Adaptation of motor imagery EEG classification model based on tensor decomposition

2014 J. Neural Eng. 11 056020

(http://iopscience.iop.org/1741-2552/11/5/056020)

IOP Publishing Journal of Neural Engineering
J. Neural Eng. 11 (2014) 056020 (14pp) doi:10.1088/1741-2560/11/5/056020

Adaptation of motor imagery EEG classification model based on tensor decomposition

Xinyang Li[1,2,4], Cuntai Guan[2], Haihong Zhang[2], Kai Keng Ang[2] and Sim Heng Ong[3]

[1] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 119613, Singapore
[2] Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632, Singapore
[3] Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119613, Singapore

E-mail: a0068297@nus.edu.sg, ctguan@i2r.a-star.edu.sg, hhzhang@i2r.a-star.edu.sg, kkang@i2r.a-star.edu.sg and eleongsh@nus.edu.sg

Received 23 April 2014, revised 17 July 2014
Accepted for publication 25 July 2014
Published 22 September 2014

Abstract

Objective. Session-to-session nonstationarity is inherent in brain–computer interfaces based on electroencephalography. The objective of this paper is to quantify the mismatch between the training model and test data caused by nonstationarity and to adapt the model towards minimizing the mismatch. *Approach*. We employ a tensor model to estimate the mismatch in a semi-supervised manner, and the estimate is regularized in the discriminative objective function. *Main results*. The performance of the proposed adaptation method was evaluated on a dataset recorded from 16 subjects performing motor imagery tasks on different days. The classification results validated the advantage of the proposed method in comparison with other regularization-based or spatial filter adaptation approaches. Experimental results also showed that there is a significant correlation between the quantified mismatch and the classification accuracy. *Significance*. The proposed method approached the nonstationarity issue from the perspective of data-model mismatch, which is more direct than data variation measurement. The results also demonstrated that the proposed method is effective in enhancing the performance of the feature extraction model.

Keywords: electroencephalograph, motor imagery, brain–computer interface

(Some figures may appear in colour only in the online journal)

1. Introduction

The nonstationarity of brain activities is an established phenomenon that has various implications in neuroscience and neuroengineering [1]. Recent neuro-imaging studies have shown that the nonstationarity may be caused by low frequency spontaneous fluctuations in brain signals that are coherent within resting state networks (RSNs) [2, 3]. As reported in [2], intrinsic brain activity of RSNs persists during task performance and contributes to variability in evoked brain responses. Other contributory factors include electrode impedance and positioning, and the subjects' different response behaviours.

The characteristics of the brain signal of a specific mental task may vary largely from trial to trial, and from session to session [4, 5]. This poses a grand challenge especially to online brain signal detection and classification such as in BCIs [6–9]; even if a BCI performs well in calibration, it may suffer from a considerable performance drop over time. For example in [10], some BCI subjects who achieved

[4] Authors to whom any correspondence should be addressed.

classification accuracies around 85% in the calibration session obtained only accuracies around 65% in later sessions. Therefore, nonstationarity and its consequences must be well addressed before BCI can be applied in real-world applications out of the laboratory [11–13].

Among a number of algorithms that have been proposed to address the nonstationary issue, one category considers improving the robustness of the model using calibration data only, such that this may translate to better generalization in processing unseen test data [14–16]. In [17] and [18], for example, cross-subject data are integrated to build subject-specific brain signal classification models with a widely-used subject-specific spatial filtering technique, i.e., common spatial pattern (CSP). In [19], CSP is modified by penalizing the nonstationary projection directions so as to minimize the effects of artefacts. In [20], the authors introduce a different penalizing term that measures the Kullback–Leibler (KL) divergence of electroencephalograph (EEG) across trials, and the learning algorithm aims to minimize within-class dissimilarities while maximizing inter-class separation.

Another category of methods investigates the actual variations across sessions and then adapts detection models accordingly. While brain signal detection algorithms usually consist of a feature extraction step and a classification step, some methods focus on the classification step and study the shift of discriminative patterns in the fixed feature space [11, 21, 22]. Studies in [11] show that the two-class motor imagery EEG classification accuracy could increase significantly among more than 90% of the subjects by using simple adaptive procedures, such as bias adaptation. Other algorithms consider variations of EEG data across sessions by incorporating data from test sessions to update the feature extraction model [23, 24]. Another adaptive approach, instead of updating the model, maps the EEG data from the evaluation session space to the training session space by a linear transformation so that differences between sessions can be reduced and the training model can work better on the transformed test data [25].

A critical question pertaining to nonstationary brain signal detection is how to construct a metric that measures this mismatch between test data and the model obtained from training data, and how to make use of the mismatch metric to guide the adaptation of feature extraction and classification algorithms. In this regard, there are two major challenges. One challenge lies in the fact that brain signal data have complex and multi-way structures. In a typical BCI system, the data structures are usually indexed by subject, trial, sensor channel, time, or frequency bin. The multi-way structures, however, may not be sufficiently explored if much simplified representations such as mean covariance matrices are used, e.g., in CSP. The other challenge is how to integrate the mismatch metric into a discriminative framework so that the former is directly relevant to the discrimination of brain signals.

This report presents a systematic attempt to quantify the data-model mismatch and use the mismatch metric as a basis for the model adaptation. We employ a tensor model to capture the multidimensional structure of EEG, because it provides a natural and convenient representation of multi-way data [26, 27]. In particular, the tensor structure is applied to the covariance matrices of EEG data so that the event-related (de) synchronization (ERD/ERS) effects of multi-trial data as well as the projection matrix can be formulated in a unified model [28]. Interpreted from a regression perspective, the residual part in this tensor model form reflects the fitness of the projection matrix in describing the ERD/ERS effects underlying the covariance matrices. Therefore, this residual error can be used to evaluate the mismatch between the feature extraction model and data.

As it is difficult to achieve the residual error minimization and the discrimination objective simultaneously, we propose a two-step approach where the residual error is estimated in the first step and then combined with the discrimination objective function in a regularized manner. For model adaptation, the major challenge in the first step lies in learning the mismatch relevant to the discriminative task without the true labels of new sessions. To this end, we adopt a semi-supervised learning approach to take the class information into consideration instead of the conventional error minimization used in regression model estimation. Then, the estimation of the mismatch metric is integrated into the discrimination objective function with a regularization approach to adapt the feature extraction model. In this way, the performance of feature extraction model can be enhanced by the adaptation toward reducing the data-model mismatch.

The performance of the proposed adaptation method is evaluated on a recorded data set from 16 subjects performing motor imagery tasks on different days. Experimental results show that there is a significant correlation between the quantified mismatch and the classification accuracy. More importantly, the classification results validate the advantage of the proposed method in comparison with other regularization-based or spatial filter adaptation approaches.

This paper is organized as follows. In section 2, spatial pattern analysis with tensor model is presented, followed by the introduction of the adaptation method based on the quantification of the mismatch between model and data. In section 3, we introduce the experimental set-up and data processing procedures for evaluating the proposed method. Section 4 presents an investigation of the correlation between the classification performance and data-model mismatch metric, and the classification results of the proposed method in a two-class motor imagery classification problem. Concluding remarks are given in section 5.

2. Spatial pattern analysis in tensor space for nonstationary signal

2.1. Spatial filtering in tensor decomposition form

For convenience, we will follow the conventional notations and definitions in the area of multi-linear algebra. Thus, in this study, tensors are denoted by calligraphic letters [26]. For the details of the definitions and notations, please refer to appendix A.

Given $X^j \in \mathbb{R}^{n_c \times n_t}$ as the time-series of EEG signal recorded from the jth trial, where n_c is the number of channels and n_t is the number of time samples, the covariance matrix of the jth trial is

$$R^j = X^j X^{jT}. \tag{1}$$

Let V be a projection matrix that maps EEG data from the scalp space to a surrogate channel space, where the resulted covariance matrix

$$\Lambda^j = V^T R^j V \tag{2}$$

is usually assumed to be diagonal for ERD/ERS feature extraction [29, 30].

Remark 1. There is evidence that multiple brain regions may cooperate during motor imagery, which would result in source interaction and connectivity [31, 32]. However, signals after projection with strengthened ERD/ERS effects are usually assumed to be independent in spatial filter design methods, e.g., CSP. Since the proposed method addresses the model adaptation issue, the diagonal covariance matrix assumption is adopted so that the regression error reflects the mismatch of the CSP spatial filters in this work.

To describe multiple trials in a unified model, we adopt the tensor model to formulate the mapping from source signals to scalp EEG data. Let $\mathcal{R}$ be a tensor including the covariance matrices of totally n_j trials as $\mathcal{R} \in \mathbb{R}^{n_c \times n_c \times n_j}$. Then, the jth frontal slice of $\mathcal{R}$ is the covariance matrix R^j for trial j, which can be written as

$$R^j = V^{-T} \Lambda^j V^{-1}. \tag{3}$$

And (3) for all trials can be formulated as

$$\mathcal{R} = \mathcal{I} \times_1 V \times_2 V \times_3 \Lambda_d + \mathcal{E}, \tag{4}$$

where $\mathcal{I} \in \mathbb{R}^{n_c \times n_c \times n_c}$ is the cubic tensor with ones along the super diagonal, and $\mathcal{E} \in \mathbb{R}^{n_c \times n_c \times n_j}$ is the tensor of residual error components. Each of the frontal slices of $\mathcal{E}$ is denoted by E^k. $\Lambda_d = \left[\lambda^1, \lambda^2, ..., \lambda^{n_j}\right]^T \in \mathbb{R}^{n_j \times n_c}$, where $\lambda^j,\ j \in 1, ..., n_j$ is the vector containing the diagonal elements of Λ^j in (3). In addition, Λ_d can be regarded as the matrix containing the variances of the signals of all trials after projection.

The objective of the discriminative spatial pattern learning is to estimate spatial filter V in (4) so that the reconstructed source signal can be classified. In CSP, the solution can be obtained as a generalized eigen-decomposition of average covariance matrices of two classes. Define $\bar{\mathcal{R}} \in \mathbb{R}^{M \times M \times 2}$ as a tensor such that $R^{(+)}$ and $R^{(-)}$ are frontal slices [28], where $R^{(+)}$ and $R^{(-)}$ are the average covariance matrices from class (+) and (−), respectively. The solution of CSP can be written in a tensor form as

$$\bar{\mathcal{R}} = \mathcal{I} \times_1 W \times_2 W \times_3 \bar{\Lambda}_d, \tag{5}$$

where W is used to denote the solution of projection matrix obtained by CSP. $\bar{\Lambda}_d = \left[\lambda^{(+)}, \lambda^{(-)}\right]^T \in \mathbb{R}^{2 \times n_c}$, where $\lambda^{(+)}$ and $\lambda^{(-)}$ are, respectively, vectors consisting of the eigenvalues of $R^{(+)}$ and $R^{(-)}$ upon the joint diagonalization.

An interesting term in (4) but absent in (5) is $\mathcal{E}$. This is the residual part of modelling which is not taken into consideration in CSP. It is often neglected in conventional spatial filter design methods, where the multi-way structure of the data is simplified by averaging covariance matrices. In [28], this non-jointly-diagonalized term has been explored and it is assumed to be related to the quality of the EEG trials. Compared with parameters that measure the data variation, the residual part $\mathcal{E}$ provides a natural data-model mismatch metric in a more direct way. In other words, the residual part $\mathcal{E}$ can be used to evaluate the performance of the spatial filter because it reflects how accurate the model is in describing the ERD/ERS process. Based on this motivation, we adopt the tensor model of the covariance matrices for the data-model mismatch metric estimation and utilize it to guide the spatial filter adaptation.

2.2. Tensor decomposition based adaptation

As the residual part $\mathcal{E}$ can be regarded as a quantification of the mismatch between model and data, the mismatch between the calibration model and test data from different sessions is of particular interest, which is formulated as

$$\mathcal{E}_{\text{te}} = \mathcal{R}_{\text{te}} - \mathcal{I} \times_1 W_{\text{tr}} \times_2 W_{\text{tr}} \times_3 \Lambda_{d,\text{te}}, \tag{6}$$

where $\mathcal{R}_{\text{te}}$ is the tensor of covariance matrices of all test trials and W_{tr} is the solution of CSP in (5) obtained from the calibration session. Then, $\Lambda_{d,\text{te}}$ contains the variances of the signals after projection, and $\mathcal{E}_{\text{te}}$ is the tensor of residual error components, i.e., the mismatch metric between the calibration model and the test data. The error part of test data, $\mathcal{E}_{\text{te}}$, is usually much larger than that of the training data, i.e.

$$\mathcal{E}_{\text{tr}} = \mathcal{R}_{\text{tr}} - \mathcal{I} \times_1 W_{\text{tr}} \times_2 W_{\text{tr}} \times_3 \Lambda_{d,\text{tr}}. \tag{7}$$

Examples will be shown in section 3.

To address the session-to-session transfer problem, W_{tr} should be adapted toward minimizing the residual error with respect to the test data while keeping power differences between classes maximized. However, it is difficult to combine the objective function that minimizes the residual error with the one maximizing the Rayleigh coefficient in CSP, as both W and Λ are dependent on each other. To this end, we propose a two-step approach where the residual error is estimated in the first place and then combined with the objective function of CSP in a regularized manner.

2.2.1. Residual error estimation. Instead of using $\mathcal{E}_{\text{te}}$ in (6), we adopt an iteration approach which is summarized in algorithm 1 to estimate the residual error. Details of the derivation of the updating equations (9) and (10) can be found in appendix B and [28]. In (6), $\Lambda_{d,\text{te}}$ corresponds to the variance features used for classification of the test data (details of variance feature extraction can be found in [15]). The estimation of $\mathcal{E}_{\text{te}}$ is not useful for the adaptation of the discrimination model, if $\Lambda_{d,\text{te}}$ is not separable. To solve this problem, we propose a semi-supervised learning approach to

evaluate the discrimination model as shown in algorithm 1. Different from the iteration in [28], the class information is addressed in the estimation of $\mathcal{E}_{\text{te}}$ to obtain the data-model mismatch metric relevant to the discriminative objective.

Algorithm 1: Estimation of residual error

Input: Training data, a batch of test EEG data w/o class label, and maximum number of iteration n_k

Output: Data-model mismatch metric $\hat{\mathcal{E}}_{te}$

begin

Train a feature extraction model based on the training data;

Obtain features of both training data and test data;

Train a classifier based on the training features;

Classify the test features to obtain the estimated label $\hat{y}$;

Initiate $\Lambda_d(0)$ as

$$\lambda^j(0) = \begin{cases} \lambda^{(+)}, & \hat{y}^j = +; \\ \lambda^{(-)}, & \hat{y}^j = -. \end{cases} \tag{8}$$

Initiate $V(0) = W_{tr}$;

while $k < n_k$ **do**

Update $V(k)$ as

$$V(k) = R_{te,(2)}\{(\Lambda_d(k-1) \odot V(k-1))^T\}^{\dagger} \tag{9}$$

where † denotes the pseudo-inverse of a matrix.

Update $\Lambda_d(k)$ as

$$\Lambda_d(k) = R_{te,(3)}\{(V(k) \odot V(k))^T\}^{\dagger} \tag{10}$$

$k = k + 1$;

Compute

$$\hat{\mathcal{E}}_{te} = \mathcal{R}_{te} - \mathcal{I} \times_1 V(k) \times_2 V(k) \times_3 \Lambda_d(k) \tag{11}$$

As shown in (5), $\bar{\Lambda}_d$ consists of $\lambda^{(+)}$ and $\lambda^{(-)}$, which are the vectors comprising, respectively, the eigenvalues of $R^{(+)}$ and $R^{(-)}$ upon joint diagonalization. Generally speaking, $\lambda^{(+)}$ and $\lambda^{(-)}$ are the centres of distributions of the training features. It is desirable that the test features are close to the corresponding centers in a class-wise way. Therefore, we adopt $\lambda^{(+)}$ and $\lambda^{(-)}$ as the references of variance features of the two classes by using pseudo labels of the test data, denoted as $\hat{y}$ as shown in (8). Upon the class-wise initialization, (9) and (10) are iterated in a data-driven manner so that this estimation process is not relying on the predicted labels totally. In other words, by combining the semi-supervised initialization and iteration procedure, we can balance the trade-off between discrimination objective and the risk of semi-supervised learning. This approach also allows that intrinsic variations remain as for each trial $V \times_2 V \times_3 \Lambda_{\text{te},d}$ are not necessarily the same, and only the residual parts that cannot be jointly diagonalized will be penalized.

2.2.2. Transformation of regularization term into positive-definite matrix. The residual error term $\hat{\mathcal{E}}_{\text{te}}$ estimated in algorithm 1 cannot be regularized directly because it may not be positive-definite, and in this case the regularization actually increases the mismatch as discussed in [19]. In this section, we introduce two methods to guarantee that the penalty term to be positive, and the results comparison and discussion will be given in the next section.

Let $\hat{E}_{\text{te}}^{\,j}$ be the jth frontal slice of $\hat{\mathcal{E}}_{\text{te}}$. To guarantee that the penalty term is positive, we consider the penalty term in the form

$$P_s(\boldsymbol{w}) = \boldsymbol{w}^T \sum_{j=1}^{n_{\text{te}}} \left(\hat{E}_{\text{te}}^{\,j} \hat{E}_{\text{te}}^{\,j^T} \right) \boldsymbol{w}, \tag{12}$$

where n_{te} is number of test trials available for adaptation and $\boldsymbol{w}$ is a spatial filter. The penalty term in (12) may fail to penalize appropriate elements of W in certain cases, as pointed out in [33, 34]. To solve this problem, we propose a novel operator $\mathcal{F}^*$. Let $E \in \mathbb{R}^{n_c \times n_c}$ be an arbitrary error term with eigen-composition

$$E = U \text{diag}(d_i) U^T, \quad i = 1, ..., n_c. \tag{13}$$

Then, we have

$$\mathcal{F}^*(E) = \sum_{i=1}^{n_c} |d_i| \begin{bmatrix} u_{i1}^2 & \dots & |u_{i1} u_{in_c}| \\ \vdots & \dots & \vdots \\ |u_{i1} u_{in_c}| & \dots & u_{in_c}^2 \end{bmatrix}, \tag{14}$$

where u_{ij} is the element of the ith column and jth row of U. Detailed discussion of operation $\mathcal{F}^*$ and its relationship with the 'flipping' method in [19] can be found in appendix C. The penalty term based on (14) is

$$P_f(\boldsymbol{w}) = \boldsymbol{w}^T \sum_{j=1}^{n_{\text{te}}} \mathcal{F}^* \left(\hat{E}_{\text{te}}^{\,j} \right) \boldsymbol{w}. \tag{15}$$

With the regularization terms, the regularized objective functions based on CSP become

$$J^{(+)}(\boldsymbol{w}) = \frac{\boldsymbol{w}^T R^{(+)} \boldsymbol{w}}{\boldsymbol{w}^T \left(R^{(+)} + R^{(-)} \right) \boldsymbol{w} + \mu P(\boldsymbol{w})}, \tag{16}$$

$$J^{(-)}(\boldsymbol{w}) = \frac{\boldsymbol{w}^T R^{(-)} \boldsymbol{w}}{\boldsymbol{w}^T \left(R^{(+)} + R^{(-)} \right) \boldsymbol{w} + \mu P(\boldsymbol{w})}, \tag{17}$$

where $\mu \in [0, 1]$ is the tuning parameter. By maximizing (16) and (17), spatial filters that respectively maximize the power of class (+) and (−) can be obtained [35].

Note that while $R^{(+)}$ and $R^{(-)}$ are computed using training data only, $P(\boldsymbol{w})$ is calculated based on a batch of unlabelled test data as presented in algorithm 1 and (12)–(14). Therefore, (16) and (17) are applied to update the spatial filters and it can be considered as adaptation. By penalizing $P(\boldsymbol{w})$ in the objective function, the residual part $\mathcal{E}$ can be minimized in the updated CSP space. Subsequently, the updated model fits the new data better, and the performance of feature extraction can be improved.

3. Experimental set-up and data description

3.1. Experimental set-up

EEGs from 27 channels were obtained using Nuamps EEG acquisition hardware with monopolar Ag/AgCl electrodes channels. The scalp map of the 27 channels being used is illustrated in figure 1. The sampling rate was 250 Hz with a resolution of 22 bits for the voltage range of ±130 mV. A

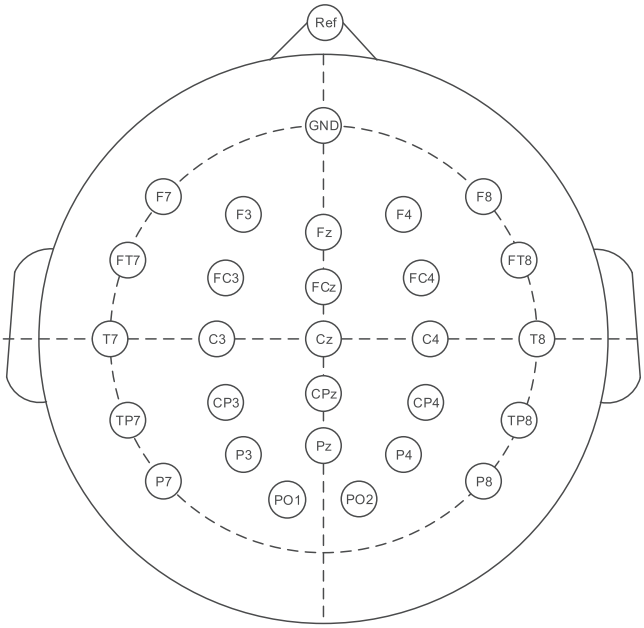

Figure 1. Scalp map of the 27 channels.

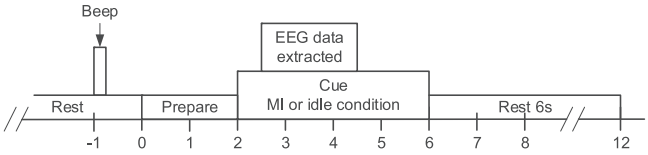

Figure 2. Time segmentation of one trial.

bandpass filter of 0.05–40 Hz was set in the acquisition hardware.

In the experiment, the training and test sessions were recorded on different days with the identical experimental setups for each subject. The training and test sessions contained 2–3 runs. Each run lasted for approximately 16 min, and comprised 40 trials of motor imagery and 40 trials of idle state. The length of each trial was 12 s, including 2 s of preparatory segment, 4 s of visual cue, and 6 s of resting, which is illustrated in figure 2. During the EEG recording process, the subjects were asked to avoid physical movement, minimize eye blinking, and perform kinaesthetic motor imagery of the chosen hand. Given that the subjects were very likely to engage in different mental activities during the idle state on different days, they were instructed to do mental counting to make the idle class EEG signal more consistent, since mental counting is a better defined paradigm than relaxing. Moreover, as a kind of mental work, it could also be more effective in contrasting motor imagery tasks [36].

3.2. Data processing and feature extraction

Since filter bank CSP (FBCSP) [37, 38] is one of the most successful feature extraction methods for motor imagery EEG classification, we implement the proposed adaptation method based on FBCSP. First, we train FBCSP and the naive Bayesian Parzen window (NBPW) classifier with the training data as in [37, 38]. Then, data from the test session is divided

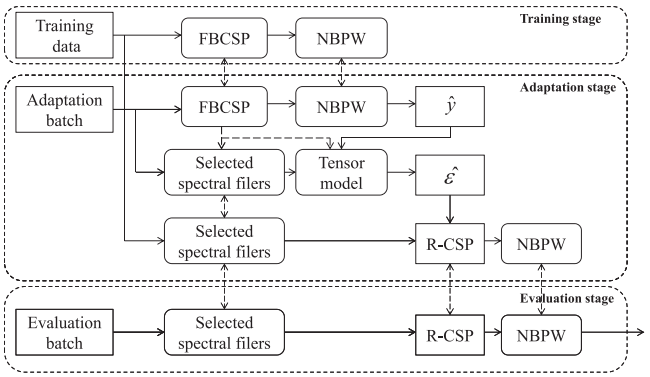

Figure 3. Flowchart of the EEG processing procedure consisting three stages: training stage, adaptation stage and evaluation stage. In the training stage, the training data are used to train FBCSP and NBPW classifier. In the adaptation stage, training model is applied to adaptation batch to obtain the pseudo label $\hat{y}$, and $\hat{\mathcal{E}}$ is estimated for the spectral bands selected in FBCSP. With the training data and $\hat{\mathcal{E}}$, the adapted feature extraction model is obtained by regularized CSP (R-CSP), and subsequently, the training features as well as the NBPW classifier are updated. In the evaluation stage, the adapted feature extraction model and NBPW classifier are applied to the evaluation batch.

equally into two batches, and, as described in section 2.2, $\hat{\mathcal{E}}_{te}$ is estimated based on the first batch of the test data and the projection matrix W_a is obtained by using different penalization terms as in section 2.2.2. Note that during the adaptation procedure the true labels of the test data are not available. This adaptation procedure is only applied to the bands selected in FBCSP for the sake of efficiency. Finally, the updated projection matrices were applied to the training data and the classifier was retrained by the updated training features. Test data from the second batch is classified by the updated model. The aforementioned processing procedures are illustrated in figure 3. For convenience of presentation, we refer the batch of test data used to estimate the error term as the adaptation batch and the rest of test data as the evaluation batch.

To compare the proposed method with other regularization based methods and adaptation methods, we implement Tikhonov (Tik) regularized CSP, spatially regularized (SP) CSP [35], unsupervised data space adaptation (DSA) [25], CSP with naive regularization (nv CSP), and stationary CSP (sCSP) [19]. For Tik and SP, we use cross-validation (CV) results of the training set to select the best regularization term, as in [35]. In DSA [25], the space adaptation matrix is calculated using the test data from the adaptation batch:

$$W_{\mathrm{DSA}} = \bar{R}_{\mathrm{te}}^{-\frac{1}{2}} \bar{R}_{\mathrm{tr}}^{\frac{1}{2}} W_{\mathrm{tr}}, \tag{18}$$

where $\bar{R}_{\mathrm{tr}}$ and $\bar{R}_{\mathrm{te}}$ are average covariance matrices of training set and adaptation batch, respectively. In nvCSP, $\bar{R}_{\mathrm{te}}$ is used as the regularization term, as shown below

$$P(\boldsymbol{w}) = \boldsymbol{w}^T \bar{R}_{\mathrm{te}} \boldsymbol{w}. \tag{19}$$

Note that for nvCSP we use the ratio between the number of the training trials and test trials to determine the regularization

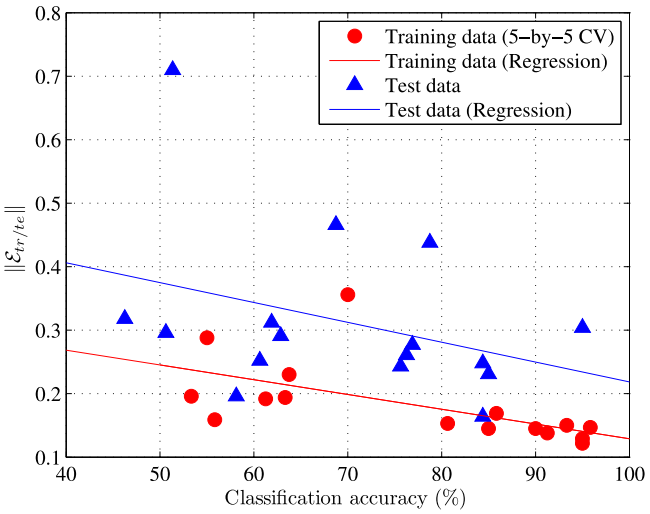

Figure 4. Relation between the residual error and classification accuracy. Each circle or triangle marks one subject. The x-axis represents the classification accuracy and the y-axis represents $\|\mathcal{E}_{\text{tr}}\|$ or $\|\mathcal{E}_{te}\|$. For both training data and test data, $\|\mathcal{E}_{\text{tr}}\|$ and $\|\mathcal{E}_{\text{te}}\|$ correlate to the classification accuracy in a negative way. Pearson's correlation test shows a significant correlation for training data with coefficient r equalling −0.60 and p-value equalling 0.01. Two regression lines are almost parallel, which indicates that the correlations are similar between the training data and the test data.

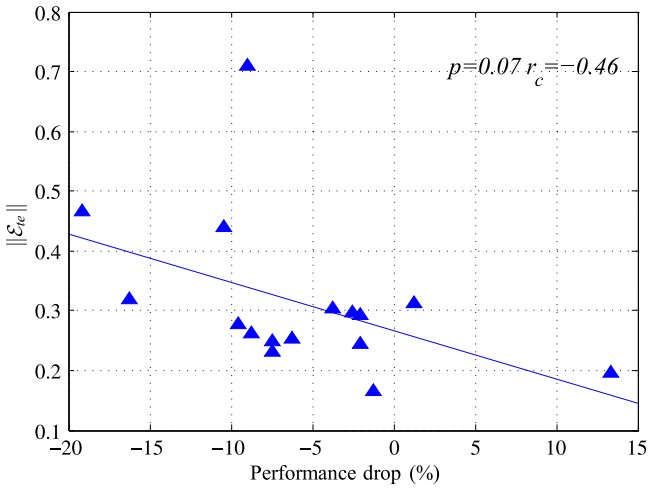

Figure 5. Correlation analysis between $\|\mathcal{E}_{\text{te}}\|$ and the difference between the CV accuracy and the BL accuracy of test data. Pearson's correlation coefficient r_c equals −0.46 with p-value as 0.07. There is a trend that with higher $\|\mathcal{E}_{\text{te}}\|$ the accuracy drop could be higher, and it is possible that the mismatch measurements for test data is better in describing the performance drop caused by the cross-session nonstationarity.

coefficient, i.e., $\mu = \frac{n_{\text{te}}}{n_{\text{tr}}}$, where n_{tr} denotes the number of training trials. For a better comparison, sCSP is implemented in an adaptive manner using data from adaptation batch

$$P(\boldsymbol{w}) = \boldsymbol{w}^T \sum_j \mathcal{F}\left(R_{\text{te}}^j - \bar{R}_{\text{tr}}\right)\boldsymbol{w}, \tag{20}$$

where $\mathcal{F}$ denotes the 'flipping' operator introduced in [19]. Moreover, to validate the necessity of algorithm 1, we use $\mathcal{E}_{\text{te}}$ in (6) as the regularization term, by substituting E_{te}^j into (12) and (15) for $\hat{E}_{\text{te}}^j$.

Since sCSP and the proposed method are used for adaptation, the CV based training set cannot be used to select μ. Thus, we choose to cross-validate the classification performance in a leave-one-subject-out manner. In particular, μ is pre-set as $\mu \in [0.1, 0.2, ..., 1]$, and for a current subject the value of μ is chosen as the one with the best average performance for the rest of the subjects. All methods are implemented with FBCSP in the same way, i.e., they are all applied to the bands selected by FBCSP.

4. Experimental results

4.1. Analysis of residual error

In this section, we investigate the residual error $\mathcal{E}$ to validate the proposed method in measuring the mismatch between the feature extraction model and data.

In particular, we perform the correlation test between $\|\mathcal{E}_{\text{tr/te}}\|$ and the classification accuracy. 5 × 5 CV accuracies are used for training data and session-to-session transfer test classification accuracies are used for test data. Figure 4 illustrates the correlation between the classification accuracy based on FBCSP and average $\|\mathcal{E}_{\text{tr/te}}\|$ of trials from the training/test set. Pearson's correlation coefficient r_c equals −0.60 for the training data with p-value as 0.01. Therefore, we can see that the accuracy for the training data significantly correlates to $\|\mathcal{E}\|$ in a negative way. The p-value for the test data is not significant (0.19) but the correlation is also negative −0.34.

For further analysis, we conduct a correlation analysis between $\|\mathcal{E}_{\text{te}}\|$ and the differences between the CV accuracy of test data and the session-to-session transfer test classification accuracy, as illustrated by figure 5. Because the CV accuracy of the test data can be deemed as the upper bound of the test classification accuracies, such differences can reflect the accuracy 'drop' caused by the session-to-session transfer. Pearson's correlation coefficient r_c equals −0.46 with p-value 0.07. As shown in figure 5, there is a trend that with higher $\|\mathcal{E}_{\text{te}}\|$ the accuracy drop could be larger. The session-to-session transfer test classification accuracy is subject to several factors including the data quality of both training set and test set, and the mismatch caused by the nonstationarity. It is possible that the mismatch measurements for test data are better in illustrating the performance drop caused by the cross-session nonstationarity ($r_c = -0.46$ V. S. $r_c = -0.34$). Given the analysis based on figures 4 and 5, $\|\mathcal{E}_{\text{te}}\|$ can reflect the performance of the computational model generally.

In addition, the change of $\hat{\mathcal{E}}_{\text{te}}$ with respect to the iteration number is also investigated because there is an iteration procedure in algorithm 1. Figure 6 shows an example of the change in $\|\hat{\mathcal{E}}_{\text{te}}\|$ during the process of iteration, where the four frequency bands are selected by mutual information for this subject. As shown in figure 6, the change in $\|\hat{\mathcal{E}}_{\text{te}}\|$ is very small after two iterations, and this trend exists for every subject. Thus, it is reasonable to run the iterations twice, and

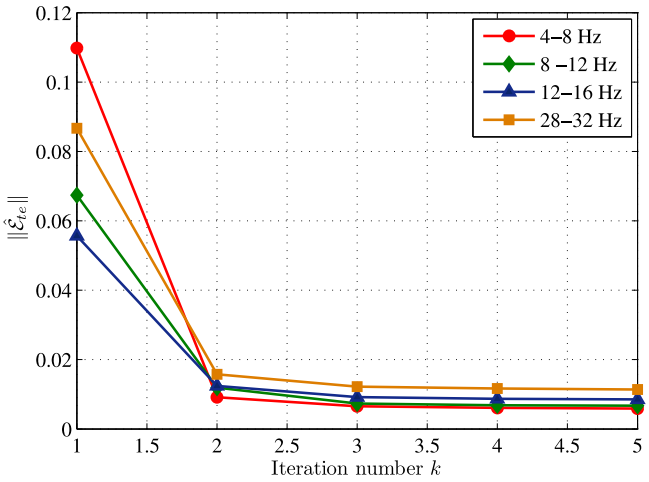

Figure 6. The change in $\|\hat{\mathcal{E}}_{\text{te}}\|$ with respect to the iteration number $k$in algorithm 1 of the best four frequency bands of subject 1. As shown in this figure, the change in $\|\hat{\mathcal{E}}_{\text{te}}\|$ becomes very small after two iterations. Thus, for the efficiency of computation, it is reasonable to run the iterations twice.

this setting is applied to all subjects to obtain the classification results in the following section.

4.2. Classification results

In this section, we present the classification results using the proposed tensor decomposition adaptation (TDA) method. Table 1 summarizes the performance of methods mentioned in section 3.2 compared with FBCSP without any adaptation or regularization as the baseline (BL). Note that all classification accuracies are based on the evaluation batch. We use the subscript s or f to indicate that (12) or (14) is used to transform E into a positive definite matrix, and E_{te} indicates that the direct differences between test data and model in (6) are used. Generally, all adaptation methods improve the performance of FBCSP while spatial-smoothing methods ('Tikhonov' and 'SP') fail to increase accuracies. Paired t-test results show that only TDA_s and TDA_f outperform the baseline in a significant way, and TDA_s achieves the highest accuracy of 74.41%, which indicates the effectiveness of the proposed methods. However, for subjects 2, 11, and 15, there is insignificant or no improvement. Since the BL accuracies for these three subjects are relatively lower, it is possible that these subjects fail to yield discriminative data, or, in other words, they can be regarded as BCI illiterate subjects [39]. Being an adaptation method, the proposed method works under the assumption that the test data itself is discriminative with possibly limited effects on illiterate subjects. Regarding the differences between TDA_s and TDA_f, one reason for the better results of TDA_s could be that $\hat{E}_{\text{te}}^{\,j}\hat{E}_{\text{te}}^{\,j^T}$ is simpler so it is closer to the original error while operation $\mathcal{F}^*(E)$ causes more changes to the error and becomes less accurate. Moreover, the iteration in algorithm 1 actually decreases $\|\hat{\mathcal{E}}_{\text{te}}\hat{\mathcal{E}}_{\text{te}}\|_F^2$, equalling $\sum_j^{n_{\text{te}}} \text{tr}\,(\hat{E}_{\text{te}}^{\,j}\hat{E}_{\text{te}}^{\,j^T})$ (appendix B), which could also be a reason that $\hat{E}_{\text{te}}^{\,j}\hat{E}_{\text{te}}^{\,j^T}$ in (12) matches TDA better.

The changes in the feature distribution between sessions are shown in figure 7. In particular, in each subfigure of figure 7, the distributions of the two-dimensional (2D) features in different sessions are plotted and the corresponding frequency bands are listed. The terms 'a-batch' and 'e-batch' are used to represent the adaption batch and the evaluation batch, respectively. Those features are the most discriminative pairs selected by mutual information in the FBCSP procedure. We can see that without adaptation, the feature distributions shift greatly. It is clearly shown that such a shift has been reduced significantly by TDA, and, subsequently, the feature distributions become more consistent across sessions. More importantly, we find that the variances of the feature distributions are also reduced by TDA, which means that the proposed method can also reduce the within-session nonstationarity.

The class-wise feature distribution is shown in figure 8 to compare the separations of features from different classes with and without adaptation. The nonlinear classification boundary in NBPW classifier is presented by the contrast of different color patterns. By comparing the top and bottom rows in figure 8, we note that the separability of the test features is improved by the proposed method for subject 5. For this subject, there is a larger overlap of the features of two classes in the case without adaptation compared to that with adaptation. For subjects 1 and 3, it can be seen that more features lie on the correct side of the classifier under the proposed method, although the improvements in the feature separability are less significant. Therefore, for some subjects (e.g., subjects 1 and 3), the reduced shifts in the average distance between training and test features contribute to the improvements of the proposed method. For some other subjects (e.g., subject 5), the proposed method is able to adapt the feature extraction model toward increasing the feature separability, which is a more meaningful adaptive behaviour.

Figures 9(a) and (b) show the change of $\|\mathcal{E}_{\text{tr/te}}\|$ with different values of tuning parameter μ. The x-axis represents the value of μ and the y-axis $\|\mathcal{E}_{\text{tr/te}}\|$. Note that in this analysis $\|\mathcal{E}_{\text{tr/te}}\|$ is calculated by substituting the projection matrix after adaptation into (6) or (7). Therefore, when $\mu = 0$, $\|\mathcal{E}_{tr/te}\|$ equals to that in (6) or (7), accordingly. The BL values are given by dotted/dashed lines. For the two sets of test data, $\|\mathcal{E}_{\text{te}}\|$ decreases first and then increases. The trends for TDA_s and TDA_f are different, the reason for which could be that after squaring the scale of the elements in the penalty terms changes greatly. Figures 9(c) and (d) show the change of accuracy with respect to μ. Comparing figures 9(a)–(d), we see that in general the lower the value of $\|\mathcal{E}_{\text{te}}\|$, the higher the accuracy. Since $\mathcal{E}$ reflects the mismatch between model and data, when a high weight is given to the penalty term, we sacrifice the fitness of that model for training data. The value of μ actually controls the balance between test data and training data. As shown in figures 9(a) and (b), $\mu = 0.1$ for TDA_f and $\mu = 0.8$ for TDA_s can be deemed as 'equilibrium' points, where the decrease of $\mathcal{E}_{\text{te}}$ is significant while $\mathcal{E}_{\text{tr}}$ is not increased greatly. This is the reason why these two parameters yield the best accuracy improvements in figures 9(c) and (d). Figures 9(e) and (f) show classification improvements with

Table 1. Session-to-session transfer classification results on the evaluation batch(%).

Subject	BL	SP	Tik	nvCSP	DSA	sCSP	$P_s(E_{te})$	$P_f(E_{te})$	TDA_s	TDA_f
1	67.50	68.75	68.75	61.25	73.50	67.50	66.25	75.00	71.25	76.25
2	58.75	55.00	47.50	68.75	60.00	53.75	56.25	56.25	56.25	56.25
3	50.63	50.63	59.49	70.89	67.09	60.76	63.29	60.76	70.89	70.89
4	71.25	71.25	71.25	61.25	83.75	86.25	78.75	77.50	80.00	87.50
5	75.00	77.50	80.00	60.00	72.50	77.50	82.50	78.75	82.50	78.75
6	82.50	82.50	82.50	77.50	81.25	82.50	81.25	81.25	81.25	82.50
7	80.00	80.00	73.75	51.25	56.25	68.75	73.75	76.25	82.50	75.00
8	93.33	93.33	93.33	95.00	93.33	95.00	96.67	95.00	96.67	95.00
9	78.75	78.75	83.75	72.50	83.75	85.00	78.75	78.75	81.25	82.50
10	65.00	63.29	65.00	51.25	62.03	58.23	63.29	61.25	73.75	63.75
11	50.00	51.25	52.50	51.25	53.75	50.00	50.00	45.00	50.00	51.25
12	78.75	77.50	78.75	77.50	77.50	76.25	80.00	81.25	85.00	80.00
13	53.95	51.25	51.25	51.25	71.25	70.00	68.75	63.75	62.50	65.00
14	71.25	71.25	71.25	80.00	80.00	76.25	73.75	76.25	75.00	72.50
15	57.50	60.00	66.25	52.50	58.75	63.75	60.00	61.25	60.00	60.00
16	73.75	75.00	75.00	76.25	81.25	77.50	77.50	80.00	77.50	76.25
Mean	69.24	69.20	69.63	66.15	72.73	71.81	71.92	71.77	74.14	73.34
p-value	—	>0.05	>0.05	>0.05	>0.05	>0.05	>0.05	>0.05	**0.0023**	**0.029**

All classification accuracies are based on the evaluation batch. FBCSP without any adaptation or regularization is used as the baseline (BL). Tikhonov (Tik) regularized CSP, spatially (SP) regularized CSP, data space adaptation (DSA), naive regularization using average covariance of the test set (nvCSP), and stationary CSP (sCSP) are introduced in section 3.2. $P_s(E_{te})$ and $P_f(E_{te})$ indicate that E_{te} in (6) is used and transformed into a positive definite matrix by (12) and (14), respectively. Similarly, TDA_s and TDA_f indicate the proposed method with (12) and (14), respectively. The significant t-test results with p-value less than 0.05 are highlighted in bold.

decrease in $||\mathcal{E}_{te}||$. In both cases, we find that improvements increase with decrease in $||\mathcal{E}_{te}||$, which is not significant in the Pearson's correlation test though. As we have discussed earlier, since the improvements are subject to both $||\mathcal{E}_{tr}||$ and $||\mathcal{E}_{te}||$, it is reasonable that such unilateral correlations are not significant.

4.3. Discussion

4.3.1. Data-model mismatch estimates. As described in section 2, the role of the regularization term of TDA can be viewed as minimizing the regression error of the model. A natural idea is to use the residual parts of the training data to regularize the model to improve model generalization. However, from the experimental study, it is found that the classification performance of such an implementation is not significantly higher than that of FBCSP without any regularization. The reason is that, since the average covariance matrices are obtained from training data, the residual parts are trivial, as shown in figure 4. Therefore, it is more effective to utilize the residual error from the test data to adapt the model. By improving the model from the perspective of fitness, the classification performance can be enhanced simultaneously. Regarding the necessity of the tensor formulation and the iteration, we have performed the adaptation using $\mathcal{E}_{te}$ in (6) and there is no significant improvement, which validates our consideration that penalizing $\mathcal{E}_{te}$ could not be effective since $\Lambda_{te,d}$ in (6) may not be discriminative.

4.3.2. Relationship between the mismatch and classification accuracy. Given the analysis based on figures 4 and 5 and the classification results, the mismatch estimates have possible different implications for different classification results. For the session-to-session results, the mismatch estimates are better in illustrating the performance drop instead of just the classification accuracies. Thus, for subjects with higher CV accuracies as well as relatively larger mismatch, reducing $||\mathcal{E}_{te}||$ yields better improvements. In contrast, for subjects with much lower CV/BL accuracies, the improvements brought by the proposed method may not be significant. Being an adaptation method based on reducing the mismatch, the proposed method is more effective when there is a performance drop caused by the cross-session nonstationarity but has limited effects for those illiterate subjects who cannot generate discriminative signals.

4.3.3. Computational complexity. For most of the regularization based methods, the most time-consuming part is related to finding the optimization parameters using CL. However, since the proposed method is designed for adaptation, such CL based on training set is meaningless.

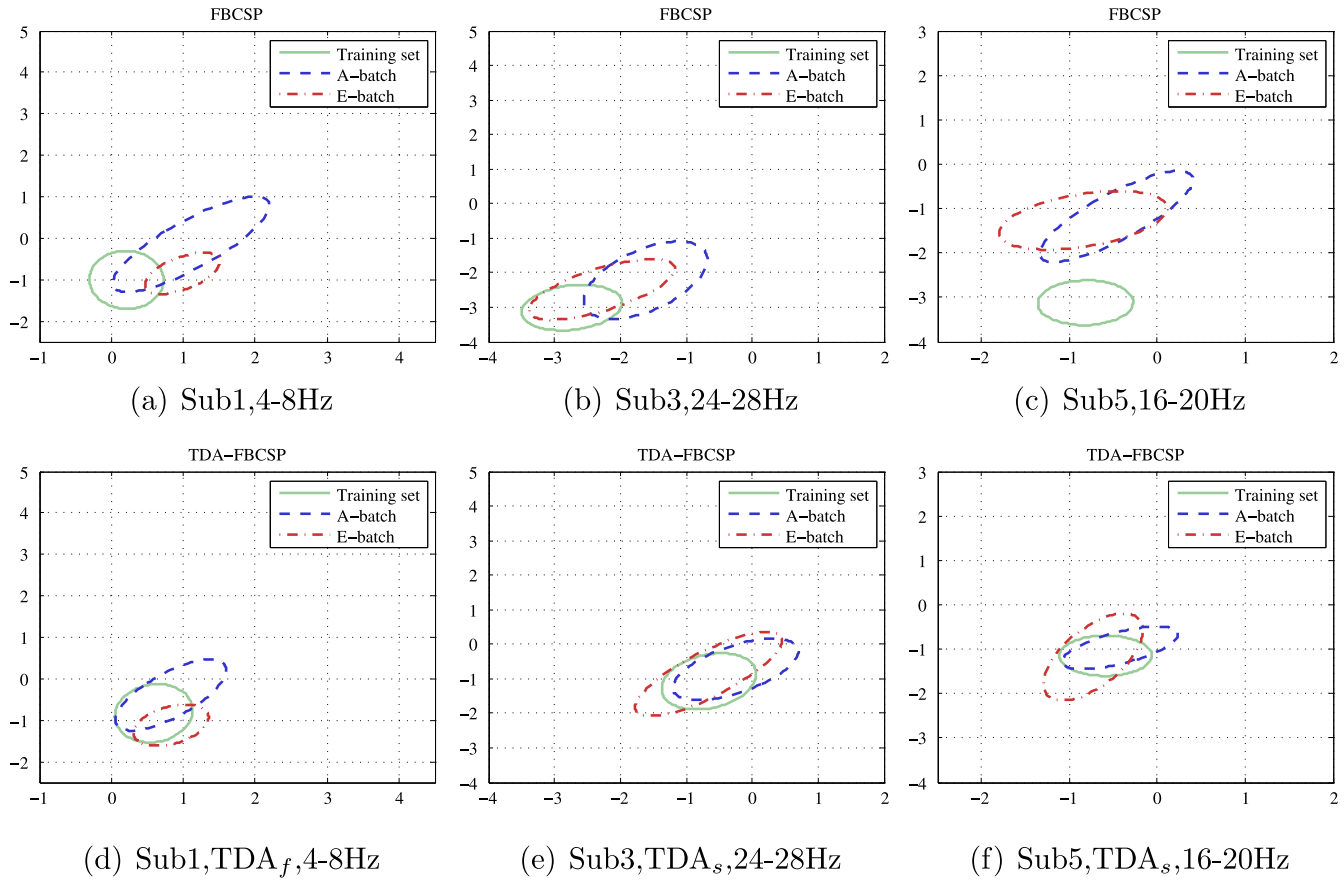

Figure 7. (a)–(c) Feature distributions of the training session and two test batches without any adaptation, where 'a-batch' and 'e-batch' are used to represent the adaption batch and the evaluation batch. (d)–(f) Corresponding feature distributions using TDA. The distances between training features and test features are smaller by using TDA.

Therefore, we adopt leave-one-out to choose different regularization terms μ. Moreover, our analysis on the relationship between $\|\mathcal{E}_{\text{tr/te}}\|$ and accuracy improvements in figure 9 also provides insights into the selection of μ by balancing $\|\mathcal{E}_{\text{te}}\|$ and $\|\mathcal{E}_{\text{tr}}\|$. For the number of iterations in estimating $\mathcal{E}_{\text{te}}$ in algorithm 1, we show that only after 2 iterations, the change of $\|\mathcal{E}_{\text{te}}\|$ becomes quite small. In addition, more iteration steps could be redundant, because we want to maintain the discriminative property of Λ_d. Therefore, we choose the number of iterations as 2, which satisfies the requirements and also reduces the computation burden. Based on the above discussion, for these two parameters there exist feasible values based on which general improvements can be achieved, although tuning the parameters for each individual subject may yield better results for certain subjects.

Moreover, we would like to address the effectiveness of the proposed method as it can be combined with FBCSP easily with low computational complexity and achieve performance improvements. For example, with four frequency bands selected for a subject in FBCSP, it takes 0.0574 s for Matlab with an off-the-shelf CPU to obtain the mismatch estimates for one trial. The rest time between trials is around 5 s and usually much longer between runs. Thus, such computational complexity is acceptable for the proposed method to be implemented online, which will be further validated through online experiment in our future work.

As described before, there exist other works to tackle the nonstationarity problem by utilizing data from other subjects [17, 18]. However, based on FBCSP, usually different frequency bands are selected for different bands, which makes such multi-subject strategies difficult to implement with FBCSP. Moreover, a generic framework is proposed in [34], in which CSP and its regularization methods are unified based on divergence. The divergence-based regularization objective function needs to be solved by a geodesic searching approach or a deflation method. Considering the computational burden combination with FBCSP, it is reasonable to focus on the adaptive or regularization objective function that could be solved by eigen-decomposition in one step. For a similar reason, the signals after projection are assumed to have diagonal covariance matrices in (3) as in CSP. Given the neuroscience findings about source connectivities, a possible extension of the proposed method could be measuring the data-model mismatch for the computational model based on convolutive sources [40].

5. Conclusion

For practical BCI systems, a computational model obtained from the training/calibration session is required to be applied

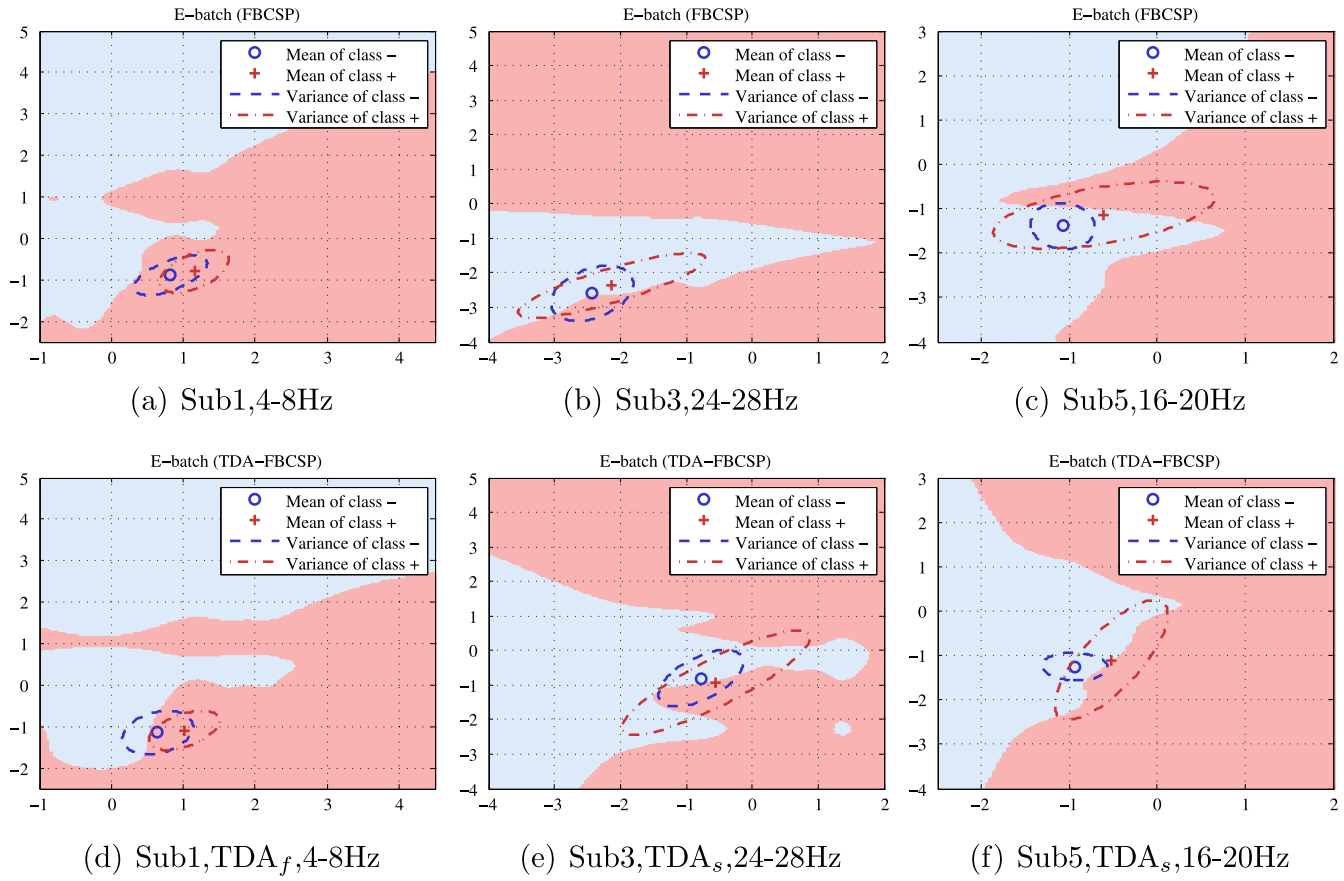

Figure 8. Visualization of the feature separation from the evaluation batch. The nonlinear classification boundary in NBPW classifier is presented by the contrast of different color patterns. (a)–(c) Feature distributions of e-batch without any adaptation. (d)–(f) Corresponding feature distributions using TDA. By employing TDA, more features fall in the corresponding side of the boundary.

to test sessions conducted on different days, while data variation between sessions often leads to the inaccuracy of the computational model. Despite the effort made on adaptive BCI, the quantification of mismatch between data and model needs to be investigated. In this work, we present a systematic attempt to quantify the mismatch between model and data, and use the mismatch metric to guide the model adaptation.

To capture the multidimensional structure of EEG, we adopt a tensor model to formulate the mapping between the variances of the source signals and covariance matrices of scalp EEG signals. The residual error of this model proves to be an effective quantification of the mismatch between model and data. Different from the conventional regression models, the mismatch metric needs to be relevant to the discrimination function. However, in adaptation, true class labels of test data are not available in this discriminative estimation of the mismatch metric. To solve this problem, the estimation is accomplished by a semi-supervised learning approach. Then, the feature extraction model can be updated accordingly toward reducing the data-model mismatch.

We implement the proposed adaptation method combined with FBCSP, which improves the session-to-session transfer classification accuracy significantly as confirmed by the statistical test. Moreover, our correlation analysis also validates the effectiveness of the proposed metric as a quantification of mismatch between model and data.

Appendix A. Notations and basic definitions

Definition 1. Tensor: a tensor, also known as a Nth-order tensor, a multidimensional array, a N-way or a N-mode, is an element of the tensor product of N vector spaces, which is a higher-order generalization of a vector (first-order tensor) and a matrix (second-order tensor), denoted as $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, where N is the order of $\mathcal{A}$. An element of $\mathcal{A}$ is denoted by $a_{i_1, i_2, \ldots, i_N}$, $1 \leqslant i \leqslant I_n$, $n = 1, \ldots, N$.

Definition 2. Tensor Slice: a tensor slice is a 2D section (fragment) of a tensor, obtained by fixing all indices except for two indices.

Definition 3. Unfolding: the n-mode unfolding of tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is denoted by $A_{(n)}$. More specifically, a tensor element $(i_1, i_2, \ldots, i_N)$ maps onto a matrix element

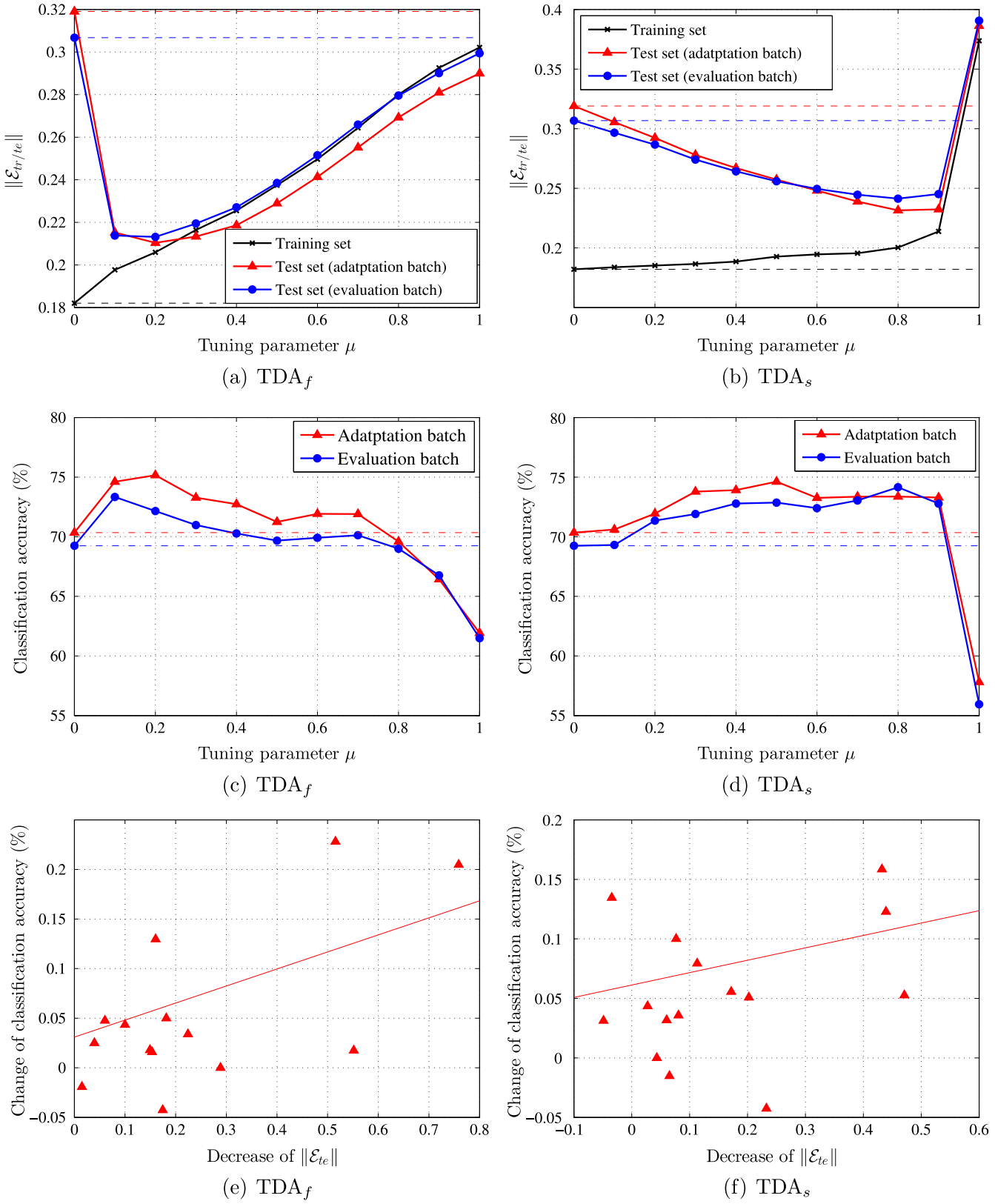

Figure 9. (a), (b) Change of $\|\mathcal{E}\|$ with respect to μ. The x-axis represents the value of μ, and the y-axis represents $\|\mathcal{E}_{\text{tr/te}}\|$ averaged across subjects. $\|\mathcal{E}_{\text{tr/te}}\|$ based on FBCSP without any adaptation are denoted with dotted-dashed lines. (c), (d) Change of accuracy with respect to μ. The x-axis represents the value of μ, and the y-axis represents accuracy averaged across subjects. (e), (f): Change of accuracy with respect to change of $\|\mathcal{E}\|$. The x-axis represents the decrease of $\|\mathcal{E}\|$, and the y-axis represents change of accuracy. Each triangle marks one subject.

(i_n, j), where

$$j = 1 + \sum_{p \neq n} (i_p - 1) J_p,$$
$$J_p = \begin{cases} 1, & \text{if } p = 1 \text{ or} \\ & \text{if } p = 2 \text{ and } n = 1; \\ \prod_{m \neq n}^{p-1} I_m, & \text{otherwise.} \end{cases} \quad (A.1)$$

Definition 4. *n*-mode product: the *n*-mode product of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ and a matrix $U \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n U$, is a tensor in $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N}$ given by

$$(\mathcal{A} \times_n U)_{i_1, i_2, \ldots, i_{n-1}, j_n, i_{n+1}, \ldots, i_N} = \sum_{i_n=1}^{I_n} a_{i_1, i_2, \ldots, i_N} u_{j_n, i_n}. \quad (A.2)$$

Remark 2. Given a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, and two matrices, $F \in \mathbb{R}^{J_n \times I_n}$ and $G \in \mathbb{R}^{J_m \times I_m}$, one has $(\mathcal{A} \times_n F) \times_m G = (\mathcal{A} \times_m G) \times_n F = \mathcal{A} \times_n F \times_m G$.

Definition 5. Khatri–Rao product: for two matrices $A = [a_1, a_2, \ldots, a_J] \in \mathbb{R}^{J_A \times J}$ and $B = [b_1, b_2, \ldots, b_J] \in \mathbb{R}^{J_B \times J}$ with the same number of columns J, their Khatri–Rao product, denoted as $\odot$, performs the following operation:

$$A \odot B = \left[\boldsymbol{vec}\left(b_1 a_1^T\right), \ldots, \boldsymbol{vec}\left(b_J a_J^T\right) \right] \in \mathbb{R}^{J_A J_B \times J}. \quad (A.3)$$

Remark 3. Given a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ and a sequence of matrices $U^n \in \mathbb{R}^{I_n \times J_n}$, $n = 1, 2, \ldots, N$, their multiplication $\mathcal{A} \times_1 U^1 \times_2 U^2 \ldots \times_N U^N$ satisfies

$$\mathcal{A} \times_1 U^1 \times_2 U^2 \ldots \times_N U^N = U^n A_{(n)} \left[U^N \odot U^{N-1} \ldots U^{n+1} \odot U^{n-1} \ldots U^1 \right]. \quad (A.4)$$

Appendix B. Derivation of the update equations

Let $J_E = \|\mathcal{E}\|_F^2$ and $E_{(3)}$ be the mode-3 unfolding of $\mathcal{E}$. Then, (4) becomes

$$E_{(3)} = R_{(3)} - \Lambda_d (V \odot V)^T. \quad (B.1)$$

substituting (B.1) into J_E, we have

$$J_E = \operatorname{tr}\left[R_{(3)} R_{(3)}^T - 2 R_{(3)} (V \odot V) \Lambda_d^T + \Lambda_d (V \odot V)^T (V \odot V) \Lambda_d^T \right] \quad (B.2)$$

differentiating (B.2) with respective to Λ_d^T, we obtain

$$\begin{aligned} \delta J_E = & \operatorname{tr}\left[-2 R_{(3)} (V \odot V) \delta \Lambda_d^T + \delta \Lambda_d (V \odot V)^T (V \odot V) \Lambda_d^T + \Lambda_d (V \odot V)^T (V \odot V) \delta \Lambda_d^T \right] \\ = & \operatorname{tr}\left[-2 R_{(3)} (V \odot V) \delta \Lambda_d^T + 2 \Lambda_d (V \odot V)^T (V \odot V) \delta \Lambda_d^T \right] \\ = & \operatorname{tr}\left[2 \left(\Lambda_d (V \odot V)^T - R_{(3)} \right) (V \odot V) \delta \Lambda_d^T \right]. \end{aligned} \quad (B.3)$$

By setting $\delta J_E = 0$, we obtain

$$\Lambda_d = R_{(3)} \left\{ (V \odot V)^T \right\}^{\dagger} \quad (B.4)$$

which is equivalent to (9) in algorithm 1. Similarly, by substituting the mode-2 unfolding of $\mathcal{E}$ into J_E, we can obtain the update equation for V, i.e., (8) in algorithm 1.

Appendix C. Comparison of different 'Flipping' methods

As pointed out in [33, 34], the 'flipping' method fails to capture relevant nonstationarity in certain cases, which is illustrated by the following example:

$$\bar{\Sigma}^{(+)} = \begin{bmatrix} 0.9 & 0.15 \\ 0.15 & 0.1 \end{bmatrix}, \quad \Sigma^{(+,1)} = \begin{bmatrix} 0.9 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}, \quad \Sigma^{(+,2)} = \begin{bmatrix} 0.9 & 0.25 \\ 0.25 & 0.1 \end{bmatrix}. \quad (C.1)$$

Suppose that $\bar{\Sigma}^{(+)}$ is the average covariance matrix of class +, and $\Sigma^{(+,1)}$ and $\Sigma^{(+,2)}$ are covariance matrices of two trials. To extract the nonstationrity between trials, the penalty matrix in sCSP with 'flipping' is

$$\Delta = \frac{1}{2} \sum_{i=1}^{2} \mathcal{F}\left(\bar{\Sigma}^{(+,k)} - \Sigma^{(+)} \right) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}. \quad (C.2)$$

Thus, the nonstationarity of the off-diagonal elements cannot be penalized. To further investigate this problem, we consider a general case where $\Delta = \Sigma - \bar{\Sigma}$ and $\Delta \in \mathbb{R}^{M \times M}$. Assume that the eigen-decomposition of Δ is

$$\Delta = U D U^T, \quad (C.3)$$

where $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M]$ are the eigenvectors and $D = \operatorname{diag}(d_i)$, $i = 1, 2 \ldots, M$, is the diagonal matrix containing corresponding eigvenvalues.

Then, the penalty term before 'flipping' is

$$\begin{aligned} \boldsymbol{w}^T\Delta\boldsymbol{w} &= \boldsymbol{w}^T\left(\sum_{i=1}^{M} d_i\boldsymbol{u}_i\boldsymbol{u}_i^T\right)\boldsymbol{w} \\ &= \sum_{i=1}^{M} d_i\boldsymbol{w}^T\boldsymbol{u}_i\boldsymbol{u}_i^T\boldsymbol{w} \\ &= \sum_{i=1}^{M} d_i \sum_{p=1}^{M}\sum_{q=1}^{M} u_{ip}u_{iq}w_{ip}w_{iq}, \end{aligned} \tag{C.4}$$

where u_{ip} or u_{ip} is the pth or the qth element in $\boldsymbol{u}_i$. The penalty term after 'flipping' is

$$\boldsymbol{w}^T\mathcal{F}(\Delta)\boldsymbol{w} = \sum_{i=1}^{M}\left|d_i\right| \sum_{p=1}^{M}\sum_{q=1}^{M} u_{ip}u_{iq}w_p w_q. \tag{C.5}$$

The reason why the 'flipping' method fails to penalize relevant nonstatioary elements is that by only taking absolute value of eigenvalue d_i some coefficients $u_{ip}u_{iq}$ would cancel each other. In the example in (C.1), assume that $\Delta^1 = \Sigma^{(+,1)} - \bar{\Sigma}^{+}$ with $\Delta^1 = U^1D^1U^{1T}$, where

$$U^1 = \begin{bmatrix} -0.707 & -0.707 \\ -0.707 & 0.707 \end{bmatrix}$$

$$D^1 = \begin{bmatrix} -0.1 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

Then, we have

$$\begin{aligned} \boldsymbol{w}^T\mathcal{F}(\Delta)\boldsymbol{w} = {} & |-0.1|\left(0.5w_1^2 + 0.1w_1w_2 + 0.5w_2^2\right) \\ & + |0.1|\left(0.5w_1^2 - 0.1w_1w_2 + 0.5w_2^2\right), \end{aligned} \tag{C.6}$$

where the coefficient of w_1w_2 is 0 after taking absolute value of eigenvalues. To avoid this, $u_{ip}u_{iq}$ should be set to be positive if it is not, as below

$$\begin{aligned} \boldsymbol{w}^T\mathcal{F}^*(\Delta)\boldsymbol{w} = {} & \sum_{i=1}^{M}\left|d_i\right| \sum_{p=1}^{M}\sum_{q=1}^{M}\left|u_{ip}u_{iq}\right| w_p w_q \\ \geqslant {} & \boldsymbol{w}^T\mathcal{F}(\Delta)\boldsymbol{w} \\ \geqslant {} & \left|\boldsymbol{w}^T\Delta\boldsymbol{w}\right| \end{aligned} \tag{C.7}$$

which is equivalent to (14).

References

[1] Kaplan A Y, Fingelkurts A A, Borisov S V and Darkhovsky B S 2005 Nonstationary nature of the brain activity as revealed by EEG/MEG: methodological, practical and conceptual challenges *Signal Process.* **85** 2109–212

[2] Fox M D, Snyder A Z, Vincent J L and Raichle M E 2007 Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior *Neuron* **56** 171–84

[3] de Pasquale F *et al* 2010 Temporal dynamics of spontaneous MEG activity in brain networks *Proc. Natl. Acad. Sci.* **107** 6040–5

[4] Blankertz B, Kawanabe M, Tomioka R, Hohlefeld F U, Nikulin V and Müller K 2008 Invariant common spatial patterns: Alleviating nonstationarities in brain–computer interfacing *Adv. Neural Inf. Process. Syst.* 113–20

[5] von Bunau P, Meinecke F C, Kiraly F J and Muller K-R 2009 Finding stationary subspaces in multivariate time series *Phys. Rev. Lett.* **103** 214101

[6] McFarland Dennis J, Miner Laurie A, Vaughan Theresa M and Wolpaw J R Mu and beta rhythm topographies during motor imagery and actual movements *Brain Topography* **12** 177–86

[7] Wolpaw J R, Birbaumer N, Heetderks W J, McFarland D J, Peckham P H, Schalk G, Donchin E, Quatrano L A, Robinson C J and Vaughanr T M 2000 Brain–computer interface technology: a review of the first international meeting *IEEE Trans. Rehabil. Eng.* **8** 164–73

[8] Guler I, Kiymik M K, Akin M and Alkan A 2001 AR spectral analysis of EEG signals by using maximum likelihood estimation *Comput. Biol. Med.* **31** 441–50

[9] Gao L, Wang J and Chen L 2013 Event-related desynchronization and synchronization quantification in motor-related EEG by kolmogorov entropy *J. Neural Eng.* **10** 036023

[10] Ang K K, Guan C, Wang C, Phua K S, Tan A H G and Chin Z Y 2011 Calibrating EEG-based motor imagery brain–computer interface from passive movement *Annual Int. Conf. of the IEEE on Engineering in Medicine and Biology Society (EMBC)* 4199–202

[11] Vidaurre C, Kawanabe M, von Bunau P, Blankertz B and Muller K R 2011 Toward unsupervised adaptation of LDA for brain–computer interfaces *IEEE Trans. Biomed. Eng.* **58** 587–97

[12] Tangwiriyasakul C, Mocioiu V, van Putten M J A M and Rutten W L C 2003 Classification of motor imagery performance in acute stroke *J. Neural Eng.* **11** 036001

[13] Tangwiriyasakul C, Verhagen R, van Putten M J A M and Rutten W L C 2002 Importance of baseline in event-related desynchronization during a combination task of motor imagery and motor observation *J. Neural Eng.* **10** 026009

[14] Ramoser H, Muller-Gerking J and Pfurtscheller G 2000 Optimal spatial filtering of single trial EEG during imagined hand movement *IEEE Trans. Rehabil. Eng.* **8** 441–6

[15] Koles Z J 1991 The quantitative extraction and toporraphic mapping of the abnormal components in the clinical EEG *Electroencephalogr. Clin. Neurophysiol.* **79** 440–7

[16] Gerkinga J M, Pfurtscheller G and Flyvbjergc H 1999 Designing optimal spatial filters for single-trial EEG classiffication in a movement task *Clin. Neurophysiol.* **110** 787–98

[17] Samek W, Meinecke F C and Müller K 2013 Transferring subspaces between subjects in brain-computer interfacing *IEEE Trans. Biomed. Eng.* **60** 2289–98

[18] Lu H, Eng H, Guan C, Plataniotis K N and Venetsanopoulos A N 2010 Regularized common spatial pattern with aggregation for EEG classification in small-sample setting *IEEE Trans. Biomed. Eng.* **57** 2936–46

[19] Samek W, Vidaurre C, Muller K and Kawanabe M 2012 Stationary common spatial patterns for brain–omputer interfacing *J. Neural Eng.* **9** 026013

[20] Arvaneh M, Guan C, Ang K K and Quek C 2013 Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface *IEEE Trans. Neural Netw. Learn. Syst.* **24** 610–9

[21] Liyanage S R, Guan C, Zhang H, Ang K K, Xu J and Lee T H 2013 Dynamically weighted ensemble classification for nonstationary EEG processing *J. Neural Eng.* **10** 036007

[22] Vidaurre C, Schlogl A, Cabeza R, Scherer R and Pfurtscheller G 2007 Study of on-line adaptive discriminant analysis for EEG-based brain–computer interfaces *IEEE Trans. Biomed. Eng.* **54** 550–6

[23] Li Y and Guan C 2006 An extended EM algorithm for joint feature extraction and classification in brain–computer interfaces *Neural Comput.* **18** 2730–61

[24] Bamdadian A, Guan C, Ang K K and Xu J 2012 Online semi-supervised learning with KL distance weighting for motor imagery-based BCI *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)* pp 2732–5
[25] Arvaneh M, Guan C, Ang K K and Quek C 2013 EEG data space adaptation to reduce inter-session non-stationarity in brain–computer interface *Neural Comput.* **25** 2146–71
[26] Cichocki A, Zdunek R, Phan A H and Amari S 2009 *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation* (New York: Wiley)
[27] Washizawa Y, Higashi H, Rutkowskiand T, Tanaka T and Cichocki A 2010 Tensor based simultaneous feature extraction and sample weighting for EEG classification *Lecture Notes in Computer Science* **6444** 26–33
[28] Tomida N, Higashi H and Tanaka T 2013 A joint tensor diagonalization approach to active data selection for EEG classification *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* pp 983–7
[29] Wu W, Chen Z, Gao S and Brown E N 2011 A hierarchical bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG *NeuroImage* **56** 1929–45
[30] Wu W, Chen Z, Gao X, Li Y, Brown E and Gao S 2014 Probabilistic common spatial patterns for multichanne lEEG analysis *IEEE Trans. Pattern Anal. Mach. Intell.* doi:10.1109/TPAMI.2014.2330598
[31] Pfurtscheller G, Neuper Ch, Andrew C and Edlinger G 1997 Foot and hand area mu rhythms *Int. J. Psychophysiology* **26** 121– 135
[32] Chen H, Yang Q, Liao W, Gong Q and Shen S 2009 Evaluation of the effective connectivity of supplementary motor areas during motor imagery using granger causality mapping *NeuroImage* **47** 1844–53
[33] Kawanabe M, Samek W, Müller K and Vidaurre C 2014 Robust common spatial filters with a maxmin approach *Neural Comput.* **26** 349–76
[34] Samek W, Kawanabe M and Muller K 2013 Divergence-based framework for common spatial patterns algorithms *IEEE Rev. Biomed. Eng.* **7** 50–72
[35] Lotte F and Guan C 2011 Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms *IEEE Trans. Biomed. Eng.* **58** 355–62
[36] Friedrich E V C, Scherer R and Neuper C 2012 The effect of distinct mental strategies on classification performance for brain–computer interfaces *Int. J. Psychophysiology* **84** 86–94
[37] Ang K K, Chin Z Y, Wang C, Guan C and Zhang H 2012 Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b *Front. Neurosci.* **6**
[38] Ang K K, Chin Z Y, Zhang H and Guan C 2012 Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs *Pattern Recognit.* **45** 2137–44
[39] Blankertz B, Sannelli C, Halder S, Hammer E M, Kübler A, Müller K, Curio G and Dickhaus T 2010 Neurophysiological predictor of SMR-based BCI performance *NeuroImage* **51** 1303–9
[40] Li X, Zhang H, Guan C, Ong S H, Ang K K and Pan Y 2013 Discriminative learning of propagation and spatial pattern for motor imagery EEG analysis *Neural Comput.* **25** 2709–33