A Unified Fisher's Ratio Learning Method for Spatial Filter Optimization

Xinyang Li, Cuntai Guan, *Senior Member, IEEE*, Haihong Zhang, *Member, IEEE*, and Kai Keng Ang, *Senior Member, IEEE*

***Abstract*— To detect the mental task of interest, spatial filtering has been widely used to enhance the spatial resolution of electroencephalography (EEG). However, the effectiveness of spatial filtering is undermined due to the significant nonstationarity of EEG. Based on regularization, most of the conventional stationary spatial filter design methods address the nonstationarity at the cost of the interclass discrimination. Moreover, spatial filter optimization is inconsistent with feature extraction when EEG covariance matrices could not be jointly diagonalized due to the regularization. In this paper, we propose a novel framework for a spatial filter design. With Fisher's ratio in feature space directly used as the objective function, the spatial filter optimization is unified with feature extraction. Given its ratio form, the selection of the regularization parameter could be avoided. We evaluate the proposed method on a binary motor imagery data set of 16 subjects, who performed the calibration and test sessions on different days. The experimental results show that the proposed method yields improvement in classification performance for both single broadband and filter bank settings compared with conventional nonunified methods. We also provide a systematic attempt to compare different objective functions in modeling data nonstationarity with simulation studies.**

***Index Terms*— BCI, electroencephalography (EEG), motor imagery, optimization, spatial filtering.**

I. Introduction

In DISCRIMINATIVE learning of electroencephalography (EEG), spatial filtering has been widely applied as preprocessing or feature extraction, especially in EEG-based brain computer interface (BCI) systems [1], [2]. The function of spatial filtering lies in enhancing the spatial resolution of EEG [3], [4]. Besides, it reduces the number of features as the number of designed spatial filters is usually much less than the number of channels in the scalp space [5]. However, it assumes the stationarity of the spatial distribution of the relevant EEG components over time in spatial filtering. Sensitive to variability in evoked brain responses or human behaviors, EEG recordings usually contain contributions of multiple varying mental processes, only a small portion of which relates to the task of interest. In fact, even resting state brain networks would cause spontaneous fluctuations in the brain signals [6]–[8]. The low signal-to-noise ratio and nonstationary nature of EEG significantly undermines the reliability of spatial filtering in enhancing the task-related processes.

Many efforts have been made to optimize spatial filter so that it is discriminative while being robust against EEG nonstationarity and artifacts contamination. For instance, common spatial pattern (CSP) analysis is one of the most effective spatial filter design methods for motor imagery discrimination [9], [10]. Given the event-related (de)synchronization (ERD/ERS) in EEG during the imagination of certain movements, CSP extracts the ERD/ERS effects by maximizing the power of the spatially filtered signals for one class while minimizing it for the other class [11], [12]. As the EEG data are in the form of multichannel time series, the covariance matrix of EEG consists of covariances between EEG signals from pairs of channels. With the classwise average of the covariance matrices, the objective function of CSP is casted in relation to Rayleigh quotient between the average covariance matrices, which is equivalent to the ratio of powers of the EEG signals from the two classes. The objective function has been used for a joint temporal-spatial analysis as an improvement of CSP [5]. Higashi and Tanaka [13] proposed the discriminative filter bank CSP (DFBCSP), where the finite impulse response filters and the associated spatial weights are obtained in a sequential manner by optimizing Rayleigh quotient. In iterative spatio-spectral patterns learning (ISSPL), Rayleigh quotient is combined with the objective function of support vector machine (SVM) to optimize spatio-spectral filters and the classifier in a sequential manner [14]. Similarly, in [15], spatial filters and time-lagged coefficient matrices in a convolutive model are jointly optimized using Rayleigh quotient to address the propagation effects in motor imagery EEG.

The Rayleigh quotient objective function is effective in terms of discrimination, while it is prone to nonstationarity. Thus, a number of regularization methods have been proposed to improve its robustness against the data nonstationarity [16]. The regularization method refers to adding a regularization term to the denominator of the Rayleigh quotient so that this term can be penalized in the objective function [17]. To achieve the invariant property of the spatial filters, stationary CSP is proposed to address nonstationary noise in a more

Manuscript received January 27, 2016; accepted July 31, 2016. Date of publication August 29, 2016; date of current version October 16, 2017.

X. Li is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, U.K. (e-mail: jiangffjiang@gmail.com).

C. Guan and K. K. Ang are with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632, and also with Nanyang Technological University, Singapore 639798 (e-mail: ctguan@i2r.a-star.edu.sg; kkang@i2r.a-star.edu.sg).

H. Zhang is with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632 (e-mail: hhzhang@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2016.2601084

general case without using additional recordings to estimate the nonstationary artifacts [18]. In particular, nonstationarity is estimated as the sum of absolute differences between the mean variance and variance of a certain trial in the projected space. In [19], the nonstationary projection directions are estimated based on the principal component analysis using cross-subject data, and then penalized in the objective function to build subject-specific spatial filters.

Arvaneh *et al.* [20] introduced a different penalizing term that measures the Kullback–Leibler divergence (KL-divergence) of the EEG distributions across trials, and subsequently, the objective function can minimize within-class dissimilarities. In [21], different divergence measurements are used as nonstationarity regularization, and it is proved that the spatial filters in CSP project the EEG data into subspaces where the KL-divergence between the data distributions from two classes is maximized. Therefore, the objective function of CSP can also be formed in a divergence-based framework. The significance of this paper lies in the fact that it is a unified framework for CSP with different kinds of regularization.

Besides Rayleigh quotient and KL-divergence, the mutual information has also been used for spatial filter optimization. In [22] and [23], multiple bandpass filters denoted as a filter bank are applied for raw EEG data, and the CSP spatial filters are calculated for each filter band. Thus, each pair of bandpass and spatial filter yields CSP features that are specific to the frequency range of the bandpass filter. Then, both spatial and temporal filters are optimized by selecting those features whose mutual information with the class labels is higher. Similarly, in the optimum spatio-spectral filtering network (OSSFN), the bandpass filters and the spatial filters are jointly optimized to maximize the mutual information between feature vector variables and the class labels via gradient searching [24].

In general, Rayleigh quotient, KL-divergence, and mutual information are conventional objective functions for spatial filter optimization, and they are often combined in different ways. As aforementioned, KL-divergence-based loss function can be used as a regularization term in the Rayleigh quotient objective function to penalize the within-class dissimilarity, while mutual information is used to select filter bands as a way of temporal filter optimization. However, their relationships with classification in the feature space have not been investigated sufficiently, especially when nonstationarity needs to be taken into consideration. Zhang *et al.* [25] established the theory linking Rayleigh quotient with Bayes classification error in the CSP feature space. In that work, the feature nonstationarity is simplified by assuming that the shape parameter of the feature distribution is independent of the spatial filters. As most of the stationary spatial filter design methods are based on regularization, the intertrial nonstationarity or the within-class dissimilarity is minimized at the cost of interclass dissimilarity, which can be regarded as one of the limitations of the regularization methods. Moreover, divergence measurements are usually applied in the covariance matrix space rather than in the feature space. The problem lies in the fact that the spatial filters with regularization usually fail to jointly diagonalize the covariance matrices, while only the diagonal elements are used as features. In other words, the feature used for classification is not consistent with the optimization objective function in the spatial filter design based on the covariance matrix. Due to the inconsistency, stationarity or discrimination for the covariance matrix cannot be fully transferred into that for the features.

TABLE I
NOMENCLATURE

X	band-passed EEG data;
n_c	number of electrodes or channels;
n_t	number of time points;
j	trial index;
i	feature dimension index;
$c \in +, -$	class index;
$\mathcal{Q}^c$	a set containing indexes of trials that belong to class c;
R	covariance matrix of EEG data;
$\bar{R}^c$	average covariance matrix of class c;
W	projection matrix for feature extraction;
$\mathbf{w}$	spatial filter, one row in W;
P	whitening matrix;
U	rotation matrix;
U_m	discriminative subspace;
m	dimension of the discriminative subspace;
J	objective function;
Λ	covariance matrix after projection;
$\mathbf{f}$	variance feature of the EEG signal after projection;
$\mathbf{d}$	feature distance;
λ	regularization parameter;
Δ_s	stationary regularization term;
k	iteration index.

To address the issue, we propose to optimize spatial filers using Fisher's ratio in the feature space. This is a unified framework for spatial filter design, as it directly addresses discrimination and stationarity in the feature space, compared with objective functions in the conventional spatial filter optimization. Given its ratio form, within-class feature dissimilarity could be minimized without sacrificing interclass discrimination, and regularization parameter selection could be avoided. Moreover, we provide a systematic attempt to compare the effectiveness of different objective functions by applying them to both feature extraction and selection with single broadband or filter bank as preprocessing. In particular, the roles that different objective functions play in nonstationarity measurement have been investigated and discussed. Based on the above discussion, we highlight the contributions of this paper as follows.

1) Fisher's ratio-based spatial filter design method is proposed.
2) Relationships between different objective functions, such as Rayleigh quotient, KL-divergence and mutual information, are discussed.
3) Efficiency of different objective functions in the nonstationarity modeling is investigated.

II. Preliminaries of Spatial Filter Design

A. Common Spatial Pattern Analysis

To make this work self-contained, in this section, we will introduce the preliminaries about the spatial filter design. Please refer to Table I for nomenclature. Let $X_j \in \mathbb{R}^{n_c \times n_t}$ be

the the bandpass filtered EEG data of trial j, where n_c and n_t are the numbers of channels and time points, respectively. As the mean of X_j is zero after the preprocessing, the covariance matrix R_j could be obtained as

$$R_j = \frac{X_j (X_j)^T}{\mathrm{tr}[X_j (X_j)^T]}. \tag{1}$$

The CSP spatial filter $\mathbf{w}$ is designed to maximize the variance of the spatially filtered signal under one condition, and given the constraint that the sum of the variances under two conditions is one, the variance under the other condition could be minimized subsequently. The objective function of CSP can be expressed in the following optimization problem:

$$\max_{\mathbf{w}} \mathbf{w}^T \bar{R}^c \mathbf{w} \quad \text{s.t. } \mathbf{w}^T (\bar{R}^+ + \bar{R}^-)\mathbf{w} = 1 \tag{2}$$

where $\bar{R}^c$ is the average covariance matrix for class c, that is

$$\bar{R}^c = \frac{1}{|\mathcal{Q}^c|} \sum_{j \in \mathcal{Q}^c} R_j, \quad c \in \{+, -\} \tag{3}$$

where $\mathcal{Q}^c$ is the set containing indices of trials that belong to class c.

B. Common Spatial Pattern Analysis in KL-Divergence Form

Given that EEG data are usually processed to be centered, the KL-divergence between two EEG data sets, i_1 and i_2, is

$$D_{kl}(R^{i_1} || R^{i_2}) = \frac{1}{2} \left(\mathrm{tr}((R^{i_2})^{-1} R^{i_1}) - \ln \left(\frac{\det R^{i_1}}{\det R^{i_2}} \right) - n_c \right) \tag{4}$$

and the symmetric KL-divergence is defined as

$$\tilde{D}_{kl}(R^{i_1} || R^{i_2}) = D_{kl}(R^{i_1} || R^{i_2}) + D_{kl}(R^{i_2} || R^{i_1}). \tag{5}$$

In [21], it is proved that the solution of (2) is equal to that of the following optimization problem, i.e., the spatial filters maximizing the symmetric KL-divergence between the two classes:

$$W = \arg\max_{W} \tilde{D}_{kl}(W^T R^+ W || W^T R^- W). \tag{6}$$

The objective function with stationary regularization term, Δ_s, can be formulated as

$$J_{kl} = (1 - \lambda) \tilde{D}_{kl}(W^T R^+ W || W^T R^- W) - \lambda \Delta_s \tag{7}$$

Δ_s may vary according to the type of nonstationarity to be minimized. Typically, it is the classwise intertrial nonstationarity, measured as the classwise average divergence between the trials and mean data distributions as follows:

$$\Delta_s = \sum_{c=\{+,-\}} \frac{1}{|\mathcal{Q}^c|} \sum_{j \in \mathcal{Q}^c} D_{kl}(W^T R_j W || W^T R^c W). \tag{8}$$

III. Spatial Filter Optimization Based on Fisher's Ratio

Most of the existing stationary approaches have a similar form as (7), where different regularization terms are used to penalize the data nonstationarity [18]–[20]. One of the limitations of such a formulation is that the data nonstationarity is minimized at the cost of the interclass dissimilarity. It usually takes cross validation to find the regularization term λ. To address both the intertrial nonstationarity and class discrimination in the spatial filter design, in this section, we will introduce the objective functions based on Fisher's ratio.

A. Objective Function Based on Fisher's Ratio

The projection matrix W in CSP can also be represented as

$$W = P^T U \tag{9}$$

where each row of W is one solution of (3), and P is the whitening matrix such that

$$P(\bar{R}^+ + \bar{R}^-)P^T = I. \tag{10}$$

Define the covariance matrix after whitening

$$\bar{\Sigma}^c = P \bar{R}^c P^T. \tag{11}$$

Each column of U in (9), $\mathbf{u}_i$, $i = 1, \ldots, n_c$, is an eigenvector of $\bar{\Sigma}^c$. Usually, only those eigenvectors $\mathbf{u}_i$ corresponding to the largest and smallest eigenvalues of $\bar{\Sigma}^c$ are used for feature extraction as follows:

$$W = P^T U_m \tag{12}$$

$$U_m = I_m U \tag{13}$$

where U_m can be regarded as the discriminative subspace, m is the dimension of feature, and I_m is a m-by-n_c identity matrix. The covariance matrix after projection can be written as

$$\Lambda_j = U_m^T \Sigma_j U_m \tag{14}$$

where

$$\Sigma_j = P R_j P^T. \tag{15}$$

Remark 1: As CSP maximizes the power of one class and minimizes it for the other class, the whitening in (10) is a necessary constraint. Since the proposed optimization framework adopts Fisher's ratio in the feature space as objective function, the whitening is no longer necessary. However, with the whitening matrix P, features could be normalized to the range around [0, 1], which is beneficial for the classification. Thus, the proposed spatial filter optimization is based on the covariance matrix after the whitening, i.e., Σ.

Then, the feature vector $\mathbf{f}_j \in \mathbb{R}^m$ for trial j extracted by the projection matrix W is

$$\mathbf{f}_j = \mathrm{diag}(\Lambda_j), \quad j = 1, \ldots, m \tag{16}$$

which contains the variances of the EEG signals after projection. Let the ith element of $\mathbf{f}_j$ be $\mathbf{f}_{j,i}$. Suppose trial i belongs

to class c, and let $\bar{\mathbf{f}}_i^c$ be the ith element of the average feature of class c, i.e., $\bar{\mathbf{f}}^c$. Then, the distance between $\mathbf{f}_{j,i}$ and $\bar{\mathbf{f}}_i^c$ is

$$\mathbf{d}_{j,i} = \mathbf{f}_{j,i} - \bar{\mathbf{f}}_i^c. \tag{17}$$

Furthermore, the within-class distance S_w is

$$S_w = \frac{1}{2}\sum_{c\in\mathcal{C}} \frac{1}{|\mathcal{Q}^c|} \sum_{j\in\mathcal{Q}^c} \frac{1}{m}\sum_{i=1}^{m} (\mathbf{d}_{j,i})^2. \tag{18}$$

Similarly, the interclass distance between the mean features of the two classes is

$$S_b = \frac{1}{m}\sum_{i=1}^{m} (\bar{\mathbf{d}}_i)^2 \tag{19}$$

where $\bar{\mathbf{d}}_i$ is the ith element of $\bar{\mathbf{d}}$, that is

$$\bar{\mathbf{d}} = \bar{\mathbf{f}}^+ - \bar{\mathbf{f}}^-. \tag{20}$$

With (18) and (19), Fisher's ratio objective function can be obtained as

$$J_{fs} = \frac{S_w}{S_b}. \tag{21}$$

Let $\mathbf{1}_n^i \in \mathbb{R}^n$ be a n-dimension vector with the ith element as 1 and other elements as 0, that is

$$\mathbf{1}_n^i = [0, 0, \ldots, 1, \ldots, 0]^T. \tag{22}$$

Then, the features can be rewritten as

$$\mathbf{f}_{j,i} = \left(\mathbf{1}_{n_c}^i\right)^T \Lambda_j \mathbf{1}_{n_c}^i \tag{23}$$

$$\bar{\mathbf{f}}_i^c = \left(\mathbf{1}_{n_c}^i\right)^T \bar{\Lambda}^c \mathbf{1}_{n_c}^i \tag{24}$$

Substituting (14), (23), and (24) into (17), the distance between $\mathbf{f}_{j,i}$ and the mean of features from the same class $\bar{\mathbf{f}}^c(i)$ is

$$\begin{aligned}\mathbf{d}_{j,i} &= \left(\mathbf{1}_{n_c}^i\right)^T (\Lambda_j - \bar{\Lambda}^c)\mathbf{1}_{n_c}^i \\ &= \left(\mathbf{1}_m^i\right)^T U_m^T \Delta\Sigma_j U_m \mathbf{1}_m^i.\end{aligned} \tag{25}$$

where

$$\Delta\Sigma_j = \Sigma_j - \bar{\Sigma}^c, \quad j \in \mathcal{Q}^c. \tag{26}$$

Let $\mathbf{1}_n^{ii} \in \mathbb{R}^{n\times n}$ be a n-by-n matrix with the element of the ith column and the ith row as 1 and other elements as 0, that is

$$\begin{aligned}\mathbf{1}_n^{ii} &= \mathbf{1}_n^i \left(\mathbf{1}_n^i\right)^T \\ &= \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & \ddots & \ldots & 0 \\ \vdots & \vdots & 1 & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}.\end{aligned} \tag{27}$$

Then, the squared distance is

$$\mathbf{d}_{j,i}^2 = (\mathbf{1}_m^i)^T \left(U_m^T \Delta\Sigma_j U_m \mathbf{1}_m^{ii} U_m^T \Delta\Sigma_j U_m\right)\mathbf{1}_m^i. \tag{28}$$

Substituting (28) into (18), we have

$$\begin{aligned}S_w = \frac{1}{2}\sum_{c\in\mathcal{C}} \frac{1}{|\mathcal{Q}^c|} \sum_{j\in\mathcal{Q}^c} \frac{1}{m}\sum_{i=1}^{m} (\mathbf{1}_m^i)^T \\ \times \left(U_m^T \Delta\Sigma_j U_m \mathbf{1}_m^{ii} U_m^T \Delta\Sigma_j U_m\right)\mathbf{1}_m^i.\end{aligned} \tag{29}$$

Similarly, the interclass distance between the mean features of the two classes can be obtained as

$$S_b = \frac{1}{m}\sum_{i=1}^{m} (\bar{\mathbf{d}}_i)^2 \tag{30}$$

$$= \frac{1}{m}\sum_{i=1}^{m} (\mathbf{1}_m^i)^T \left(U_m^T \Delta\Sigma U_m \mathbf{1}_m^{ii} U_m^T \Delta\Sigma U_m\right)\mathbf{1}_m^i \tag{31}$$

where

$$\Delta\Sigma = \bar{\Sigma}^+ - \bar{\Sigma}^-. \tag{32}$$

Substituting (29) and (30) into (21), we can obtain Fisher's ratio objective function of U_m as

$$\hat{U}_m = \arg\min_{U_m} J_{fs}(U_m) \quad \text{s.t. } U_m^T U_m = I. \tag{33}$$

Remark 2: Although Fisher's ratio in (21) in linear discriminative analysis could be regarded as a kind of Rayleigh quotient, the subspace optimization in (33) is actually not in Rayleigh quotient form when (29) and (30) are substituted into (21). In this paper, Rayleigh quotient objective function refers to the Rayleigh quotient used in CSP, which is equivalent to the ratio between the means of the two classes in the feature space. The difference between (21) and CSP lies in that CSP only considers the interclass distance and Fisher's ratio in (21) maximizes interclass distance while minimizing within-class distances.

After optimizing U_m using (33), the projection matrix can be obtained by substituting $\hat{U}_m$ into (9).

The optimization of the discriminative subspace is accomplished by using gradient descent on the manifold of orthogonal matrices, which is shown in Algorithm 1. The details of the subspace optimization approach can be found in [21] and [26].

B. Filter Bank Fisher's Ratio Spatial Filtering

In this paper, we evaluate the proposed spatial filter optimization method by applying it to EEG signals bandpassed by single broadband filter, and a filter bank that consists of multiple filters [22], [23]. Moreover, given the spatial filter optimization based on Fisher's ratio, we propose to conduct feature selection also using Fisher's ratio. In the rest parts of the paper, the combination of filter bank and Fisher's ratio is denoted as filter bank Fisher's ratio spatial pattern analysis (FB-FSSP). The flowchart of the EEG signal processing procedures is shown in Fig. 1, where the objective functions used in FBCSP and FB-FSSP for spatial filtering, feature selection, and classification steps are annotated for comparison. With Fisher's ratio (21) as the objective function, the proposed spatial optimization is consistent with the features in (16) used for feature selection and classification. Compared with FBCSP, the proposed FB-FSSP is a more unified framework also in the sense that Fisher's ratio objective function is used in all the three steps in EEG processing.

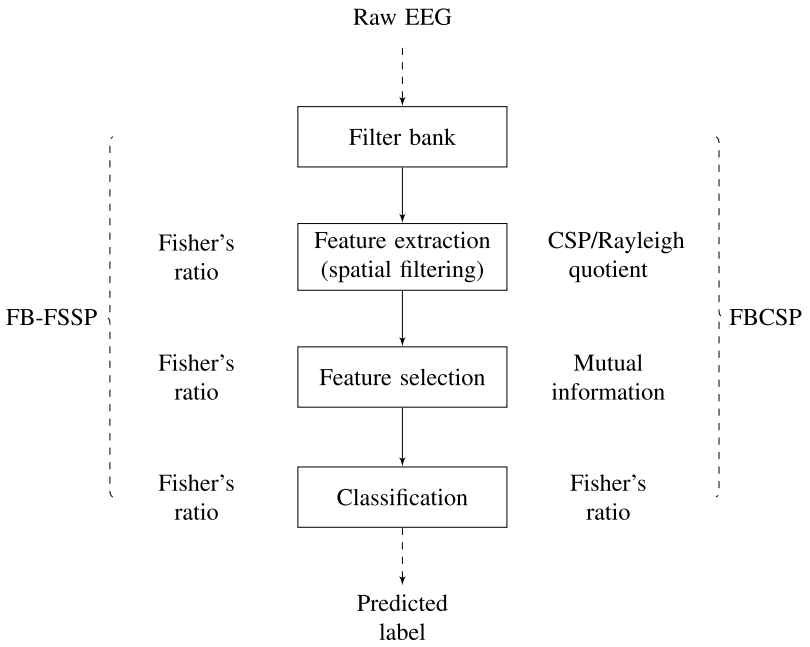


Fig. 1. Flowchart of the EEG signal processing procedures in FB-FSSP and FBCSP.

Algorithm 4 Subspace Optimization Through Gradient Searching

Input: training data with labels;
Output: $\hat{U}_m$.
begin
Compute whitening matrix P;
Compute whitened covariance matrices Σ_j and $\bar{\Sigma}^c$;
Initialize U_m^k as the discriminative subspace in the CSP solution in Eq. (9) to Eq. (12) with $k = 0$.
repeat:
Compute the covariance matrices after rotation Λ^k or features $\mathbf{f}^k$;
With Λ^k or $\mathbf{f}^k$, compute the gradient matrix ∇_{U_m} with respect to $J(U_m)$ and obtain H as:

$$H = \begin{pmatrix} 0 & \nabla_{U_m} \\ -\nabla_{U_m}^T & 0 \end{pmatrix};$$

Given $t_u = [0.9^5, 0.9^6, \ldots, 0.9^{10}]$, estimate the optimal step size by line search such that

$$\hat{t}_u = \arg\min_{t_u} J_{fs}(U_m^k \exp(t_u H))$$

Update the rotation matrix and the covariance matrix as below

$$U_m^{k+1} = (U_m^k) \exp(\hat{t}_u H)$$

Increase the iteration index as $k \rightarrow k + 1$.
until convergence.
end

IV. Experimental Study

A. Experimental Setup and Data Processing

EEGs from the full 27 channels were obtained using Nuamps EEG acquisition hardware with unipolar Ag/AgCl electrodes channels. The sampling rate was 250 Hz with a resolution of 22 b for the voltage range of ±130 mV. A bandpass filter of 0.05–40 Hz was set in the acquisition hardware.

The data set contains 16 subjects who attended two parts of the experiment on separate days. In the first part, there were one motor imagery session and one passive movement session, each of which contained two runs. During the motor imagery session, the data were recorded from subjects performing kinaesthetic motor imagery of the chosen hand or background idle condition. During the passive movement session, EEG data were collected from the subjects with passive movement of the chosen hand performed by a haptic knob robot or performing similar background idle condition. Each run lasted for approximately 16 min and comprised 40 trials of motor imagery or passive movement, and 40 trials of idle state. In the second part, there was one motor imagery session consisting of 2–3 runs. During the EEG recording process, the subjects were asked to avoid physical movement and eye blinking. Thus, there are 80 trials per class in the training session and 80–120 trials per class in the test session, yielding totally 160 training trials and 160–240 test trials for each subject. The details of the experimental setup can be found in [27].

With 27 channels of EEG, $n_c = 27$. For each trial, the EEG from the time segment of 0.5–2.5 s after the cue is extracted, resulting in $n_t = 500$. Single or multiple bandpass filters are applied to the time segment. In particular, for single band setting, the EEG segment is bandpassed by a broadband-pass filter of 4–40 Hz, and then spatial filtering as feature extraction is applied with two pairs of the spatial filters used. For multiple band setting, we construct nine temporal bandpass filters, $4-8, 8-12, \ldots,$ 36–40 Hz, which have been proved to cover the frequency range with the most distinctive ERD/ERS effects [22], [23]. Similarly, spatial filtering with two pairs

TABLE II
CLASSIFICATION RESULTS OF USING DIFFERENT FEATURE SELECTION METHODS (%)

Subject	Training Acc.			Test Acc.		
	MI_0	KL	FS	MI_0	KL	FS
1	80.63	88.13	85.63	50.42	52.08	57.08
2	89.17	87.50	95.00	60.42	56.67	51.25
3	84.17	90.83	91.67	52.92	61.67	68.33
4	94.17	95.00	94.17	71.25	57.50	71.25
5	95.00	95.00	90.83	58.75	59.17	60.83
6	97.50	97.50	96.67	83.33	76.67	85.00
7	97.50	91.67	97.50	78.75	58.33	72.92
8	98.33	99.17	98.33	96.25	95.83	96.25
9	94.38	93.13	94.38	69.17	69.58	69.58
10	85.00	86.67	90.83	65.00	60.00	60.42
11	69.17	95.83	92.50	51.25	50.00	51.67
12	92.50	91.67	92.50	79.58	75.83	79.58
13	73.13	83.13	72.50	75.63	49.58	50.00
14	90.63	93.75	94.38	94.38	75.00	74.58
15	91.88	93.75	89.38	91.88	62.92	67.50
16	90.63	92.50	92.50	90.63	71.67	79.58
mean	88.98	92.20	92.03	67.27	64.09	68.49
median	91.25	92.81	92.50	67.08	60.83	68.96

of the spatial filters is used for each of the nine frequency bands, yielding totally 36 features. For feature selection, eight features are selected based on mutual information, Fisher's ratio, or KL-divergence. Finally, feature vectors are classified using a linear discriminant analysis classifier.

B. Comparing Different Objective Functions in Feature Selection

In this section, we compare the different objective functions in feature selection. To avoid the regularization coefficient λ in (7), we construct a loss function as the ratio of interclass KL-divergence and the within-class KL-divergence, that is

$$\tilde{J}_{kl} = \frac{\sum_{c=+,-} \frac{1}{|\mathcal{Q}^c|} \sum_{j \in \mathcal{Q}^c} D_{kl}(W^T R_j W || W^T R^c W)}{\tilde{D}_{kl}(W^T R^+ W || W^T R^- W)}. \tag{34}$$

The classification results of features selected by different methods are summarized in Table II, where "MI_0" indicates that mutual information based on Parzen window is used for feature selection, "FS" indicates that (21) is used, and "KL" indicates that (34) is used. As shown in Table II, "FS" yields the better results than that of "MI_0," which is regarded as the baseline method. Note that the results in Table II are based on training and testing the model both by motor imagery EEG.

Fig. 2 shows Fisher's ratio and the KL-divergence ratio (34) of different bands, averaged across all subjects. In Fig. 2, the x-axis represents the starting frequency of each bandpass filter, and the blue and red lines represent Fisher's ratio and the KL-divergence ratio, respectively. Despite the fluctuation in Fisher's ratio, both Fisher's ratio and the KL-divergence ratio are lower in the frequency range 8–20 Hz, which is known to be the frequency range with stronger ERD/ERS effects. The results of feature selection show the effectiveness of Fisher's ratio and it is feasible to optimize the spatial filter using it as the objective function.

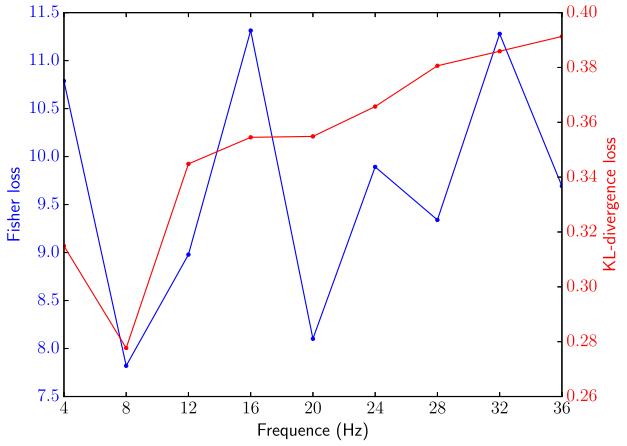


Fig. 2. Comparison between Fisher's ratio and the KL-divergence ratio.

TABLE III
CLASSIFICATION RESULTS USING SINGLE BROAD BAND (%)

Subject	Training Acc.			Test Acc.		
	CSP	MI_s	FS	CSP	MI_s	FS
1	84.38	88.75	88.75	57.08	66.67	66.67
2	70.83	81.67	81.67	51.25	57.92	55.00
3	67.50	90.83	94.17	57.50	55.83	58.75
4	87.50	94.17	94.17	52.50	51.67	51.67
5	90.00	93.33	93.33	50.83	51.25	51.25
6	90.83	94.17	96.67	65.00	67.92	81.67
7	78.33	86.67	86.67	70.00	72.08	72.08
8	99.17	98.33	98.33	95.00	96.67	96.67
9	88.75	90.00	90.00	72.50	70.00	70.00
10	70.00	78.33	81.67	61.67	59.58	58.33
11	60.83	74.17	76.67	47.50	48.75	52.50
12	90.83	93.33	93.33	81.25	78.33	78.33
13	66.88	67.50	68.13	50.83	50.42	50.00
14	90.00	91.88	91.88	72.50	77.08	77.08
15	85.00	85.00	85.63	64.58	65.42	64.58
16	85.00	90.00	90.00	70.00	70.83	70.83
mean	81.61	87.38	88.19	63.75	65.03	65.96
median	85.00	90.00	90.00	63.13	66.04	65.63

C. Spatial Filter Optimization

In this section, we evaluate the performance of the spatial filter optimization method proposed in Section III. For a more comprehensive comparison, we also implement the subspace optimization using mutual information as the objective function. To reduce the computational complexity, the mutual information is calculated using single Gaussian function in the optimization process instead of the Parzen window used in FBCSP, and the details of the spatial filter optimization can be found in Appendix A. To differentiate the mutual information calculation methods, we use MI_s as the notation of the simplified mutual information objective function. During the subspace optimization procedure described in Algorithm 1, the subspace dimension m is set as half of the number of channels, i.e., $m = 0.5n_c$. Upon the completion of the optimization, two pairs of spatial filters are selected via sorting them according to Fisher's ratio of each feature dimension.

Table III shows the classification results of applying different spatial filter optimization methods proposed to EEG filtered by a single broadband. Note that the results in Table III are based on training and testing the model both by motor

TABLE IV
CLASSIFICATION RESULTS USING FILTER BANK (%)

	Training Acc.				Test Acc.			
Subject	FBCSP	$MI_s(MI_0)$	FB-FSSP1	FB-FSSP2	FBCSP	$MI_s(MI_0)$	FB-FSSP1	FB-FSSP2
M1	80.63	79.38	88.75	90.63	50.42	52.08	61.25	61.67
M2	89.17	90.83	95.83	94.17	60.42	58.75	50.00	54.17
M3	84.17	89.17	91.67	91.67	52.92	63.75	66.25	69.17
M4	94.17	95.83	95.83	96.67	71.25	54.58	71.25	77.08
M5	95.00	95.83	94.17	94.17	58.75	58.75	66.25	66.25
M6	97.50	95.83	96.67	96.67	83.33	81.25	84.17	84.17
M7	97.50	97.50	97.50	97.50	78.75	79.17	73.33	75.00
M8	98.33	98.33	98.33	98.33	96.25	95.42	95.83	95.00
M9	94.38	95.00	94.38	94.38	69.17	70.00	68.33	68.33
M10	85.00	85.00	90.00	90.00	65.00	62.08	59.58	59.58
M11	69.17	80.83	94.17	90.83	51.25	46.67	51.67	50.42
M12	92.50	93.33	93.33	92.50	79.58	80.83	80.42	75.00
M13	73.13	75.00	76.25	76.25	49.58	48.33	48.75	48.75
M14	90.63	88.75	95.63	95.00	75.00	71.25	75.83	75.42
M15	91.88	87.50	92.50	91.88	62.92	58.75	71.25	68.33
M16	90.63	90.00	90.63	91.88	71.67	56.25	81.66	79.17
P1	87.50	85.63	85.63	90.63	53.75	61.67	61.67	61.67
P2	85.83	93.33	93.33	94.17	47.08	48.33	48.33	54.17
P3	57.50	94.17	94.17	91.67	50.00	55.00	55.00	69.17
P4	95.00	96.67	96.67	96.67	57.92	65.00	65.00	77.08
P5	95.00	96.67	96.67	94.17	57.92	55.00	55.00	66.25
P6	95.00	96.67	96.67	96.67	75.83	75.00	75.00	84.17
P7	97.50	98.33	98.33	97.50	83.33	81.67	81.67	75.00
P8	98.33	97.50	97.50	98.33	79.58	77.08	77.08	95.00
P9	96.88	98.13	98.13	94.38	61.67	62.08	62.08	68.33
P10	81.67	90.83	90.83	90.00	50.42	58.33	58.33	59.58
P11	77.50	90.00	90.00	90.83	52.50	53.75	53.75	50.42
P12	95.83	95.83	95.83	92.50	75.42	74.58	74.58	75.00
P13	88.75	91.88	91.88	76.25	52.50	49.17	49.17	48.75
P14	97.50	98.75	98.75	95.00	76.25	55.00	55.00	75.42
P15	88.13	93.75	93.75	91.88	57.08	60.42	60.42	68.33
P16	88.13	90.63	90.63	91.88	67.92	75.00	75.00	79.17
mean	89.06	92.09	93.57	93.52	62.45	63.91	66.03	65.89
median	91.25	93.54	94.17	94.17	57.92	61.04	65.63	65.63
***p*-value**	-	**0.02**	**0.003**	**0.008**	-	0.28	**0.01**	**0.02**

imagery EEG. It is found that both spatial filter optimization methods outperform single band CSP, and FS yields the best average classification accuracy.

Table IV shows the classification results of applying different optimization methods to EEG filtered by filter bank with different feature selection methods. For mutual information-based spatial filer optimization, MI_0 is used for feature selection. Moreover, FB-FSSP is implemented in two manners, which are denoted by FB-FSSP1 and FB-FSSP2 in Table IV. In FB-FSSP1, Fisher's ratio spatial filter optimization is applied to EEG signals bandpassed by filter bank of all nine frequency bands at the training stage, followed by feature extraction and feature selection. In FB-FSSP2, FBCSP is applied at the first place, and Fisher's ratio spatial filter optimization is only applied to the selected bands for feature extraction. Thus, compared with FB-FSSP1, the computational complexity of FB-FSSP2 is lower. Moreover, "M" is used to denote using motor imagery data as the training data, and "P" to denote using the passive movement data as the training data in Table IV.

As shown in Table IV, the spatial filter optimization methods based on Fisher's ratio outperform both the baseline and the mutual information spatial filter optimization with higher average and median values. Moreover, the significance of the improvements is validated by paired t-test with $\boldsymbol{p}$-values for both "FB-FSSP1" and "FB-FSSP2" below 0.05. For both settings, the training and test data were recorded on different days, resulting in the significant session-to-session transfer data variation. In particular, the proposed method could still capture the common sensorimotor patterns despite the data-type transfer. Hence, it can found that the proposed method yields a better model generalization and is robust against both the session-to-session and the data-type transfer.

The comparison results are illustrated more clearly in Fig. 3, where the x-axis represents the results of FBCSP, the y-axis represents that of the other three methods in Table IV, and each dot marks one subject. Therefore, the more dots above the $x = y$ line, the more improvements achieved by the proposed method compared to FBCSP. It is found that both FB-FSSP1 and FB-FSSP2 achieve significant improvements for training data. For the test data, it still can be seen that FB-FSSP1 and FB-FSSP2 yield improvements although they are not as significant as that of the training data.

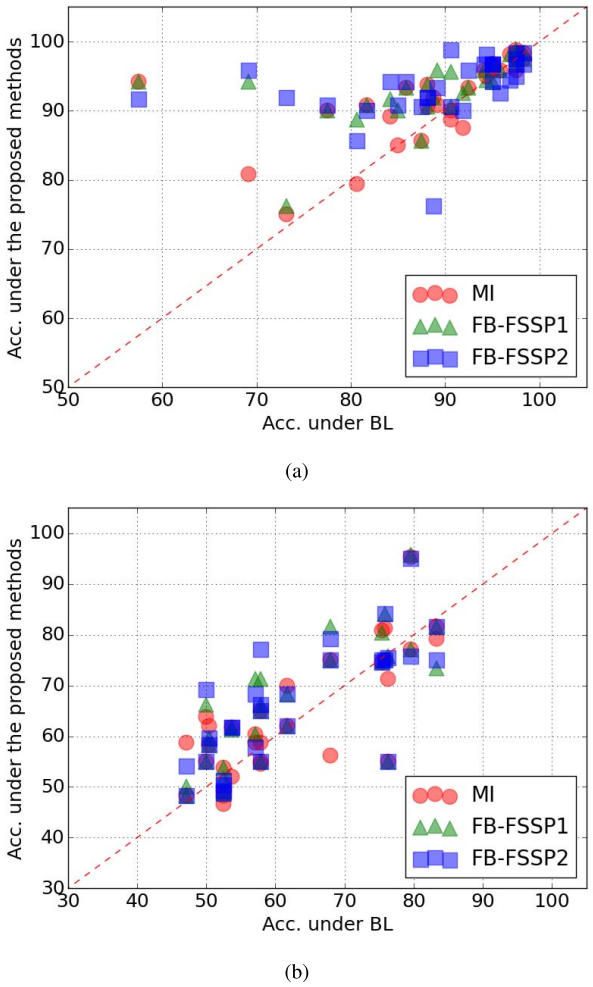


Fig. 3. Classification results comparison. (a) Training data. (b) Test data.

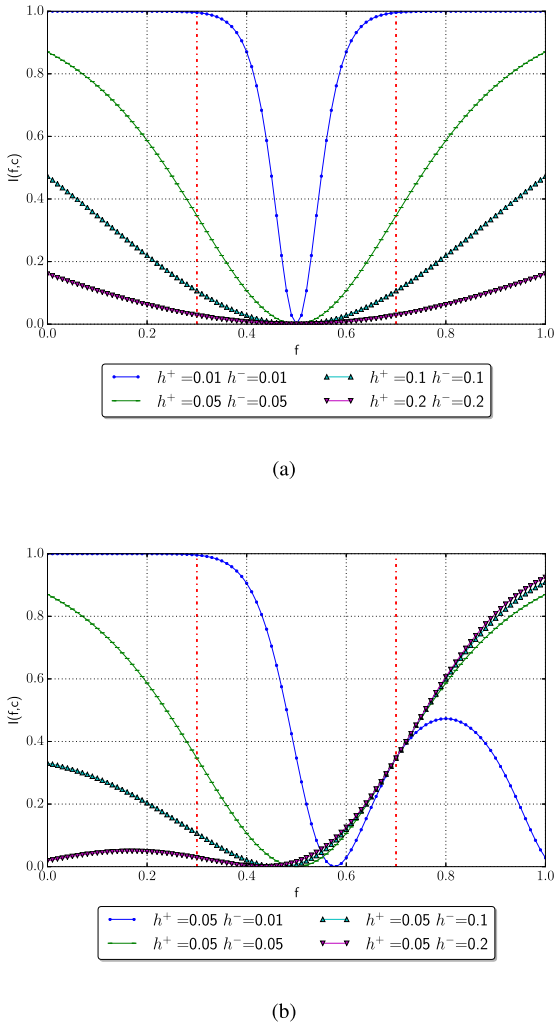


Fig. 4. Mutual information as a function of f.

D. Discussion

1) Nonstationarity in Feature Space and Covariance Matrix Space: In this section, we investigate the objective functions in measuring the data nonstationarity. In Fisher's ratio objective function, the within-class feature distances are calculated in a supervised manner as a measurement of data nonstationarity. Although mutual information is not explicitly used to address the nonstationarity issue when first adopted in FBCSP, feature distances are used to calculate the mutual information. For further analysis, we conduct a simulation study to investigate the relationship between feature distance and mutual information using the mutual information calculation in Appendix A.

Considering a set of 1-d feature f, let the mean features of class $+$ and $-$ be

$$\bar{f}^- = 0.3$$
$$\bar{f}^+ = 0.7 \tag{35}$$

which is a typical pair of averaged CSP features given the constraint $\bar{f}^- + \bar{f}^+ = 1$ in the optimization objective function (2). Assuming that $p(c) = 0.5$, we could calculate the mutual information between the class variable c and a 1-d feature f, $I(f, c)$, by using (36) to (41). The relationship between the feature f and mutual information $I(f, c)$ is shown in Fig. 4, where the x-axis represents f, the y-axis represents $I(f, c)$, and the mean features of both classes are represented by two red dashed-dotted lines. $I(f, c)$ based on different values of h^c are represented by different types of lines. $I(\mathbf{f}, c)$ calculated with the same h^c in (41) for both classes is shown in Fig. 4(a). It can be seen that symmetric about $f = 0.5$, $I(f, c)$ archives the minimal value when $f = 0.5$ and the larger the distance between f and 0.5, the higher the mutual information. In Fig. 4(b), we show the examples of $I(f, c)$ when h^c are different for the two classes. It can be found that although the minimum is no longer $f = 0.5$, generally the mutual information is higher if the distance between f and 0.5 is larger. Different from the minimization of the within-class distances in Fisher's ratio objective function, in the mutual information objective function, the distance between a feature and the feature boundary is maximized. With different values of h^c for different classes, it is likely that by maximizing mutual information, the feature is to be close to the center of one certain class. Moreover, as the calculation of mutual information is unsupervised, it cannot be guaranteed that the feature is closer to the center of the correct class.

While Fisher's ratio and mutual information are objective functions formulated in feature space, Rayleigh coefficient and KL-divergence can be regarded as that in covariance

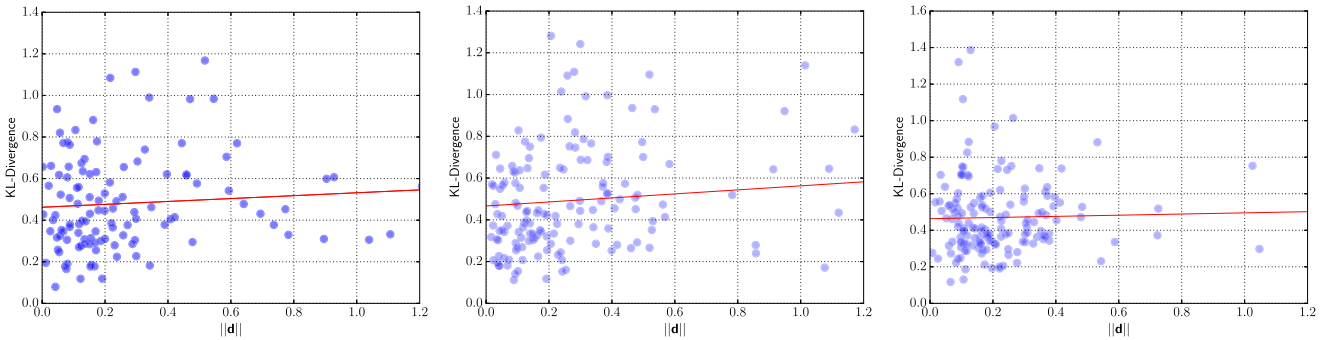


Fig. 5. Relationship between KL-divergence and ||**d**||. (a) Subject 7. (b) Subject 13. (c) Subject 14.

matrix space. When the nonstationarity is not considered in the objective function, maximizing Rayleigh coefficient or KL-divergence is equivalent to maximizing the mean interclass distance due to the joint diagonalization of the average covariance matrices from the two classes. When the nonstationarity is addressed by the regularization term in (8) based on (4), the solution can no longer jointly diagonalize the covariance matrices and the off-diagonal elements also contribute to the objective function in (7). However, only diagonal elements are used as features for classification, which means that there is a gap between the features optimized in the spatial filter design and that those used for classification.

To further investigate such inconsistency, the relationship between KL-divergence and feature distance is shown in Fig. 5, where each dot marks one trial from the training data of the subject indicated in the figure. The x-axis represents $||\mathbf{d}||$ in (17) and y-axis represents the KL-divergence between a trial and the average EEG distribution from the same class after projection based on CSP, i.e., $D_{kl}(WR_jW^T||WR^cW^T)$ with $j \in \mathcal{Q}^c$. Moreover, linear regression between the KL-divergence and $||\mathbf{d}||$ is conducted, the result of which is also shown in Fig. 5. It can be seen that the correlation between KL-divergence and feature distance is not very clear, and the resultant correlation coefficients are not significant in any of the three examples. Given that only diagonal elements are used as features, the KL-divergence objective function is less directly related to classification accuracy in feature space. This could explain why using KL-divergence to select features is not as effective as Fisher's ratio and mutual information. Yet, it is possible that addressing the nonstationarity in the covariance matrix space is less prone to overfitting problem compared with Fisher's ratio objective function, as Fisher's ratio spatial filter optimization achieves more significant improvements in training accuracy than that in test accuracy.

2) Feature Distribution: In our previous work in [25], it is shown that the CSP features from two classes can be modeled as two Gamma distributions. It is proved that the Bayes error could be minimized by maximizing the Rayleigh quotient of the covariance matrices of the two classes under the assumptions that: 1) the two Gamma distributions have the same shape parameter and 2) the shape parameter is a constant independent of the spatial filter. By simplifying the relationship between the shape parameter and the spatial filter, the within-class dissimilarity has not been addressed sufficiently. Thus, in this paper, we approach to the problem with Fisher's ratio so that both interclass and within-class dissimilarity could be taken into consideration. Yet, the assumption that the features follow normal distribution could be one limitation of the proposed method, and we expect that there would be a more comprehensive Bayes learning of the spatial filter.

3) Optimization: We investigate the subjects for whom the proposed method fails to achieve improvements. It is found that for some subjects, the subspace optimization converges at the very beginning of the searching, which yields the results similar to FBCSP. Given that U_m is initialized as the discriminative subspace of the CSP solution, it is possible that the CSP solution is the optimized one, or there is a local minimum so that the searching cannot continue. In our future work, we would conduct more analysis to improve the subspace searching.

4) Relationship With Other Methods: In OSSFN proposed in [24], mutual information is used as the objective function of both temporal and spatial filters design by optimizing coefficient vectors of a group of subspace spatial filters, i.e., CSP or FBCSP spatial filters. In this paper, we propose to search the subspace on manifold with Fisher's ratio or mutual information as objective function. Moreover, OSSFN adopts a "deflation" approach by optimizing the coefficient vectors one by one, which is proved to be less effective than the subspace gradient searching on manifolds [21].

The temporal filters are also optimized using SVM objective function and Rayleigh coefficient objective function in ISSPL and DFBCSP, respectively, while in FBCSP, a soft-optimization approach is adopted by selecting the bands with higher mutual information [13], [14], [22], [23]. Compared with searching globally optimized temporal filters, the computational complexity of the soft-optimization approach is much lower with comparable results. Thus, in this paper, we follow FBCSP by optimizing the temporal filters using the selection strategy. Moreover, in existing joint-temporal-spatial filter optimization methods, usually different objective functions are used for temporal and spatial filter optimization, feature selection, and classification. In the proposed method

shown in Fig. 1, Fisher's ratio is used for spatial filter optimization, feature selection, and classification, which leads to a more unified model.

V. Conclusion

For practical BCI systems, how to optimize spatial filters to extract discriminative EEG features and be robust against EEG nonstationarity is one of the most challenging issues in spatial filter design. Given various objective functions and joint-temporal-spatial analysis methods, the relationship between different optimization methods and feature classification has not been sufficiently investigated.

In this paper, we propose a novel spatial filter design method, which directly addresses both interclass and within-class feature dissimilarity with Fisher's ratio as the objective function. The proposed method is a unified framework as the spatial filter optimization is directly formulated in the feature space, and the inconsistency between spatial filtering and feature extraction could be avoided. In addition, the proposed method does not require regularization parameter selection, which needs to be conducted by cross validation in regularization-based stationary spatial filter design. We implement the proposed method on both single broadband filter and filter bank with feature selection, and it is shown that Fisher's ratio objective function improves classification accuracy for both spatial filter design and feature selection. We also present a systematic attempt to compare it with different objective functions used for feature extraction and feature selection. With experimental and simulation studies, we discuss the advantages and disadvantages of different objective functions.

Appendix
Mutual Information Objective Function

In [22] and [23], mutual information objective function is used for soft optimization by selecting the best spatial filters yielding the highest mutual information. In this paper, we propose to optimize the subspace U_m to maximize the mutual information between class label variables c and feature variable $\mathbf{f}$ as shown in (36), that is

$$I(\mathbf{f}, c) = H(c) - H(c|\mathbf{f}) \tag{36}$$

where

$$H(c) = \sum_{c\in\mathcal{C}} p(c)\log_2 p(c) \tag{37}$$

and

$$H(\mathcal{C}|\mathbf{f}) = -\frac{1}{m}\frac{1}{|\mathcal{Q}^c|}\sum_{i=1}^{m}\sum_{j\in\mathcal{Q}^c}\sum_{c\in\mathcal{C}} p(c|\mathbf{f}_{j,i})\log(p(c|\mathbf{f}_{j,i})) \tag{38}$$

$p(c|\mathbf{f}_{j,i})$, which is the conditional probability of class c given feature $\mathbf{f}_{j,i}$, can be computed as

$$p(c|\mathbf{f}_{j,i}) = \frac{p(\mathbf{f}_{j,i}|c)P(c)}{\sum_{c=+,-} p(\mathbf{f}_{j,i}|c)}. \tag{39}$$

The conditional probability of $\mathbf{f}_{j,i}$ given class c, $p(\mathbf{f}_{j,i}|c)$, can be estimated using Gaussian function as

$$\begin{aligned} p(\mathbf{f}_{j,i}|c) &= \varphi\left(\mathbf{f}_{j,i} - \bar{\mathbf{f}}_i^c, h_i^c\right) \\ &= \varphi\left(\mathbf{d}_{j,i}^c, h_i^c\right) \end{aligned} \tag{40}$$

where

$$\varphi(\mathbf{d}, h) = \frac{1}{\sqrt{2\pi}} e^{-\frac{||\mathbf{d}||}{h^2}}. \tag{41}$$

The smoothing parameter h_i^c can be calculated as

$$h_i^c = \left(\frac{4}{3|\mathcal{Q}^c|}\right)^{0.2}\sigma_i^c \tag{42}$$

where σ_i^c is the standard deviation of the ith dimension of features belonging to class c.

Thus, with (36)–(42), the subspace U_m can be optimized using the following optimization objective function:

$$\hat{U}_m = \arg\max_{U_m} J_{mi}(U_m) \quad \text{s.t. } U_m^T U_m = I \tag{43}$$

where

$$J_{mi}(U_m) = I(\mathbf{f}, c). \tag{44}$$

Remark 3: To reduce the computation complexity of calculating the gradient, we adopt a simplified way to estimate mutual information with $p(\mathbf{f}_{j,i}|c)$ shown in (40) compared with the mutual information calculation in [22] and [23].

References

[1] H. Cecotti, M. P. Eckstein, and B. Giesbrecht, "Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2030–2042, Nov. 2014.

[2] F. Qi, Y. Li, and W. Wu, "RSTFC: A novel algorithm for spatio-temporal filtering and classification of single-trial EEG," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3070–3082, Dec. 2015.

[3] J. R. Wolpaw *et al.*, "Brain-computer interface technology: A review of the first international meeting," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 164–173, Jun. 2000.

[4] V. Gandhi, G. Prasad, D. Coyle, L. Behera, and T. M. McGinnity, "Quantum neural network-based EEG filtering for a brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 278–288, Feb. 2014.

[5] J. Lu, K. Xie, and D. J. McFarland, "Adaptive spatio-temporal filtering for movement related potentials in EEG-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 4, pp. 847–857, Jul. 2014.

[6] M. D. Fox, A. Z. Snyder, J. L. Vincent, and M. E. Raichle, "Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior," *Neuron*, vol. 56, no. 1, pp. 171–184, Oct. 2007.

[7] F. de Pasquale *et al.*, "Temporal dynamics of spontaneous MEG activity in brain networks," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 13, pp. 6040–6045, 2010.

[8] A. Y. Kaplan, A. A. Fingelkurts, A. A. Fingelkurts, S. V. Borisov, and B. S. Darkhovsky, "Nonstationary nature of the brain activity as revealed by EEG/MEG: Methodological, practical and conceptual challenges," *Signal Process.*, vol. 85, no. 11, pp. 2190–2212, Nov. 2005.

[9] Z. J. Koles, "The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG," *Electroencephalogr. Clin. Neurophysiol.*, vol. 79, no. 6, pp. 440–447, Dec. 1991.

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

[11] G. Pfurtscheller and A. Aranibar, "Evaluation of event-related desynchronization (ERD) preceding and following voluntary self-paced movement," *Electroencephalogr. Clin. Neurophysiol.*, vol. 46, no. 2, pp. 138–146, Feb. 1979.

[12] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, 2006.

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

[14] 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, Jun. 2008.

[15] X. Li, H. Zhang, C. Guan, S. H. Ong, K. K. Ang, and Y. Pan, "Discriminative learning of propagation and spatial pattern for motor imagery EEG analysis," *Neural Comput.*, vol. 25, no. 10, pp. 2709–2733, 2013.

[16] X. Li, C. Guan, H. Zhang, K. K. Ang, and S. H. Ong, "Adaptation of motor imagery EEG classification model based on tensor decomposition," *J. Neural Eng.*, vol. 11, no. 5, p. 056020, 2014.

[17] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–362, Feb. 2011.

[18] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain–computer interfacing," *J. Neural Eng.*, vol. 9, no. 2, p. 026013, Feb. 2012.

[19] W. Samek, F. C. Meinecke, and K.-R. Müller, "Transferring subspaces between subjects in brain–computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 8, pp. 2289–2298, Aug. 2013.

[20] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 610–619, Apr. 2013.

[21] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Rev. Biomed. Eng.*, vol. 7, pp. 50–72, Apr. 2014.

[22] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, no. 39, 2012.

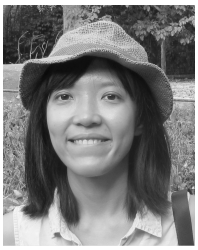
[23] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Mutual information-based selection of optimal spatial–temporal patterns for single-trial EEG-based BCIs," *Pattern Recognit.*, vol. 45, no. 6, pp. 2137–2144, Jun. 2012.

[24] H. Zhang, Z. Y. Chin, K. K. Ang, C. Guan, and C. Wang, "Optimum spatio-spectral filtering network for brain–computer interface," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 52–63, Jan. 2011.

[25] H. Zhang, H. Yang, and C. Guan, "Bayesian learning for spatial filtering in an EEG-based brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1049–1060, Jul. 2013.

[26] M. D. Plumbley, "Geometrical methods for non-negative ICA: Manifolds, Lie groups and toral subalgebras," *Neurocomputing*, vol. 67, pp. 161–197, Aug. 2005.

[27] K. K. Ang, C. Guan, C. Wang, K. S. Phua, A. H. G. Tan, and Z. Y. Chin, "Calibrating EEG-based motor imagery brain-computer interface from passive movement," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug./Sep. 2011, pp. 4199–4202.



Xinyang Li received the Ph.D. degree from the NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore, in 2014, and the B.E. degree from Beihang University, Beijing, China, in 2010.

She was a Research Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. She is currently a Senior Research Officer with the BCI-NE Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K. Her current research interests include brain computer interface, machine learning, and neuro-technology.

Cuntai Guan (SM'03) received the Ph.D. degree in electrical and electronic engineering from Southeast University, Nanjing, China, in 1993.

He is currently a Principal Scientist and the Department Head with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. He is also the Co-Director of the Rehabilitation Research Institute of Singapore, and an Adjunct Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He was the founder and the Director of the Brain-computer Interface Laboratory in the Institute for Infocomm Research and several Medtech research programmes and departments with the Institute for Infocomm Research. His current research interests include neural and biomedical signal processing, machine learning and pattern recognition, neural and cognitive process and its clinical applications, brain–computer interface algorithms, systems and applications, neural image processing, medical system, and device research and development.

Haihong Zhang (S'01–M'04) received the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2005, and the master's and bachelor's degrees in electronics engineering from the University of Science and Technology of China, Hefei, China, and Hefei Technological University, Hefei, in 1997 and 2000, respectively.

He is currently the Lab Head for Neural Signal and Sensing with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His current research interests include pattern recognition, brain-computer interface, and computational neuroscience.

Kai Keng Ang (GS'05–M'07–SM'15) received the B.A.Sc. (Hons.) and Ph.D. degrees in computer engineering from Nanyang Technological University, Singapore.

He is currently the Head of Brain-Computer Interface Laboratory and a Senior Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. He is also an Adjunct Assistant Professor with the School of Computer Science and Engineering, Nanyang Technological University. His current research interests include brain-computer interfaces, computational intelligence, machine learning, pattern recognition, and signal processing.