Research Article

Single-Trial EEG Classification via Common Spatial Patterns with Mixed Lp- and Lq-Norms

Qian Cai,¹ Weiqiang Gong¹,² Yue Deng,³ and Haixian Wang³

¹School of Statistics and Mathematics, Nanjing Audit University, Nanjing 211815, Jiangsu, China

²Nanjing Les Information System Technology Company, Ltd., Nanjing 210007, Jiangsu, China

³Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science & Medical Engineering, Southeast University, Nanjing 210096, Jiangsu, China

Correspondence should be addressed to Weiqiang Gong; 1036264411@qq.com and Haixian Wang; hxwang@seu.edu.cn

Received 24 November 2020; Revised 27 April 2021; Accepted 15 May 2021; Published 3 June 2021

Academic Editor: Guillermo Cabrera Guerrero

Copyright © 2021 Qian Cai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As a multichannel spatial filtering technique, common spatial patterns (CSP) have been successfully applied in brain-computer interfaces (BCI) community based on electroencephalogram (EEG). However, it is sensitive to outliers because of the employment of the L2-norm in its formulation. It is beneficial to perform robust modelling for CSP. In this paper, we propose a robust framework, called CSP-Lp/q, by formulating the variances of two EEG classes with Lp- and Lq-norms ($0 < p$ and $q < 2$) separately. The method CSP-Lp/q with mixed Lp- and Lq-norms takes the class-wise difference into account in formulating the sample dispersion. We develop an iterative algorithm to optimize the objective function of CSP-Lp/q and show its monotonicity theoretically. The superiority of the proposed CSP-Lp/q technique is experimentally demonstrated on three real EEG datasets of BCI competitions.

1. Introduction

The new communication channel between brain and machine is widely studied in the brain-computer interface (BCI) community based on electroencephalogram (EEG) [1]. It has potentially large application in the domains of biomedical engineering, rehabilitation, education, and entertainment [2]. The central issue in BCI is to accurately decode brain activities. It involves discriminative feature extraction of the brain signals [3].

In fact, the EEG signals are of multichannel time series. They have high temporal resolution. A large number of machine learning algorithms are used to perform the extraction of discriminative EEG features, where the samples are the EEG signals collected at different time points [4, 5]. For the brain activity of motor imagery (MI), there exists a phenomenon called event-related synchronization/desynchronization (ERS/ERD) [6]. That is, the brain energy in the corresponding region of the cerebral cortex increases in the μ -band (8–13 Hz) and the β -band (13–30 Hz), while

the rhythmic energy in the contralateral region significantly decreases [7, 8]. Based on the ERS/ERD principle, the technique of common spatial patterns (CSP) carries out spatial filtering so as to reflect the difference between two EEG classes [9–11]. By maximizing or minimizing the ratio of spatially filtered variance of the two EEG classes, a lot of literature address the application or development of CSP. For example, local temporal CSP (LTCSP) [12], comprehensive CSP (cCSP) [13], and spatial-temporal filter [14]. Some methods have been developed based on sparse learning for EEG analysis. For example, the optimization of spatial patterns using multiview learning within CSP [15], prediction of antidepressant response based on EEG signature [16], sparse Bayesian extreme learning machine for EEG classification [17], and sparse group representation model [18]. Also, some methods address the problem of channel selection, say, based on correction [19] and bispectrum [20]. Recently, a novel method based on CSP (called CSP-LBP), which uses logarithmic band power (LBP) instead of variance to form the objective function, is developed

to extract features [21]. CSP-LBP re-distributes the features so that they can be easily classified. The frequency domain CSP (FDCSP) employs the samples in the frequency domain as input signals of CSP [22]. The transformed data reduce the signal complexity, and some statistical features such as mean, median, maximum, and skewness are extracted. The classification results by the artificial neural network verify the effectiveness of FDCSP.

In the case of clean data, the classical CSP-based methods exhibit satisfactory performance in discriminative feature extraction. However, the EEG data is usually polluted by electromyogram (EMG), electrooculogram (EOG), and spikes. It is, therefore, important to resort to a robust avenue. We have ever developed L1-norm-based CSP (CSP-L1) [23] for the purpose of robust modelling, in which the sample dispersion is formulated with the L1-norm. The L1-norm is then used as internal feature selection [24]. In contrast with the L2-norm, the L1-norm has the robust property from the perspective of statistical theory [25]. The L2-norm involves the square operation, which would enlarge the negative effect of outliers associated with big deviations. In order to balance the big deviations, the resulting projection is likely to be distorted. The L1-norm, however, uses the absolute value operation that could reduce the bad impact of outliers since the dominance of the big deviations is weakened. The L1-norm is then generalized to the Lp-norm scenario for CSP (CSP-Lp) [26] as well as in machine learning [27]. More importantly, the two classes of EEG signals are collected under two different mental conditions, and different mental activities may reflect different intrinsic properties of EEG signals. Specifically, for the motor imageries of left and right hands, the brain demonstrates contralateral effect. The EEG signals recorded under the two mental conditions are related with different cerebral hemispheres and may have different expressions of neurophysiological information. It may be suggested unnecessarily to use the same norm to measure the band powers (i.e., dispersions of samples) of the two classes of EEG signals. It is expected to extract discriminative features by considering individual class of EEG signals specifically. If using two different norms to measure the dispersions of the two classes of samples, there may still be room to promote the capacity of spatial filtering.

In this paper, we consider further generalizing the filtering performance of CSP-L1 by modelling individual class of EEG signals. Specifically, we define the dispersions of the two class of samples in the objective function by using the Lp- and Lq-norms ($0 < p$ and $q < 2$). We term the proposed approach as CSP-Lp/q. It is noted that, in the field of machine learning, the mixed norms have shown great advantage in classification recently [28]. The CSP-Lp/q method aims to find the optimal projection vectors by optimizing the ratio of the Lp-norm dispersion of one class to the Lq-norm dispersion of the other class. The features are then extracted using the mixed Lp/q-norm dispersions. With the features, the classification is implemented via Fisher linear discriminant analysis (LDA) [29]. That is, the feature vectors produced on

training trials are regarded as input samples, by which the objective function of the Fisher criterion is evaluated. According to the principle of maximizing the ratio of between-class to within-class scatter, LDA seeks optimal projection vectors. The testing trials are then discriminated into the class that is nearest with the testing trials in terms of LDA-projected features. We point it that the proposed CSP-Lp/q approach has the following three advantages. (1) CSP-Lp/q incorporates robust modelling by defining the dispersions of the two classes of samples with different robustness-associated norms. (2) CSP-Lp/q is a generalization of CSP, CSP-L1, and CSP-Lp. Mathematically, if taking $p = q = 2$ and $p = q = 1$, then the conventional CSP and CSP-L1 methods occur, respectively. Likewise, CSP-Lp is a special case of CSP-Lp/q by setting $p = q$. (3) An iterative procedure is given to optimize the objective function of CSP-Lp/q.

We organize the rest of this paper as follows. We briefly review the classical CSP and CSP-L1 methods in Section 2. In Section 3, the CSP-Lp/q method, as well as its iterative procedure, is presented. The multiple spatial filters' extension and feature extraction are also outlined in this section. Section 4 illustrates the experimental results on three real EEG datasets. Then, the paper is concluded in Section 5.

2. Conventional CSP and Its L1-Norm-Based Version

The basic idea of the classical CSP is to contrast the difference of two EEG classes by using the spatial filtering technique. Assume that $\mathbf{X}^l \in \mathbb{R}^{s \times n}$ ($l = 1, \dots, T_x$) and $\mathbf{Y}^r \in \mathbb{R}^{s \times n}$ ($r = 1, \dots, T_y$) are the EEG trials of the two classes, where s represents the number of sensors adopted in the experiment, n is the number of samples recorded within each epoch, and T_x and T_y are the numbers of trials of the two classes, respectively. We assume that EEG signals are already band-pass filtered, centered, and scaled.

Denote the whole trials of the two classes by $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^{T_x})$ and $\mathbf{Y} = (\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^{T_y})$, and relabel the columns (i.e., normalized samples) of $\mathbf{X}$ and $\mathbf{Y}$ as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_x}) \in \mathbb{R}^{s \times N_x}$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{N_y}) \in \mathbb{R}^{s \times N_y}$, respectively. Note that $N_x = n \times T_x$ and $N_y = n \times T_y$. Then, the covariance matrices of the two classes are simply formulated as $\mathbf{V}_x = (1/T_x)\mathbf{X}\mathbf{X}^T$ and $\mathbf{V}_y = (1/T_y)\mathbf{Y}\mathbf{Y}^T$, where T is the transpose operator.

Mathematically, the problem of seeking the spatial filter ω with CSP boils down to the optimization of the objective function given by

$$J_{\text{CSP}}(\omega) = \frac{\omega^T \mathbf{V}_x \omega}{\omega^T \mathbf{V}_y \omega} = \frac{\|\omega^T \mathbf{X}\|_2^2 / T_x}{\|\omega^T \mathbf{Y}\|_2^2 / T_y}, \quad (1)$$

where $\|\cdot\|_2$ denotes the L2-norm. It turns out to be an eigenvalue decomposition problem: $\mathbf{V}_x \omega = \lambda \mathbf{V}_y \omega$. The few leading eigenvectors are associated with the largest or smallest eigenvalues, onto which the projected variances could produce discriminative features.

While CSP defines its objective function based on the L2-norm, the CSP-L1 approach, by contrast, utilizes the L1-norm to define the objective function given by

$$J_{\text{CSP-L1}}(\omega) = \frac{\|\omega^T \mathbf{X}\|_1 / T_x}{\|\omega^T \mathbf{Y}\|_1 / T_y} = \frac{\sum_i^{N_x} |\omega^T \mathbf{x}_i| / T_x}{\sum_j^{N_y} |\omega^T \mathbf{y}_j| / T_y}, \quad (2)$$

where $\|\cdot\|_1$ denotes the L1-norm which is the sum of the absolute values of entries. Due to the indifference of the absolute function at the zero point, the optimization of CSP-L1 is usually implemented by the iterative algorithm.

3. CSP with Mixed Lp- and Lq-Norms Optimization

In the presence of outliers, CSP-L1 exhibits advantage due to the robust modelling property of the L1-norm. Further, it is a good practice to employ the Lp-norm for robust modelling since the L1-norm may not be the optimal norm selection for robust modelling and its generalized Lp-norm is more flexible for accommodating varying EEG signals [26]. Due to the fact that EEG signals are collected under different mental conditions and may embody different intrinsic expression of neurophysiological information, it is beneficial to model the individual class of EEG signals specifically. Based on the basic neurophysiological principle, we propose using the Lp-norm to measure one class and the Lq-norm to measure the other class in defining the objective function as follows:

$$J(\omega) = \frac{\|\omega^T \mathbf{X}\|_p^p / T_x}{\|\omega^T \mathbf{Y}\|_q^q / T_y} = \frac{\sum_i^{N_x} |\omega^T \mathbf{x}_i|^p / T_x}{\sum_j^{N_y} |\omega^T \mathbf{y}_j|^q / T_y}, \quad (3)$$

where $\|\cdot\|_p$ and $\|\cdot\|_q$ denote the Lp-norm and the Lq-norm, respectively. The Lp-norm of a vector $\mathbf{u} = (u_1, u_2, \dots, u_d)^T$ is defined as $\|\mathbf{u}\|_p = (\sum_{m=1}^d |u_m|^p)^{1/p}$. We refer to (3) as mixed Lp- and Lq-norms-based CSP and denote it by CSP-Lp/q. Clearly, in the case of $p = q = 1$, CSP-Lp/q turns out to be CSP-L1. So, CSP-L1 is a special case of CSP-Lp/q. Since the optimization (both maximization and minimization) of CSP-Lp/q is not straightforward due to the mixed norms, in the following, we design an iterative algorithm under the framework of bound optimization to calculate the filter bases.

3.1. Algorithmic Implementation. The maximization of the objective function of CSP-Lp/q is presented as follows.

- (1) Filter initialization: let t be the iterative index and set $t = 0$, and take any s -dimensional vector with unit length as the initial spatial filter $\omega(t)$.
- (2) Polarity definition: define two polarity functions $a_i(t)$ and $b_j(t)$ as

$$a_i(t) = \text{sgn}(\omega^T(t) \mathbf{x}_i), \quad (4)$$

$$b_j(t) = \text{sgn}(\omega^T(t) \mathbf{y}_j), \quad (5)$$

where $\text{sgn}(\cdot)$ is the sign function.

- (3) Gradient computation: compute the gradient vector as

$$\mathbf{g}(t) = \frac{p \sum_{i=1}^{N_x} a_i(t) |\omega^T(t) \mathbf{x}_i|^{p-1} \mathbf{x}_i}{\sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^p} - \frac{q \sum_{j=1}^{N_y} b_j(t) |\omega^T(t) \mathbf{y}_j|^{q-1} \mathbf{y}_j}{\sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q}. \quad (6)$$

Here, in case of $\omega^T(t) \mathbf{x}_i = 0$ or $\omega^T(t) \mathbf{y}_j = 0$, we then add a small turbulence to $\omega^T(t)$.

- (4) Filter update: update $\omega(t)$ by

$$\omega(t+1) = \omega(t) + \delta \mathbf{g}(t), \quad (7)$$

where δ is a learning rate parameter which takes a small positive value. Since only the direction of the basis vector is of interest, $\omega(t+1)$ is normalized to unit length.

- (5) Convergence inspection: if $\omega(t+1)$ does not substantially augment the value of the objective function, i.e., $(J(\omega(t+1)) - J(\omega(t))) / J(\omega(t))$ is less than a threshold ε , then break the run and set $\omega^* = \omega(t+1)$. Otherwise, set $t \leftarrow t+1$, and return to Step 2.

- (6) Filter determination: output the spatial filter ω^* .

It is seen that the algorithmic complexity is $O(s(N_x + N_y))$ per iteration. In reality, the total number of iterations and the specific EEG signals would determine the final computational complexity.

3.2. Algorithmic Justification. The algorithmic procedure provided above is justified by the following theorem.

Theorem 1. *With the iterative procedure for CSP-Lp/q, the objective function $J(\omega(t))$ is monotonically increased by each step of iteration t .*

We establish the theorem from the perspective of bound optimization [30, 31]. The bound optimization is a very powerful and general optimization framework. By using the bound optimization, we need to coin a surrogate function $Q(\omega | \omega(t))$, which satisfies the requirement that the value of the function $J(\omega) - Q(\omega | \omega(t))$ is smallest at the point $\omega(t)$. Then, for the surrogate function $Q(\omega | \omega(t))$, the next basis $\omega(t+1)$ is updated such that

$$Q(\omega(t+1) | \omega(t)) \geq Q(\omega(t) | \omega(t)). \quad (8)$$

Therefore, the bound optimization is an iterative procedure. Its convergence could be theoretically confirmed. Specifically, the monotonicity of the objective function $J(\omega)$ is presented below [30]:

$$\begin{aligned}
& J(\omega(t+1)) \\
&= Q(\omega(t+1) | \omega(t)) + (J(\omega(t+1)) - Q(\omega(t+1) | \omega(t))) \\
&\geq Q(\omega(t+1) | \omega(t)) + (J(\omega(t)) - Q(\omega(t) | \omega(t))) \quad (9) \\
&\geq Q(\omega(t) | \omega(t)) + (J(\omega(t)) - Q(\omega(t) | \omega(t))) \\
&= J(\omega(t)).
\end{aligned}$$

To achieve a surrogate function for maximizing $J(\omega)$, the following two lemmas are needed.

Lemma 1 (see [32]). Letting $p \geq 1$, for two vectors $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{v} \in \mathbb{R}^d$, one has the inequality

$$\|\mathbf{u}\|_p^p \geq p(|\mathbf{v}|^{p-1} \circ \text{sgn}(\mathbf{v}))^T \mathbf{u} + (1-p)\|\mathbf{v}\|_p^p, \quad (10)$$

and the equality holds when taking $\mathbf{u} = \mathbf{v}$, where the operations $|\cdot|^{p-1}$ and $\text{sgn}\{\cdot\}$ are applied to the elements of a vector and $\circ$ performs the element-wise product between two vectors.

Lemma 2 (see [33]). Letting $0 < p < 2$, for vector $\mathbf{u} \in \mathbb{R}^d$ and vector $\mathbf{v} \in \mathbb{R}^d$ with nonzero elements, one has the variational inequality

$$\|\mathbf{u}\|_p^p \leq \frac{p}{2} \mathbf{u}^T \text{diag}(|\mathbf{v}|^{p-2}) \mathbf{u} + \left(1 - \frac{p}{2}\right) \|\mathbf{v}\|_p^p. \quad (11)$$

If $\mathbf{u} = \mathbf{v}$, the equality in the above expression is established, where $\text{diag}(|\mathbf{v}|^{p-2})$ is the diagonal matrix with $|\mathbf{v}|^{p-2}$ being its diagonal elements.

We coin a surrogate function as

$$Q(\omega | \omega(t)) = \frac{\omega^T \mathbf{c}(t) + \mathbf{d}(t)}{\omega^T \mathbf{C}(t) \omega + \mathbf{e}(t)}, \quad (12)$$

where $\mathbf{c}(t) = p \sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^{p-1} a_i(t) \mathbf{x}_i$, $\mathbf{d}(t) = (1-p) \sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^p$, $\mathbf{C}(t) = (q/2) \sum_{j=1}^{N_y} \mathbf{y}_j \mathbf{y}_j^T / |\omega^T(t) \mathbf{y}_j|^{2-q}$, and $\mathbf{e}(t) = (1-q/2) \sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q$. The function $Q(\omega | \omega(t))$ can be served as a valid surrogate function due to the fact that $J(\omega) - Q(\omega | \omega(t)) \geq 0$ and the equality holds when taking $\omega = \omega(t)$. The fact is demonstrated as follows.

We compare the numerator and the denominator of $J(\omega)$ with the ones of $Q(\omega | \omega(t))$, respectively. For simplicity, we drop out the two constants T_x and T_y in $J(\omega)$ since they do not affect the computation of the basis vector. Firstly, we calculate the numerator. By letting $\mathbf{u} = \mathbf{X}^T \omega$ and $\mathbf{v} = \mathbf{X}^T \omega(t)$ and substituting them into the inequality of Lemma 1, we have that

$$\begin{aligned}
& \sum_i^{N_x} |\omega^T \mathbf{x}_i|^p = \|\mathbf{X}^T \omega\|_p^p, \\
& \geq p \left(|\mathbf{X}^T \omega(t)|^{p-1} \circ \text{sgn}(\mathbf{X}^T \omega(t)) \right)^T \mathbf{X}^T \omega \\
& \quad + (1-p) \|\mathbf{X}^T \omega(t)\|_p^p \quad (13) \\
& = p \sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^{p-1} a_i(t) \omega^T \mathbf{x}_i + \mathbf{d}(t) \\
& \quad + (1-p) \sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^p = \omega^T \mathbf{c}(t) + \mathbf{d}(t).
\end{aligned}$$

Clearly, when taking $\omega = \omega(t)$, i.e., $\mathbf{u} = \mathbf{v}$, the equality in (13) appears. Secondly, we go on to consider the denominator. By letting $\mathbf{u} = \mathbf{Y}^T \omega$ and $\mathbf{v} = \mathbf{Y}^T \omega(t)$ and substituting them into the inequality of Lemma 2, it follows that

$$\begin{aligned}
0 & \leq \sum_j^{N_y} |\omega^T \mathbf{y}_j|^q = \|\mathbf{Y}^T \omega\|_q^q \\
& \leq \frac{q}{2} \omega^T \mathbf{Y} \text{diag}(|\mathbf{Y}^T \omega(t)|^{q-2}) \mathbf{Y}^T \omega \\
& \quad + \left(1 - \frac{q}{2}\right) \|\mathbf{Y}^T \omega(t)\|_q^q \quad (14) \\
& = \frac{q}{2} \sum_{j=1}^{N_y} \frac{\omega^T \mathbf{y}_j \mathbf{y}_j^T \omega}{|\omega^T(t) \mathbf{y}_j|^{2-q}} + \left(1 - \frac{q}{2}\right) \sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q \\
& = \omega^T \mathbf{C}(t) \omega + \mathbf{e}(t).
\end{aligned}$$

Again, the equality in (14) arises when taking $\omega = \omega(t)$, i.e., $\mathbf{u} = \mathbf{v}$. Combining (13) and (14), we have that

$$J(\omega) = \frac{\sum_i^{N_x} |\omega^T \mathbf{x}_i|^p}{\sum_j^{N_y} |\omega^T \mathbf{y}_j|^q} \geq \frac{\omega^T \mathbf{c}(t) + \mathbf{d}(t)}{\omega^T \mathbf{C}(t) \omega + \mathbf{e}(t)} = Q(\omega | \omega(t)), \quad (15)$$

and $J(\omega) = Q(\omega | \omega(t))$ when taking $\omega = \omega(t)$. This shows that the function $Q(\omega | \omega(t))$ given in (12) is a valid surrogate function.

As guaranteed by the principle of the bound optimization, to maximize $J(\omega)$, it suffices to iteratively augment the surrogate function $Q(\omega | \omega(t))$ at each iteration t . Differentiating $Q(\omega | \omega(t))$ with respect to ω , we obtain the gradient value at the point $\omega(t)$ given by

$$\begin{aligned}
& \frac{\partial Q(\omega | \omega(t))}{\partial \omega} \Big|_{\omega=\omega(t)} \\
&= \frac{\mathbf{c}(t) (\omega^T \mathbf{C}(t) \omega + \mathbf{e}(t)) - 2(\omega^T \mathbf{c}(t) + \mathbf{d}(t)) \mathbf{C}(t) \omega}{(\omega^T \mathbf{C}(t) \omega + \mathbf{e}(t))^2} \Big|_{\omega=\omega(t)}, \\
&= \frac{p \sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^{p-1} a_i(t) \mathbf{x}_i}{\sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q} \\
& \quad - \frac{\left(\sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^p \right) q \sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^{q-1} b_j(t) \mathbf{y}_j}{\left(\sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q \right)^2}, \\
&\propto \frac{p \sum_{i=1}^{N_x} a_i(t) |\omega^T(t) \mathbf{x}_i|^{p-1} \mathbf{x}_i}{\sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^p} \\
& \quad - \frac{q \sum_{j=1}^{N_y} b_j(t) |\omega^T(t) \mathbf{y}_j|^{q-1} \mathbf{y}_j}{\sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q}, \quad (16)
\end{aligned}$$

which exactly is $\mathbf{g}(t)$ given in (6). That is, $\mathbf{d}(t)$ is the vector of the gradient-ascending direction of $Q(\omega | \omega(t))$ at the point $\omega(t)$. Consequently, by calculating the basis $\omega(t+1)$ according to (7), we have $Q(\omega(t+1) | \omega(t)) \geq Q(\omega(t) | \omega(t))$, which implies that update step (7) satisfies condition (8).

Note that, in the above derivation, we in fact require that $p \geq 1$ and $\omega^T(t)\mathbf{y}_j \neq 0$ ($j = 1, 2, \dots, N_y$) (refer to the

applicable conditions of Lemmas 1 and 2). In the case $0 < p < 1$, we would additionally require $\omega^T(t)\mathbf{x}_i \neq 0$ ($i = 1, 2, \dots, N_x$). Then, we directly differentiate the objective function $J(\omega)$ with respect to ω and obtain the gradient value at the point $\omega(t)$ given by

$$\begin{aligned}
 \frac{\partial J(\omega)}{\partial \omega} \Big|_{\omega=\omega(t)} &= \frac{\partial}{\partial \omega} \left(\frac{\sum_{i=1}^{N_x} (\text{sgn}(\omega^T \mathbf{x}_i) \omega^T \mathbf{x}_i)^p}{\sum_{j=1}^{N_y} (\text{sgn}(\omega^T \mathbf{y}_j) \omega^T \mathbf{y}_j)^q} \right) \Big|_{\omega=\omega(t)} \\
 &= \frac{\sum_{i=1}^{N_x} p |\omega^T \mathbf{x}_i|^{p-1} (\text{sgn}'(\omega^T \mathbf{x}_i) \omega^T \mathbf{x}_i + \text{sgn}(\omega^T \mathbf{x}_i)) \mathbf{x}_i}{\sum_{j=1}^{N_y} |\omega^T \mathbf{y}_j|^q} \Big|_{\omega=\omega(t)} \\
 &\quad - \frac{\left(\sum_{i=1}^{N_x} |\omega^T \mathbf{x}_i|^p \right)}{\left(\sum_{j=1}^{N_y} |\omega^T \mathbf{y}_j|^q \right)^2} \\
 &\quad \times \sum_{j=1}^{N_y} q |\omega^T \mathbf{y}_j|^{q-1} (\text{sgn}'(\omega^T \mathbf{y}_j) \omega^T \mathbf{y}_j + \text{sgn}(\omega^T \mathbf{y}_j)) \mathbf{y}_j \Big|_{\omega=\omega(t)} \\
 &= \frac{p \sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^{p-1} a_i(t) \mathbf{x}_i}{\sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q} \\
 &\quad - \frac{\left(\sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^p \right) q \sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^{q-1} b_j(t) \mathbf{y}_j}{\left(\sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q \right)^2} \\
 &\propto \frac{p \sum_{i=1}^{N_x} a_i(t) |\omega^T(t) \mathbf{x}_i|^{p-1} \mathbf{x}_i}{\sum_{i=1}^{N_x} |\omega^T(t) \mathbf{x}_i|^p} - \frac{q \sum_{j=1}^{N_y} b_j(t) |\omega^T(t) \mathbf{y}_j|^{q-1} \mathbf{y}_j}{\sum_{j=1}^{N_y} |\omega^T(t) \mathbf{y}_j|^q},
 \end{aligned} \tag{17}$$

where we use the fact that the derivative of the sign function at the nonzero point is zero. The theorem is thus established.

Therefore, the theorem suggests that the iterative algorithm would converge to a locally maximum point. The globally maximum value of the objective function is possibly reached if different initializations are tried. Particularly, the initialization could be set as the basis vector of the conventional CSP method. The optimal basis vector corresponding to the maximal value of the objective function is thus recorded.

3.3. Multiple Spatial Filters' Extension. Suppose $\omega_1, \omega_2, \dots, \omega_k$ are the first k spatial filters obtained. The $(k+1)$ th spatial filter ω_{k+1} is selected such that

$$\omega_{k+1} = (\mathbf{I}_s - \mathbf{W}_k \mathbf{W}_k^T) \bar{\omega}_k, \tag{18}$$

where $\mathbf{I}_s$ is an s -dimensional identity matrix, $\mathbf{W}_k$ is a matrix formulated as $\mathbf{W}_k = (\omega_1, \omega_2, \dots, \omega_k)$, and $\bar{\omega}_k$ is an s -dimensional coefficient vector. That is, ω_{k+1} lies in the orthogonally complementary space which is spanned by

$\omega_1, \omega_2, \dots, \omega_k$. Substituting (18) into the objective function $J(\omega)$, we have that $\bar{\omega}_k$ can be likely calculated by the bound optimization technique with the deflated samples:

$$\begin{aligned}
 \mathbf{x}_i &\leftarrow (\mathbf{I}_s - \mathbf{W}_k \mathbf{W}_k^T) \mathbf{x}_i, \\
 \mathbf{y}_j &\leftarrow (\mathbf{I}_s - \mathbf{W}_k \mathbf{W}_k^T) \mathbf{y}_j.
 \end{aligned} \tag{19}$$

Then, the basis vector ω_{k+1} is specified according to (18) using $\bar{\omega}_k$ obtained. Consequently, the bases matrix $\mathbf{W}_k$ is renewed by padding $\mathbf{W}_k$ as $\mathbf{W}_{k+1} = (\mathbf{W}_k, \omega_{k+1})$, where ω_{k+1} is normalized such that its length is one. Due to the fact that

$$\mathbf{I}_s - \mathbf{W}_{k+1} \mathbf{W}_{k+1}^T = (\mathbf{I}_s - \omega_{k+1} \omega_{k+1}^T) (\mathbf{I}_s - \mathbf{W}_k \mathbf{W}_k^T), \tag{20}$$

the deflation procedure could be performed based on the previous results.

The iterative algorithm for obtaining K spatial filters maximizing $J(\omega)$ is formally summarized in Algorithm 1. The K counterpart spatial filters $\omega'_1, \omega'_2, \dots, \omega'_K$ maximizing $1/J(\omega)$ are likewise performed.

Input: Training data $\mathbf{x}_i (i = 1, 2, \dots, N_x)$ and $\mathbf{y}_j (j = 1, 2, \dots, N_y)$, norm quantities pp and qq , and number of spatial filters KK .

Output: Spatial filters $\omega_1, \omega_2, \dots, \omega_K$.

- (1) Let $k = 0$, and set $\mathbf{B}$ as the s -dimensional identity matrix.
- (2) Compute the coefficient vector $\bar{\omega}_k$ on the samples $\mathbf{x}_i (i = 1, 2, \dots, N_x)$ and $\mathbf{y}_j (j = 1, 2, \dots, N_y)$ according to the iterative procedure in Section 3.1.
- (3) Compute the spatial filter as $\omega_{k+1} = \mathbf{B}\bar{\omega}_k / \|\mathbf{B}\bar{\omega}_k\|_2$.
- (4) Deflate the samples as $\mathbf{x}_i \leftarrow (\mathbf{I}_s - \omega_{k+1}\omega_{k+1}^T)\mathbf{x}_i$,
 $\mathbf{y}_j \leftarrow (\mathbf{I}_s - \omega_{k+1}\omega_{k+1}^T)\mathbf{y}_j$,
 and update $\mathbf{B}$ as $\mathbf{B} \leftarrow (\mathbf{I}_s - \omega_{k+1}\omega_{k+1}^T)\mathbf{B}$.
- (5) If $k + 1 < K$, then let $k \leftarrow k + 1$ and return to Step 2. Otherwise, terminate the run and output the spatial filters $\omega_1, \omega_2, \dots, \omega_K$.

ALGORITHM 1: Iterative algorithm of CSP-Lp/q for multiple spatial filters.

3.4. Feature Extraction. We apply the multiple spatial filters to the EEG samples so as to extract features. That is, the set of orthonormal spatial filters $\omega_1, \omega_2, \dots, \omega_K$ maximizing the objective function $J(\omega)$ and another set of orthonormal spatial filters $\omega'_1, \omega'_2, \dots, \omega'_K$ maximizing the objective function $1/J(\omega)$ are used as projection bases. The contrast of difference between the two classes is formulated with respect to the mixed Lp- and Lq-norms. Therefore, for one single EEG trial $\mathbf{Z}$, the feature vector is formed as

$$\mathbf{f} = \left(\|\omega_1^T \mathbf{Z}\|_p^p, \dots, \|\omega_K^T \mathbf{Z}\|_p^p, \|(\omega'_1)^T \mathbf{Z}\|_q^q, \dots, \|(\omega'_K)^T \mathbf{Z}\|_q^q \right)^T, \quad (21)$$

where $\mathbf{f}$ is a $2K$ -dimensional feature vector.

4. Experiments

We examine the performance of a family of CSP-based methods, i.e., CSP-Lp/q with varying p and q which include CSP, CSP-L1, CSP-Lp, CSP-LBP, and FDCSP. The experiments are performed on three datasets of BCI competitions: datasets III (a) and IV (a) of BCI competition III and dataset II (a) of BCI competition IV.

4.1. EEG Datasets. Three datasets of BCI competitions, which are publicly available, are employed to evaluate the performance of the proposed CSP-Lp/q approach in EEG classification. In the datasets, 17 subjects were cued to perform motor imageries of hands and tongue activities, and EEG signals were simultaneously recorded. In this paper, we are interested in binary classification (i.e., left versus right hand or right hand versus foot motor imageries).

4.1.1. Dataset III (a) of BCI Competition III. The EEG signals collected from three persons ($s1, s2$, and $s3$) were included in this dataset [34]. There were 60 electrodes employed in total, and the sampled frequency was of 250 Hz. The numbers of total EEG trials for $s1, s2$ and $s3$ were 180, 120, and 120, respectively. The total trials for each subject were equally divided into training and testing trials. That is, the numbers of training (also the testing) samples were 90, 60, and 60 for $s1, s2$, and $s3$, respectively. The purpose was to classify the testing trials into left and right hands' motor imageries.

4.1.2. Dataset IV (a) of BCI Competition III. This dataset [34] was composed of five persons, i.e., aa, al, av, aw, and ay, with 118 electrodes. There were 280 EEG trials for each subject, and the numbers of the training trials for the five subjects were, respectively, 168, 224, 84, 56, and 28. The rest trials were used as the testing trials. We were interested in classifying right hand and foot motor imageries.

4.1.3. Dataset II (a) of BCI Competition IV. The EEG signals in this dataset were generated from nine persons (A01E, ..., A09E), who were cued to perform the activity of four different motor imageries, i.e., left hand, right hand, foot, and tongue [35]. Twenty two electrodes were placed on the scalp to measure EEG signals. For each subject, there were 288 trials in each session with 72 trials per condition. In our experiment, one session of trials was used as training data, while the unlabelled trials in the other session as testing data. Consequently, the number of the training (also the testing) was 144 for each subject since we only perform the classification of the left and the right hand motor imageries.

4.2. Preprocessing and Settings of Experiments. A fifth-order Butterworth filter with the frequency band between 8 Hz and 35 Hz (i.e., cover μ and β bands) was applied to the EEG signals. For the second dataset, EEG time series were windowed in the range of 0.5 s to 3.75 s relative to the visual stimulus, while for the other two datasets, the interval was set as 0.5 s to 2.5 s.

The initial projection bases of CSP-Lp/q were set as the corresponding solutions of the conventional CSP method. In the iteration procedure of CSP-Lp/q, the learning rate parameter δ varied from $1e-6$ to $9e-4$. The convergence threshold ε was set as $1e-5$. As commonly suggested in [9], we adopted three pairs of projection bases (i.e., six components) to filter the EEG samples in carrying out CSP and CSP-Lp/q. As confirmed by the experiments, the classification accuracy would not increase if the number of components became large. By using LDA, the six-dimensional feature vectors (which were produced by CSP-Lp/q on the training trials) were used as samples to train LDA. Consequently, by maximizing the objective function of the Fisher criterion, we obtained one optimal basis vector, onto which the feature vectors were projected. The resulting one-dimensional scalars were then fed into the nearest-neighbor

FIGURE 1: The classification accuracies of CSP-Lp/q with different values of p and q (i.e., the X- and Y-axes, respectively) on the 17 subjects of the three datasets, where the color bars indicate classification accuracies. The 17 subfigures sequentially correspond to subjects $s1, \dots, s3, aa, \dots, ay$, and $A01E, \dots, A09E$.

classifier. That is, the labels of the testing trials were endowed as the labels of the nearest training trials.

The parameters p and q were determined by using the strategy of two-fold cross-validation. That was, for each combination of p and q , the filters were obtained based on the training set (subset of the total training set), and then, the classification accuracy was evaluated on the validation set (the rest of the total training set). This procedure was repeated two times, and in each time, the validation set was different. We used the average of the classification rates across the two folds as classification accuracy. We selected the combination of p and q that resulted in the maximum average classification accuracy. Then, we re-computed the filters using all the two folds with the optimal combination

determined. With the filters obtained, the generalization capacity was estimated on the testing trials, in which the testing process was not concerned with the training of the filters and the parameters.

4.3. Outliers Simulation. To highlight the classification performance of the proposed technique in the situation of outliers with varying p and q , artificial outliers were introduced into the training samples. Specifically, we design the randomly generated outlier from an s -dimensional Gaussian distribution $N(\mu + 3\sigma, \Sigma)$, where μ and σ were the quantities of channel-wise mean and standard deviation of all the training samples, respectively, and Σ was the

TABLE 1: The classification accuracies (%) of CSP-Lp/q with different values of p and q (i.e., the column and row, respectively) on subject s1.

	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
0.2	97.78	98.89	97.78	97.78	98.89	97.78	95.56	95.56	95.56	94.44
0.4	97.78	97.78	98.89	97.78	98.89	98.89	95.56	95.56	95.56	94.44
0.6	97.78	97.78	96.67	97.78	98.89	98.89	97.78	97.78	95.56	94.44
0.8	97.78	96.67	95.56	96.67	97.78	97.78	98.89	98.89	98.89	97.78
1.0	96.67	96.67	96.67	96.67	96.67	96.67	96.67	98.89	97.78	98.89
1.2	96.67	95.56	95.56	96.67	96.67	96.67	96.67	97.78	96.67	97.78
1.4	96.67	95.56	95.56	96.67	96.67	96.67	96.67	95.56	96.67	96.67
1.6	96.67	96.67	94.4	96.67	96.67	96.67	96.67	95.56	96.67	96.67
1.8	95.56	95.56	96.67	96.67	96.67	96.67	96.67	95.56	96.67	96.67
2.0	96.67	96.67	95.56	96.67	96.67	96.67	96.67	96.67	95.56	95.56

FIGURE 2: Examples of spatial filters obtained by different CSP methods on subject aw, where the two filters corresponding to the largest (upper row) and the smallest eigenvalues (lower row) are shown. (a) CSP; (b) CSP-Lp/q ($p = 0.4$ and $q = 0.8$); (c) CSP-L1; (d) CSP-LBP.

covariance matrix of the training samples. These artificial outlier points were imposed to the original EEG training samples at randomly selected positions. For each incorporation, the experiments were repeated ten times. The mean classification accuracy was figured.

4.4. Experimental Results. We investigated the classification accuracy of the proposed CSP-Lp/q approach with varying p and q in the range $0 < p$ and $q \leq 2$ on the three datasets. Particularly, CSP-Lp/q becomes the conventional CSP method when taking $p = q = 2$ and turns out to be CSP-L1 in the case of $p = q = 1$. For visual perception, the classification accuracies of CSP-Lp/q with different values of p and q on all the 17 subjects of the three datasets are shown in Figure 1, where the step length of p and q was set as 0.1. To clearly see the difference between classification accuracies with different p and q values, the quantities on subject s1 are shown in Table 1 as an example. It was seen that most of the classification accuracies exceeded 70% on the three datasets,

which demonstrated the effectiveness of the CSP-based spatial filtering techniques. From Figure 1, it was observed that, in most cases, the highest classification accuracy of CSP-Lp/q was obtained when p and q took different values. It thus demonstrated the necessity of the utilization of different norms in modelling different classes of EEG samples. It was also noted that, in most cases, the optimal values of p and q were less than 1. It may imply that, for EEG signals, small norms were preferred so as to accommodate outliers accompanied. For visual perception, examples of spatial filters obtained by different CSP methods are delineated in Figure 2, where FDCSP was not shown since the frequency quantities were used as input data.

The classification accuracies of CSP-Lp/q, as well as CSP, CSP-LBP, FDCSP, and CSP-L1, on the 17 subjects are shown in Table 2, where CSP-Lp/q was compared with the benchmark method CSP and with three state-of-the-art methods: CSP-LBP [21], FDCSP [22], and CSP-L1 [23]. For different subjects, the classification results obtained were quite different, which further illustrated individual

TABLE 2: The classification accuracies (%) of CSP-Lp/q with the corresponding optimal combinations of (p, q) , as well as CSP, CSP-LBP, FDCSP, and CSP-LI, on the 17 subjects of the three datasets.

Dataset	III (a) of BCI competition III				IV (a) of BCI competition III				II (a) of BCI competition IV			
Subject	s1	s2	s3	Average	aa	al	av	aw	ay	Average	A04E	A05E
CSP	97.78	70.00	96.67	88.15	60.71	100.00	66.33	91.07	70.63	77.75		
CSP-LBP	97.78	70.00	96.67	88.15	60.71	100.00	66.33	94.64	75.40	79.42		
FDCSP	100.00	70.00	100.00	90.00	62.50	100.00	68.37	92.86	78.57	80.46		
CSP-LI	96.67	63.33	100.00	86.67	69.64	100.00	67.35	91.96	79.37	81.66		
CSP-Lp/q	97.78(0.4, 0.1)	80.00(0.4, 0.2)	100.00(0.1, 0.1)	92.59	71.79(0.9, 0.3)	100.00(0.1, 0.1)	73.47(0.5, 0.1)	94.64(0.4, 0.8)	81.75(1.2, 0.9)	84.33		
Dataset	III (a) of BCI competition III				IV (a) of BCI competition III				II (a) of BCI competition IV			
Subject	A01E	A02E	A03E	A04E	A05E	A06E	A07E	A08E	A09E	Average		
CSP	84.72	51.39	94.44	77.78	68.06	61.11	75.00	94.44	87.50	77.16		
CSP-LBP	84.72	51.39	94.44	77.78	69.44	62.52	81.94	94.44	87.50	78.24		
FDCSP	86.11	52.78	94.44	77.78	70.83	65.28	79.17	98.61	88.89	79.32		
CSP-LI	84.72	56.94	94.44	79.17	76.39	65.28	80.56	95.83	88.89	80.25		
CSP-Lp/q	86.11(0.6, 0.8)	65.00(0.8, 0.8)	94.44(1.2, 0.7)	77.78(0.1, 0.2)	79.17(0.1, 0.2)	66.67(0.1, 0.1)	80.56(0.8, 0.8)	100.00(0.1, 0.1)	90.28(0.4, 0.2)	82.22

FIGURE 3: The classification accuracies of the five methods (i.e., CSP, CSP-LBP, FDCSP, CSP-L1, and CSP-Lp/q) with increasing occurrence frequencies of outliers on the first dataset. The first three subfigures sequentially correspond to subjects s_1 , s_2 , and s_3 , and the last subfigure is the average of the classification accuracies across the three subjects.

FIGURE 4: The classification accuracies of the five methods (i.e., CSP, CSP-LBP, FDCSP, CSP-L1, and CSP-Lp/q) with increasing occurrence frequencies of outliers on the second dataset. The first five subfigures sequentially correspond to subjects aa , al , av , aw , and ay , and the last subfigure is the average of the classification accuracies across the five subjects.

difference in the BCI domain. It was straightforward that the average classification accuracy of CSP-Lp/q was higher than that of CSP (i.e., $p = q = 2$) and CSP-L1 (i.e., $p = q = 1$) since CSP and CSP-L1 were the two special cases of CSP-Lp/q. Also, CSP-Lp/q was superior than CSP-LBP and FDCSP. Moreover, the classification accuracies of 7 out of 17 subjects have reached more than 90% when using CSP-Lp/q.

The experimental comparison of training times on subject s_1 costed by the five methods (i.e., CSP-Lp/q, CSP, CSP-LBP, FDCSP, and CSP-L1) were 91s, 13s, 42s, 78s, and

43s, respectively. Here, the experiments were run on the platform of Lenovo ThinkPad with 2.90 GHz CPU and 8.00 GB RAM. It was not surprising that the iterative algorithms consumed more training times. Especially, the CSP-Lp/q method needed additional time to train the parameters p and q .

Statistical analysis was implemented to investigate the significance of performance differences between different methods. We adopted the nonparametric Friedman test [36] to meet the purpose since the classification accuracies were

FIGURE 5: The classification accuracies of the five methods (i.e., CSP, CSP-LBP, FDCSP, CSP-L1, and CSP-Lp/q) with increasing occurrence frequencies of outliers on the third dataset. The first nine subfigures sequentially correspond to subjects A01E through A09E, and the last subfigure is the average of the classification accuracies across the nine subjects.

reported on the same subjects and then were correlated. During the test, the classification accuracies produced by each method were regarded as a group, and blocks were formed with different subjects. With the rank statistic of the test, it was indicated that the performance differences between different methods were significant ($p < 0.05$). Further, the Friedman rank sum test [37] was conducted to perform the post hoc multiple comparison. It was found that the proposed CSP-Lp/q method exceeded the other approaches ($p < 0.05$).

The classification accuracies of the five methods (i.e., CSP, CSP-LBP, FDCSP, CSP-L1, and CSP-Lp/q) in the case of increasing occurrence frequencies of outliers on the three datasets are presented in Figures 3–5. Due to individual difference, the classification accuracies varied on different

subjects when outliers were introduced. From these figures, it was seen that the classification accuracies of most methods declined with the increased frequency of outliers. Clearly, on almost all subjects, the conventional CSP method and CSP-LBP had the sharpest decrease. That is, the two methods were influenced by the outliers badly. Due to the introduction of frequency, FDCSP was better than CSP and CSP-LBP. However, CSP, CSP-LBP, and FDCSP are essentially based on the L2-norm, which is sensitive to outliers. By contrast, the classification accuracy of CSP-L1 had relatively mild decrease, which demonstrated the effect of the L1-norm in suppressing the outliers. CSP-L1 is based on the L1-norm, which, however, is a special case of CSP-Lp/q (i.e., $p = q = 1$). Generally, among the five methods, CSP-Lp/q had relative superior classification accuracies on the three

datasets. In these situations, the decreasing range was small, and the classification accuracies were still satisfactory on most subjects. Note that the optimal combination of p and q may vary in different occurrence frequencies of outliers. Again, the advantage of selecting appropriate norms for robust modelling was illustrated.

5. Conclusion

In this paper, we proposed a new approach, called CSP-Lp/q, by using the mixed Lp- and Lq-norms for modelling two classes of EEG signals individually. The objective function of the CSP-Lp/q approach was defined by reformulating the sample dispersion of one class with the Lp-norm, while the other class with the Lq-norm. It took the class difference into account fully when defining sample dispersion. Thanks to the mix-norms adopted, the CSP-Lp/q approach was a robust generalization of CSP, CSP-L1, and CSP-Lp. The classification performance of the proposed CSP-Lp/q technique was demonstrated by the experiments on three datasets of BCI competitions.

One limitation of the proposed method was that the optimal combination of p and q in CSP-Lp/q may depend on the data on hand. This was the issue of model selection. A practical avenue was to use the strategy of cross validation to determine them. It was an interesting but challenging issue to set it analytically. Another limitation was that, compared with the traditional CSP method, the implementation of the proposed method was in an iterative way. The determination of the step length δ in the gradient-ascent procedure, however, was not a trivial issue since inappropriate setting would make the iteration trapped into local optimization. We are currently studying the element-wise update with the close-form solution. Further, the proposed method could be extended into the online (incremental) version, which is useful for the online BCI system. Finally, we pointed out that the proposed technique of adopting Lp- and Lq-norms was a general framework for neuroscience decoding. It could be applied to other relative spatial patterns, say, discriminative spatial patterns [38].

Data Availability

All data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61773114 and University Synergy Innovation Program of Anhui Province under Grant GXXT-2020-015.

References

- [1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.
- [2] B. He, B. Baxter, B. J. Edelman, C. C. Cline, and W. W. Ye, "Noninvasive brain-computer interfaces based on sensorimotor rhythms," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 907–925, 2015.
- [3] J. D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, no. 7, pp. 523–534, 2006.
- [4] H.-J. Hwang, S. Kim, S. Choi, and C.-H. Im, "EEG-based brain-computer interfaces: a thorough literature survey," *International Journal of Human-Computer Interaction*, vol. 29, no. 12, pp. 814–826, 2013.
- [5] H. Yuan and B. He, "Brain-computer interfaces using sensorimotor rhythms: current state and future perspectives," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1425–1435, 2014.
- [6] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.
- [7] G. Pfurtscheller and A. Aranibar, "Evaluation of event-related desynchronization preceding and following voluntary self-paced movement," *Electroencephalography and Clinical Neurophysiology*, vol. 46, no. 2, pp. 138–146, 1979.
- [8] A. Schnitzler, S. Salenius, R. Salmelin, V. Jousmäki, and R. Hari, "Involvement of primary motor cortex in motor imagery: a neuromagnetic study," *NeuroImage*, vol. 6, no. 3, pp. 201–208, 1997.
- [9] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008.
- [10] G. Dornhege, M. Krauledat, K.-R. Müller, and B. Blankertz, "General signal processing and machine learning tools for BCI," in *Toward Brain-Computer Interfacing*, pp. 207–233, MIT Press, Cambridge, MA, USA, 2007.
- [11] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda, "Recipes for linear analysis of EEG," *Neuroimage*, vol. 28, no. 2, pp. 326–341, 2005.
- [12] H. Wang and W. Zheng, "Local temporal common spatial patterns for robust single-trial EEG classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 2, pp. 131–139, 2008.
- [13] H. Wang and D. Xu, "Comprehensive common spatial patterns with temporal structure information of EEG data: minimizing nontask related EEG component," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2496–2505, 2012.
- [14] A. Jiang, J. Shang, X. Liu, Y. Tang, H. K. Kwan, and Y. Zhu, "Efficient CSP algorithm with spatio-temporal filtering for motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 4, pp. 1006–1016, 2020.
- [15] Y. Jiao, T. Zhou, L. Yao, G. Zhou, X. Wang, and Y. Zhang, "Multi-view multi-scale optimization of feature representation for EEG classification improvement," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 14, pp. 2589–2597, 2020.
- [16] W. Wu, Y. Zhang, J. Jiang et al., "An electroencephalographic signature predicts antidepressant response in major depression," *Nature Biotechnology*, vol. 38, pp. 439–447, 2020.
- [17] Z. Jin, G. Zhou, D. Gao, and Y. Zhang, "EEG classification using sparse Bayesian extreme learning machine for brain-

- computer interface,” *Neural Computing and Applications*, vol. 32, pp. 6601–6609, 2020.
- [18] Y. Jiao, Y. Zhang, X. Chen et al., “Sparse group representation model for motor imagery EEG classification,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 631–641, 2019.
 - [19] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki, “Correlation-based channel selection and regularized feature optimization for MI-based BCI,” *Neural Networks*, vol. 118, pp. 262–270, 2019.
 - [20] J. Jin, C. Liu, I. Daly et al., “Bispectrum-based channel selection for motor imagery based brain-computer interfacing,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2153–2163, 2020.
 - [21] M. Aljalal, R. Djemal, and S. Ibrahim, “Robot navigation using a brain computer interface based on motor imagery,” *Journal of Medical and Biological Engineering*, vol. 39, no. 4, pp. 508–522, 2019.
 - [22] P. K. Saha, M. A. Rahman, and M. N. Mollah, “Frequency domain approach in CSP based feature extraction for EEG signal classification,” in *Proceedings of the 2nd International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox’s Bazar, Bangladesh, February 2019.
 - [23] H. Wang, Q. Tang, and W. Zheng, “L1-norm-based common spatial patterns,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 3, pp. 653–662, 2012.
 - [24] J. Jin, R. Xiao, I. Daly, Y. Miao, X. Wang, and A. Cichocki, “Internal feature selection method of CSP based on L1-norm and Dempster-Shafer theory,” *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
 - [25] Q. Ke and T. Kanade, “Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 739–746, San Diego, CA, USA, June 2005.
 - [26] N. Fang and H. Wang, “Generalization of local temporal correlation common spatial patterns using Lp-norm ($0 < p < 2$),” in *Proceedings of the International Conference on Neural Information Processing*, pp. 769–777, Guangzhou, China, November 2017.
 - [27] N. Kwak, “Principal component analysis by Lp-norm maximization,” *IEEE Transactions on Cybernetics*, vol. 44, no. 5, pp. 594–609, 2014.
 - [28] Q. Ye, L. Fu, Z. Zhang, H. Zhao, and M. Naiem, “Lp- and Ls-norm distance based robust linear discriminant analysis,” *Neural Networks*, vol. 105, pp. 393–404, 2018.
 - [29] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
 - [30] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, “Sparse multinomial logistic regression: fast algorithms and generalization bounds,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, 2005.
 - [31] K. Lange, D. Hunter, and I. Yang, “Optimization transfer using surrogate objective functions,” *Journal of Computational and Graphical Statistics*, vol. 9, pp. 1–59, 2000.
 - [32] J. Wang, “Generalized 2-D principal component analysis by Lp-norm for image analysis,” *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 792–803, 2016.
 - [33] R. Jenatton, G. Obozinski, and F. Bach, “Structured sparse principal component analysis,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Zakopane, Poland, June 2010.
 - [34] B. Blankertz, K.-R. Müller, D. J. Krusienski et al., “The BCI competition III: validating alternative approaches to actual BCI problems,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 153–159, 2006.
 - [35] M. Naeem, C. Brunner, R. Leeb, B. Graimann, and G. Pfurtscheller, “Seperability of four-class motor imagery data using independent components analysis,” *Journal of Neural Engineering*, vol. 3, pp. 208–216, 2006.
 - [36] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
 - [37] R. Eisinga, T. Heskes, B. Pelzer, and M. Te Grotenhuis, “Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers,” *BMC Bioinformatics*, vol. 18, no. 1, p. 68, 2017.
 - [38] X. Liao, D. Yao, D. Wu, and C. Li, “Combining spatial filters for the classification of single-trial EEG in a finger movement task,” *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 821–831, 2007.