Active Data Selection for Motor Imagery EEG Classification

Naoki Tomida, *Student Member, IEEE*, Toshihisa Tanaka*, *Senior Member, IEEE*, Shunsuke Ono, *Student Member, IEEE*, Masao Yamagishi, *Member, IEEE*, Hiroshi Higashi, *Member, IEEE*

***Abstract*—Rejecting or selecting data from multiple trials of electroencephalography (EEG) recordings is crucial. We propose a sparsity-aware method to data selection from a set of multiple EEG recordings during motor-imagery tasks, aiming at brain machine interfaces (BMIs). Instead of empirical averaging over sample covariance matrices for multiple trials including low-quality data, which can lead to poor performance in BMI classification, we introduce weighted averaging with weight coefficients that can reject such trials. The weight coefficients are determined by the ℓ_1-minimization problem that lead to sparse weights such that almost zero-values are allocated to low-quality trials. The proposed method was successfully applied for estimating covariance matrices for the so-called common spatial pattern (CSP) method, which is widely used for feature extraction from EEG in two-class classification. Classification of EEG signals during motor imagery was examined to support the proposed method. It should be noted that the proposed data selection method can be applied to a number of variants of the original CSP method.**

***Index Terms*—brain-machine interfaces, sparsity-aware signal processing, ℓ_1-norm, motor-imagery, electroencephalography (EEG)**

I. Introduction

The brain machine interface (BMI) is a challenging application of signal processing, machine learning, and neuroscience [1]. Such interfaces capture brain activities associated with mental tasks and external stimuli and enable nonmuscular communication and a control channel for conveying messages and commands to the external world [1]–[5]. A noninvasive BMI uses recordings of brain activities such as electroencephalogram (EEG), magnetoencephalogram (MEG), and functional magnetic response imaging (fMRI). Because of its simplicity of device and high temporal resolution, using EEG is the most practical for engineering applications [6], [7].

A crucial technique for enabling BMIs associated with motor-imagery (MI-BMI) [8], [9] is efficient decoding around the motor-cortex, which leads to practical biomedical applications in rehabilitation and neuroprosthesis [10]–[13]. For instance, real and imaginary movements of hands and feet evoke a change in the so-called *mu rhythm* in different brain regions [2], [3]. Therefore, by accurately capturing these changes from EEG in the presence of measurement noise and spontaneous components related to other brain activities, we can classify the EEG signal associated with imagination of different motor actions such as hand, arm, or foot movement.

A well known method to extract brain activity for MI-BMI is the common spatial pattern (CSP) [1], [14], [15]. CSP is a set of spatial weight coefficients corresponding to each electrode in a multichannel EEG. These coefficients are determined from measured EEG data in such a way that the variances of the signal extracted by the spatial weights differ greatly between two tasks (e.g. left and right hand movement imageries). These weights can also be regarded as a spatial filter that projects observed EEG signals onto the optimal space used to classify the observed data to a class corresponding to a subject's cerebral status. Several variants of the CSP have been proposed such as Common Spatio-Spectral Pattern (CSSP) [16], Common Sparse Spectral Spatial Pattern (CSSSP) [17], SPECtrally weighted CSP (SPEC-CSP) [18], [19], iterative spatio-spectral patterns learning (ISSPL) [20], Filter Bank CSP (FBCSP) [21], Discriminative Filter Bank CSP (DFBCSP) [22], Common Spatio-Time-Frequency Patterns (CSTFP) [23], divergence-based method [24], and augmented complex CSP [25].

A common manipulation for this CSP family is to estimate the true covariance matrices in two different tasks of observed signals. To increase estimation accuracy, EEG signals (training data) are observed several times (called trials) for the same task, which yields empirical covariance matrices called *within-trial* covariance matrices. These matrices of all trials are then simply averaged. This is due to an implicit assumption that an EEG corresponding to the same task should be a (wide-sense) stationary process. However, simply averaging all trials can lead to poor estimation of the covariance matrices mainly due to the following reasons. First, the feature signal can be influenced by the user's concentration. Second, the observed EEG can be contaminated by non-stationary artifacts such as eye and muscle movement. We call a trial leading to heavily contaminated EEG a *low-quality* trial. It is crucial to eliminate low-quality trials from a dataset used for obtaining a more accurate CSP.

In this paper, we propose a method for estimating the true covariance matrix of each task not by the simple average of but by a weighted average of within-trial covariance matrices. To evaluate quality of trials, within-trial covariance matrices are

N. Tomida, S. Ono, and M. Yamagishi are with the Department of Communications and Computer Engineering, Tokyo Institute of Technology, 2-12-1-S3-60, Ookayama, Meguro-ku, Tokyo, 152-8550, Japan (e-mail: {tomida, ono, myamagi}@sp.ce.titech.ac.jp). T. Tanaka is with the Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology, 2-24-16, Nakacho, Koganei-shi, Tokyo, 184-8588, Japan (e-mail: tanakat@cc.tuat.ac.jp). H. Higashi is with the Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1, Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi, 441-8580, Japan (e-mail: higashi@tut.jp). T. Tanaka and H. Higashi are also affiliated with RIKEN Brain Science Institute, Wako-shi, Saitama, 351-0198, Japan.

This work was supported in part by JSPS KAKENHI Grant Number 24360146.

approximately jointly diagonalized. The underlying assumption behind this diagonalization is that the residue resulting from the diagonalization with respect to a low-quality trial is large. This idea of weighted averaging is related to our previous work [26], where a weighted ℓ_2-norm minimization with residue improved classification accuracy. Moreover, to increase estimation accuracy of the covariance matrices, efficient approaches are covariance shrinkage and reduced rank estimation (see [27], [28], for instance). However, our aim with this paper is to determine trial weights to reject low-quality trials. To this end, the residue is involved in an ℓ_1-norm term to design sparse weights in such a way that a larger residue yields almost zero weight to reject low-quality trials. A convex optimization problem to find the sparse weights is introduced and an iterative algorithm for solving this problem is developed.

Notations: The following terminology, notations, and mathematical operations are used throughout the paper. A matrix is denoted by a capital bold letter, e.g., $\boldsymbol{A}$ and the (i,j)-th entry of matrix $\boldsymbol{A}$ and the jth column vector are respectively denoted by $[\boldsymbol{A}]_{i,j}$ and $[\boldsymbol{A}]_{:,j}$. A matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$ is called *positive (semi) definite* if $\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} (\geqslant) > 0$ for all nonzero $\boldsymbol{u} \in \mathbb{R}^M$. The ℓ_p-norm of $\boldsymbol{x} \in \mathbb{R}^N$ is defined as $\|\boldsymbol{x}\|_p := \left(\sum_{i=1}^N |x_i|^p\right)^{1/p}$, where x_i is the ith entry of $\boldsymbol{x}$.

II. Common Spatial Pattern (CSP) in Terms of Joint Diagonalization

Before discussing the proposed method, we summarize the CSP, which is obtained as a generalized eigenvectors of a pair of two covariance matrices. In other words, the CSP is a result of joint diagonalization.

Let $\boldsymbol{X}^k \in \mathbb{R}^{M\times N}$ be a matrix consisting of M channel signals with N samples at the kth trial. The CSP is a topological pattern derived from scalp EEG given as vector $\boldsymbol{v} \in \mathbb{R}^M$, which minimizes the in-class variance of a signal extracted by a linear combination of $\boldsymbol{X}^k$ [14], [15]. In general, each channel signal in $\boldsymbol{X}^k$ is band-limited by a bandpass filter that passes the frequency components related to the target brain activity. The components of $\boldsymbol{X}^k$ are denoted by $\boldsymbol{X}^k = [\boldsymbol{x}_1^k, \ldots, \boldsymbol{x}_N^k]$, where $\boldsymbol{x}_n^k \in \mathbb{R}^M$, and n is the time index $(n = 1, \ldots, N)$. The sample mean of the observed signal is given by $\boldsymbol{\mu}^k = (1/N)\sum_{n=1}^N \boldsymbol{x}_n^k$. Then, the sample variance of the extracted signal of $\boldsymbol{X}^k$ is given by

$$\sigma^2(\boldsymbol{X}^k, \boldsymbol{v}) = \frac{1}{N}\sum_{n=1}^{N} |\boldsymbol{v}^\top(\boldsymbol{x}_n^k - \boldsymbol{\mu}^k)|^2. \tag{1}$$

Let $\mathfrak{C}_1$ and $\mathfrak{C}_2$ be the training data containing the signals observed at all trials belonging to classes (tasks) 1 and 2, respectively, such that $\mathfrak{C}_1 \cap \mathfrak{C}_2 = \varnothing$. Let K_d be the number of elements in class d $(d = 1, 2)$. The CSP of class 1 (resp. 2) is given as the maximizer (resp. minimizer) of the following generalized Rayleigh quotient [14], [15];

$$J(\boldsymbol{v}) = \frac{\boldsymbol{v}^\top \boldsymbol{S}_1 \boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{S}_2 \boldsymbol{v}} \tag{2}$$

where $\boldsymbol{S}_d$ $(d = 1, 2)$ is given as

$$\boldsymbol{S}_d = \frac{1}{K_d}\sum_{k\in\mathfrak{C}_d} \boldsymbol{S}^k, \tag{3}$$

and $\boldsymbol{S}^k \in \mathbb{R}^{M\times M}$ is the *within-trial* covariance matrix for the kth trial given as

$$\boldsymbol{S}^k := \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n^k - \boldsymbol{\mu}^k)(\boldsymbol{x}_n^k - \boldsymbol{\mu}^k)^\top. \tag{4}$$

Note that the solution of (2) is given by the generalized eigenvector corresponding to the smallest generalized eigenvalue of the generalized eigenvalue problem described as

$$\boldsymbol{S}_1 \boldsymbol{v} = \lambda \boldsymbol{S}_2 \boldsymbol{v}. \tag{5}$$

It should be noted that solving (5) is equivalent to finding a matrix, denoted by $\boldsymbol{V}$, jointly diagonalizing both $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$:

$$\boldsymbol{V}^\top \boldsymbol{S}_1 \boldsymbol{V} = \boldsymbol{\Lambda}_1, \ \boldsymbol{V}^\top \boldsymbol{S}_2 \boldsymbol{V} = \boldsymbol{\Lambda}_2, \tag{6}$$

where $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are diagonal matrices.

III. Trial Selection with Sparse Weights for Covariance Matrices

Ideally, $\boldsymbol{S}^k$ in (4) is invariant over trials up to noise since it is a result of the same mental task. This motivates the simple arithmetic averaging given in (3). However, as mentioned above, the observed EEG is highly trial-variant even for the same mental task. Moreover, the measurement environment of EEG (electronic noise, electrode impedance, etc.) always varies. Thus, we soften (3) and consider the weighted average defined as

$$\boldsymbol{S}_d^* = \sum_{k\in\mathfrak{C}_d} w_k \boldsymbol{S}^k, \tag{7}$$

where w_k is the weight coefficient at the kth trial and holds

$$\sum_{k\in\mathfrak{C}_d} w_k = 1, \ w_k \geq 0. \tag{8}$$

We define the weight vector consisting of the weights of all trials as

$$\boldsymbol{w} := [w_1, \ldots, w_{K_d}]^\top \in \mathbb{R}^{K_d}.$$

Note that $\boldsymbol{w}$ is included by $\mathcal{C}_H \cap \mathcal{C}_N$, where

$$\mathcal{C}_H := \{\boldsymbol{w} \in \mathbb{R}^{K_d} \mid \boldsymbol{w}^\top \boldsymbol{1}_{K_d} = 1\},$$
$$\mathcal{C}_N := \{\boldsymbol{w} \in \mathbb{R}^{K_d} \mid w_k \geq 0, \ \forall k\},$$

and $\boldsymbol{1}_{K_d}$ is the vector of ones of size K_d. Note that, in the CSP, $w_k = 1/K_d$ in the above equation. The underlying idea behind the weighted average is illustrated in Fig. 1. Under the constraints, the positive semi-definiteness, which is the inherent property of covariance matrices, is guaranteed.

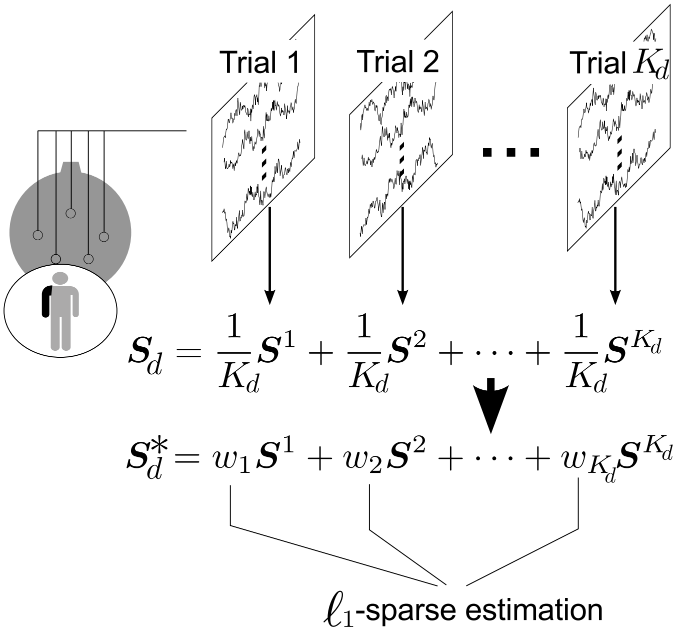

Trial 1
Trial 2
Trial K_d
$S_d = \frac{1}{K_d}S^1 + \frac{1}{K_d}S^2 + \cdots + \frac{1}{K_d}S^{K_d}$
$S_d^* = w_1 S^1 + w_2 S^2 + \cdots + w_{K_d} S^{K_d}$
ℓ_1-sparse estimation

Fig. 1: Covariance matrix of EEG for motor-imagery task is estimated by weighted averaging over within-trial covariance matrices

A. Cost Function Promoting Sparsity

Thus, the underlying problem is to find a sparse set of w_k that can reject low-quality trials.

Since the CSP is designed for binary classification, we have to determine the weights in (7) corresponding to each class. For simplicity, we design the weight of K_d trial matrices $\boldsymbol{S}^k$ $(k \in \mathfrak{C}_d)$ for a single class d.

To find trial weights eliminating low-quality trials, we introduce a weighted ℓ_1-norm term with positive scalar q_k:

$$J_{\ell_1}(\boldsymbol{w}) := \frac{1}{\sum_{k\in\mathfrak{C}_d} q_k} \sum_{k\in\mathfrak{C}_d} q_k |w_k| = \frac{1}{\mathrm{tr}(\boldsymbol{D}_1)} \left\|\boldsymbol{D}_1 \boldsymbol{w}\right\|_1 ,$$

where $\boldsymbol{D}_1$ is defined as

$$\boldsymbol{D}_1 := \mathrm{diag}\left[q_1, \ldots, q_{K_d}\right] \in \mathbb{R}^{K_d \times K_d}.$$

Note that parameter q_k is chosen to evaluate the quality of each trial. The process of choosing this parameter is discussed in the next subsection.

Assume that the ideal covariance matrix exists close to the covariance matrix obtained by simple averaging. To evaluate this, we define

$$\begin{aligned} J_{\ell_2}(\boldsymbol{w}) &:= \frac{1}{2\sum_{k\in\mathfrak{C}_d} \|\boldsymbol{S}^k\|_F^2} \left\| \sum_{k\in\mathfrak{C}_d} \left(\frac{1}{K_d} - w_k\right) \boldsymbol{S}^k \right\|_F^2 \\ &= \frac{1}{2\mathrm{tr}(\boldsymbol{G})} (\boldsymbol{w} - \boldsymbol{1}_{K_d}/K_d)^\top \boldsymbol{G} (\boldsymbol{w} - \boldsymbol{1}_{K_d}/K_d), \end{aligned}$$

where $\boldsymbol{G}$ is defined as

$$\boldsymbol{G} := \begin{bmatrix} \mathrm{tr}\left[\boldsymbol{S}^1 \boldsymbol{S}^{1\top}\right] & \cdots & \mathrm{tr}\left[\boldsymbol{S}^1 \boldsymbol{S}^{K_d\top}\right] \\ \vdots & \ddots & \vdots \\ \mathrm{tr}\left[\boldsymbol{S}^{K_d} \boldsymbol{S}^{1\top}\right] & \cdots & \mathrm{tr}\left[\boldsymbol{S}^{K_d} \boldsymbol{S}^{K_d\top}\right] \end{bmatrix} \in \mathbb{R}^{K_d \times K_d}.$$

Following the above discussion, the proposed optimization problem is given, with positive parameter α to control the sparsity, as

$$\min_{\boldsymbol{w}\in\mathcal{C}_H\cap\mathcal{C}_N} \quad \alpha J_{\ell_1}(\boldsymbol{w}) + J_{\ell_2}(\boldsymbol{w}). \tag{9}$$

B. Trial Quality Deduced from Approximate Joint Diagonalization

1) Quantification of Trial Quality: If the observed EEG is stationary over the trial up to noise, $\boldsymbol{S}^k$ $(k \in \mathfrak{C}_1 \cup \mathfrak{C}_2)$ should be diagonalized in the same way, even though $\boldsymbol{S}_d$ is substituted with any $\boldsymbol{S}^k$ in (6). However, as mentioned above, this assumption is not true. Hence, we do not consider the exact joint diagonalization but an approximate joint diagonalization given by

$$\boldsymbol{S}^k = \boldsymbol{U}\boldsymbol{\Sigma}^k\boldsymbol{U}^\top + \boldsymbol{E}^k \ (k \in \mathfrak{C}_1 \cup \mathfrak{C}_2), \tag{10}$$

where $\boldsymbol{U}$ is a common factor and $\boldsymbol{\Sigma}^k$ and $\boldsymbol{E}^k$ are respectively a diagonal matrix and an error matrix at the kth trial. If the desired EEG is not observed at some trial, the covariance matrix of the EEG should be the outliers. In this situation, residues resulting from the diagonalization of the covariance matrices should be large.

Therefore, we detect trials, which are not jointly diagonalized well, to assign those trials to smaller weights. That is, we regard trials, where the Frobenius norms of $\boldsymbol{E}^k$, $\|\boldsymbol{E}^k\|_F$ are large, as low-quality trials and impose small weights on those trials. Thus, we simply choose

$$q_k := \|\boldsymbol{E}^k\|_F.$$

2) Approximate Joint Diagonalization by FFDIAG: Various approaches to the approximate joint diagonalization algorithm can be considered. In this paper, we use the Fast Frobenius Diagonalization (FFDIAG) algorithm [29]. This iterative algorithm attempts to solve the following optimization problem.

$$\min_{\boldsymbol{B}\in\mathbb{R}^{M\times M}} \ \mathcal{F}(\boldsymbol{B}) := \sum_{k\in\mathfrak{C}_1\cup\mathfrak{C}_2} \sum_{i\neq j} \left[\boldsymbol{B}\boldsymbol{S}^k\boldsymbol{B}^\top\right]_{i,j}^2 .$$

The FFDIAG algorithm for the above minimization problem is summarized in Algorithm 1, which yields a common factor $\boldsymbol{B}$ such that

$$\boldsymbol{S}^k = \boldsymbol{B}^{-1}\boldsymbol{\Lambda}^k(\boldsymbol{B}^{-1})^\top + \boldsymbol{B}^{-1}\boldsymbol{R}^k(\boldsymbol{B}^{-1})^\top,$$

where $\boldsymbol{\Lambda}^k$ and $\boldsymbol{R}^k$ are respectively a diagonal matrix and an off-diagonal matrix at the kth trial. We adopt this decomposition as the joint diagonalization in (10), i.e., we adopt

$$\boldsymbol{U} = \boldsymbol{B}^{-1}, \ \boldsymbol{\Sigma}^k = \boldsymbol{\Lambda}^k, \ \boldsymbol{E}^k = \boldsymbol{B}^{-1}\boldsymbol{R}^k(\boldsymbol{B}^\top)^{-1}.$$

C. Iterative Optimization Method

The optimization problem described in (13) can be solved using projected gradient methods with a simplex projection $\boldsymbol{P}_{\mathcal{C}_H\cap\mathcal{C}_N}$ [30]. However, in this case, it takes a great deal of time to converge to the optimal solution due to the ill-condition of $\boldsymbol{G}$ in J_{ℓ_1}. To avoid this situation, we apply the Alternating-Direction Method for Multipliers (ADMM) [31]–[33].

Algorithm 1 Approximate Joint Diagonalization by using FFDIAG

```
Input $\boldsymbol{S}^k \ (k \in \mathfrak{C}_1 \cup \mathfrak{C}_2)$.
$\boldsymbol{A}_{(1)} = \boldsymbol{0}$, $\boldsymbol{B}_{(1)} = \boldsymbol{I}$.
repeat
    1. Compute $\boldsymbol{A}_{(n)}$ as follows:
        $y_{ij} = \sum_k [\boldsymbol{S}^k_{(n)}]_{j,j} [\boldsymbol{S}^k_{(n)}]_{i,j}, \ z_{ij} = \sum_k [\boldsymbol{S}^k_{(n)}]_{i,i} [\boldsymbol{S}^k_{(n)}]_{j,j},$
        $[\boldsymbol{A}_{(n)}]_{i,j} = \frac{z_{ij} y_{ji} - z_{ii} y_{ij}}{z_{jj} z_{ii} - z_{ij}^2}, \ [\boldsymbol{A}_{(n)}]_{j,i} = \frac{z_{ij} y_{ij} - z_{jj} y_{ji}}{z_{jj} z_{ii} - z_{ij}^2}.$
        if $\|\boldsymbol{A}_{(n)}\|_F > \theta$ then
            $\boldsymbol{A}_{(n)} = \frac{\theta}{\|\boldsymbol{A}_{(n)}\|_F} \boldsymbol{A}_{(n)}$.
        end if
    2. $\boldsymbol{B}_{(n+1)} = (\boldsymbol{I} + \boldsymbol{A}_{(n)}) \boldsymbol{B}_{(n)}$.
    3. Normalize columns of $\boldsymbol{B}_{(n)}$.
    4. $\boldsymbol{S}^k_{(n+1)} = (\boldsymbol{I} + \boldsymbol{A}_{(n)}) \boldsymbol{S}^k_{(n)} (\boldsymbol{I} + \boldsymbol{A}_{(n)})^\top$.
until converged.
$\boldsymbol{B} = \boldsymbol{B}_{(n+1)}$.
Store the diagonal part of $\boldsymbol{S}^k_{(n+1)}$ in $\boldsymbol{\Lambda}^k$.
Store the offdiagonal part of $\boldsymbol{S}^k_{(n+1)}$ in $\boldsymbol{R}^k$.
$\boldsymbol{E}^k = \boldsymbol{B}^{-1} \boldsymbol{R}^k (\boldsymbol{B}^{-1})^\top$.
Output $\boldsymbol{B}$ and $\boldsymbol{E}^k \ (k \in \mathfrak{C}_1 \cup \mathfrak{C}_2)$.
```

In the ADMM, we consider the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^N, \boldsymbol{z} \in \mathbb{R}^M} & f(\boldsymbol{w}) + g(\boldsymbol{z}), \\ \text{subject to } & \boldsymbol{L}\boldsymbol{w} - \boldsymbol{z} = \boldsymbol{0}, \end{aligned} \tag{11}$$

where f and g are *proper lower semicontinuous convex*[1], i.e., $f \in \Gamma_0(\mathbb{R}^N)$, $g \in \Gamma_0(\mathbb{R}^M)$, and a linear operator $\boldsymbol{L} \in \mathbb{R}^{M \times N} \backslash \{\boldsymbol{O}\}$ satisfies a mild condition [33].

Assume that $\boldsymbol{L}$ in (11) has full column-rank. For (11), the ADMM consists of minimizing $\mathcal{L}_\gamma$ over $\boldsymbol{w}$ and over $\boldsymbol{z}$, and updating the Lagrange multiplier $\boldsymbol{d}$.

$$\begin{cases} \boldsymbol{w}_{(n+1)} = \underset{\boldsymbol{w} \in \mathbb{R}^N}{\operatorname{argmin}} \ \mathcal{L}_\gamma(\boldsymbol{w}, \boldsymbol{z}_{(n)}, \boldsymbol{d}_{(n)}) \\ \boldsymbol{z}_{(n+1)} = \underset{\boldsymbol{z} \in \mathbb{R}^M}{\operatorname{argmin}} \ \mathcal{L}_\gamma(\boldsymbol{w}_{(n+1)}, \boldsymbol{z}, \boldsymbol{d}_{(n)}) \, , \\ \boldsymbol{d}_{(n+1)} = \boldsymbol{d}_{(n)} + (\boldsymbol{L}\boldsymbol{w}_{(n+1)} - \boldsymbol{z}_{(n+1)}) \end{cases} \tag{12}$$

where $\mathcal{L}_\gamma$ is the augmented Lagrangian of index $\gamma \in (0, \infty)$ defined by

$$\mathcal{L}_\gamma(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{d}) = f(\boldsymbol{w}) + g(\boldsymbol{z}) + \frac{1}{\gamma} \boldsymbol{d}^\top (\boldsymbol{L}\boldsymbol{w} - \boldsymbol{z}) + \frac{1}{2\gamma} \|\boldsymbol{L}\boldsymbol{w} - \boldsymbol{z}\|_2^2,$$

where $\boldsymbol{d} \in \mathbb{R}^M$ and γ are a Lagrange multiplier and a positive scalar parameter, respectively. With the ADMM, the effect of the ill-condition can be reduced to a certain degree by properly tuning a parameter γ appearing in the steps given as in (12).

To apply the ADMM to the constrained minimization problem, we rewrite (9) with the indicator function as the following unconstrained optimization problem.

$$\min_{\boldsymbol{w} \in \mathbb{R}^{K_d}} \ \alpha \boldsymbol{w}^\top \frac{\boldsymbol{D}_1}{\operatorname{tr}(\boldsymbol{D}_1)} \boldsymbol{1}_{K_d} + J_{\ell_2}(\boldsymbol{w}) + \iota_{\mathcal{C}_H}(\boldsymbol{w}) + \iota_{\mathcal{C}_N}(\boldsymbol{w}), \tag{13}$$

where $\iota_{\mathcal{C}_H}$ and $\iota_{\mathcal{C}_N}$ denote the *indicator functions*[2]. Note that in the above problem, the second linear term $\boldsymbol{w}^\top \frac{\boldsymbol{D}_1}{\operatorname{tr}(\boldsymbol{D}_1)} \boldsymbol{1}_{K_d}$ is a replacement of the ℓ_1-norm in $J_{\ell_1}(\boldsymbol{w})$ since $\boldsymbol{w}$ is constrained in $\mathcal{C}_N$. The steps of the above algorithm are shown in Algorithm 2, which is derived by adopting the ADMM algorithm with

$$\begin{aligned} f(\boldsymbol{w}) &:= \alpha \boldsymbol{w}^\top \frac{\boldsymbol{D}_1}{\operatorname{tr}(\boldsymbol{D}_1)} \boldsymbol{1}_{K_d} + J_{\ell_2}(\boldsymbol{w}) + \iota_{\mathcal{C}_H}(\boldsymbol{w}), \\ g(\boldsymbol{z}) &:= \iota_{\mathcal{C}_N}(\boldsymbol{z}), \end{aligned}$$

and $\boldsymbol{L} = \boldsymbol{I} \in \mathbb{R}^{K_d \times K_d}$ in (11).

IV. Experimental Results

Two experiments are conducted to support the proposed method. The first one is an experiment in artificial situation to confirm whether weights corresponding to low-quality trials (nonstationary data) become relatively small values or zeros by using the proposed method. The other one is an experiment of classification of EEG signals during motor imagery to show performance in accuracy with the proposed method.

A. EEG Data Description

We used dataset IVa from BCI competition III and dataset 1 from BCI competition IV, which were public datasets provided by Fraunhofer FIRST (Intelligent Data Analysis Group) and Campus Benjamin Franklin of the Charité - University Medicine Berlin (Department of Neurology, Neurophysics Group) [35], [36], respectively. Aside from the public datasets, we recorded the EEG of motor-imagery (called dataset *JK-HH 1*). The experiment for obtaining JK-HH 1 was approved by the research ethics committee of Tokyo University of Agriculture and Technology.

1) Dataset IVa: This public dataset consists of EEG signals during right hand and right foot motor-imageries. The EEG signals from 118 channels at positions of the extended international 10/20-system were recorded from five subjects assigned labels *aa*, *al*, *av*, *aw*, and *ay*. The measured signal was bandpass-filtered with a passband of 0.05–200 Hz then digitized at 1000 Hz with 16 bits (0.1 μV). In the experiment, visual cues told the subject which imagery task (left hand, right hand, or right foot) should be performed. The cue was indicated for 3.5 seconds and the subject performed the motor imagery for this period. The resting interval between two trials was randomized from 1.75–2.25 seconds. Only EEG trials for right hand and right foot were provided.

We also applied a bandpass filter whose passband was 7–30 Hz and downsampled to 100 Hz to this dataset. The dataset

[1] A function $f : \mathbb{R}^N \to (-\infty, \infty]$ is called proper lower semicontinuous convex if $\operatorname{dom}(f) := \{\boldsymbol{x} \in \mathbb{R}^N \mid f(\boldsymbol{x}) < \infty\} \neq \varnothing$, $\operatorname{lev}_{\leq \alpha}(f) := \{\boldsymbol{x} \in \mathbb{R}^N \mid f(\boldsymbol{x}) \leq \alpha\}$ is closed for every $\alpha \in \mathbb{R}$, and $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ and $\lambda \in (0,1)$, respectively [34]. The set of all proper lower semicontinuous convex functions in $\mathbb{R}^N$ is denoted by $\Gamma_0(\mathbb{R}^N)$.

[2] A subset $\mathcal{C} \subset \mathbb{R}^N$ is called *convex* if for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and $\lambda \in (0,1)$, $\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathcal{C}$. For a given nonempty closed convex subset $\mathcal{C} \subset \mathbb{R}^N$, the *indicator function* $\iota_{\mathcal{C}} \in \Gamma_0(\mathbb{R}^N)$ is defined by

$$\iota_{\mathcal{C}}(\boldsymbol{x}) := \begin{cases} 0 & (\boldsymbol{x} \in \mathcal{C}) \\ \infty & (\boldsymbol{x} \notin \mathcal{C}) \end{cases},$$

and the *metric projection* onto $\mathcal{C}$ is the mapping $\boldsymbol{P}_{\mathcal{C}} : \mathbb{R}^N \to \mathcal{C} : \boldsymbol{x} \mapsto \operatorname{argmin}_{\boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{y}\|_2$. The metric projection is also described, for any $\gamma \in (0, \infty)$, as $\boldsymbol{P}_{\mathcal{C}}(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{y} \in \mathbb{R}^N} \iota_{\mathcal{C}}(\boldsymbol{y}) + \frac{1}{2\gamma} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, which is a particular case of the *proximity operator* [34].

Algorithm 2 Solver for optimization problem in (13)

```
Initialize $\boldsymbol{w}_{(1)}$, $\boldsymbol{z}_{(1)}$, and $\boldsymbol{d}_{(1)}$.
repeat
    1a. $\xi = \frac{\mathbf{1}_{K_d}^\top (\gamma \frac{\boldsymbol{G}}{\mathrm{tr}(\boldsymbol{G})} + \boldsymbol{I})^{-1} \left[ (\boldsymbol{z}_{(n)} - \boldsymbol{d}_{(n)}) + \gamma \left( \frac{\boldsymbol{G}\mathbf{1}_{K_d}}{K_d \mathrm{tr}(\boldsymbol{G})} - \alpha \frac{\boldsymbol{D}_1 \mathbf{1}_{K_d}}{\mathrm{tr}(\boldsymbol{D}_1)} \right) \right] - 1}{\gamma \mathbf{1}_{K_d}^\top (\gamma \frac{\boldsymbol{G}}{\mathrm{tr}(\boldsymbol{G})} + \boldsymbol{I})^{-1} \mathbf{1}_{K_d}}.$   (14)
    1b. $\boldsymbol{w}_{(n+1)} = \left( \gamma \frac{\boldsymbol{G}}{\mathrm{tr}(\boldsymbol{G})} + \boldsymbol{I} \right)^{-1} \left[ \boldsymbol{z}_{(n)} - \boldsymbol{d}_{(n)} + \gamma \left( \frac{\boldsymbol{G}\mathbf{1}_{K_d}}{K_d \mathrm{tr}(\boldsymbol{G})} - \alpha \frac{\boldsymbol{D}_1 \mathbf{1}_{K_d}}{\mathrm{tr}(\boldsymbol{D}_1)} - \xi \mathbf{1}_{K_d} \right) \right].$   (15)
    2. $\boldsymbol{z}_{(n+1)} = \boldsymbol{P}_{\mathcal{C}_N} \left( \boldsymbol{w}_{(n+1)} + \boldsymbol{d}_{(n)} \right)$. (See (20) for a specific form.)   (16)
    3. $\boldsymbol{d}_{(n+1)} = \boldsymbol{d}_{(n)} + \boldsymbol{w}_{(n+1)} - \boldsymbol{z}_{(n+1)}$.   (17)
until converged.
```

for each subject consisted of signals of 140 trials per class. The signal in each trial was extracted 3.5 seconds after the visual cue.

2) Dataset 1: This public dataset consists of EEG signals during two motor-imageries, which were selected from three classes; left hand, right hand, and foot (side chosen by the subject; optionally also both feet). The EEG signals were recorded from four subjects assigned labels *a*, *b*, *f*, and *g*. The signals from 59 EEG channels were measured, which were most densely distributed over sensorimotor areas. The measured signal was bandpass-filtered with a passband of 0.05–200 Hz then digitized at 1000 Hz with 16 bits (0.1 μV). Additionally, the data passed through the low-pass filter (Chebyshev Type II filter of order 10 with stopband ripple of 50 dB down and stopband edge frequency of 49 Hz) then downsampled at 100 Hz (calculating the mean of blocks of 10 samples). During each experiment, visual cues were displayed for a period of 4.0 seconds during which the subject was instructed to perform the cued motor imagery task (left hand, right hand, or right foot). These periods were interleaved with 2.0 seconds of blank screen and 2.0 seconds with a fixation cross shown in the center of the screen.

We also applied a bandpass filter whose passband was 7–30 Hz to this data. The dataset for each subject consisted of signals of 100 trials per class. The signal in each trial was extracted 4.0 seconds after the visual cue.

3) Dataset JK-HH 1: This original dataset consists of EEG signals during two motor-imageries, right hand and foot. They were recorded from five (5 males; averaged age 23.2 with SD 1.6) subjects assigned labels *sa*, *sb*, *sc*, *sd*, and *se*. During the recording, the subjects performed the motor-imagery tasks instructed by a visual cue. The cue was given by an arrow on an LCD screen. The right and down arrows instructed the subjects to perform the motor imagery tasks of the right hand and the foot, respectively. The subjects performed the tasks repeatedly with an interval of around 3 seconds. The EEG signals were recorded with Ag/AgCl active electrodes (g.LADYbird, g.LADYbirdGND, and g.GAMMAearclip produced by Guger Technologies) and a power supply (g.GAMMAbox produced by Guger Technologies). There were 29 electrodes, which were placed at F3, Fz, F4, FC5, FC3, FC6, FCz, FC2, FC4, FC6, T7, C5, C3, C1, Cz, C2, C4, C6, T8, CP5, CP3, CP1, CPz, CP2, CP4, CP6, P3, Pz, and P4 (the positions are represented by the notation of the International 10-10 system [37]). The signals observed from the electrodes were amplified using a bio-amplifier (MEG-6116 produced by Nihon Kohden). The amplifier analog-filtered the signals with a passband of 0.5–100 Hz. The signals through the amplifier were sampled using an A/D converter (AIO-163202F-PE produced by Contec) with a sampling rate of 256 Hz. The converted signals were recorded with the Data Acquisition Toolbox, which is one of the toolboxs of MATLAB (MathWorks). We also applied to this dataset a Butterworth lowpass filter, whose cutoff frequency was 50 Hz and filter order was 4, and downsampled to 128 Hz.

We also applied to this dataset a bandpass filter whose passband was 7–30 Hz. The dataset for each subject consisted of signals of 100 trials per class. The signal in each trial was extracted 4.0 seconds after the visual cue.

B. Confirmation of Sparsity in Artificial Situation

The following numerical experiments were conducted to confirm whether the weights corresponding to the low-quality trials (nonstationary data) are almost zeros when the proposed method is applied.

1) Simulation Scenario: Suppose in this simulation that the observed dataset of an EEG in class d consists of K_d trials, where K_0 trials out of K_d trials (i.e., $K_0 < K_d$) are wide-sense stationary observed with white Gaussian noise and the remaining $K_d - K_0$ trials are non-stationary with different covariance matrices.

This scenario was implemented similar to [38], [39] as follows. As a reference signal, we used a signal corresponding to each class that was chosen randomly out of the dataset of subject *al*. We assumed $\overline{\boldsymbol{X}}_d \in \mathbb{R}^{M \times N}$ as the pure EEG signal in class d. Based on the signal, we produced K_d trials $\boldsymbol{Y}_d^k \in \mathbb{R}^{M \times N}$ $(k = 1, \ldots, K_d)$ as follows.

$$\boldsymbol{Y}_d^k = \begin{cases} \overline{\boldsymbol{X}}_d + \boldsymbol{N}_1^k & (k = 1, \ldots, K_0) \\ \overline{\boldsymbol{X}}_d + \boldsymbol{N}_1^k + \boldsymbol{N}_2^k & (k = K_0 + 1, \ldots, K_d) \end{cases},$$

where $\boldsymbol{N}_1^k \in \mathbb{R}^{M \times N}$ denotes Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma_1^2 \boldsymbol{I})$ and $\boldsymbol{N}_2^k \in \mathbb{R}^{M \times N}$ stands for outlier noise generated from a normal mixture distribution [40] such as

$$[\boldsymbol{N}_2^k]_{:,n} \sim (1 - \epsilon)\delta_{\mathbf{0}} + \epsilon \mathcal{N}(\mathbf{0}, \sigma_2^2 \boldsymbol{I}),$$

where n $(n = 1, \ldots, N)$ is a time index, $\delta_{\mathbf{0}}$ denotes a point mass distribution located at zero, and $\epsilon > 0$ is the occurrence probability [39].

TABLE I: Number of trials of which coefficients were zero, $\sharp(w_k = 0)$ in both classes when $K_0 = 111$ (one non-stationary trial) and $K_0 = 102$ (ten non-stationary trials)

	$K_0 = 111$	$K_0 = 102$
$\sharp(w_k = 0)$ in class 1	1	10
$\sharp(w_k = 0)$ in class 2	1	10

2) Results: We set $K_d = 112$, and chose $\sigma_1 = 1.0$, $\sigma_2 = 1.0 \times 10^3$ for the noise distributions and $\epsilon = 1.0 \times 10^{-1}$ for the occurrence probability. Table I lists the resulting zero weight coefficients for both classes when $K_0 = 111$ (one non-stationary trial) and $K_0 = 102$ (ten non-stationary trials). In the parameter settings of Algorithm 2, we chose $\alpha = 2.0 \times 10^{-1}$ and $\gamma = 1.0 \times 10^{-3}$ in both cases.

C. Two-Class EEG Classification

1) Parameter Settings: The following three types of CSP are used for feature extraction of motor-imagery EEG:

- A CSP with the empirically averaged covariance matrix, as in (3).
- A CSP with the weight-averaged covariance matrix, as in (7), with a simple weighting technique:

$$w_k = \eta \left\| \boldsymbol{E}^k \right\|_F^{-1} \quad (k \in \mathfrak{C}_d), \tag{18}$$

  where η is a constant for normalization such that $1 = \sum_{k \in \mathfrak{C}_d} w_k$. The underlying idea is to simply give a small weight corresponding to a large residue (a low qulity trial).
- A CSP with the weight-averaged covariance matrix, as in (7), with the proposed sparsity-aware estimation method.

It should be noted that more recent CSP-based methods could be used in the experiments; however, the aim with this study was to show the effectiveness of the proposed data selection/rejection method, and that the choice of CSP was not the issue.

We defined the following feature vector as the output of feature extraction using CSP. Although the solution of (2) is given by the eigenvector corresponding to the largest eigenvalue in (5), we can use the other eigenvectors for classification [41]. The M eigenvectors can be obtained by solving (5) as $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_M$, where $\hat{\boldsymbol{v}}_i$ is the eigenvector corresponding to the ith smallest eigenvalue of (5). We used the $2r$ eigenvectors to form the feature vector, denoted by $\boldsymbol{y}$, for classification of unlabeled data, $\boldsymbol{X}$.

$$\begin{aligned} \boldsymbol{y} = [&\sigma^2(\boldsymbol{X}, \hat{\boldsymbol{v}}_1), \ldots, \sigma^2(\boldsymbol{X}, \hat{\boldsymbol{v}}_r), \\ &\sigma^2(\boldsymbol{X}, \hat{\boldsymbol{v}}_{M-r+1}), \ldots, \sigma^2(\boldsymbol{X}, \hat{\boldsymbol{v}}_M)]^\top \in \mathbb{R}^{2r}. \end{aligned} \tag{19}$$

The feature is classified with linear discriminant analysis (LDA) [42].

2) Results: In Table II, we list the classification accuracy by CSP with the following weighting techniques: (i) the simple average ($w_k = 1/K_d$), (ii) the weighted average with $w_k \sim \left\| \boldsymbol{E}^k \right\|_F^{-1}$, and (iii) the weighted average with the proposed sparse weights. The results of the proposed data selection method are the highest classification accuracy for each subject among the accuracies obtained with several α. The results were obtained by conducting 5-fold cross validation (CV). In the table, $\sharp(w_k \neq 0)$ stands for the average number of trials of which coefficients were not zero. In other words, $\sharp(w_k \neq 0)$ was the average number of selected trials from the dataset. In all cases, for simplicity of comparison, the number of the associated spatial weights r in (19) was fixed to 3. For every parameter α, we chose $\gamma = 1.0 \times 10^5$.

To see the sensitivity of parameter α, we measured the classification accuracy for varying sparsity parameter α for each subject in each dataset, as shown in Fig. 2. The more α increased, the more sparsity was promoted. In the case of the smallest α in the figures was 10^{-7}, the term of the ℓ_1-norm in the cost function could be virtually ignored; therefore, we observed that the resulting weights were identical. As also shown in Fig. 2, there was no common trend in the change in classification accuracy by the parameter among the subjects. The results suggest that the number of low-quality samples was different among the subjects.

V. Discussion and Conclusion

The main contribution of this paper was to establish new methods for selecting or rejecting trials. A weight coefficient was assigned to each within-trial covariance matrix to measure the *quality* of the trial, and a sparse set of weights was determined by the ℓ_1 optimization problem.

The experiments to confirm sparsity in an artificial situation have shown that the trials, where the residues yielded by joint diagonalization were large, correspond to the low-quality trials. As expected, only non-stationary trials (assumed to be low-quality trials) were weighted with (almost) zero, and the others were quite uniquely weighted, as shown in Table. I.

The results of classification accuracy shown in Table II exhibit the advantage of the proposed method. Detailed discussion is given in the following. First of all, the simple non-sparse weights determined by the error matrices obtained in joint diagonalization do not help to improve the classification accuracy. This implies that the error matrices should be utilized for designing weight coefficients in more sophisticated ways.

In contrast, the proposed sparse weights led to noticeable classification results. Subject *av* showed a large improvement in accuracy by more than 7 % with the proposed trial rejection method. It is well known in the BCI community that this dataset of *av* always shows poor classification performance with variants of CSP. From this table, the average number of non-zero weights was 98.2, which implies that 126 trials out of 224 were rejected by the ℓ_1 optimization. This fact suggests that the dataset of Subject *av* contains many low-quality trials.

Next, note that even with the standard CSP, Subject *aw* showed a high accuracy of 97.66 %, and no trials were rejected with the proposed method. This implies that the dataset of this subject includes stationary signals.

On the other hand, Subject f showed a small improvement of 1.00 %, even though a large number of trials was rejected, i.e., only 62.8 trials out of 160 were selected on average. This may contradict the above argument that a dataset consists of stationary trials. However, as observed in Fig. 2b, the accuracies of Subject f appeared inconsistent over parameter α. In other words, the value of $\sharp(w_k \neq 0)$ did not mean

TABLE II: Classification accuracy [%] from 5-fold cross validation. Highest classification accuracies for each subject among the accuracies obtained with several α are listed. The figures with ± denote standard deviation. $\sharp(w_k \neq 0)$ stands for average number of trials of which coefficients were not zero. $K_1 + K_2$ denotes number of trials in both classes. We regard trial weight coefficients of less than 10^{-5} as zeros.

Method Subject		Common Spatial Pattern (CSP) Method with several weighting techniques				
		$(w_k = 1/K_d)$	$(w_k \sim 1/\left\lVert \boldsymbol{E}^k \right\rVert_F)$	$(w_k$: ℓ_1 sparse weights$)$	$\sharp(w_k \neq 0)$	$(K_1 + K_2)$
dataset IVa	*aa*	75.71 ±12.66	76.43 ±11.32	80.36 ±12.81	94.4	(224)
	al	93.57 ± 2.99	93.93 ± 5.14	95.36 ± 4.48	143.8	(224)
	av	63.21 ± 5.14	65.36 ± 2.71	71.07 ± 7.08	98.2	(224)
	aw	97.86 ± 1.96	95.71 ± 2.04	97.86 ± 1.96	224.0	(224)
	ay	92.86 ± 3.79	93.21 ± 4.07	93.57 ± 2.99	214.8	(224)
	Ave.	84.64	84.93	87.64	155.0	(224)
dataset 1	*a*	66.00 ± 9.78	66.50 ± 6.75	71.50 ± 9.75	133.2	(160)
	b	71.50 ± 5.18	67.50 ± 4.33	75.00 ± 4.68	160.0	(160)
	f	88.50 ± 6.75	89.50 ± 6.94	89.50 ± 4.47	62.8	(160)
	g	89.00 ± 4.87	79.50 ± 6.47	90.00 ± 3.54	159.8	(160)
	Ave.	78.75	75.75	81.25	129.0	(160)
JK-HH 1	*sa*	83.50 ± 6.02	78.50 ± 7.62	83.50 ± 6.02	160.0	(160)
	sb	56.50 ± 3.79	54.00 ± 6.27	62.00 ± 5.42	148.0	(160)
	sc	47.50 ±10.00	49.00 ± 6.75	56.50 ± 6.98	105.0	(160)
	sd	49.00 ± 9.12	48.00 ±15.45	56.50 ± 7.42	85.6	(160)
	se	85.50 ±10.37	85.50 ± 9.25	87.00 ±11.00	131.8	(160)
	Ave.	64.40	63.00	69.10	126.1	(160)

the quality of trials in the dataset. Unlike Subject *f*, some subjects exhibited a clear relation between α and accuracy. For instance, Subject *aw* showed that increased sparsity led to decreased accuracy.

The experiment of two EEG classification showed that the proposed method is effective. What we would like to emphasize is that the proposed method can be applied to variants of CSP. By introducing the ℓ_1 norm to the cost function, we can obtain the sparse weights, which lead to the rejection of low-quality trials.

Even though introducing sparse weight coefficients improved classification accuracy, an important question arose: Does a zero weight really correspond to a low-quality trial? This paper established how to select or reject trials from a dataset based on the sparse ℓ_1 optimization. The established method should be verified through a psychophisiological experiment in which a subject is randomly distracted during a motor-imagery mental task and the distribution of weight coefficients derived based on the proposed method is evaluated. This important problem will be addressed in the future. The analysis with such EEG data in which low-quality trials are on purpose might help in developing a model for evaluating the quality of EEG signals. We then can discuss the issue of bias caused by ℓ_1 regularization in the solution [43].

The limitation of the proposed method is that we have to choose the regularization parameter. The simplest way to choose the parameter is using a CV method with a dataset. When the learning and test data are separated out of the dataset for CV and the number of low-quality data in the learning data is significantly different in each CV, the choice of the parameter by CV might not work well. Therefore, we need a method for estimating an appropriate regularization parameter for each dataset or subject.

In this paper, we did not discuss the problem of how to reject a low-quality test sample to be classified. However, how to reject low-quality samples by real-time processing is crucial for practical use of BMI. We will address this problem by expanding the proposed method.

Moreover, the concept of the proposed method could be extended as a method for rejecting each sample instead of each trial. To reject samples, we need to design $N \times (K_1 + K_2)$ coefficients as the weights. This can lead to additional computational cost compared to that for finding the trial weights. Additionally, a joint diagonalization with matrices whose rank is 1 would be unstable. In this case, we need another method to estimate $\{q_k\}_{k=1}^{N \times (K_1+K_2)}$. Although we should solve these problems for extending the proposed method, the proposed method can be used as a framework for rejecting samples not only for rejecting trials.

Appendix A
Derivation of Steps in Algorithm 2

We derive Steps 1 and 2 in Algorithm 2. Step 1 is derived by solving the following optimization problem (also see (12)).

$$\begin{aligned}\boldsymbol{w}_{(n+1)} &= \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^{K_d}} \ \mathcal{L}_\gamma(\boldsymbol{w}, \boldsymbol{z}_{(n)}, \boldsymbol{d}_{(n)}) \\ &= \operatorname*{argmin}_{\boldsymbol{w} \in \mathcal{C}_H} \ \alpha \boldsymbol{w}^\top \frac{\boldsymbol{D}_1}{\mathrm{tr}(\boldsymbol{D}_1)} \boldsymbol{1}_{K_d} + J_{\ell_2}(\boldsymbol{w}) \\ &\qquad + \frac{1}{2\gamma} \| \boldsymbol{z}_{(n)} - \boldsymbol{w} - \boldsymbol{d}_{(n)} \|_2^2 .\end{aligned}$$

Using a multiplier, we define the Lagrangian as

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) := &\frac{1}{2} \left(\boldsymbol{w} - \boldsymbol{1}_{K_d}/K_d \right)^\top \frac{\boldsymbol{G}}{\mathrm{tr}(\boldsymbol{G})} \left(\boldsymbol{w} - \boldsymbol{1}_{K_d}/K_d \right) \\ &+ \alpha \boldsymbol{w}^\top \frac{\boldsymbol{D}_1}{\mathrm{tr}(\boldsymbol{D}_1)} \boldsymbol{1}_{K_d} + \frac{1}{2\gamma} \| \boldsymbol{z}_{(n)} - \boldsymbol{w} - \boldsymbol{d}_{(n)} \|_2^2 \\ &+ \xi (\boldsymbol{w}^\top \boldsymbol{1}_{K_d} - 1)\end{aligned}$$

with ξ being the Lagrange multiplier for the constraint set $\mathcal{C}_H$. Taking a gradient of $\mathcal{L}(\boldsymbol{w})$ with respect to $\boldsymbol{w}$, we obtain the requirement

$$\begin{aligned}\nabla \mathcal{L}(\boldsymbol{w}) = &\frac{\boldsymbol{G}}{\mathrm{tr}(\boldsymbol{G})} (\boldsymbol{w} - \boldsymbol{1}_{K_d}/K_d) + \alpha \frac{\boldsymbol{D}_1 \boldsymbol{1}_{K_d}}{\mathrm{tr}(\boldsymbol{D}_1)} \\ &+ \frac{1}{\gamma} (\boldsymbol{w} + \boldsymbol{d}_{(n)} - \boldsymbol{z}_{(n)}) + \xi \boldsymbol{1}_{K_d} \\ = &\,\boldsymbol{0}.\end{aligned}$$

Thus, the solution is obtained as (15). Plugging this solution into the constraint $\boldsymbol{w}^\top \mathbf{1}_{K_d} = 1$ leads to

$$\begin{aligned}
&\mathbf{1}_{K_d}^\top \boldsymbol{w}_{(n+1)} \\
&= \mathbf{1}_{K_d}^\top \left(\gamma \frac{\boldsymbol{G}}{\mathrm{tr}(\boldsymbol{G})} + \boldsymbol{I}\right)^{-1} \\
&\quad \times \left[\boldsymbol{z}_{(n)} - \boldsymbol{d}_{(n)} + \gamma \left(\frac{\boldsymbol{G}\mathbf{1}_{K_d}}{K_d \mathrm{tr}(\boldsymbol{G})} - \alpha \frac{\boldsymbol{D}_1 \mathbf{1}_{K_d}}{\mathrm{tr}(\boldsymbol{D}_1)} - \xi \mathbf{1}_{K_d}\right)\right] \\
&\quad - \gamma \xi \mathbf{1}_{K_d}^\top \left(\gamma \frac{\boldsymbol{G}}{\mathrm{tr}(\boldsymbol{G})} + \boldsymbol{I}\right)^{-1} \mathbf{1}_{K_d} \\
&= 1,
\end{aligned}$$

which readily results in (14). Note that strict discussion about the Lagrange method is written in, e.g., [34, Proposition 26.11].

Step 2 is derived as follows.

$$\begin{aligned}
\boldsymbol{z}_{(n+1)} &= \underset{\boldsymbol{z} \in \mathbb{R}^{K_d}}{\mathrm{argmin}} \ \mathcal{L}_\gamma(\boldsymbol{w}_{(n+1)}, \boldsymbol{z}, \boldsymbol{d}_{(n)}) \\
&= \underset{\boldsymbol{z} \in \mathbb{R}^{K_d}}{\mathrm{argmin}} \ \iota_{\mathcal{C}_N}(\boldsymbol{z}) + \frac{1}{2\gamma} \|\boldsymbol{z} - (\boldsymbol{w}_{(n+1)} + \boldsymbol{d}_{(n)})\|_2^2 \\
&= \boldsymbol{P}_{\mathcal{C}_N}(\boldsymbol{w}_{(n+1)} + \boldsymbol{d}_{(n)}),
\end{aligned}$$

where $\boldsymbol{P}_{\mathcal{C}_N}$ is the metric projection onto $\mathcal{C}_N$ defined by

$$\boldsymbol{P}_{\mathcal{C}_N} : \mathbb{R}^{K_d} \to \mathcal{C}_N : [\boldsymbol{x}]_i \mapsto \begin{cases} 0 & \text{if } [\boldsymbol{x}]_i \leq 0 \\ [\boldsymbol{x}]_i & \text{otherwise} \end{cases}. \tag{20}$$

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their insightful comments and constructive suggestion. The authors would also like to thank Prof. Isao Yamada of Tokyo Institute of Technology and members of his laboratory for their useful comments and proofreading on this paper.

REFERENCES

[1] G. Dornhege, J. d. R. Millan, T. Hinterberger, D. McFarland, and K.-R. Müller, *Toward Brain-Computer Interfacing*. Cambridge, MA: MIT Press, 2007.

[2] S. Sanei, *Adaptive Processing of Brain Signals*. Hoboken, NJ: John Wiley & Sons, April 2013.

[3] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[4] B. He, S. Gao, H. Yuan, and J. R. Wolpaw, "Brain–computer interfaces," in *Neural Engineering*. Springer, 2013, pp. 87–151.

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

[6] D. J. McFarland and J. R. Wolpaw, "Brain-computer interface operation of robotic and prosthetic devices," *Computer*, vol. 41, no. 6, pp. 52–56, 2008.

[7] C. Zhang, Y. Kimura, H. Higashi, and T. Tanaka, "A simple platform of brain–controlled mobile robot and its implementation by SSVEP," in *Proc. 2012 Int. Joint Conf. Neural Netw. (IJCNN 2012)*, 2012, 7 pages.

[8] D. J. McFarland and J. R. Wolpaw, "Brain-computer interfaces for communication and control," *Commun. ACM*, vol. 54, no. 5, pp. 60–66, 2011.

[9] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans," *Nat. Acad. Sci.*, vol. 101, no. 51, pp. 17 849–17 854, 2004.

[10] Y. Liu, M. Li, H. Zhang, H. Wang, J. Li, J. Jia, Y. Wu, and L. Zhang, "A tensor-based scheme for stroke patients' motor imagery EEG analysis in BCI–FES rehabilitation training," *Journal of Neuroscience Methods*, vol. 222, no. 0, pp. 238 – 249, 2014.

[11] K. K. Ang and C. Guan, "Brain-computer interface in stroke rehabilitation," *Journal of Computing Science and Engineering*, vol. 7, no. 2, pp. 139–146, 2013.

[12] K. K. Ang, C. Guan, K. Sui Geok Chua, B. T. Ang, C. Kuah, C. Wang, K. S. Phua, Z. Y. Chin, and H. Zhang, "Clinical study of neurorehabilitation in stroke using EEG-based motor imagery brain-computer interface with robotic feedback," in *Proc. 32nd Annu. Int. Conf. IEEE Eng. Med. Bio. Soc. (EMBC 2010)*, 2010, pp. 5549–5552.

[13] G. Prasad, P. Herman, D. Coyle, S. McDonough, and J. Crosbie, "Using motor imagery based brain-computer interface for post-stroke rehabilitation," in *Neural Engineering, 2009. NER '09. 4th International IEEE/EMBS Conference on*, April 2009, pp. 258–262.

[14] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clin. Neurophysiol.*, vol. 110, no. 5, pp. 787–798, 1999.

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

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

[17] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 11, pp. 2274–2281, 2006.

[18] R. Tomioka, G. Dornhege, G. Nolte, B. Blankertz, K. Aihara, and K. R. Müller, "Spectrally weighted common spatial pattern algorithm for single trial EEG classification," Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan, Tech. Rep., 2006.

[19] R. Tomioka and K.-R. Müller, "A regularized discriminative framework for EEG analysis with application to brain–computer interface," *NeuroImage*, vol. 49, no. 1, pp. 415–432, 2010.

[20] W. Wu, X. Gao, B. Hong, and S. Gao, "Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL)," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 6, pp. 1733–1743, 2008.

[21] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. 2008 Int. Joint Conf. Neural Netw. (IJCNN 2008)*, 2008, pp. 2390–2397.

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

[23] ——, "Common spatio-time-frequency patterns for motor-imagery based brain machine interfaces," *Computational Intelligence and Neuroscience*, vol. 2013, 2013.

[24] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomed. Eng.*, vol. 7, pp. 50–72, 2013.

[25] C. Park, C. C. Took, and D. P. Mandic, "Augmented Complex Common Spatial Patterns for Classification of Noncircular EEG From Motor Imagery Tasks," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 22, no. 1, pp. 1–10, 2014.

[26] N. Tomida, H. Higashi, and T. Tanaka, "A joint tensor diagonalization approach to active data selection for EEG classification," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2013)*, 2013, pp. 983–987.

[27] D. Bartz and K.-R. Müller, "Generalizing Analytic Shrinkage for Arbitrary Covariance Structures," in *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 1869–1877.

[28] N. Tomida, M. Yamagishi, I. Yamada, and T. Tanaka, "A reduced rank approach for covariance matrix estimation in EEG signal classification," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Bio. Soc. (EMBC 2014)*, 2014, pp. 668–671.

[29] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller, "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation," *Journal of Machine Learning Research*, vol. 5, pp. 801–818, 2004.

[30] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l1-ball for learning in high dimensions." in *ICML*, ser. ACM International Conference Proceeding Series, vol. 307. ACM, 2008, pp. 272–279.

[31] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comput. and Math. Appli.*, vol. 2, no. 3, pp. 19–40, 1976.

[32] J. Eckstein and D. P. Bertsekas, “On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators,” *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.

[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[34] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[35] B. Blankertz, K.-R. Müller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlögl, G. Pfurtscheller, J. del R. Millan, M. Schröder, and N. Birbaumer, “The BCI competition III: Validating alternative approaches to actual BCI problems,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, 2006.

[36] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, “Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms,” *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993–1002, 2004.

[37] R. Oostenveld and P. Praamstra, “The five percent electrode system for high-resolution eeg and erp measurements,” *Clinical Neurophysiology*, vol. 112, no. 4, pp. 713–719, 2001.

[38] H. Wang, Q. Tang, and W. Zheng, “L1-norm-based common spatial patterns,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 653–662, 2012.

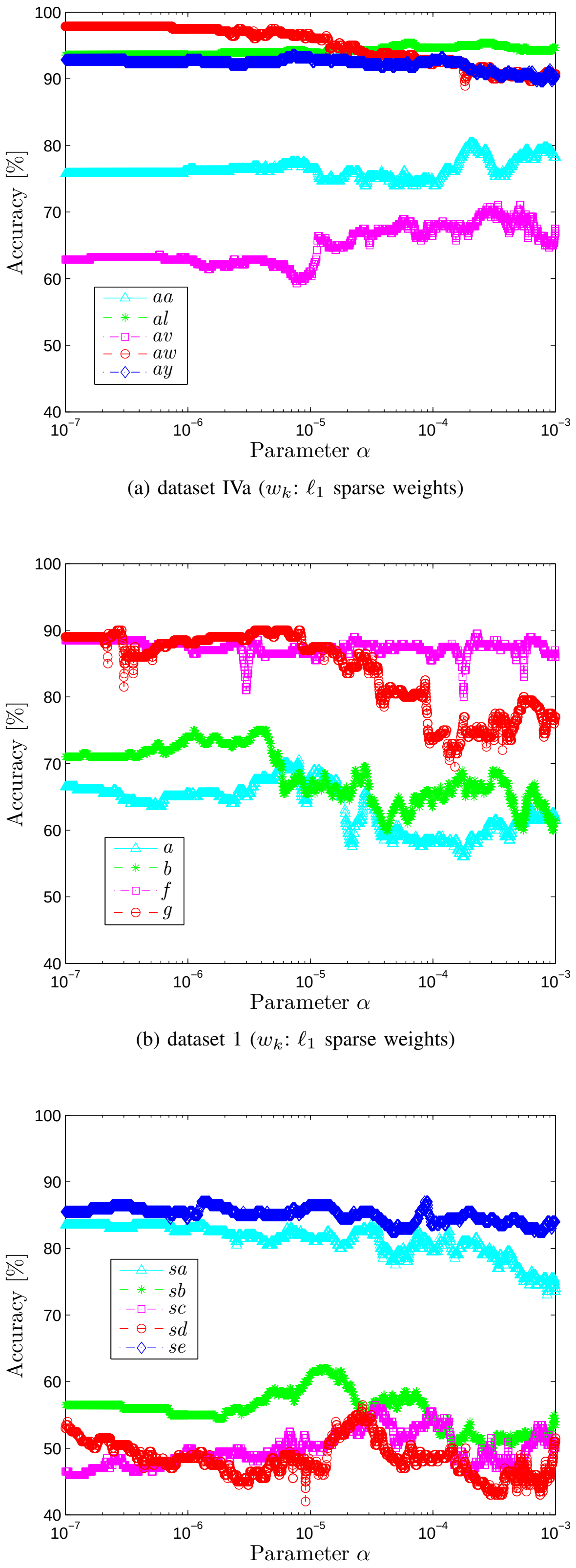
[39] X. Yong, R. K. Ward, and G. E. Birch, “Robust common spatial patterns for EEG signal preprocessing,” in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Bio. Soc. (EMBC 2008)*, 2008, pp. 2087–2090.

[40] R. Maronna, D. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*, ser. Wiley Series in Probability and Statistics. Wiley, 2006.

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

[42] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

[43] D. H. Dini and D. P. Mandic, “Exploiting sparsity in widely linear estimation,” in *Proc. of the Tenth International Symposium on Wireless Communication Systems (ISWCS 2013)*, 2013, pp. 1–5.



(a) dataset IVa (w_k: ℓ_1 sparse weights)

(b) dataset 1 (w_k: ℓ_1 sparse weights)

(c) JK-HH 1 (w_k: ℓ_1 sparse weights)

Fig. 2: Classification accuracy for varying parameter α for each subject in datasets IVa and JK-HH 1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2014.2358536, IEEE Transactions on Biomedical Engineering

Naoki Tomida (S'13) received the B.E. degree in electrical and electronic engineering from Tokyo University of Agriculture and Technology (TUAT), Japan, in 2013.

Since 2013, he has been a master course student in the Department of Communication and Computer Engineering at the Tokyo Institute of Technology, Japan. His research interests include signal processing, brain and biomedical signal processing, machine learning, matrix and tensor factorization, and convex optimization.

Toshihisa Tanaka received the B.E., the M.E., and the Ph.D. degrees from the Tokyo Institute of Technology in 1997, 2000, and 2002, respectively. From 2000 to 2002, he was a JSPS Research Fellow. From October 2002 to March 2004, he was a Research Scientist at RIKEN Brain Science Institute. In April 2004, he joined Department of Electrical and Electronic Engineering, the Tokyo University of Agriculture and Technology, where he is currently an Associate Professor. In 2005, he was a Royal Society Visiting Fellow at the Communications and Signal Processing Group, Imperial College London, U.K. From June 2011 to October 2011, he was a visiting faculty member in Department of Electrical Engineering, the University of Hawaii at Manoa.

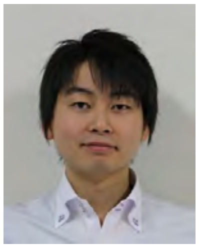
His research interests include image and signal processing, statistical signal processing and machine learning, brain and biomedical signal processing, and adaptive systems. He is a co-editor of Signal Processing Techniques for Knowledge Extraction and Information Fusion (with Mandic, Splinger), 2008.

He served as a guest editor of special issues in journals including Neurocomputing. He served as an associate editor of IEICE Transactions on Fundamentals. He was a chair of the Technical Committee on Biomedical Signal Processing, APSIPA. He is a senior member of IEEE, and a member of IEICE and APSIPA.

Shunsuke Ono (S'11) received the B.E. degree in computer science and the M.E. degree in communications and computer engineering from the Tokyo Institute of Technology, in 2010 and 2012, respectively, where he is currently pursuing the Ph.D. degree with the Department of Communications and Computer Engineering.

He is a Research Fellow with the Japan Society for the Promotion of Science. His current research interests are in signal and image processing, convex optimization, and inverse problems.

Mr. Ono received the Best Paper Award in 2014 and the Young Researchers' Award in 2013 from the Institute of Electronics, Information and Communication Engineers (IEICE).



Masao Yamagishi (M'12) received the B.E., M.E., and Ph.D. degrees from Tokyo Institute of Technology, Japan, in 2007, 2008, and 2012, respectively. From September to December 2012, he was a Visiting Researcher at the Technical University of Munich, Germany. He is currently an Assistant Professor in the Department of Communications and Computer Engineering, Tokyo Institute of Technology, Japan. His research interests include mathematical signal processing, adaptive filtering, convex optimization, and inverse problems.

From April 2009 to March 2012, he was a recipient of the Research Fellowship of the Japan Society for the Promotion of Science (JSPS). He received the Young Researcher Award from the Institute of Electrical, Information and Communication Engineers (IEICE) of Japan in 2010.

Hiroshi Higashi received the B.E., M.E., and Ph.D. degrees from Tokyo University of Agriculture and Technology (TUAT), Japan, in 2009, 2011, and 2013, respectively.

From 2011–2012, he was a Junior Research Associate with the Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, Japan. From 2012–2014, he was a Research Fellow of Japan Society for the Promotion of Science, Japan. He is currently an Assistant Professor in the Department of Computer Science and Engineering, Toyohashi University of Technology, Japan. He is a visiting researcher in Brain Science Institute, RIKEN, Japan. His research interests are in brain and biomedical signal processing.