A JOINT TENSOR DIAGONALIZATION APPROACH TO ACTIVE DATA SELECTION FOR EEG CLASSIFICATION

Naoki Tomida, Hiroshi Higashi, and Toshihisa Tanaka

Department of Electrical and Electronic Engineering
Tokyo University of Agriculture and Technology
2-24-16, Nakacho, Koganei-shi, Tokyo, 184-8588, Japan

ABSTRACT

We present a novel method based on joint tensor diagonalization for selecting or weighting electroencephalogram (EEG) data to estimate the covariance matrices to accurately find common spatial pattern (CSP). CSP and its variants need a pair of covariance matrices of two different tasks, which are obtained as the average over trials. This trial average can affect the accurate estimation of covariance matrices and cause the decrease of classification accuracy in brain machine interfaces (BMIs) due to the non-stationarity of EEG or experimental environments. We focus on the fact that finding CSP is equivalent to joint diagonalization of a pair of covariance matrices, and extend it to joint diagonalization of data tensor at each trial to determine importance of each trial. Numerical experiment of motor imagery (MI) classification supports the proposed algorithm is effective.

Index Terms— Brain machine interfaces, EEG signal processing, common spatial pattern, tensor algebra, joint diagonalization

1. INTRODUCTION

Brain machine interfacing (BMI) is a challenging application of signal processing, machine learning, and neuroscience [1]. BMIs capture brain activities associated to mental tasks and external stimuli, and realize non-muscular communication and control channel for conveying messages and commands to the external world [1, 2, 3]. Basically, noninvasive measurement devices such as electroencephalogram (EEG), magnetoencephalogram (MEG), and functional magnetic response imaging (fMRI) are widely used to observe the brain activities. Among them, because of its simplicity and low cost, EEG is a practical measurement device for use in engineering applications [4, 5].

Efficient decoding around motor-cortex is a crucial technique for realization of BMI associated with motor-imagery (MI-BMI) [6, 7], rehabilitation, and so forth. For instance, it is also known that the real and imaginary movements of hands and feet evoke the change of the so-called mu rhythm in different brain regions [2, 3]. Therefore, by accurately capturing these changes features 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 action such as hand, arm, or foot movement.

A well-known approach to extract the brain activity for MI-BMI is the so-called common spatial pattern (CSP) [1, 8, 9]. CSP is a set of spatial weight coefficients corresponding to each electrode in a multichannel EEG. These coefficients are determined from training data in such a way that the variances of the signal extracted by the spatial weights differ between two tasks (e.g. left and right hand movement imageries) as much as possible. These weights can also regarded as a spatial filter which projects observed EEG signals onto the optimal space used to classify the observed data to a class corresponding to subject's cerebral status. Several variants of CSP has been proposed such as common spatio-spectral pattern (CSSP) [10], spectrally weighted CSP (SPEC-CSP) [11], iterative spatio-spectral patterns learning (ISSPL) [12], filter bank CSP (FBCSP) [13] and discriminative filter bank CSP (DFBCSP) [14, 15, 16].

To obtain these spatial patterns, it is necessary to estimate the covariance matrices of observed signals. To increase the accuracy of the estimation, usually, EEG signals (training data) are observed several times (called trials) for the same task, and the covariance matrices of all trials are simply averaged. However, equally 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 user's concentration. Second, the observed EEG is contaminated by non-stationary artifacts such as eye and muscle movement. Therefore the covariance matrices at different trials can differ from each other, although the EEG for the same task should be (wide-sense) stationary process. Heavily contaminated EEG data, that we call "low-quality trials" in the rest of this paper, should be removed from the training data set to design the spatial weights.

In this paper, we propose a new method for detecting and weighting the low-quality trials by using joint tensor diagonalization [17, 18, 19]. We focus on the fact that the standard CSP is given as a generalized eigenvector of a pair of covariance matrices, and indeed this idea leads to exact joint matrix diagonalization. To solve the above problem, we extend this idea to approximate joint diagonalization of data tensors at trials, where we expect that the off-diagonal residue with respect to a low-quality trial is large. For such a trial, a smaller weight is assigned in the weighted average of covariance matrices over all trials. We formulate this idea as a regularized joint tensor diagonalization problem and derive an iterative algorithm to find the solution.

1.1. Notations and Tensor Algebra

The following terminology, notation, and mathematical operations for tensors (multi-way data) are used throughout the paper [20].

A *tensor* is defined as a quantity with multiple, usually more than three indexes. A tensor is denoted by a calligraphic capital letter e.g., $\mathcal{A}$ and the $(i_1, i_2, \ldots, i_N)$-th element of tensor $\mathcal{A}$ is denoted by $a_{i_1,i_2,\ldots,i_N} (1 \leq i_1 \leq I_1,\ 1 \leq i_2 \leq I_2,\ \ldots,\ 1 \leq i_N \leq I_N)$, where i_n $(n = 1, \ldots, N)$ is the index of the nth *mode*.

This work is supported in part by JSPS KAKENHI Grant Number 24360146. Higashi and Tanaka are also affiliated with RIKEN Brain Science Institute, Saitama, Japan.

An I_j-dimensional vector obtained by fixing all modes except the jth mode is called the *fiber* of the jth mode. The *unfolding matrix* of the jth mode is the matrix of size I_j by $(\prod_{k\neq j} I_k)$ denoted by $\boldsymbol{A}_{(j)}$ laying all possible fibers of the jth mode. *Tensor-matrix multiplication* is introduced as folllows. Suppose $\boldsymbol{a}$ be a fiber vector of $\mathcal{A}$. Given $\boldsymbol{B}$ of size m by I_j, the jth multiplication $\mathcal{A} \times_j \boldsymbol{B}$ is defined as the replacement of all possible fibers $\boldsymbol{a}$ by $\boldsymbol{Ba}$. Note that the dimension of the jth mode of $\mathcal{A} \times_j \boldsymbol{B}$ is m.

The *Frobenius norm* of a tensor $\mathcal{A}$ is defined as $\|\mathcal{A}\|_F^2 = \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_N=1}^{I_N}|a_{i_1,i_2,\dots,i_N}|^2$. The *Matrix Hadamard, Kronecker, and Khatri-Rao products* [20] are denoted by $\circledast$, $\otimes$, and $\odot$, respectively. Note that given matrices $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_N] \in \mathbb{R}^{L\times N}$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_N] \in \mathbb{R}^{M\times N}$, $\boldsymbol{A}\odot\boldsymbol{B} = [\boldsymbol{a}_1\otimes\boldsymbol{b}_1, \boldsymbol{a}_2\otimes\boldsymbol{b}_2, \dots, \boldsymbol{a}_N\otimes\boldsymbol{b}_N] \in \mathbb{R}^{LM\times N}$.

The *all-ones vector* of size M is defined as $\mathbf{1}_M \stackrel{\text{def}}{=} [1,\dots,1]^\top \in \mathbb{R}^M$.

2. COMMON SPATIAL PATTERN — REVIEW

Let $\boldsymbol{X}^{(k)} \in \mathbb{R}^{M\times N}$ be a matrix consisting of M channel signals with N samples at kth trial. CSP is given as a spatial weight vector, $\boldsymbol{v} \in \mathbb{R}^M$, minimizing the in-class variance of a signal extracted by linear combination of $\boldsymbol{X}^{(k)}$ [8, 9]. In general, each channel signal in $\boldsymbol{X}^{(k)}$ is band-limited by a bandpass filter which passes the frequency components related to the target brain activity. Denote the components of $\boldsymbol{X}^{(k)}$ by $\boldsymbol{X}^{(k)} = [\boldsymbol{x}_1^{(k)},\dots,\boldsymbol{x}_N^{(k)}]$, where $\boldsymbol{x}_n^{(k)} \in \mathbb{R}^M$ and n is the time index ($n = 1,\dots,N$). The time mean of the observed signal is given by $\boldsymbol{\mu}^{(k)} = (1/N)\sum_{n=1}^N \boldsymbol{x}_n^{(k)}$. Then, the time variance of the extracted signal of $\boldsymbol{X}^{(k)}$ is given by

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

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 $\mathfrak{C}_1 \cap \mathfrak{C}_2 = \emptyset$. CSP of class c ($c = 1, 2$) is given as the weight vector $\boldsymbol{v}_c$ that is the solution of the following optimization problem [8, 9];

$$\min_{\boldsymbol{v}} \ \frac{1}{K_c}\sum_{k\in\mathfrak{C}_c}\sigma^2(\boldsymbol{X}^{(k)},\boldsymbol{v}), \quad \text{subject to} \ \sum_{d=1,2}\frac{1}{K_d}\sum_{k\in\mathfrak{C}_d}\sigma^2(\boldsymbol{X}^{(k)},\boldsymbol{v}) = 1, \quad (2)$$

where K_d is the number of elements in class d. In terms of covariance matrices, (2) can be rewritten as

$$\min_{\boldsymbol{v}} \quad \boldsymbol{v}^\top\boldsymbol{S}_c\boldsymbol{v}, \quad \text{subject to} \quad \boldsymbol{v}^\top(\boldsymbol{S}_1+\boldsymbol{S}_2)\boldsymbol{v} = 1, \quad (3)$$

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

$$\boldsymbol{S}_d = \frac{1}{K_d}\sum_{k\in\mathfrak{C}_d}\boldsymbol{S}^{(k)}, \quad (4)$$

and $\boldsymbol{S}^{(k)} \in \mathbb{R}^{M\times M}$ is the sample covariance matrix at trial k given as

$$\boldsymbol{S}^{(k)} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n^{(k)} - \boldsymbol{\mu}^{(k)})(\boldsymbol{x}_n^{(k)} - \boldsymbol{\mu}^{(k)})^\top. \quad (5)$$

Note that the solution of (3) is given by the generalized eigenvector corresponding to the smallest generalized eigenvalue of the generalized eigenvalue problem described as

$$\boldsymbol{S}_c\boldsymbol{v} = \lambda(\boldsymbol{S}_1+\boldsymbol{S}_2)\boldsymbol{v}. \quad (6)$$

3. SAMPLE WEIGHTING BY JOINT TENSOR DIAGONALIZATION

Ideally, $\boldsymbol{S}^{(k)}$ is trial- or sample-invariant up to noise. This motivates the simple arithmetic averaging given as in (4). However, as mentioned before, 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 propose to soften (4) and consider the weighted average defined as

$$\hat{\boldsymbol{S}}_d = \sum_{k\in\mathfrak{C}_d} w_k\boldsymbol{S}^{(k)}, \quad \text{subject to} \quad \sum_{k\in\mathfrak{C}_d} w_k = 1, \quad (7)$$

where w_k is the weight coefficient at trial k. Note that, in the CSP, $w_k = 1/K_d$ in the above equation. The main problem of the proposed method is to find w_k that can remove the low-quality trials. To this end, we focus on the fact that solving (6) is equivalent to joint diagonalization for $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. If the observed EEG is trial-invariant, $\boldsymbol{S}^{(k)}$ at all trials should be exactly jointly diagonalized. However, as mentioned, this assumption is not true. So, we try to find trials preventing the covariance matrix from being jointly diagonalized and set the corresponding weight small. To formulate this idea, we exploit the tensor notation for (6). Define $\boldsymbol{\mathcal{S}}_t \in \mathbb{R}^{M\times M\times 2}$ as a tensor such that $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are the frontal slices. Define $\boldsymbol{V} = [\boldsymbol{v}_1,\dots,\boldsymbol{v}_M] \in \mathbb{R}^{M\times M}$ and $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2]^\top \in \mathbb{R}^{2\times M}$, where $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are vectors consisting of eigenvalues of $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ respectively. Then, (6) can be rewritten as a tensor form:

$$\boldsymbol{\mathcal{S}}_t = \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{V} \times_2 \boldsymbol{V} \times_3 \boldsymbol{\Lambda}, \quad (8)$$

where $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{M\times M\times M}$ is the cubic tensor with ones along the super-diagonal. Note that this is a tensor notation for joint diagonalization for $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$.

Next, we extend $\boldsymbol{\mathcal{S}}_t$ to a tensor including the covariance matrices of all trials defined as $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{M\times M\times(K_1+K_2)}$, which has frontal slices $\boldsymbol{S}^{(k)}$, $k = 1,\dots,K_1+K_2$. We can consider the following decomposition:

$$\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} + \boldsymbol{\mathcal{E}}, \quad \boldsymbol{A} = \boldsymbol{B}, \quad (9)$$

where $\boldsymbol{A}$ and $\boldsymbol{B} \in \mathbb{R}^{M\times M}$ are arbitrary common factor matrices , $\boldsymbol{C} \in \mathbb{R}^{(K_1+K_2)\times M}$ is an arbitrary matrix and $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{M\times M\times(K_1+K_2)}$ is the residue that has frontal slices $\boldsymbol{E}^{(k)}$, $k = 1,\dots,K_1+K_2$. This can be regarded as a tensor diagonalization of $\boldsymbol{\mathcal{S}}$. By adopting (9), we determine the quality at every trial from information about the error tensor $\boldsymbol{\mathcal{E}}$. We regard the kth trial data, where $\|\boldsymbol{E}^{(k)}\|_F^2$ is greater, as low-quality trials and impose a smaller weight on it. We define the weight vector consisting of weights for all trials as

$$\boldsymbol{w} \stackrel{\text{def}}{=} [w_1,\dots,w_{K_1}, w_{K_1+1},\dots,w_{K_1+K_2}]^\top \in \mathbb{R}^{K_1+K_2}. \quad (10)$$

Then, let us consider the following cost function to find the trial weights;

$$J_1 = \frac{1}{2M^2(K_1+K_2)}\|\boldsymbol{\mathcal{W}} \circledast (\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C})\|_F^2, \quad (11)$$

where $\boldsymbol{\mathcal{W}}$ is a tensor representation of $\boldsymbol{w}$ defined as

$$\boldsymbol{\mathcal{W}} \stackrel{\text{def}}{=} \mathbf{1}_M \circ \mathbf{1}_M \circ \boldsymbol{w} \in \mathbb{R}^{M\times M\times(K_1+K_2)}, \quad (12)$$

where $\circ$ denotes the outer product. Assuming that $\hat{\boldsymbol{S}}_d$ in (7) is close to $\boldsymbol{S}_d$ in (4), we introduce a data-fidelity term;

$$J_2 = \frac{1}{2M^2}\left\|\sum_{k=1}^{K_1+K_2}\left(\frac{1}{K_k} - w_k\right)\boldsymbol{S}^{(k)}\right\|_F^2, K_k = \begin{cases} K_1 \ (1 \le k \le K_1) \\ K_2 \ (K_1+1 \le k \le K_1+K_2) \end{cases}. \quad (13)$$

Finally, the regularized optimization problem is given as:

$$\min_{\boldsymbol{w},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}} \quad J(\boldsymbol{w},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) = J_1 + \alpha J_2,$$
$$\text{subject to} \quad \sum_{k=1}^{K_1} w_k = \sum_{k=K_1+1}^{K_1+K_2} w_k = 1, \tag{14}$$

where α is a regularization parameter.

This optimization problem is solved by the proposed steps.

Step 0. Initialize $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$.

Step 1. For $\boldsymbol{A}$ and $\boldsymbol{B}$ fixed, $\boldsymbol{C} = \boldsymbol{S}_{(3)}\{(\boldsymbol{B}\odot\boldsymbol{A})^\top\}^\dagger$.

Step 2. For $\boldsymbol{A}$ and $\boldsymbol{C}$ fixed, $\boldsymbol{B} = \boldsymbol{S}_{(3)}\{(\boldsymbol{C}\odot\boldsymbol{A})^\top\}^\dagger$.

Step 3. Set $\boldsymbol{A} = \boldsymbol{B}$.

In these steps, $\cdot^\dagger$ denotes the pseudo-inverse matrix and recall that $\boldsymbol{S}_{(3)}$ is the unfolding matrix of the 3rd mode of $\boldsymbol{\mathcal{S}}$. We iterate through Step 1 to Step 3 until $\|\mathcal{E}\|_F^2$ in (9) becomes smaller than a certain value or the fixed maximum number of iteration is reached. After convergence, we find $\boldsymbol{w}$. Since cost function J is a quadratic form with respect to $\boldsymbol{w}$, the optimal $\boldsymbol{w}^*$ can be directly found as

$$\boldsymbol{w}^* = \left(\frac{1}{M^2(K_1+K_2)}\boldsymbol{D}_e + \frac{\alpha}{M^2}\boldsymbol{G}\right)^{-1}\left(\frac{\alpha}{M^2}\boldsymbol{G}\boldsymbol{Z}\boldsymbol{k} + \boldsymbol{\zeta}\right). \tag{15}$$

See Appendix A for the derivation of Step 1, Step 2, and (15) and definitions for $\boldsymbol{D}_e$, $\boldsymbol{G}$, $\boldsymbol{Z}$, $\boldsymbol{k}$, and $\boldsymbol{\zeta}$.

4. EXPERIMENT OF TWO EEG CLASSIFICATION

We conducted the experiment of classification of EEG signal during motor imagery to show the performance of the proposed method. When the proposed method was used for feature extraction of EEG signals, the weight for trials determined by the proposed method was used for calculating covariance matrices for the CSP spatial weights. For comparison, the result of the standard CSP that uses equally averaged covariance matrices is also shown.

As the output of feature extraction using CSP, we defined the following feature vector. Although the solution of (3) is given by the eigenvector corresponding to the smallest eigenvalue in (6), we can use the other eigenvectors for classification [21]. The M eigenvectors can be obtained by solving (6) 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 (6). We used the $2r$ eigenvectors to form the feature vector, denoted by $\boldsymbol{y}$, for classification of unlabeled data, $\boldsymbol{X}$:

$$\boldsymbol{y} = [\sigma^2(\boldsymbol{X},\hat{\boldsymbol{v}}_1),\ldots,\sigma^2(\boldsymbol{X},\hat{\boldsymbol{v}}_r),\sigma^2(\boldsymbol{X},\hat{\boldsymbol{v}}_{M-r+1}),\ldots,\sigma^2(\boldsymbol{X},\hat{\boldsymbol{v}}_M)]^\top. \tag{16}$$

For classification, $\boldsymbol{y}$ is input to a classifier, linear discriminant analysis (LDA) [22].

4.1. Data Description

We used dataset IVa from BCI competition III [23], which was provided by Fraunhofer FIRST (Intelligent Data Analysis Group) and Campus Benjamin Franklin of the Charité - University Medicine Berlin (Department of Neurology, Neurophysics Group) [24]. This dataset consists of EEG signals during right hand and right foot motor-imageries. The EEG signals were recorded from five subjects labeled *aa*, *al*, *av*, *aw*, and *ay*. 118 EEG channels were measured at positions of the extended international 10/20-system. The measured signal was bandpass-filtered with the passband of 0.05–200 Hz, and then digitized at 1000 Hz with 16 bits (0.1 μV). During each 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.

Table 1. Classification accuracy [%] given by 5×5 cross validation. The figure with ± represents the standard deviation (S.D.). In the proposed method, we show the highest classification accuracy at each subject among the accuracies obtained with several α.

Subjects	*aa*	*al*	*av*	*aw*	*ay*	Ave.
Simple Ave.	75.71	93.57	63.21	97.86	92.86	84.64
(S.D.)	±12.7	±2.99	±5.14	±1.96	±3.78	
Weighted Ave.	80.36	95.36	71.07	97.86	93.57	87.64
(S.D.)	±14.7	±3.70	±4.95	±1.96	±3.48	

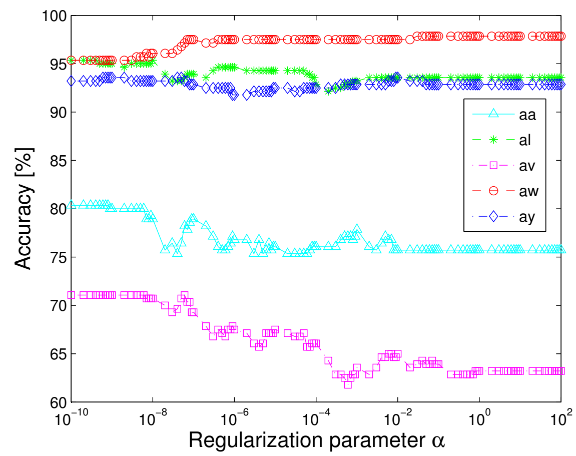
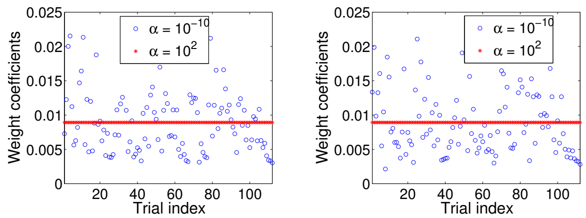


Fig. 1. Classication accuracy for varying regularization parameter, α, in subject *aa, al, av, aw*, and *ay*.

In this experiment, we furthermore applied to this data a bandpass filter whose passband is 7-30 Hz and downsampled to 100 Hz. The dataset for each subject consisted of signals of 140 trials per class. The signal in each trial is extracted from the period of 3.5 seconds after a visual cue.

4.2. Result

In Table 1, we show classification accuracy in CSP with proposed method and the standard CSP. In both cases, for the sake of simplicity of comparison, the number of the associated spatial weights, r, in (16) is fixed to 3. The results were obtained by conducting 5×5 cross validation. It should be noted that for *av*, the proposed method increases the classification accuracy with the standard CSP by 7.8%. Figure 1 shows the classification accuracies at each subject to varying α from 10^{-10} to 10^2 order. It is observed that in the parameter's range of about 10^{-10}-10^{-6}, the classification accuracy tends to be high. We can also observe that around 10^2 the proposed method is almost identical to CSP. Figure 2 shows an example of the weight coefficients for each trial in subject *av*. We can observe that the weight, w_k, approach $1/K_k$ as α becomes greater.



(a) Class 1 (b) Class 2

Fig. 2. Weight coefficients of subject *av*

It should be emphasized that the proposed method can be applied to the variants of CSP. Therefore, we applied the proposed method to the standard CSP in this experiment for the sake of simplicity.

5. CONCLUSION

In this paper, utilizing joint tensor diagonalization, we proposed the novel method that selects or weights trials to accurately find CSP. The experimental results have suggested that the proposed averaging method with CSP is effective in classification of the MI-BMI.

A. DERIVATION OF THE UPDATE EQUATIONS

The Lagrangian $\hat{J}$ of (14) is given as

$$\hat{J} = J_1 + \alpha J_2 + \beta\left(1 - \sum_{k=1}^{K_1} w_k\right) + \gamma\left(1 - \sum_{k=K_1+1}^{K_1+K_2} w_k\right), \quad (17)$$

where β and γ are Lagrange multipliers. The update equation in Step 1 is derived as follows. The first term of (17) is modified as

$$\begin{aligned} J_1 =& \frac{1}{2M^2(K_1+K_2)} \|\boldsymbol{\mathcal{W}} \circledast (\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C})\|_F^2 \\ =& \frac{1}{2M^2(K_1+K_2)} \mathrm{tr}\left[\{\boldsymbol{W}_{(3)} \circledast \boldsymbol{S}_{(3)} - \boldsymbol{W}_{(3)} \circledast \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top\}\right. \\ & \left.\times \{\boldsymbol{W}_{(3)} \circledast \boldsymbol{S}_{(3)} - \boldsymbol{W}_{(3)} \circledast \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top\}^\top\right] \\ =& \frac{1}{2M^2(K_1+K_2)} \mathrm{tr}\left[(\boldsymbol{W}_{(3)} \circledast \boldsymbol{S}_{(3)})(\boldsymbol{W}_{(3)} \circledast \boldsymbol{S}_{(3)})^\top\right. \\ & -2(\boldsymbol{W}_{(3)} \circledast \boldsymbol{S}_{(3)})\left\{\boldsymbol{W}_{(3)}^\top \circledast (\boldsymbol{B} \odot \boldsymbol{A})\boldsymbol{C}^\top\right\} \\ & \left.+ \{\boldsymbol{W}_{(3)} \circledast \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top\}\left\{\boldsymbol{W}_{(3)}^\top \circledast (\boldsymbol{B} \odot \boldsymbol{A})\boldsymbol{C}^\top\right\}\right], \end{aligned} \quad (18)$$

where tr[·] denotes the trace of a matrix. By using the above equation, the differentiation of $\hat{J}$ with respect to $\boldsymbol{C}$ is given as

$$\begin{aligned} \delta\hat{J} =& \frac{1}{2M^2(K_1+K_2)} \mathrm{tr}\left[-2(\boldsymbol{W}_{(3)} \circledast \boldsymbol{S}_{(3)})\left\{\boldsymbol{W}_{(3)}^\top \circledast (\boldsymbol{B} \odot \boldsymbol{A})\delta\boldsymbol{C}^\top\right\}\right. \\ & + \{\boldsymbol{W}_{(3)} \circledast \delta\boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top\}\left\{\boldsymbol{W}_{(3)}^\top \circledast (\boldsymbol{B} \odot \boldsymbol{A})\boldsymbol{C}^\top\right\} \\ & \left.+ \{\boldsymbol{W}_{(3)} \circledast \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top\}\left\{\boldsymbol{W}_{(3)}^\top \circledast (\boldsymbol{B} \odot \boldsymbol{A})\delta\boldsymbol{C}^\top\right\}\right] \\ =& \frac{1}{2M^2(K_1+K_2)} \mathrm{tr}\left[-2(\boldsymbol{W}_{(3)} \circledast \boldsymbol{S}_{(3)})\left\{\boldsymbol{W}_{(3)}^\top \circledast (\boldsymbol{B} \odot \boldsymbol{A})\delta\boldsymbol{C}^\top\right\}\right. \\ & \left.+2\{\boldsymbol{W}_{(3)} \circledast \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top\}\left\{\boldsymbol{W}_{(3)}^\top \circledast (\boldsymbol{B} \odot \boldsymbol{A})\delta\boldsymbol{C}^\top\right\}\right] \\ =& \frac{1}{M^2(K_1+K_2)} \mathrm{tr}\left[\{\boldsymbol{W}_{(3)} \circledast \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top - \boldsymbol{W}_{(3)} \circledast \boldsymbol{S}_{(3)}\}\right. \\ & \left.\times \left\{\boldsymbol{W}_{(3)}^\top \circledast (\boldsymbol{B} \odot \boldsymbol{A})\delta\boldsymbol{C}^\top\right\}\right]. \end{aligned} \quad (19)$$

Thus, setting the above formula equal to zero, we obtain a following equation:

$$\boldsymbol{W}_{(3)} \circledast \{\boldsymbol{S}_{(3)} - \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top\} = \boldsymbol{0}. \quad (20)$$

From the above equation, we obtain $\boldsymbol{C}$ in Step 1. Indeed, this derivation is similar to the individual differences in scaling (INDSCAL) [18, 19]. The update equation in Step 2 is proved as well as Step 1.

The equation (15) is proved as follows. Again, (17) becomes

$$\begin{aligned} \hat{J} =& \frac{1}{2M^2(K_1+K_2)} \|\boldsymbol{\mathcal{W}} \circledast \boldsymbol{\mathcal{E}}\|_F^2 + \frac{\alpha}{2M^2}\left\|\sum_{k=1}^{K_1+K_2}\left(\frac{1}{K_k} - w_k\right)\boldsymbol{S}^{(k)}\right\|_F^2 \\ & + \beta\left(1 - \sum_{k=1}^{K_1} w_k\right) + \gamma\left(1 - \sum_{k=K_1+1}^{K_1+K_2} w_k\right) \\ =& \frac{1}{2M^2(K_1+K_2)} \sum_{k=1}^{K_1+K_2} w_k^2 \|\boldsymbol{E}^{(k)}\|_F^2 \\ & + \frac{\alpha}{2M^2} \sum_{k=1}^{K_1+K_2} \sum_{l=1}^{K_1+K_2} \left(\frac{1}{K_k} - w_k\right)\left(\frac{1}{K_l} - w_l\right) \mathrm{tr}\left[\boldsymbol{S}^{(k)}\boldsymbol{S}^{(l)\top}\right] \\ & + \beta\left(1 - \sum_{k=1}^{K_1} w_k\right) + \gamma\left(1 - \sum_{k=K_1+1}^{K_1+K_2} w_k\right). \end{aligned} \quad (21)$$

Define

$$\boldsymbol{D}_e = \mathrm{diag}\left[\left\|\boldsymbol{E}^{(1)}\right\|_F^2, \ldots, \left\|\boldsymbol{E}^{(K_1+K_2)}\right\|_F^2\right] \in \mathbb{R}^{(K_1+K_2)\times(K_1+K_2)},$$

$$\boldsymbol{G} = \begin{bmatrix} \mathrm{tr}\left[\boldsymbol{S}^{(1)}\boldsymbol{S}^{(1)\top}\right] & \cdots & \mathrm{tr}\left[\boldsymbol{S}^{(1)}\boldsymbol{S}^{(K_1+K_2)\top}\right] \\ \vdots & \ddots & \vdots \\ \mathrm{tr}\left[\boldsymbol{S}^{(K_1+K_2)}\boldsymbol{S}^{(1)\top}\right] & \cdots & \mathrm{tr}\left[\boldsymbol{S}^{(K_1+K_2)}\boldsymbol{S}^{(K_1+K_2)\top}\right] \end{bmatrix} \in \mathbb{R}^{(K_1+K_2)\times(K_1+K_2)},$$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{1}_{K_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1}_{K_2} \end{bmatrix} \in \mathbb{R}^{(K_1+K_2)\times 2},\ \boldsymbol{k} = \begin{bmatrix} 1/K_1 \\ 1/K_2 \end{bmatrix} \in \mathbb{R}^2, \text{ and } \boldsymbol{\zeta} = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \in \mathbb{R}^2.$$

Considering the properties of the trace and a covariance matrix, we can easily show that $\boldsymbol{G}$ is a real symmetric matrix. With $\boldsymbol{D}_e$, $\boldsymbol{G}$, $\boldsymbol{k}$, and $\boldsymbol{\zeta}$, (21) can be rewritten as follows:

$$\begin{aligned} \hat{J} =& \frac{1}{2M^2(K_1+K_2)} \boldsymbol{w}^\top \boldsymbol{D}_e \boldsymbol{w} + \frac{\alpha}{2M^2}(\boldsymbol{w} - \boldsymbol{Z}\boldsymbol{k})^\top \boldsymbol{G}(\boldsymbol{w} - \boldsymbol{Z}\boldsymbol{k}) \\ & - \boldsymbol{w}^\top \boldsymbol{Z}\boldsymbol{\zeta} + \beta + \gamma. \end{aligned} \quad (22)$$

The gradient of $\hat{J}$, is obtained as

$$\frac{\partial \hat{J}}{\partial \boldsymbol{w}} = \frac{1}{M^2(K_1+K_2)} \boldsymbol{D}_e \boldsymbol{w} + \frac{\alpha}{M^2}\boldsymbol{G}(\boldsymbol{w} - \boldsymbol{Z}\boldsymbol{k}) - \boldsymbol{Z}\boldsymbol{\zeta}. \quad (23)$$

Setting the above formula equal to zero, we obtain (15). The constraints given in (14) is rewritten as $\boldsymbol{Z}^\top \boldsymbol{w} = \boldsymbol{1}_2$, which yields with (15), Lagrange multipliers β and γ are found as follows:

$$\boldsymbol{1}_2 = \boldsymbol{Z}^\top \boldsymbol{w} = \boldsymbol{Z}^\top(\boldsymbol{H}\boldsymbol{Z}\boldsymbol{K}\boldsymbol{1}_2 + \boldsymbol{F}\boldsymbol{Z}\boldsymbol{\zeta}). \quad (24)$$

It follows that

$$\boldsymbol{\zeta} = (\boldsymbol{Z}^\top \boldsymbol{F}\boldsymbol{Z})^{-1}(\boldsymbol{I} - \boldsymbol{Z}^\top \boldsymbol{H}\boldsymbol{Z}\boldsymbol{K})\boldsymbol{1}_2, \quad (25)$$

where $\boldsymbol{F}$, $\boldsymbol{H}$, and $\boldsymbol{K}$ are defined as

$$\boldsymbol{F} = \left(\frac{1}{M^2(K_1+K_2)}\boldsymbol{D}_e + \frac{\alpha}{M^2}\boldsymbol{G}\right)^{-1} \in \mathbb{R}^{(K_1+K_2)\times(K_1+K_2)},$$

$$\boldsymbol{H} = \boldsymbol{F}\boldsymbol{G} \in \mathbb{R}^{(K_1+K_2)\times(K_1+K_2)}, \text{ and } \boldsymbol{K} = \frac{\alpha}{M^2}\begin{bmatrix} 1/K_1 & 0 \\ 0 & 1/K_2 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

6. 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 and J. Chambers, *EEG Signal Processing*. Chichester, UK: Wiley, 2007.

[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] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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," Dept. Math. Eng., Univ. Tokyo, Tokyo, Japan, Tech. Rep., 2006.

[12] 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.

[13] 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, pp. 2390–2397.

[14] H. Higashi and T. Tanaka, "Optimal design of a bank of spatio-temporal filters for EEG signal classification," in *Proc. 33rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC 2011)*, 2011, pp. 6100–6103.

[15] ——, "Classification by weighting for spatio-frequency components of EEG signal during motor imagery," in *Proc. 2011 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2011)*, 2011, pp. 585–588.

[16] ——, "Simultaneous design of FIR filter banks and spatial patterns for EEG signal classification," *IEEE Trans. Biomed. Eng.*, 2013 (to appear).

[17] D. P. Mandic, M. Golz, A. Kuh, D. Obradovic, and T. Tanaka, *Signal Processing Techniques for Knowledge Extraction and Information Fusion*. New York, NY: Springer, 2008.

[18] T. G. Kolda and B. W. Bader, "Tensor decompositons and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[19] T. M. Selee, T. G. Kolda, W. P. Kegelmeyer, and J. D. Griffin, "Extracting clusters from large datasets with multiple similarity measures using IMSCAND," in *CSRI Summer Proc. 2007*, 2007, pp. 87–103.

[20] A. Cichocki, R. Zduneck, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Chichester, UK: Wiley, 2009.

[21] 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.

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

[23] B. Blankertz, K.-R. Müller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. SchlUgl, 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.

[24] 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.