal for brain-computer interfacing," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 5, pp. 725–736, Sep. 2015.

[21] C. Brunner, "SCoT: A Python toolbox for EEG source connectivity," Inst. Knowl. Discovery, Graz Univ. Technol., Graz, Austria, 2014. [Online]. Available: http://www.bbci.de/competition/iv/

[22] C. D. Tesche, M. A. Uusitalo, R. J. Ilmoniemi, M. Huotilainen, M. Kajola, and O. Salonen, "Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources," *Electroencephalogr. Clin. Neurophysiol.*, vol. 95, no. 3, pp. 189–200, Sep. 1995.

[23] M. A. Uusitalo and R. J. Ilmoniemi, "Signal-space projection method for separating MEG or EEG into components," *Med. Biol. Eng. Comput.*, vol. 35, no. 2, pp. 135–140, 1997.

[24] V. Zarzoso and P. Comon, "Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size," *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 248–261, Feb. 2010.

[25] *Robust ICA Matlab Package*. Accessed: Feb. 2, 2018. [Online]. Available: http://www.i3s.unice.fr/~zarzoso/robustica.html

[26] J. Olias, R. Martín-Clemente, M. A. Sarmiento-Vega, and S. Cruces, "EEG signal processing in MI-BCI applications with improved covariance matrix estimators," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 895–904, May 2019. doi: 10.1109/TNSRE.2019.2905894.

[27] H. Ramoser, J. Müler-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[28] J. R. Schott, "Testing for elliptical symmetry in covariance-matrix-based analyses," *Statist. Probab. Lett.*, vol. 60, no. 4, pp. 395–404, Dec. 2002.

[29] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2000.

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

[31] L. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[32] *T-SNE Matlab Package*. Accessed: Mar. 28, 2019. [Online]. Available: https://lvdmaaten.github.io/tsne/code/tSNE_matlab.zip

[33] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, 4th ed. Hoboken, NJ, USA: Wiley, 2013.