Biomedical Signal Processing and Control 10 (2014) 313–321

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Short communication

L1-norm based discriminative spatial pattern for single-trial EEG classification

Qin Tang, Jing Wang, Haixian Wang*

Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast University, 2 Sipailou Road, Nanjing 210096, Jiangsu, PR China

ARTICLE INFO

Article history:
Received 7 May 2012
Received in revised form 4 December 2012
Accepted 21 December 2012
Available online 14 January 2013

Keywords:
Spatial filtering
EEG
L1-norm
Discriminative spatial pattern (DSP)

ABSTRACT

Spatial filtering provides an efficient method for single-trial EEG classification and has been widely used in EEG-based brain computer interfaces. However, scalp-recorded EEG signals are usually very noisy since they could be contaminated by various outliers, such as EOG or EMG artifacts. The outliers may seriously distort the performance of spatial filters. To solve this problem, we propose a new robust spatial filtering algorithm, namely DSP-L1, which is L1-norm based discriminative spatial pattern (DSP). Compared with the conventional DSP, DSP-L1 takes advantage of the robust L1-norm modeling that expects to perform better in suppressing the effect of outliers. Computationally, an iterative approach is introduced to find the spatial filters of DSP-L1. Experimental results on two EEG data sets of motor movements demonstrate the efficiency of the proposed method.

1. Introduction

EEG-based brain computer interfaces (BCIs) provide a communication channel for patients with severe neuromuscular disabilities by directly translating human intentions into control signals for outside services [1–3]. One crucial part of a BCI system is the accurate and efficient pattern recognition for various mental tasks. However, the scalp-recorded EEG samples generally suffer a low signal-to-noise-ratio (SNR) due to some accompanying physiological activities, such as unconscious jaw clenching and swallowing, electrooculogram (EOG) and electromyogram (EMG) [4–7], as well as non-physiological sources, such as 50 Hz power-line noise [7]. Particularly in a clinical setting, it seems more difficult to obtain outlier-free data sets [7,8]. In some occasions, the outlier-contaminated data can be manually discarded during an offline analysis by visual inspection, while they are unavoidable in online applications. If the contaminated signals are rejected during specific time periods, the online system may not be applicable [7]. Therefore, developing robust machine learning algorithms to handle outliers for EEG signal pattern recognition is necessary in the implementation of a stable BCI system.

In recent years, spatial filtering algorithms are widely used in EEG signal processing and proven to be extremely efficient in single-trial analysis [2,9,10]. Discriminative spatial pattern (DSP) is one of such methods developed for the classification of EEG signals based on movement related potentials (MRPs) components, which are typical brain activities during motor movement task. Specifically, MRPs are non-oscillatory potential shifts (0–7 Hz) starting with a steep negative slope prior to finger movement onset and reaching a negative peak approximately 100 ms after movement onset [9,11–13]. The basic idea of DSP comes from two-dimensional Fisher linear discriminant analysis (2DLDA) [14], which separates two classes by maximizing the between-class separation and minimizing the within-class separation. Particularly, to focus on MRPs-based EEG classification, subject-specific optimization of time and frequency are required before the spatial filtering.

However, DSP method focuses on low-frequency signals which happen to be the frequency range of artifacts such as EOG activity [5,7]. The negative effect of noise is likely to be highlighted by the L2-norm employed in the formulation of DSP, giving a rise to the possibility of distorting the spatial filters of DSP. Given the fact that the L1-norm is more robust with respect to noise than the L2-norm [15–18], we propose an L1-norm based DSP, named DSP-L1, to suppress the potential influence of noise. Due to the introduction of the L1-norm, the computation of DSP-L1 is far from trivial. To address this computational issue, we use a simple and effective iterative algorithm to obtain the spatial directions of DSP-L1. With the L1-norm, it is expected that DSP-L1 has better performance than DSP. We demonstrate the advantage of the proposed DSP-L1 over DSP on both EEG data artificially added with multivariate outliers and real EEG data that is badly contaminated by EOG or EMG.

* Corresponding author.
E-mail addresses: hxwang@seu.edu.cn, haixian_wang@hotmail.com (H. Wang).

1746-8094/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bspc.2012.12.004

The remainder of the paper is organized as follows: In Section 2, we first review the conventional DSP formulation and then propose the DSP-L1 method. The experimental results of EEG single-trial classification are presented in Section 3. Discussion is given in Section 4. Finally, Section 5 concludes the paper.

2. Methods

2.1. Discriminative spatial patterns (DSP)

In this study, we consider the situation with only two classes. For each of the subject, an aggregation of multi-trials in both classes is used to compute optimal spatial filters. Assume that $X_j(i) \in R^{C\times K}$ is the EEG data of the ith trial from class $j(j \in \{+, -\})$, where C is the number of electrodes and K is the number of samples (i.e., recording time points). The within-class scatter matrix and the between-class scatter matrix are respectively given by

$$S_W = \sum_j \sum_{i=1}^{n_j} (X_j(i) - M_j)(X_j(i) - M_j)^T, \tag{1}$$

$$S_B = \sum_j n_j (M - M_j)(M - M_j)^T, \tag{2}$$

where T denotes the transpose operator, n_j is the number of trials in class j, M_j is the mean of the trials from class j defined as $M_j = 1/n_j \sum_{i=1}^{n_j} X_j(i)$, and M is the mean of all the trials.

The objective function of DSP is given by

$$J_{\mathrm{DSP}}(\Omega) = \frac{\det(\Omega^T S_B \Omega)}{\det(\Omega^T S_W \Omega)}, \tag{3}$$

where $\det(\cdot)$ denotes matrix determinant. DSP aims to find the optimal discriminative transformation matrix $\Omega_{\mathrm{DSP}}(\Omega_{\mathrm{DSP}} \in C \times D, D \le C)$ maximizing class means and minimizing within class scatters, where D is the number of selected spatial filters [9,19]. Solution to Ω_{DSP} can be achieved by solving the generalized eigenvalue problem $S_W\sigma_l = \lambda_l S_B \sigma_l$, where λ_l denotes eigenvalue and $\sigma_l \in R^C$ is the corresponding eigenvector. Ω_{DSP} is given by the eigenvectors of the first D largest eigenvalues, i.e., $\Omega_{\mathrm{DSP}} = \{\sigma_1, \sigma_2, \ldots, \sigma_D\}$.

It should be pointed out that, in the case that S_W is singular, a regularization parameter α (say $\alpha = 0.1$) is introduced to modify S_W by $S_W \leftarrow (1-\alpha)S_W + \alpha I_C$, where I_C is the C-dimensional identity matrix.

2.2. L1-norm based discriminative spatial patterns (DSP-L1)

The formulation of DSP is based on L2-norm. To see this point, we substitute (1) and (2) into (3), resulting in

$$J_{\mathrm{DSP}}(\sigma_l) = \frac{\sum_j n_j ||\sigma_l^T (M - M_j)||_2^2}{\sum_j \sum_{i=1}^{n_j} ||\sigma_l^T (X_j(i) - M_j)||_2^2}, \tag{4}$$

where $||\cdot||_2$ denotes L2-norm. Since $M - M_j$ and $X_j(i) - M_j$ are fixed given training data, we use $Y_j = M - M_j$ and $Z_{ij} = X_j(i) - M_j$ to simplify the expression. For a robust modeling [15–18], we replace the L2-norm in DSP with the L1-norm and propose the objective function of DSP-L1 as

$$J_{\text{DSP-L1}}(\omega_d) = \frac{\sum_j n_j ||\omega_d^T Y_j||_1}{\sum_j \sum_{i=1}^{n_j} ||\omega_d^T Z_{ij}||_1}, \tag{5}$$

Table 1
Algorithmic procedure of DSP-L1.

Input: $Y_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jm}, \ldots, Y_{jK}), Z_{ij} = (Z_{ij1}, Z_{ij2}, \ldots, Z_{ijm}, \ldots, Z_{ijK}), Y_{jm} \in R^C, Z_{ijm} \in R^C$
Initial vector $\omega_1(0) \in R^C$ with unit length, step size η chosen from $\{\eta_1, \eta_2, \ldots, \eta_r\}$, iterative number t and convergence threshold ε.
Output: Optimal spatial filter ω_1^* maximizing (5).

DSP-L1 Algorithm. At iteration t, the objective function (5) is rewritten as

$$J_{\text{DSP-L1}}(\omega_1(t)) = \frac{\sum_j n_j ||\omega_1(t)^T Y_j||_1}{\sum_j \sum_{i=1}^{n_j} ||\omega_1(t)^T Z_{ij}||_1}, \tag{6}$$

While

(a) Define two polarity functions to remove the absolute value operation in (6)

$$p_{jm}(t) = \begin{cases} 1 & \omega_1(t)^T Y_{jm} \ge 0 \\ -1 & \omega_1(t)^T Y_{jm} < 0 \end{cases} \tag{7}$$

$$q_{ijm}(t) = \begin{cases} 1 & \omega_1(t)^T Z_{ijm} \ge 0 \\ -1 & \omega_1(t)^T Z_{ijm} < 0 \end{cases} \tag{8}$$

(b) Let

$$d(\omega_1(t)) = \frac{\sum_j n_j \sum_{m=1}^{K} p_{jm}(t) Y_{jm}}{\sum_j n_j \sum_{m=1}^{K} \left|\omega_1(t)^T Y_{jm}\right|} - \frac{\sum_j \sum_{i=1}^{n_j} \sum_{m=1}^{K} q_{ijm}(t) Z_{ijm}}{\sum_j \sum_{i=1}^{n_j} \sum_{m=1}^{K} \left|\omega_1(t)^T Z_{ijm}\right|}$$

(c) For $h = 1 : r$

$$\omega_1^{(h)}(t+1) = \omega_1(t) + \eta_h d(\omega_1(t))$$

End
(d) $\omega_1^*(t+1) := arg\max J_{\text{DSP-L1}}(\omega_1^{(h)}(t+1)), \quad h \in \{1, 2, \ldots, r\}$
(e) If $J_{\text{DSP-L1}}(\omega_1^*(t+1)) - J_{\text{DSP-L1}}(\omega_1(t)) < \varepsilon$

$\omega_1^* = \omega_1(t)$, break
Else

$$\omega_1^*(t+1) \leftarrow \frac{\omega_1^*(t+1)}{||\omega_1^*(t+1)||_2}, \quad \omega_1(t) \leftarrow \omega_1^*(t+1), \quad \text{and} \quad t \leftarrow t+1$$

End while

where $||\cdot||_1$ denotes L1-norm, ω_d is a discriminative spatial filter in the L1-norm modeling, i.e., the dth spatial filter of $\Omega_{\text{DSP-L1}}$ (the transformation matrix for DSP-L1, i.e., $\Omega_{\text{DSP-L1}} = (\omega_1, \omega_2, \ldots, \omega_D)$). Due to the introduction of the L1-norm, the optimization problem of (5) is far from trivial. Motivated by the recent research in solving L1-norm-based optimization problem [15–18], we present an iterative algorithm to find the first optimal spatial projection ω_1, and then extend to multiple ones. The iterative steps are listed in Table 1.

2.2.1. Justification

With the iterative procedure, we guarantee that the objective function (6) is increased monotonously, i.e., in each iterative step $J_{\text{DSP-L1}}(\omega_1(t+1)) > J_{\text{DSP-L1}}(\omega_1(t))$ always holds. Following the basic idea of [15–18], we present the justification as follows.

We rewrite (6) as

$$J_{\text{DSP-L1}}(\omega_1(t)) = \frac{\sum_j n_j \omega_1(t)^T \varphi_j(t)}{\sum_j \sum_{i=1}^{n_j} \left[(1/2)(\omega_1(t)^T \mu_{ij}(t) \omega_1(t)) + (1/2)\gamma_{ij}(t)\right]}, \tag{9}$$

where $\varphi_j(t) = \sum_{m=1}^{K} p_{jm}(t) Y_{jm}$, $\mu_{ij}(t) = \sum_{m=1}^{K} (Z_{ijm} Z_{ijm}^T)/(|\omega_1(t)^T Z_{ijm}|)$, $\gamma_{ij}(t) = \sum_{m=1}^{K} |\omega_1(t)^T Z_{ijm}|$.

Since (9) is nondifferentiable, we introduce a vicarious function $L(\xi)$ as

$$L(\xi) = \ln\left(\frac{\sum_j n_j \xi^T \varphi_j(t)}{\sum_j \sum_{i=1}^{n_j}\left[(\xi^T \mu_{ij}(t)\xi) + \gamma_{ij}(t)\right]}\right). \tag{10}$$

Note that in (10), $\varphi_j(t)$, $\mu_{ij}(t)$, $\gamma_{ij}(t)$ are fixed at iteration t while only ξ is variable. The gradient of $L(\xi)$ with respect to ξ is calculated as

$$G(\xi) = \frac{\sum_j n_j \varphi_j(t)}{\sum_j n_j \xi^T \varphi_j(t)} - \frac{\sum_j \sum_{i=1}^{n_j} \mu_{ij}(t)\xi}{\sum_j \sum_{i=1}^{n_j}(1/2)\left[(\xi^T \mu_{ij}(t)\xi) + \gamma_{ij}(t)\right]}. \tag{11}$$

Replacing ξ with $\omega_1(t)$, we have that

$$G(\omega_1(t)) = \frac{\sum_j n_j \sum_{m=1}^{K} p_{jm}(t) Y_{jm}}{\sum_j n_j \sum_{m=1}^{K} |\omega_1(t)^T Y_{jm}|} - \frac{\sum_j \sum_{i=1}^{n_j} \sum_{m=1}^{K} q_{ijm}(t) Z_{ijm}}{\sum_j \sum_{i=1}^{n_j} \sum_{m=1}^{K} |\omega_1(t)^T Z_{ijm}|}. \tag{12}$$

Here, (12) is the exact quantity $d(\omega_1(t))$ in Table 1, which points to the increasing direction of (10) at the point $\omega_1(t)$. By (10), we have that $L(\omega_1(t+1)) > L(\omega_1(t))$, i.e.

$$\frac{\sum_j n_j \omega_1(t+1)^T \varphi_j(t)}{(1/2)\sum_j \sum_{i=1}^{n_j}\left[(\omega_1(t+1)^T \mu_{ij}(t)\omega_1(t+1)) + \gamma_{ij}(t)\right]} > \frac{\sum_j n_j \omega_1(t)^T \varphi_j(t)}{(1/2)\sum_j \sum_{i=1}^{n_j}\left[(\omega_1(t)^T \mu_{ij}(t)\omega_1(t)) + \gamma_{ij}(t)\right]}. \tag{13}$$

Clearly, the right part of (13) is identical with (9), i.e., $J_{\text{DSP-L1}}(\omega_1(t))$. On the other hand, the numerator in the left part of (13) can be rewritten as

$$\begin{aligned}\sum_j n_j \omega_1(t+1)^T \varphi_j(t) &= \sum_j n_j \omega_1(t+1)^T \sum_{m=1}^{K} p_{jm}(t) Y_{jm} \\ &\le \sum_j n_j \omega_1(t+1)^T \sum_{m=1}^{K} p_{jm}(t+1) Y_{jm} \\ &= \sum_j n_j \omega_1(t+1)^T \varphi_j(t+1).\end{aligned} \tag{14}$$

The inequality in (14) holds due to the fact that $p_{jm}(t+1)$ is designed for $\omega_1(t+1)$. In other words, $\omega_1(t+1)^T \varphi_j(t)$ may be negative value while $\omega_1(t+1)^T \varphi_j(t+1)$ is always nonnegative. The denominator in the left part of (13) can be rewritten as

$$\begin{aligned}&\frac{1}{2}\sum_j \sum_{i=1}^{n_j}[(\omega_1(t+1)^T \mu_{ij}(t)\omega_1(t+1)) + \gamma_{ij}(t)] \\ &= \sum_j \sum_{i=1}^{n_j}\left[\frac{1}{2}\sum_{m=1}^{K}\frac{(\omega_1(t+1)^T Z_{ijm})^2}{|\omega_1(t)^T Z_{ijm}|} + \frac{1}{2}\sum_{m=1}^{K}|\omega_1(t)^T Z_{ijm}|\right] \\ &\ge \sum_j \sum_{i=1}^{n_j}\min_{\zeta \in R_+^K}\left[\frac{1}{2}\sum_{m=1}^{K}\frac{(\omega_1(t+1)^T Z_{ijm})^2}{|\zeta_m|} + \frac{1}{2}||\zeta||_1\right] \\ &= \sum_j \sum_{i=1}^{n_j}||\omega_1(t+1)^T Z_{ij}||_1,\end{aligned} \tag{15}$$

where $\zeta = (\zeta_1, \zeta_2, ..., \zeta_K)^T$. The last inequality in (15) is presented according to the fact [20] that, for any vector $\phi = (\phi_1, \phi_2, ..., \phi_K)^T$, there is the equality $||\phi||_1 = \min_{\zeta \in R_+^K}(1/2)\sum_{m=1}^{K}(\phi_m^2/|\zeta_m|) + (1/2)||\zeta||_1$ and the minimum is uniquely reached when $\zeta_m = |\phi_m|$, $\forall m \in \{1, 2, ..., K\}$.

Combining (13)–(15), we have that

$$\begin{aligned}J_{\text{DSP-L1}}(\omega_1(t)) &= \frac{\sum_j n_j \omega_1(t+1)^T \varphi_j(t)}{(1/2)\sum_j \sum_{i=1}^{n_j}\left[(\omega_1(t+1)^T \mu_{ij}(t)\omega_1(t+1)) + \gamma_{ij}(t)\right]} \\ &\le \frac{\sum_j n_j \omega_1(t+1)^T \varphi_j(t+1)}{\sum_j \sum_{i=1}^{n_j}||\omega_1(t+1)^T Z_{ij}||_1} = J_{\text{DSP-L1}}(\omega_1(t+1)).\end{aligned} \tag{16}$$

Thus, the justification is done.

It should be pointed out that the step size η should be carefully selected. In this study, we select η from 1e−5 to 1e−2. Since the initial vector $\omega_1(0)$ can be set arbitrarily, we employ the solution of DSP as the initial vector, which seems to be more likely to produce a global maximum value.

2.2.2. Extension to multiple spatial filters

After the computation of ω_1, we proceed to find other directions ω_d ($d \in Z_+$, $2 \le d \le D$) as follows.

First we will seek ω_2 which is in the orthogonally complementary direction of ω_1, i.e., the constraint $\omega_1^T \omega_2 = 0$ holds. Therefore, we define an equation

$$\omega_2 = (I_c - \omega_1 \omega_1^T)\beta_1, \tag{17}$$

where $\beta_1 \in R^C$. Then, substituting (17) into (5) and setting Y_j as $Y_j^{(1)}$, Z_{ij} as $Z_{ij}^{(1)}$, we find the optimal β_1 to maximize

$$J_{\text{DSP-L1}}(\beta_1) = \frac{\sum_j n_j ||\beta_1^T Y_j^{(2)}||_1}{\sum_j \sum_{i=1}^{n_j}||\beta_1^T Z_{ij}^{(2)}||_1}, \tag{18}$$

where $Y_j^{(2)} = (I_c - \omega_1 \omega_1^T) Y_j^{(1)}$, $Z_{ij}^{(2)} = (I_c - \omega_1 \omega_1^T) Z_{ij}^{(1)}$. The optimal β_1 can be obtained by following the steps in Table 1. Specifically, Y_j and Z_{ij} are replaced by $Y_j^{(2)}$ and $Z_{ij}^{(2)}$ respectively, and initial vector $\beta_1(0)$ is set as the secondary spatial direction of DSP. Thus ω_2 are obtained.

In the process of finding ω_k, we require that $\omega_1^T \omega_{k+1} = 0$, $\omega_2^T \omega_{k+1} = 0, \ldots, \omega_k^T \omega_{k+1} = 0$. Let $\Omega_k = (\omega_1, \omega_2, ..., \omega_k)$ and we define $\omega_{k+1} = (I_c - \Omega_k \Omega_k^T)\beta_k$. Then β_k can be obtained by maximizing

$$J_{\text{DSP-L1}}(\beta_k) = \frac{\sum_j n_j ||\beta_k^T Y_j^{(k+1)}||_1}{\sum_j \sum_{i=1}^{n_j}||\beta_k^T Z_{ij}^{(k+1)}||_1}, \tag{19}$$

where $Y_j^{(k+1)} = (I_c - \Omega_k \Omega_k^T) Y_j^{(k)}$ and $Z_{ij}^{(k+1)} = (I_c - \Omega_1 \Omega_k^T) Z_{ij}^{(k)}$.

2.3. Classification

For each trial $X_j(i)$, the mean values (in terms of the samples in time) of each projection after the transformation matrices of DSP or DSP-L1 are calculated as features for classification [9,11], i.e.,

$$f_{\text{DSP}} = \text{mean}(\Omega_{\text{DSP}}^T X_j(i)), \quad f_{\text{DSP}} \in R^D. \tag{20}$$

Likewise,

$$f_{\text{DSP-L1}} = \text{mean}(\Omega_{\text{DSP-L1}}^T X_j(i)), \quad f_{\text{DSP-L1}} \in R^D. \tag{21}$$

The selection of the number of spatial filters for both DSP and DSP-L1 will be discussed in Section 3.

3. Experiments

In this section, we evaluate the performance of DSP-L1 in the EEG classification task of two motor movements. Two EEG data

sets are used in our experiments. Data set 1 is from BCI Competition 2003 [21]. We do the classification task as the competitors did, i.e., spatial filters and classifier learned from the training data are used to predict the labels of the testing data. Besides, to investigate the robustness of DSP and DSP-L1, multivariate outliers are introduced into the training data. Data set 2 is provided by [22–24], consisting of EEG data from 30 subjects (the data of 109 volunteers are available online and we use the first 30 ones, i.e., S001-S030, in this study). Since the data of some subjects are badly contaminated by EOG or EMG artifacts (as illustrated in Section 3.4), we investigate the performance of DSP-L1 in suppressing the influence of those real outliers. Finally for comparison purpose, the classification accuracies obtained from DSP-L1 algorithm are compared with that of DSP algorithm, some LDA variants and other relevant methods under the same experimental setting.

3.1. EEG data sets for evaluation

Data set 1 is the data set IV of the BCI Competition 2003, which was recorded from one normal subject pressing corresponding keys with left or right fingers in a self-chosen order and timing (self-paced 1 s). For classification task of left vs. right finger movement, 316 trials with labels and 100 unlabeled trials were provided for training and testing respectively. Each trial contained time segment of 500 ms length ending at 130 ms before the key press onset. The data were recorded using 28 EEG channels mainly covering sensorimotor cortices at positions of the international 10/20-system. In our experiment, we use the data that were down-sampled at 100 Hz and band-pass filtered between 0. 5 Hz and 200 Hz.

Data set 2 was recorded using BCI2000 instrumentation system [23] with 64-channel. Each subject was required to perform 14 experimental runs including 2 one-minute baseline runs and 12 runs of two-minute motor or imagery movement. In this study, we focus on the third, seventh and eleventh runs, where subjects performed left or right fist open and close movement task. Specifically, a target appeared on either the left or the right side of the screen for 4.1 s while the subjects opened and closed corresponding fist until the target disappeared, and then the subjects relaxed for about 4.1 s. Left or right targets were presented in a randomized order and an annotation file with the target presenting time and type was available for each run. The data was sampled at 160 HZ and the scalp locations for the 64 electrodes were also downloadable online.

3.2. Preprocessing

For both data sets, a band-pass filter with cutoff frequencies 0.5 Hz and 7 Hz is applied to capture MRPs shift as recommended in [9,12]. For data set 1, the time segment located from −330 ms to −130 ms, which is expected to contain the most discriminative MRPs, are selected for feature extraction as in [9,13]. As for each trial in data set 2, we compute common average reference and correct baseline according to the average EEG data of the two baseline runs. Considering the specific experimental paradigm of data set 2, i.e., only the time when the visual instructions (targets) presented is given, while the time of movement onset is unknown, we find it difficult to define the starting time of MRPs accurately. Previous studies [25–27] suggested that the starting time for MRPs varied greatly across subjects. We follow the criterion in [26] by inspecting the significant current sink to select the optimal time-window for each subject. Specifically, we calculate the mean current source density (CSD) values every 100 ms from 500 ms to 800 ms after the target occurs on three electrodes, i.e., C3, CP3, CP5, which cover the major sensorimotor area. The time segment resulting in the largest absolute mean CSD values are selected for the following classification process. That is to say, the length of the time window is fixed at 100 ms for all the subjects but located at different time positions within 500 ms to 800 ms after the targets appear. It should be noted that results may be further improved using some time window selection strategies as mentioned in [2].

3.3. Introducing outliers

For data set 1, an outlier stimulation method [28] derived from multivariate normal distribution model

$$c = N_P(\mu + 3\tau, \Sigma) \tag{22}$$

is introduced to investigate the robustness of DSP-L1, where μ and τ are the mean and standard deviation vector of all the channels of the entire training sample respectively, and Σ is the covariance matrix of all the training trials. Outliers are added to the training data at random selected time position. Parameter θ is used to control the occurrence probability of the introduced outliers, i.e., the proportion of samples added with outliers, and it varies among $\{0, 5\%, 10\%, 15\%, 20\%, 25\%, 30\%, 35\%, 40\%, 45\%, 50\%\}$. At each occurrence probability, experiments are repeated 20 times for average classification accuracy.

3.4. Observing real outliers

In this subsection, S001 and S004 from data set 2 are illustrated to show the effect of real EOG or EMG artifacts.

Fig. 1(a) shows the average time courses of S001 under two conditions (i.e., left/right motor movement) over trials, where one can observe the large absolute values centering at anterior head regions around 700 ms after target occurs. It indicates possible presence of strong EOG activities, which generally have high-amplitude than normal EEG signals and most stress over the anterior head regions [7]. As shown in Fig. 1(b), EEG signals of S004 are contaminated by EMG artifacts, which have a stronger amplitude at the temporal area, and can be distinguished from EOG artifacts by localization [4,7]. Due to the fact that EMG artifacts have a broad frequency distribution and mainly focus on 20-30 Hz and 40-80 Hz, the amplitude of EMG artifacts can be reduced by the band-pass filtering process in preprocess to some degree. While EOG artifacts are well known to be low-frequency and they remain high amplitude after preprocessing. In addition, both Fig. 1(a) and (b) imply that MRPs present a maximum amplitude at the vertex in the steep negative slopes and a greater effect in pre-frontal region just as the finding in [25].

3.5. Evaluation and results

All the algorithms involved in our study are used as feature extraction methods and the performance of those methods are evaluated in terms of classification accuracy on the testing data. Besides, the linear support vector machine (SVM) embedded in the MatlabR2010b toolbox is used as classifier.

To well illustrate the iterative process of the proposed method, we give an example on the training data of data set 1 as shown in Fig. 2. Initial vectors of DSP-L1 are set as the corresponding spatial filters of DSP after scaled to unit length. The convergence threshold ε is set as 0.001. It shows that the value of DSP-L1 objective function ($J_{\text{DSP-L1}}$) enhances with each step of iteration when $d = 1, 4, 8, 10$. The number of iterations varies for different spatial filters. We also find that with introduced outliers, the number of iterations increases for all the spatial filters, especially suboptimal ones. It makes sense as larger deviation may require more iteration steps for a robust spatial filter when outliers occur.

Fig. 3 shows classification results of DSP and DSP-L1 with increasing occurrence frequency of outliers for data set 1. Brain mappings of the first spatial filters with varied θ for both methods are illustrated in Fig. 4. On the other hand, Fig. 5 shows the

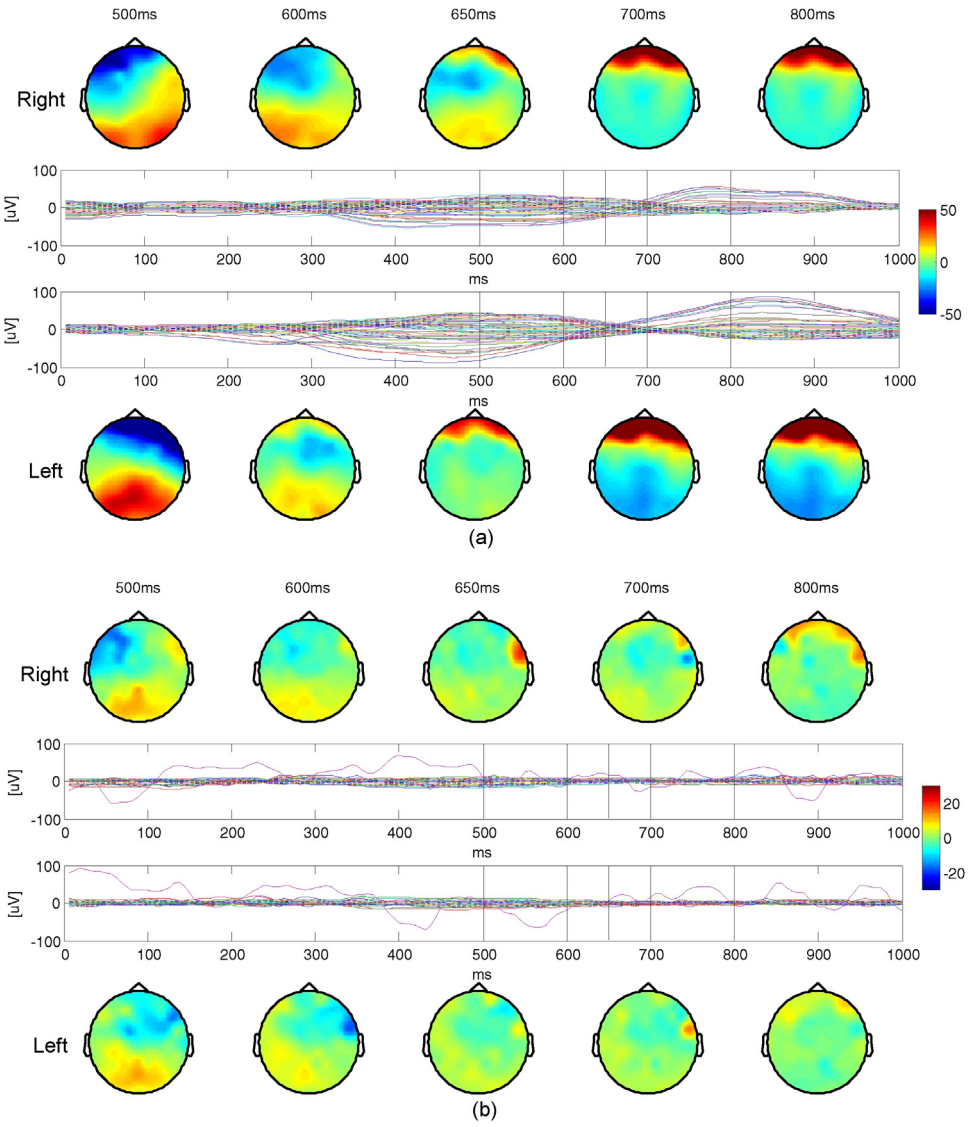

Fig. 1. Averaged time courses under two conditions (i.e., left/right motor movement) over trials. (a) S001. (b) S004.

performance of DSP and DSP-L1 with different values of D, i.e., the number of spatial filters.

For data set 2, 20×5-fold cross-validations are employed on the EEG data of each subject. Specifically, in each procedure, the training and testing data would be reorganized 5 times, i.e., all the trials (45 trials from 3 runs) for one subject are divided into 5 parts, where each part is used once as testing data while the rest parts are used as training data. This procedure would be repeated 20 times. The average classification accuracy of the 20×5-fold is computed as the final result. In this study, we compare the robustness of proposed DSP-L1 with DSP, regularized LDA (RLDA) [29,30], uncorrelated LDA (ULDA) [31,32], orthogonal LDA (OLDA) [33], locality preserving projections (LPP) [34], L1-norm based common spatial pattern (CSP-L1) [16], local discriminative spatial patterns (LDSP) [11], two-dimension uncorrelated LDA (2DULDA) and two-dimension orthogonal LDA (2DOLDA). Default values are adopted for the regularization type and parameter in RLDA and a leave-one-out cross-validation strategy is employed to find appropriate parameter κ (κ-nearest neighbors) for LPP and LDSP as in [11]. Table 2 reports the average classification accuracies and the corresponding standard deviations over 30 subjects for all the methods when the number of spatial filters is set as 1, 4, 8 or 10. Particularly,

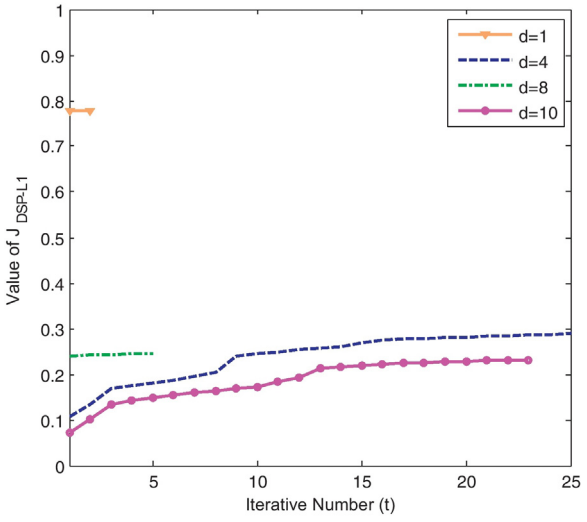

Fig. 2. A plot of the convergence of $J_{\text{DSP-L1}}$ versus iterative number t for the first, 4th, 8th and 10th spatial filters of $\Omega_{\text{DSP-L1}}$.

Table 2
Average classification accuracies (%) and corresponding standard deviations over 30 subjects produced by RLDA, OLDA, ULDA, LPP, 2DOLDA, 2DULDA, CSP-L1, DSP and DSP-L1.

Methods	Number of spatial filters			
	1	4	8	10
RLDA	70.87 ± 8.2	–	–	–
OLDA	70.87 ± 8.2	–	–	–
ULDA	70.87 ± 8.2	–	–	–
LPP	54.92 ± 10.3	64.84 ± 11.8	67.76 ± 8.2	67.97 ± 8.1
2DOLDA	71.48 ± 6.9	67.58 ± 6.6	67.25 ± 7.0	67.52 ± 6.3
2DULDA	71.48 ± 6.9	68.90 ± 7.4	67.34 ± 7.2	66.36 ± 6.7
CSP-L1	57.71 ± 9.6	65.12 ± 10.7	67.01 ± 10.1	67.77 ± 10.2
DSP	72.34 ± 6.9	69.35 ± 7.4	67.74 ± 7.3	66.85 ± 6.7
DSP-L1	**72.38** ± 6.9	**73.03** ± 7.8	**74.47** ± 8.4	**75.50** ± 9.4

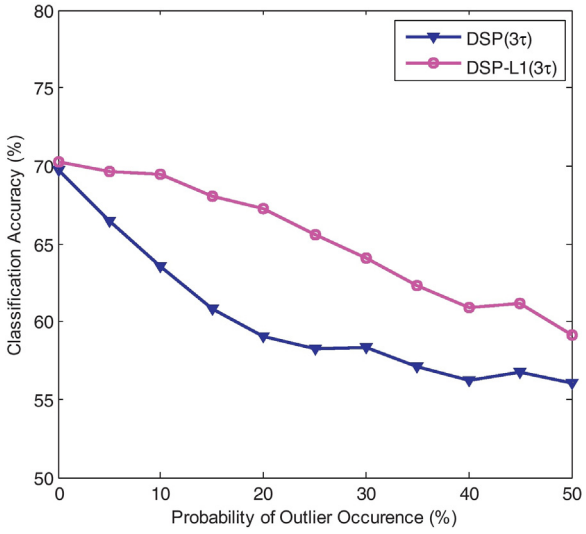

Fig. 3. Comparing performance of DSP and DSP-L1 for data set 1 with increasing occurrence frequency of outliers. Average classification accuracies over the range of $D \in \{1, 4, 8, 10\}$ are used as the final results.

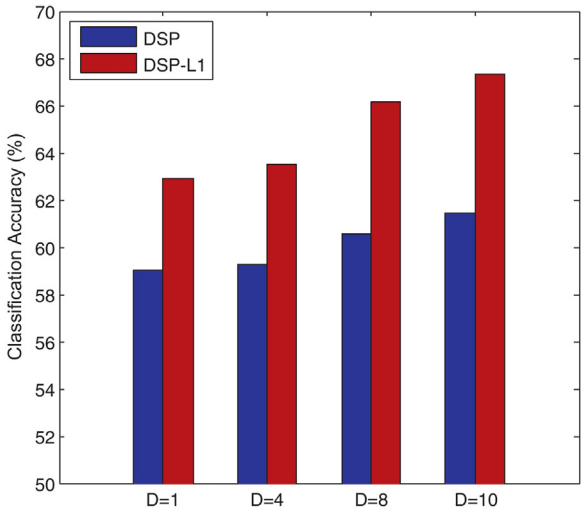

Fig. 5. Average classification accuracies over all values of θ using DSP and DSP-L1 with different value of D, i.e., the number of spatial filters, for data set 1.

the results produced by DSP and DSP-L1 methods are compared by scatter plots in Fig. 6(a)–(d) for different numbers of spatial filters.

4. Discussion

Data set 1 is used to investigate how DSP-L1 and DSP perform in resisting impact of introducing outliers. From Fig. 3, we can see that DSP-L1 outperforms DSP in general, and outstanding superiority is shown at $\theta = 15\%$ by 8%. Less difference is found in classification accuracies with increasing occurrence frequencies of outliers. It may due to the reason that when large part of the signals (>50%) are contaminated by outliers, both methods learn spatial filters from outliers instead of useful information. We also find the smaller deviation of introduced outliers has no big influence on classification results. Specifically, we control the deviation of outliers by altering 3τ in (22) to τ. The classification results with τ are nearly identical with that of 3τ at certain probability of outliers occurrence. Fig. 4 illustrates some brain mappings of the first spatial filters when $\theta \in \{0, 15\%, 25\%, 50\%\}$. Though they are messy at first sight, the spatial filters of DSP-L1 seem to be more helpful to distinguish the

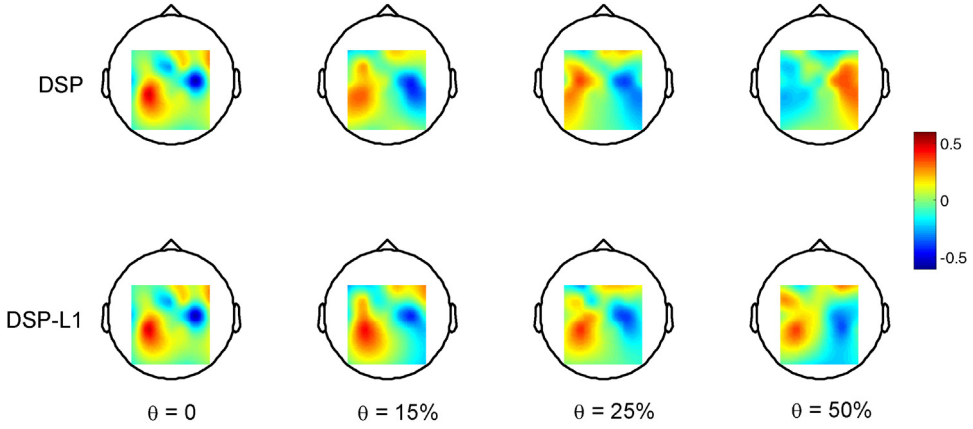

Fig. 4. Comparing brain mappings of the first spatial filters learnt by DSP and DSP-L1 on data set 1 when $\theta = \{0, 15\%, 25\%, 50\%\}$.

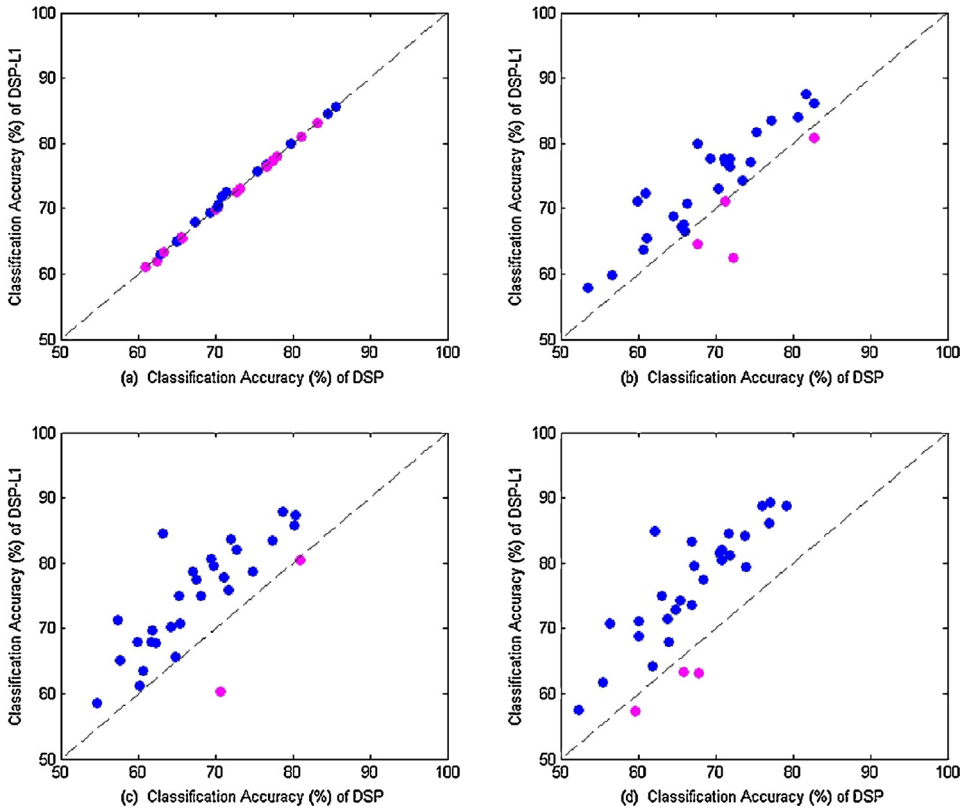

Fig. 6. Scatter plots comparing the classification accuracies of DSP and DSP-L1 with different number of spatial filters for data set 2. For the subjects where DSP-L1 outperforms DSP are represented by blue dots, and depicted by magenta dots otherwise. (a) $D = 1$, (b) $D = 4$, (c) $D = 8$, (d) $D = 10$. The dash line indicates $y = x$.

differences in two motor movements in the case with introduced outliers comparing to those of DSP.

On the other hand, from Fig. 5, we find that DSP-L1 has more improvement over DSP with large number of spatial filters (i.e., $D = 8$ and $D = 10$ in this experiment). It seems that DSP-L1 produces better minor spatial filters instead of principle ones comparing with DSP. This can be explained by the fact that outliers are similar to noise in the sense that they both have high-frequency components, which can only be captured by minor basis vectors instead of principal ones [17]. Since the initial vectors of DSP-L1 are set as the basis vectors of DSP, the iterative process, i.e., the process to find spatial filters of DSP-L1, can be seen as the process to modify the basis vectors of DSP toward the optimal L1-norm dispersion, which is more robust to outliers.

For data set 2, we compare the proposed DSP-L1 method with DSP, RLDA, ULDA, OLDA, LPP, LDSP and CSP-L1 in terms of feature extraction method for EEG classification.

Table 2 shows that DSP-L1 produces better average classification accuracies than DSP over 30 subjects with different number of spatial filters. Scatter plots in Fig. 6(a)–(d) further demonstrate the superiority of DSP-L1 with large number of spatial filters. Two-sample t-test indicates statistical significance improvement of DSP-L1 over DSP when $D = 8$ and $D = 10$ with $p = 0.02$ and $p = 0.00$ at the 5% level, respectively. It proves efficiency of proposed DSP-L1 in withstanding the impact of natural occurring outliers.

RLDA, ULDA and OLDA are proposed to deal with high-dimensional, under sampled problem of classical LDA [31,33]. Specifically, RLDA uses a scaled multiple of the identity matrix to obtain nonsingular scatter matrix. ULDA computes the optimal transformation matrix by firstly removing the null space of the total scatter matrix and the singularity problem is avoided implicitly. Discriminative spatial filters of OLDA are obtained by a further orthogonalization step on those of ULDA. They have been applied in various applications successfully [31–33]. We point out that the LDA variants (i.e., RLDA, ULDA and OLDA in this study) are one-dimensional methods, i.e., the spatio-temporal matrix for one trial is required to be scanned as a long vector before performing feature extraction. They are not appropriate in MRPs-based single-trial EEG classification as they fail to capture the differences in the amplitude of MRPs. To make the comparison fair and conduct MRPs-based EEG classification, we make a modification to form 2DULDA and 2DOLDA.

By the results in Table 2, we can see the superiority of DSP-L1 over other methods with increasing number of spatial filters. It is known that LDA and its variants are potentially constrained to $c - 1$ (c is the number of classes, that is, 2 in this study) features as limited by the rank of scatter matrices. Besides, more features may not provide extra information upon the first spatial filter. However, DSP-L1 is free from this restriction and discriminates two classes by learning more spatial filters. It is worth to point out that DSP-L1 is a robust modeling of DSP with regards to MRPs-based EEG classification, while RLDA, ULDA and OLDA aim to improve classification capability of the classical LDA from some aspects like regularization or orthogonalization of filters. When applying in MRPs-based EEG classification, 2DULDA and 2DOLDA result in less improvement comparing with the conventional DSP. This demonstrates the advantage of DSP-L1 over LDA variants as a robust modeling in single-trial EEG classification. In addition, Table 3 gives the classification accuracies of S001 and S004 (both are illustrated to be badly contaminated by outliers in Section 3.4). It shows that DSP-L1 results in relatively better classification accuracies over other methods in specific cases.

LDSP and LPP are manifold modeling developed in the field of machine learning and proven to be helpful in classification via the using of the intrinsic local property of data [11,34]. LDSP is not tested due to the aborted use of a leave-one-out strategy in finding

Table 3
Classification accuracies (%) of S001 and S004 using RLDA, ULDA, OLDA, LPP, 2DOLDA and 2DULDA, CSP, DSP and DSP-L1 methods with increasing number of spatial filters.

Subjects	Methods	Number of spatial filters				
		1	4	8	10	Mean
S001	RLDA	73.15	–	–	–	73.15
	OLDA	73.15	–	–	–	73.15
	ULDA	73.15	–	–	–	73.15
	LPP	73.72	71.98	74.69	74.57	73.74
	2DOLDA	75.24	68.38	71.09	70.35	71.27
	2DULDA	75.24	66.66	71.09	61.29	68.57
	CSP-L1	59.00	71.53	74.20	72.09	69.21
	DSP	75.48	67.66	63.19	62.21	67.13
	DSP-L1	**75.77**	**79.86**	**84.61**	**84.93**	**81.29**
S004	RLDA	70.36	–	–	–	70.36
	OLDA	70.36	–	–	–	70.36
	ULDA	70.36	–	–	–	70.36
	LPP	50.94	60.44	62.91	64.94	58.81
	2DOLDA	79.06	70.15	73.28	70.15	73.16
	2DULDA	**79.06**	73.68	73.00	72.03	74.44
	CSP-L1	49.03	67.91	75.26	74.96	66.79
	DSP	76.64	71.86	72.01	70.46	72.74
	DSP-L1	76.74	**77.59**	**83.64**	**81.61**	**79.89**

Table 4
Comparison of the computational time (s) for all the proposed/utilized methods on the training data of data set 1 with $\theta = 0.05$ and $D = 1$. The second row is the corresponding classification accuracies (%).

	RLDA	OLDA	ULDA	LPP	2DOLDA	2DULDA	CSP-L1	DSP	DSP-L1
Computational time (s)	0.62	0.37	0.37	40.41	1.89	1.80	1.21	**0.03**	0.81
Classification accuracy (%)	57.10	57.10	57.10	55.77	66.35	66.35	58.08	66.35	**69.80**

the appropriate parameter κ with this small sample setting (SSS). LPP is designed for unsupervised learning essentially [11], and may be not necessarily appropriate for the supervised discrimination.

CSP-L1 is a robust version of common spatial pattern (CSP), which has been widely used in EEG classification based on event-related synchronization and desynchronization (ERS/ERD) phenomena [2]. Specifically, ERD/ERS is the attenuation/increase effect in rhythmic brain activity over sensorimotor cortex during both actual and imagined movement. It covers modulation of μ rhythm (around 10 Hz) and β rhythm (around 20 Hz) [2,9]. Since ERD/ERS effects directly reflect the temporal variance change of EEG data, CSP is capable to catch this kind of feature by maximizing the variance of spatially filtered signal under one class while minimizing it for the other class. Refs. [9,12] have pointed out that MRPs and ERD/ERS show different spatio-temporal activation patterns in motor movements. From Table 2, we find that performance of CSP-L1 is not comparable with that of DSP-L1 due to the inappropriate application of CSP-L1 in MRPs-based EEG classification.

Computationally, Table 4 lists the training time of all the involved methods when applying to data set 1.

5. Conclusion

In this paper, we present a new algorithm, named DSP-L1, to extract robust spatial filters from EEG data. Derived from the formulation of DSP, DSP-L1 utilizes L1-norm based robust modeling to further improve performance in withstanding the impact of outliers. Computationally, we propose an effective iterative approach to learn the spatial filters of DSP-L1. Experimental results on two EEG data sets, in which one is artificially added with multivariate outliers and the other contains the data that are badly contaminated by EOG or EMG, demonstrate the effectiveness of the proposed method. Finally, we point out that other neurophysiological features in motor movement, such as ERD, can be combined to design a better BCI system with higher classification accuracies.

Acknowledgments

The authors would like to thank Berlin BCI groups for providing the BCI competition datasets, the developers of the BCI2000 instrumentation for the EEG Motor Movement dataset on www.PhysioNet.org, and the anonymous referees for the helpful comments, which improve the quality of the paper. This work was supported in part by the National Natural Science Foundation of China under Grant 61075009, in part by the Natural Science Foundation of Jiangsu Province under Grant BK2011595, in part by the Program for New Century Excellent Talents in University of China, and in part by the Qing Lan Project of Jiangsu Province.

References

[1] H. Ramoser, J. Muller-Gerking, G. Pfurtscheller, Optimal spatial filtering of single trial EEG during imagined hand movement, IEEE Transactions on Rehabilitation Engineering 8 (2000) 441–446.
[2] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K.R. Muller, Optimizing spatial filters for robust EEG single-trial analysis, IEEE Signal Processing Magazine 25 (2008) 41–56.
[3] B. Blankertz, M. Kawanabe, R. Tomioka, F. Hohlefeld, V. Nikulin, K.R. Müller, Invariant common spatial patterns: alleviating nonstationarities in brain-computer interfacing, in: Advances in Neural Information Processing Systems 20 507 (2007).
[4] I. Goncharova, D.J. McFarland, T.M. Vaughan, J.R. Wolpaw, EMG contamination of EEG: spectral and topographical characteristics, Clinical Neurophysiology 114 (2003) 1580–1593.
[5] M. Krauledat, G. Dornhege, B. Blankertz, K.R. Müller, Robustifying EEG data analysis by removing outliers, Chaos and Complexity Letters 2 (2007) 259–274.
[6] M. Woolrich, Robust group analysis using outlier inference, Neuroimage 41 (2008) 286–301.
[7] M. Fatourechi, A. Bashashati, R.K. Ward, G.E. Birch, EMG and EOG artifacts in brain computer interface systems: a survey, Clinical Neurophysiology 118 (2007) 480–494.
[8] M. Grosse-Wentrup, C. Liefhold, K. Gramann, M. Buss, Beamforming in noninvasive brain–computer interfaces, IEEE Transactions on Biomedical Engineering 56 (2009) 1209–1219.
[9] X. Liao, D. Yao, D. Wu, C. Li, Combining spatial filters for the classification of single-trial EEG in a finger movement task, IEEE Transactions on Biomedical Engineering 54 (2007) 821–831.

[10] Y. Wang, Y.T. Wang, T.P. Jung, Translation of EEG spatial filters from resting to motor imagery using independent component analysis, PloS One 7 (2012) e37665.
[11] H. Wang, J. Xu, Local discriminative spatial patterns for movement-related potentials-based EEG classification, Biomedical Signal Processing and Control 6 (2011) 427–431.
[12] G. Dornhege, B. Blankertz, G. Curio, K.R. Muller, Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms, IEEE Transactions on Biomedical Engineering 51 (2004) 993–1002.
[13] Y. Wang, Z. Zhang, Y. Li, X. Gao, S. Gao, F. Yang, BCI competition 2003-data set IV: an algorithm based on CSSD and FDA for classifying single-trial EEG, IEEE Transactions on Biomedical Engineering 51 (2004) 1081–1086.
[14] M. Li, B. Yuan, 2D-LDA: a statistical linear discriminant analysis for image matrix, Pattern Recognition Letters 26 (2005) 527–532.
[15] H. Wang, W. Zheng, Fisher discriminanta analysis with L1-norm, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, submitted for publication.
[16] H. Wang, Q. Tang, W. Zheng, L1-norm-based common spatial patterns, IEEE Transactions on Biomedical Engineering 59 (2012) 653–662.
[17] X. Li, Y. Pang, Y. Yuan, L1-norm-based 2DPCA, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40 (2010) 1170–1175.
[18] N. Kwak, Principal component analysis based on L1-norm maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 1672–1680.
[19] X. Liao, D.H. Yao, D. Wu, C.Y. Li, Combining spatial filters for the classification of, single-trial EEG in a finger movement task, IEEE Transactions on Biomedical Engineering 54 (2007) 821–831.
[20] R. Jenatton, G. Obozinski, F. Bach, Structured sparse principal component analysis, in: International Conference on Artificial Intelligence and Statistics (AISTATS), 2010.
[21] B. Blankertz, K.R. Muller, G. Curio, T.M. Vaughan, G. Schalk, J.R. Wolpaw, A. Schlogl, C. Neuper, G. Pfurtscheller, T. Hinterberger, The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials, IEEE Transactions on Biomedical Engineering 51 (2004) 1044–1051.
[22] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (2000) 215–220.
[23] G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, J.R. Wolpaw, BCI2000: a general-purpose brain–computer interface (BCI) system, IEEE Transactions on Biomedical Engineering 51 (2004) 1034–1043.
[24] http://www.bci2000.org/
[25] M. Kukleta, M. Lamarche, Steep early negative slopes can be demonstrated in pre-movement bereitschaftspotential, Clinical Neurophysiology 112 (2001) 1642–1649.
[26] R. Cui, D. Huter, W. Lang, L. Deecke, Neuroimage of voluntary movement: topography of the Bereitschaftspotential, a 64-channel DC current source density study, Neuroimage 9 (1999) 124–134.
[27] B. Blankertz, G. Curio, K.R. Muller, Classifying single trial EEG: towards brain computer interfacing, in: Advances in Neural Information Processing Systems 14 (2001) 157–164.
[28] X. Yong, R.K. Ward, G.E. Birch, Robust common spatial patterns for EEG signal preprocessing, in: The 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, British Columbia, Canada (2008) 2087–2090.
[29] D. Cai, X. He, J. Han, SRDA: an efficient algorithm for large-scale discriminant analysis, IEEE Transactions on Knowledge and Data Engineering 20 (2008) 1–12.
[30] D. Cai, Spectral regression: a regression framework for efficient regularized subspace learning, PhD Dissertation, Graduate College of the University of Illinois at Urbana-Champaign (2009).
[31] J. Ye, Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, Journal of Machine Learning Research 6 (2005) 483–502.
[32] J. Ye, R. Janardan, Q. Li, H. Park, Feature extraction via generalized uncorrelated linear discriminant analysis, IEEE Transactions on Knowledge and Data Engineering 18 (2006) 1312–1322.
[33] J. Ye, T. Xiong, Computational and theoretical analysis of null space and orthogonal linear discriminant analysis, Journal of Machine Learning Research 7 (2006) 1183–1204.
[34] X. He, P. Niyogi, Locality preserving projections, in: Advances in Neural Information Processing Systems 16 (2003).