Simultaneously optimizing spatial spectral features based on mutual information for EEG classification

Jianjun Meng, Lin Yao, Xinjun Sheng, Dingguo Zhang and Xiangyang Zhu

Abstract—High performance of brain-computer interface (BCI) needs efficient algorithms to extract discriminative features from raw electroencephalography (EEG) signals. In this paper, we present a novel scheme to extract spatial spectral features for motor imagery based BCI. The learning task is formulated by maximizing the mutual information between spatial spectral features (MMISS) and class labels, by which a unique objective function directly related to Bayes classification error is optimized. The spatial spectral features are assumed to follow a parametric Gaussian distribution, which has been validated by the normal distribution Mardia’s test, and under this assumption the estimation of mutual information (MI) is derived. We propose a gradient based alternative and iterative learning algorithm to optimize the cost function and derive the spatial and spectral filters simultaneously. The experimental results on dataset IVa of BCI competition III and dataset IIa of BCI competition IV show that the proposed MMISS is able to efficiently extract discriminative features from motor imagery-based EEG signals to enhance the classification accuracy compared to other existing algorithms.

Index Terms—Brain-Computer Interface (BCI), mutual information, spatial spectral feature, filter optimization

I. INTRODUCTION

A Brain-computer interface (BCI) provides a potential way for paralyzed patients to restore lost motor function which has been impaired by devastating neuromuscular disorders [1]. Motor imagery based BCI is a promising implementation, in which users perform the imagination or mental rehearsal of motor actions without actually moving [2]. During the imagination of hand and foot movements, the suppression or augmentation of cortical rhythmic activity shows distinct spatial asymmetry in the regions of motor or somatosensory cortex. This neurophysiological phenomenon called event-related desynchronization (ERD) and event-related synchronization (ERS) accompanying real and imagined body part movement lays the foundation for classifying motor imagery EEG signals [3]. However, the rhythm activity derived by noninvasive EEG recorded from the scalp is blurred due to volume conduction and is easily contaminated by other sources and artifacts [4]. Hence, advanced signal processing algorithms are required for high performance of BCI.

It’s well known that feature extraction, including spatial filters and temporal/spectral filters, plays a vital role in distinguishing patterns of different motor imaginations1. Common spatial pattern (CSP) is a highly successful spatial filtering technique for this purpose [5], [6]. This technique uses covariance analysis to extremely amplify the class differences in spatial scale, but neglects the frequency information which is important for portraying rhythmic activities [3]. To compensate for this, wide or narrow band temporal/spectral filters are often applied to the raw EEG signals before spatial filtering. In this study, we consider the bandpass temporal and spectral filters as the same filters and choose to use the spectral filters for consistency.

However, the frequency band is generally subject specific. To avoid the exhausting work of selecting a frequency band manually for each subject, the combined optimization of spatial and spectral filters has gained much attention for improving the performance of a BCI system. In recent years, great effort has been devoted to this area and can be categorized into three schemes. 1) The first scheme contains the embedded solutions such as common spatial spectral pattern (CSSP) [7] and its improved version of common sparse spectral spatial pattern (CSSSP) [8]. The core of their ideas is that embed one and several time-delayed signals into raw EEG signals and then use the CSP algorithm to optimize spatial and finite impulse response (FIR) filters simultaneously. Recently, Higashi et al [9] proposed a new algorithm of optimizing FIR filter banks and spatial patterns sequentially (named DFBCSP). They designed several orthogonal spectral filters and obtained local maximums of cost function by sequential optimization. 2) The second category optimizes the bandpass filters equivalently in the frequency domain, and thus these filters are called spectral filters. Different cost functions are used to optimize the spatial and spectral filters iteratively by spectrally weighted common spatial pattern (SWCSP) [10]. Similarly, iterative spatial spectral pattern learning (ISSPL) [11] optimizes the spectral filters by a classifier for the purpose of good generalization performance. 3) The third scheme utilizes several filter banks and selects a reduced set of features from pre-defined narrow bands. The sub-band common spatial pattern and filter bank common spatial pattern (FB CSP) [12], [13] belong to this category. This delicate algorithm (FB CSP) with its multi-class extension achieved the best performance on both Datasets Ia and Ib in the BCI Competition IV [14], [15]. Later, Zhang et al. [16] proposed an optimum spatio-spectral filtering network (OSSFN) to optimize spatial filters by maximizing nonparametric based mutual information.

1The bandpass filters applied or optimized in the time domain or frequency domain are named as temporal and spectral filter respectively and they are functionally equivalent in our context.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2014.2345458, IEEE Transactions on Biomedical Engineering

In this section, the CSP method and other related works are comparatively reviewed.

\[ \text{TABLE I} \]

<table>
<thead>
<tr>
<th>Method</th>
<th>Formulations and Notations</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSP</td>
<td>[ \max_w E_{y(i)=1}[\sigma^2(w^TX)] ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{s.t. } \sum_{\omega=1}^{2} E_{y(i)=\omega}[\sigma^2(w^TX)]=1 ]</td>
</tr>
<tr>
<td>CSSP</td>
<td>[ \max_w E_{y(i)=1}[\sigma^2(w^TX^T)] ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{s.t. } \sum_{\omega=1}^{2} E_{y(i)=\omega}[\sigma^2(w^TX_i^T)] = 1 ]</td>
</tr>
<tr>
<td>CSSPP</td>
<td>[ \max_{\omega} \max_w { \sum_{\tau=0}^{P-1} \sum_{j=1}^{N} \omega(j)h(j+\tau)</td>
</tr>
<tr>
<td></td>
<td>[ \text{s.t. } \sum_{j=1}^{N} \omega(i)C_i^2(X_i^T) = \text{max non-negative regularization coefficient} ]</td>
</tr>
<tr>
<td>DFBCSP</td>
<td>[ \max_{w,b_j,k} { \sum_{\omega=1}^{2} \omega(i)E_{y(i)=\omega}[\sigma^2(X_i^Tw+b_j,k)] } ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{s.t. } \omega(i)\omega(i)</td>
</tr>
<tr>
<td>SWCSP</td>
<td>[ \max_w { \sum_{\omega=1}^{2} \omega(i)E_{y(i)=\omega}[\sigma^2(X_i^Tw+b)] } ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{s.t. } \text{Fisher's criterion with constraint} ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{Note: } \sigma^2(\cdot) \text{ is a variance based function of spatial and temporal filter} ]</td>
</tr>
<tr>
<td>ISSPL</td>
<td>[ \max_w { \sum_{\omega=1}^{2} \omega(i)E_{y(i)=\omega}[\sigma^2(X_i^Tw+b)] } ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{s.t. } \text{maximally margin classifier} ]</td>
</tr>
<tr>
<td>FBCSP</td>
<td>[ {h_1, \ldots , h_F }, \sum_{\omega=1}^{2} E_{y(i)=\omega}[\sigma^2(w^TX_i(h_k))] ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{Select best features by feature selection method} ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{Note: } X_i(h_k) \text{ means the block signal } X_i \text{ is filtered by} ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{bandpass filter } h_k \text{ (the } k\text{th filter bank).} ]</td>
</tr>
<tr>
<td>OSSFN</td>
<td>[ {h_{opt}, w_{opt}} = \max (I(A, \Omega)) ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{Note: } I(A, \Omega) \text{ is the mutual information between feature set } A ]</td>
</tr>
<tr>
<td></td>
<td>[ \text{and class label set } \Omega ]</td>
</tr>
</tbody>
</table>

Note: \( \sigma^2(\cdot) \) means the variance of a feature vector or a variance based function. \( E_{y(i)=\omega}[\cdot] \) means the expectation of a feature vector, the class label \( Y(i) \) of which is \( \omega \) and \( i \) denotes the feature vector belongs to \( i\text{th trial.} \)

A brief comparison of various algorithms is listed in Table I. Throughout the paper, the spatial filter is denoted by the bold lower-case letter \( w \). Because in the previous researches, the bandpass filters applied or optimized in the time domain and frequency domain are explicitly or implicitly named as temporal filters and spectral filters respectively, they are denoted by the bold lower-case letters \( h \) and \( b \), accordingly. The element of a temporal filter is denoted by \( h(j) \), where \( j \) is a positive integer. Note that \( h_1 \) means the \( j\text{th} \) temporal filter (sometimes \( k \) is used) rather than a component. A block of raw EEG signals is organized in a matrix form and denoted as an upper case letter (with a subscript) \( X_i \in R^{M \times N} \), where \( i \) means the block belongs to \( i\text{th trial,} \ M \) is the number of
channels and $N$ is the number of samples in the block. The delayed signals are denoted as $X_i(\tau) \in R^{M \times N}$, which means the samples are delayed by time $\tau$. $T$ is the transpose of a vector or a matrix. Other specific notation is explained in the table.

CSP: Aim to maximize one class covariance while minimize the other class covariance. Its equivalent optimization problem is to maximize one specific class covariance with the normalization constraint of two class covariance. The solution is given by the generalized eigenvalue decomposition [5], [6].

CSSP: Double the channels of raw EEG sample by concatenating its one time-delayed signals into the original ones. Similar to CSP, the solution is also obtained by the generalized eigenvalue decomposition. The CSSP derives both spatial filters and several first order FIR filters simultaneously [7].

CSSSP: Combine several time-delayed signals with raw EEG sample. With the assumption that the signals are approximately stationary in short time, the objective function is simplified and the spatial filters and a common high order FIR filter can be solved simultaneously. In order to avoid overfitting, the regularization coefficient $C$ is applied to this problem when the number of coefficients in FIR filter is relatively large. The optimization technique like gradient or line-search method is suggested to solve this problem [8].

DFBCSV: Design high order FIR filters for raw EEG signals and then spatial filters are applied. The DFBCSV maximizes the expectation of a variance based feature for one specific class under the normalization constraint of expectation of variance based features for both classes. The optimization is conducted by solving two subproblems of generalized eigenvalue decomposition sequentially and alternatively [9].

SWCSP: Transform the linear time-invariant bandpass filters from the time domain to the frequency domain. The SWCSP optimizes the spatial filters using the CSP algorithm and the spectral filters by Fisher’s criterion (with positive coefficients constraint) sequentially and iteratively [10].

ISSPL: Solve the bandpass filters in the frequency domain. Similar to SWCSP, the ISSPL optimizes the spatial filters by the CSP technique. However, the spectral filters are equivalently solved by a max-margin machine. The algorithm of ISSPL iteratively solves the spatial and spectral filters until the stop condition is satisfied [11].

FBCSV: Construct non-overlapped filter banks and filter the raw EEG signals by these multiple bandpass filters. Then CSP algorithm is used to obtain spatial filters from each bandpass filtered EEG signals. Feature selection based on the best individual mutual information is employed to select the best subset of features [13].

OSSFN: Bandpass raw EEG signals by multiple filter banks (similar to FBCSV) firstly. Derive the spatial filters by maximizing the mutual information of features and class labels. The mutual information is estimated by the nonparametric probability density function (kernel density estimation). Finally, the features in the optimum frequency band are assumed to have the maximum mutual information [16].

III. MAXIMIZING MUTUAL INFORMATION OF SPATIAL SPECTRAL FEATURES (MMISS)

Although CSP is an efficient algorithm to extract discriminative features of two-class motor imagination, the direct connection between Bayes classification error and CSP is only established based on the assumption that the EEG signals conditioned on any class follow a Gaussian distribution recently [19]. Whether the combination of spatial and bandpass filters such as CSSP, CSSSP and DFBCSV hold the connection is still unknown. For SWCSP and ISSPL, experimental results show that the iteration usually helps to improve the classification accuracy, however, the convergence of iteration is not always guaranteed. FBCSP and OSSFN take advantage of fixed filter banks, but the flexibility of subject specific spectral filters might be lost at the same time.

In this section, we introduce the MMISS to extract discriminative spatial and spectral features by maximizing mutual information which is directly related to Bayes classification error. Instead of solving a bandpass filter $h$ in the time domain, we compute it equivalently in the frequency domain like SWCSP and ISSPL. The main purpose of this algorithm is to find the spatial filter $w$ and spectral filter $b$ iteratively and simultaneously.

A. Design optimization formulation

In this study, we consider the log-power features based on spatial and spectral filtered EEG signals. The procedure of transforming raw EEG signals $X_i = [x_i(1), \cdots, x_i(N)]$ into the feature vector $a_i$ and related cost function with respect to spatial filters $W$ and spectral filter $b$ comprise the following procedures.

1) Bandpass filtering: Suppose $B \in R^{N \times N}$ is a linear time-invariant bandpass filter ($B$ is used instead of $h$ introduced before because $B$ is a matrix here. The bandpass filter and spectral filter are used interchangeably according to the meaning of the context.) which can be formulated as a circulant matrix. To derive the interested rhythm activity of EEG signals, the bandpass filter is necessarily applied to the raw EEG signals. We then have the filtered signals:

$$Z_i = X_iB$$

2) Spatial filtering: A linear projection (named spatial filters) is used to reduce the dimension of bandpass filtered EEG signals $Z_i$. The spatial and bandpass filtered EEG signal is formulated as:

$$Y_i = W^TZ_i$$

where $W = [w_1, w_2, \cdots, w_{n_t}] \in R^{M \times n_t}$, $n_t$ is the number of spatial filters.

3) Log-power feature: The mean power of a block of EEG signals after spatial and bandpass filtering is used as the log-power feature.

$$a_i = log(diag(\frac{1}{N}W^TX_iBB^TX_i^TW))$$

where the function of $diag(\cdot)$ returns a column vector formed by the diagonal elements of $(\cdot)^T$ if it is a square matrix and returns a diagonal matrix with the elements of $(\cdot)^T$ when it is
a vector. Note that the logarithm operation is used for numerical consideration. Since we use the assumption of Gaussian probability density function in the following, the log operation moderates the scalability of the exponential operation. On the other hand, the similar log operation is adopted in the CSP algorithm to approximate normal distribution of the data.

4) Fourier transformed log-power feature: As stated before, we solve the bandpass filter in the frequency domain instead of directly solving it in the time domain, therefore we name this bandpass filter as spectral filter. It is known that a circulant matrix can be diagonalized by the discrete Fourier transform [22]. For the bandpass filter B, which is circulant, the diagonalization is achieved by

\[ F^T B = \text{diag}(b_1, \ldots, b_N)F^T \]

where \( F \in \mathbb{C}^{N \times N} \) is the Fourier matrix, \( \mathbb{C} \) means the set of all complex numbers and denote \( b = [b_1, \ldots, b_N]^T \). Then the spatial spectral feature is derived by

\[
\begin{align*}
\mathbf{a}_i &= \log(\text{diag}(\frac{1}{N}W^TX_iBBX_i^TW)) \\
&= \log(\text{diag}(\frac{1}{N}W^TX_iFF^TX_iBB^TFF_i^TW)) \\
&= \log(\text{diag}(\frac{1}{N}W^T\tilde{X}_i\text{diag}(b_1^2, \ldots, b_N^2)\tilde{X}_i^TW)) 
\end{align*}
\]

where \( \tilde{X}_i \) denotes the discrete Fourier transformed data matrix.

5) Optimization objective: Denote the set of feature vectors and the set of class labels by \( \mathcal{A} = \{a_i\} \) and \( \Omega = \{\omega\} \), respectively. Mutual information between the set of feature vectors and class labels are used to formulate the objective function for optimizing spatial and spectral filters.

\[
\max_{\mathbf{A}, \Omega} I(\mathcal{A}, \Omega) = H(\mathcal{A}) - \sum_{\omega \in \Omega} H(\mathcal{A}|\omega)P(\omega) 
\]

where \( H(\mathcal{A}) \) and \( H(\mathcal{A}|\omega) \) are the entropy of feature vector and the conditional entropy of feature vector given a particular class \( \omega \) (e.g. \( \omega = 1 \) or \( \omega = 2 \) represents the left hand or right hand motor imagery movement), respectively.

The mutual information has been proved to have connection with minimum Bayes classification error via lower and upper bounds [23]. Maximizing the mutual information of \( \mathcal{A} \) and \( \Omega \) implies minimizing the error probability. Recently, studies of linear feature extraction by maximizing mutual information show the power of this method [21], [24]. Hence, we choose (6) to be our objective function.

Since it is difficult to simultaneously find all parameters of \( W \) and \( \mathbf{b} \), we consider two subproblems of optimizing spatial and spectral filters alternatively and iteratively [25].

### B. Optimize spatial filter

The first subproblem is to optimize \( W \) while fixing \( \mathbf{b} \). Define

\[
\tilde{R}_{xi} = \tilde{X}_i\text{diag}(b_1^2, \ldots, b_N^2)\tilde{X}_i^T
\]

Recall that \( W = [\mathbf{w}_1 \mathbf{w}_2 \ldots \mathbf{w}_{n_l}] \in \mathbb{R}^{M \times n_l} \) is a linear projection matrix that we want to find (also called spatial filters). Then the spatial spectral feature is written as

\[
\mathbf{a}_i = \log(\text{diag}(\frac{1}{N}W^T\tilde{R}_{xi}W))
\]

We consider the mutual information between the feature vector variable \( \mathcal{A} \) and the class label variable \( \Omega \). Then the objective function (6) can be written as (fixing \( \mathbf{b} \))

\[
\max_W I(\mathcal{A}, \Omega) = H(\mathcal{A}) - \sum_{\omega \in \Omega} H(\mathcal{A}|\omega)P(\omega)
\]

In order to estimate \( H(\mathcal{A}) \) and \( H(\mathcal{A}|\omega) \), we assume that the spatial spectral features conditioned on any class \( p(\mathbf{a}|\omega) \) follow the Gaussian distributions.

\[
p(\mathbf{a}|\omega) = (2\pi)^{-\frac{m_p}{2}}|\Psi_{\omega}|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathbf{a}^T\Psi_{\omega}^{-1}\mathbf{a}}
\]

where \( \mathbf{r} \) denotes the term \( \mathbf{a} - \mathbf{\mu}_\omega \), \( \mathbf{\mu}_\omega = \frac{1}{n_\omega}\sum_{k=1}^{n_\omega} a_k \). \( \Psi_{\omega} = \frac{1}{n_\omega}\sum_{k=1}^{n_\omega}(a_k - \mathbf{\mu}_\omega)(a_k - \mathbf{\mu}_\omega)^T \). \( n_\omega \) is the number of trials in the training data belonging to class \( \omega \). \( \mathbf{\mu}_\omega \) and \( \Psi_{\omega} \) are the mean vector and covariance matrix for specific class \( \omega \), respectively. Since the number of samples for feature vector \( a_k \) is often limited, \( \Psi_{\omega} \) usually takes a diagonal matrix form. Then the entropy of feature vector given class \( \omega \) is expressed by

\[
H(\mathcal{A}|\omega) = -E[\log(p(\mathbf{a}|\omega))] = \frac{1}{2}\log((2\pi e)^{n_p}|\Psi_{\omega}|)
\]

The marginal distribution \( p(\mathbf{a}) \), however, does not follow a Gaussian distribution since \( p(\mathbf{a}) = \sum_{\omega=1}^{2}P(\omega)p(\mathbf{a}|\omega) \). It’s a mixture of Gaussian distribution. The entropy of feature vector \( \mathcal{A} \) can be viewed as an expectation of the function \( \log p(\mathbf{a}) \) [26]. Then the entropy can be estimated by samples of feature vector \( a \).

\[
H(\mathcal{A}) = -E[\log(p(\mathbf{a}))] \simeq -\frac{1}{n_a}\sum_{k=1}^{n_a} \log(p(a_k))
\]

where \( n_a \) is the number of empirical samples of feature vector \( a \) (usually equals to the number of trials) and \( p(\mathbf{a}_k) = \sum_{\omega=1}^{2}P(\omega)p(a_k|\omega) \).

The gradient based optimization technique is used to solve the cost function of maximizing mutual information. Then we solve the gradient of mutual information \( I(\mathcal{A}, \Omega) \) with respect to the spatial filter \( \mathbf{w}_l \)

\[
\nabla_{\mathbf{w}_l} I(\mathcal{A}, \Omega) = \nabla_{\mathbf{w}_l} H(\mathcal{A}) - \sum_{\omega \in \Omega} \nabla_{\mathbf{w}_l} H(\mathcal{A}|\omega)P(\omega)
\]

where

\[
\nabla_{\mathbf{w}_l} H(\mathcal{A}|\omega) = \frac{\partial(\frac{1}{2}\log((2\pi e)^{n_p}|\Psi_{\omega}|))}{\partial_{\mathbf{w}_l}}
\]

The last equality is due to \( \frac{\partial \Psi_{j,l}}{\partial_{\mathbf{w}_l}} \) is a function of \( \mathbf{w}_l \) if and only if \( j = l \). Then we turn to solve

\[
\nabla_{\mathbf{w}_l} H(\mathcal{A}|\omega) = \frac{1}{2}\log((2\pi e)^{n_p}|\Psi_{\omega}|)) = \frac{1}{2}\log(\frac{1}{n_\omega}\sum_{k=1}^{n_\omega}(a_k - \mathbf{\mu}_\omega)^2)
\]

\[
\frac{\partial \Psi_{j,l}}{\partial_{\mathbf{w}_l}} = \frac{2}{n_\omega}\sum_{k=1}^{n_\omega}(a_k - \mathbf{\mu}_\omega)^2
\]

\[
\frac{\partial \Psi_{j,l}}{\partial_{\mathbf{w}_l}} = \frac{1}{n_\omega}\sum_{k=1}^{n_\omega}(a_k - \mathbf{\mu}_\omega)^2 - \frac{1}{n_\omega}\sum_{k=1}^{n_\omega}(a_k - \mathbf{\mu}_\omega)(a_k - \mathbf{\mu}_\omega)^T
\]

Diagonal form of the covariance matrix for each class implies uncorrelation among each element of feature vector \( a_k \). This operation is similar to the independency assumption for Naive Bayesian, which is one of the most effective inductive learning.
where \( \frac{\partial \mathbf{w}_l}{\partial w_i} = \frac{1}{n} \sum_{j=1}^{n} \frac{\partial a_{i,j}^{l}}{\partial w_i} = \frac{1}{n} \sum_{j=1}^{n} 2 \frac{\partial \Phi_{k,j}^{l}}{\partial \mathbf{e}_j^{\omega}} \mathbf{w}_l \), hence substitution of the above equalities to Eq. (15) gives,

\[
\frac{\partial \Psi_{l}}{\partial \mathbf{w}_l} = \frac{2}{n} \sum_{k=1}^{n} \left( a_{k,l} - \mathbf{a}_l^{\omega} \right) \left( 2 \frac{\partial \Phi_{k,l}^{l}}{\partial \mathbf{e}_k^{\omega}} - \frac{1}{n} \sum_{j=1}^{n} 2 \frac{\partial \Phi_{k,j}^{l}}{\partial \mathbf{e}_j^{\omega}} \mathbf{w}_l \right)
\]

Therefore,

\[
\nabla_{\mathbf{w}_l} H(\mathbf{A}|\omega) = \frac{2}{n} \sum_{k=1}^{n} a_{k,l}^{\omega} - \mathbf{a}_l^{\omega} \mathbf{w}_l
\]

The derivation of \( \nabla_{\mathbf{w}_l} H(\mathbf{A}) \) is performed similarly. With the gradient information, our iterative optimization algorithm updates a spatial filter by

\[
\mathbf{w}_l^{(iter+1)} = \mathbf{w}_l^{(iter)} + \lambda \nabla_{\mathbf{w}_l} I(\mathbf{A}, \Omega)
\]

\( \lambda \) is the step size. Given a proper initial value of \( \mathbf{w}_l^{(0)} \), the gradient based iterative algorithm usually outputs a better spatial filter after iteration steps. Note that, since the cost function of (9) is not usually a convex function, the gradient-based optimization procedure may reach a local maximum instead of the global one. This implies that the initial value of \( \mathbf{w}_l^{(iter)} \) is relatively important for the optimization problem. Here we use the spatial filters given by the CSP algorithm to be the initial values.

One spatial filter can be obtained by the above procedure. For simultaneous optimization of all the spatial filters in \( \mathbf{W} \), we propose a gradient-based method by considering partial derivatives with respect to a joint vector, which is formed by concatenating all the spatial filters together. The details of the method are provided in Appendix A.

C. Optimize spectral filter

The second subproblem is to optimize \( \mathbf{b} \) while fixing \( \mathbf{W} \). Similarly the objective function (6) can be written as

\[
\max_{\mathbf{b}} I(\mathbf{A}, \Omega) = H(\mathbf{A}) - \sum_{\omega} H(\mathbf{A}|\omega)P(\omega)
\]

Now consider the spatial spectral feature extractor (5)

\[
\mathbf{a}_i = \log(\text{diag}(\frac{1}{N} \mathbf{W}^T \mathbf{X}_i \text{diag}([b_1^2, \cdots, b_N^2]) \mathbf{X}_i^T \mathbf{W}))
\]

Let \( a_{i,l} \) be a specific component of \( \mathbf{a}_i \), which can be expressed by

\[
a_{i,l} = \log(\frac{1}{N} \mathbf{w}_l^T \mathbf{X}_i \text{diag}([b_1^2, \cdots, b_N^2]) \mathbf{X}_i^T \mathbf{w}_l)
\]

\[
= \log(\frac{1}{N} \mathbf{w}_l^T \mathbf{X}_i \text{diag}(\mathbf{b}^2) \mathbf{X}_i^T \mathbf{w}_l)
\]

\[
= \log(\frac{1}{N} \mathbf{b}^T \Lambda_{i,l} \mathbf{b})
\]

where the wave line \( \mathbf{X} \) denotes the conjugate of a complex value and the diag() function stands for the diagonal vector of a matrix. Here \( \Lambda_{i,l} \) is the diagonal matrix where the diagonal elements are \( \text{diag}(\mathbf{w}_l^T \mathbf{X}_i \mathbf{w}_l^T) \). \( \mathbf{b} = [b_1, \cdots, b_N]^T \) and \( \beta = \text{diag}(\mathbf{b} \mathbf{b}^T) \). From equation (20) we can see it has similar formulation to (7). Consequently, we can solve the spectral filter by similar optimization algorithm. The gradient of mutual information \( I(\mathbf{A}, \Omega) \) with respect to spectral filter \( \mathbf{b} \) is

\[
\nabla_{\mathbf{b}} I(\mathbf{A}, \Omega) = \nabla_{\mathbf{b}} H(\mathbf{A}) - \sum_{\omega} \nabla_{\mathbf{b}} H(\mathbf{A}|\omega)P(\omega)
\]

where \( \nabla_{\mathbf{b}} H(\mathbf{A}|\omega) = \frac{\partial \log(2\pi e\|\psi_i^l\|)}{\partial \mathbf{b}} \mathbf{a}_l^{\omega} \mathbf{a}_l^{\omega} - \mathbf{a}_l^{\omega} \mathbf{w}_l \), Then we turn to solve

\[
\frac{\partial \Psi_{l}}{\partial \mathbf{b}} = \frac{2}{n} \sum_{k=1}^{n} (a_{k,l} - \mathbf{a}_l^{\omega}) \frac{2\lambda k,l}{N e^k} - \frac{1}{n} \sum_{j=1}^{n} \frac{2\lambda j,l}{N e^j} \mathbf{b}
\]

The derivation of \( \nabla_{\mathbf{b}} H(\mathbf{A}) \) is performed similarly. With the gradient information, our iterative optimization algorithm updates a spectral filter by

\[
\mathbf{b}^{(iter+1)} = \mathbf{b}^{(iter)} + \lambda \nabla_{\mathbf{b}} I(\mathbf{A}, \Omega)
\]

Note that, the coefficients of the spectral filter \( \mathbf{b} \) are usually formed in a high dimensional vector. In order to make the solution robust and efficient, a regularization term can be introduced to achieve the sparse solution [8] or the smooth solution [27]. To avoid the additional hyperparameter introduced by a regularization term, the similar subspace gradient-based learning algorithm is used to solve the spectral filter \( \mathbf{b} \) in this study. For more details, please refer to Appendix B.

D. Iteration and Convergence

We alternatively optimize \( \mathbf{W} \) and \( \mathbf{b} \) by solving the optimization problem of (9) and (19), respectively. Given an initial value \( \mathbf{W}^{(k)}, \mathbf{b}^{(k)} \), the value of mutual information is denoted as \( \tilde{I}(\mathbf{W}^{(k)}, \mathbf{b}^{(k)}) = \tilde{I}_1(\mathbf{W}^{(k)}|\mathbf{b}^{(k)}) \). Derive \( \mathbf{W}^{(k+1)} \) by maximizing Eq. (9) and the value of mutual information is updated to be \( \tilde{I}_1(\mathbf{W}^{(k+1)}|\mathbf{b}^{(k)}) \). After optimizing the spatial filters, we sequentially optimize the spectral filter. Similarly, derive \( \mathbf{b}^{(k+1)} \) by maximizing Eq. (19) when \( \mathbf{W}^{(k+1)} \) is fixed and the value of mutual information is updated to be \( \tilde{I}_2(\mathbf{b}^{(k+1)}|\mathbf{W}^{(k+1)}) \). The update of spatial filters \( \mathbf{W} \) and spectral filters \( \mathbf{b} \) are performed sequentially and iteratively. The initial spatial filters for the first iteration can be set to accommodate a broad frequency band, e.g., 7-32Hz. The stop criterion for iteration can be chosen by the user, e.g., stopping when the mutual information gain between two consecutive iteration is less than a preset threshold \( \varepsilon \) \( (\varepsilon = 1 \times 10^{-4}) \) or a maximum number of iterations has been achieved. In this paper, we iterate the optimization of spatial and spectral filters 5 times or until the mutual information gain between \( \tilde{I}_1 \) and \( \tilde{I}_2 \) is less than \( \varepsilon \).

Suppose \( C \) is an upper bound of \( I(\mathbf{A}, \Omega) \) in our problem. We give the following convergence analysis of the cost function in the alternating iteration. Because EEG signals are complex, the cost function \( I(\mathbf{A}, \Omega) \) is usually sophisticated. In this case, we make an assumption that an initial value \( (\mathbf{W}^{(0)}, \mathbf{b}^{(0)}) \), which is near the local maximum if it exists, has been given. Note that, this can usually be achieved by the CSP algorithm which has been mentioned at the end of section B.

Proposition 1: If an initial value \( (\mathbf{W}^{(0)}, \mathbf{b}^{(0)}) \) which is near a uniquely local maximum is given, the cost function
is monotonically increasing or remains the same by iterations until it reaches the local maximum. That is,
\[0 \leq \tilde{I}(W^{(k)}, b^{(k)}) \leq \tilde{I}(W^{(k+1)}, b^{(k+1)}) < C\]  
(24)
where \(W^{(k)}\) and \(b^{(k)}\) are the spatial filters and spectral filter to be optimized at the \(k\)th iteration.

**Proof:** The upper bound of \(I(A, \Omega)\) is derived as follows. \(I(A, \Omega) = H(\Omega) - H(\Omega|A) \leq H(\Omega) = - \sum_{\omega} P(\omega) \log P(\omega) = 1\) for the classification of a two-class motor imagination with equal probability. Hence, \(1 \leq C < +\infty\).

Note that, \(\tilde{I}(W^{(k)}, b^{(k)}) = \tilde{I}_1(W^{(k)}|b^{(k)}) = \tilde{I}_2(b^{(k)}|W^{(k)}) \geq 0\). Derive \(W^{(k+1)}\) by maximizing (9) when \(b^{(k)}\) is fixed. Given the initial value \((W^{(0)}, b^{(0)})\) and the assumption of a unique local maximum, we can find a uniquely attained maximum by solving (9) with an additional constraint that Eq. (9) is continuously differentiable with respect to \(W^{(k)}\). Then the relationship \(0 \leq \tilde{I}_1(W^{(k)}|b^{(k)}) \leq \tilde{I}_2(b^{(k)}|W^{(k)}) \leq C\) hold, because the gradient-based learning algorithm for the first subproblem (9) achieves its local maximum by \(W^{(k+1)}\). At the same time, it means the relationship yields \(0 \leq \tilde{I}(W^{(k)}, b^{(k)}) \leq \tilde{I}(W^{(k+1)}, b^{(k)}) \leq C\). Similarly, \(0 \leq \tilde{I}_1(W^{(k+1)}|b^{(k)}) \leq \tilde{I}_2(b^{(k)}|W^{(k+1)}) \leq C\) can be obtained by the gradient-based learning algorithm for the second subproblem (19). As a result, it holds that \(0 \leq \tilde{I}(W^{(k)}, b^{(k)}) \leq \tilde{I}(W^{(k+1)}, b^{(k)}) \leq \tilde{I}(W^{(k+1)}, b^{(k+1)}) \leq C\) (See Section 2.7 in [25]). This completes the proof of proposition 1.

IV. EXPERIMENTS AND RESULTS

In this section, experiments on two public datasets are performed: BCI Competition III Dataset IVa [28] and BCI Competition IV Dataset IIa. Comparison of the proposed algorithm against the existing algorithms are presented. The same preprocessing and hyperparameters setting are applied to both datasets for fair comparison.

A. Preprocessing and Selection of Hyperparameters

The implementations for each method are as follows. A homogeneous setting of channels, time window and the number of spatial filters is adopted for all the comparative methods for fair comparison. Other hyperparameters for each method are finely tuned according to previous reports.

- **CSP:** Fourth order bandpass filter (Butterworth, 7-32Hz) is applied to raw EEG signals before the CSP algorithm is implemented. The two eigenvectors corresponding to eigenvalues from both ends of the eigenvalue spectrum are used as spatial filters.

- **CSSP:** The raw EEG signals are band-pass filtered by a 7-32Hz Butterworth filter as preprocessing. The time delay \(\tau\) is chosen out of \(\{1, 2, \ldots, 15\}\). It is optimized by using 10 x 10 CV in the training data. Here, the time interval of \([0.5, 2.5]\) is inappropriate for the CSSP algorithm, since twice the channel number is larger than the number of samples in the time window. The covariance matrices consequently cause an ill-conditioned eigenvalue problem. Hence, the time interval of \([0.5, 3.5]\) is chosen for this method.

- **DBFCS:** Two spatial filters are derived by maximizing the variance of filtered right hand imagery EEG signals under the constraint of normalization. The order of the FIR filter is set to 40. The iteration stops when the error of the cost function between successive iterations is less than \(10^{-4}\).

- **SWCSP:** The iteration number in alternating optimization is two since two iterations give good classification accuracy [10]. Similarly, two top eigenvectors corresponding to the largest and smallest eigenvalues are used as spatial filters. The spectral filter is designed according to Fisher’s criterion with positive coefficient constraint. \(p' = 0, q' = 1\)

- **ISSPL:** Two spatial filters (one spatial filter per class) are solved for ISSPL. The initial spectral filters for the first iteration are set to a broad frequency range 7-32Hz. The stop criterion is set to two iterations of alternative optimizations according to previous results [11]. The regularization parameter \(C\) is chosen out of 20 candidates log-linearly spaced between \(10^{-1}\) and \(10^4\) by \(10 \times 10\) cross-validation on the training set.

- **FCBSP:** The filter banks are designed to cover the broad frequency band of 4-40Hz, which consists of 9 non-overlapping temporal band-pass filters that cover a bandwidth of 4Hz each. All filters are fourth order Butterworth filters. We choose two pairs of CSP features in each band to construct the pool of CSP features (four eigenvectors corresponding to the two largest and two smallest eigenvalues). According to the suggestions in [14], two pairs of CSP features usually give high classification accuracy. Then these CSP features from multiple bands are selected by the mutual information-based best individual feature (MIBIF) algorithm. The number \((k)\) of best individual features is set to 1.

- **OSSFN:** Construct the same 9 non-overlapping band-pass filters as FBCSP1 that covers the possible EEG rhythms of motor imagery, then derive a discriminative spatial filter subspace from each band by the CSP algorithm (obtain initial values for optimization of OSSFN). Optimize the two spatial filters in each band by maximizing the mutual information proposed in [16]. Select the optimal frequency band and spatial filters by the maximum mutual information criterion.

- **MMISS:** The feature vector defined in (5) is used. Two spatial filters are derived by simultaneous optimization of all spatial filters, i.e., equation (25) and (27). The subspace \(U\) is constructed by the first two pairs of CSP spatial filters. The initial spectral filter \(b\) is set to a broad frequency band 7-32Hz. In our implementation, we perform the line search procedure to test the \(\lambda\) values (in Eq. (18) and Eq. (23) ) in the range of \([1,1]\) with an interval of 0.01 until a local maximum of \(I(A, \Omega)\) is reached. The iteration stops when the mutual information gain between two successive alternative iterations becomes less than \(\varepsilon = 10^{-4}\) or the number of alternative iterations is larger than 5.

B. BCI Competition III Dataset IVa

1) Data description: The dataset IVa from BCI competition III are used in our experiments [28]. These EEG data sets were recorded from five healthy subjects (aa, al, av, aw and ay).
118 electrodes were placed for each subject and the sampling rate in this experiment was 1000Hz. During each trial, the subject was given visual cues for 3.5s, during which the three motor imageries should be performed: left hand, right hand and right foot. Only EEG trials for right-hand and right-foot movements were provided for analysis. The presentation of target cues was intermitted by periods of random length from 1.75 to 2.25 seconds. During the random periods, the subject could relax. 140 trials were collected for each subject and each task.

In our experiment, the EEG data are down-sampled to 100Hz for use. The signals in the time interval of [0.5,2.5]s are analyzed for each trial. The time when a visual cue is presented is the beginning time (0s).

2) Results: In this section, ten-fold cross-validation is used to assess the performance of extracted features. All the 280 trials of dataset IVa for each subject are used to perform cross-validation. In each fold, nine parts of data are used for training data, the remaining one part is left for testing data. Then tenfold cross-validation is repeated 10 times and the accuracies are averaged to get the mean result of $10 \times 10$ fold cross-validation. In order to ensure a fair comparison among the competing algorithms, the same training and testing partitions are used for evaluation of each method in the cross-validation.

First, classification accuracy of the proposed method is compared with the conventional methods described in section II. Spatial filters for a particular example are illustrated to show the consistency between the learned filters and prior neurophysiological knowledge. We then discuss the amplitude characteristic of spectral filter in each method. Finally, the behaviors of convergence for the proposed alternative and iterative algorithm in MMISS are presented.

a) Classification Accuracy: In the following classification experiment, discriminative features are extracted from EEG signals by each method separately. For simplicity and fair comparison, only one pair of spatial filters and one spectral filter are solved by each method. Then support vector machine (SVM) with the best parameters is used to classify the feature vectors [29]. The Gaussian radial basis function (RBF) is chosen as the kernel function and the hyperparameters such as cost value and parameter in kernel function are selected from $\{0.5, 1, 2, 4, 8\}$ and $\{0.0625, 0.125, 0.25, 0.5, 1, 2\}$ by the nested cross-validation. Note that more than one pair of features for each method might further improve the classification accuracy, however, the simplification for comparison doesn’t affect the major conclusions.

The mean accuracy and standard deviation of the $10 \times 10$ CV for each subject and each method is listed in Table II. The proposed method outperforms the conventional spatial spectral methods in terms of mean classification accuracy under the same comparative condition. The standard deviation of the CV test for MMISS is almost the smallest (similar to FBCSP) on average over five subjects.

The optimal spatial filters for each method are illustrated in Fig. 1. We choose a specific example of subject ‘av’ to show the spatial filters of each method, which correspond to the smallest/largest eigenvalues. The coefficients of each filter are normalized to keep the sum of squares of the elements equal to one. For this subject, the spatial filters obtained by MMISS exhibit the most physiologically interpretable topography, in which the near mid-central vertex and the left hemisphere discriminate the foot and right hand imagery, respectively. This possibly explains the higher classification accuracy than other algorithms. However, for other methods except FBCSP, the topography generally disagrees with the prior physiological knowledge. For other subjects, most of the methods output similar topography which isn’t shown due to paper space requirements.

b) Spectral filters: To demonstrate differences among spectral filters learned by all the comparative methods. One specific training dataset of CV test is picked up for subject ‘av’. We show the normalized amplitude characteristics of the compared

<table>
<thead>
<tr>
<th>Method</th>
<th>Subject</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSP</td>
<td>aa</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>al</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>av</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>aw</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>81.0</td>
</tr>
<tr>
<td>CSSP</td>
<td>aa</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>al</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>av</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>aw</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>81.0</td>
</tr>
<tr>
<td>DFBCSP</td>
<td>aa</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>al</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>av</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>aw</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>81.0</td>
</tr>
<tr>
<td>SWCSP</td>
<td>aa</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>al</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>av</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>aw</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>81.0</td>
</tr>
<tr>
<td>ISSPL</td>
<td>aa</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>al</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>av</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>aw</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>81.0</td>
</tr>
<tr>
<td>FBCSP</td>
<td>aa</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>al</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>av</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>aw</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>81.0</td>
</tr>
<tr>
<td>OSSFN</td>
<td>aa</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>al</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>av</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>aw</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>81.0</td>
</tr>
<tr>
<td>MMISS</td>
<td>aa</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>al</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>av</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>aw</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>81.0</td>
</tr>
</tbody>
</table>

Fig. 1. The most significant spatial filters extracted by each of the comparative methods. The spatial filters are derived from a specific data set of subject ‘av’.

(a) CSP (b) CSSP (c) DFBCSP (d) SWCSP (e) ISSPL (f) FBCSP (g) OSSFN (h) MMISS
filters for this subject in Fig. 2. These filters are optimized from the time domain (e.g. CSSP in (b) and DFBBCSP in (c)) or the frequency domain (e.g. SWCSP in (d), ISSPL in (e) and the proposed MMISS in (h)). Some of them are picked up from an initial filter banks (e.g. FBCSP in (f), and OSSFN in (g)). All the methods have extracted the important components of higher alpha (10-16Hz) rhythms for this specific example (the spectrum of the filter selected by OSSFN concentrates on higher beta rhythm). This implies that these methods can learn or select a bandpass spectral filter which is effective for classification. However, there are distinct differences among the shape of spectral filters, which means different discriminative information is captured by different methods.

C. BCI Competition IV Dataset IIa

1) Data description: These data consisted of EEG signals recorded from nine subjects of four motor imagery task, including left and right hand, foot and tongue. Two sessions were conducted on two different days, comprising 288 trials for each session. Altogether, each session contains 72 trials for each task. The EEG data were sampled with 250Hz and acquired by 22 electrodes. Because one channel contains invalid signals in session two for some subjects, only 21 out of 22 EEG channels are used in this experiment. Additionally, for each subject there are different amounts of trials which are contaminated by artifacts. These trials are excluded from the data. We consider the signals between 0.5s and 2.5s after the onset of the visual cue for evaluation.

2) Results: a) Cross-validation Results of Multi-class classification:
First of all, we combine data in two sessions into one and thus there are more than 200 trials for each binary classification. Five-fold cross-validation is used to assess the performance of all the algorithms for this data set because some subjects contain only about 200 valid trials for binary classification. Multi-class extensions to all the algorithms are investigated and pair-wise approach which discriminates the derived features by each pair of classes is adopted in this experiment. For the four classes of motor imagery in the data set Ia, $4 \times (4 - 1)/2 = 6$ binary classifiers are required to discriminate each pair of classes. The preprocessing and selection of hyperparameters are the same as those described in part A. The mean kappa values of the $5 \times 5$-fold cross-validation for each subject and the average kappa value for all the nine subjects on this combined dataset of two sessions are shown in Table III. Note that only the classification results during the time interval of $[0.5, 2.5]$s are computed and reported in this table. The results show that the MMISS with pair-wise approach yields the best performance in terms of averaged mean kappa value (0.573). The Wilcoxon signed ranks test for comparison of two classifiers and the Friedman test with the corresponding post-hoc tests for comparison of more classifiers over multiple data sets are applied to perform the statistical analysis. They are nonparametric equivalent of the repeated measure ANOVA and do not need to assume the population of multiple data sets follows normal distribution [30], [31]. Since the size of data sets in our experiments are small (five subjects plus nine subjects), well below 30, the Friedman’s test and Wilcoxon signed ranks test are more appropriate. Friedman’s test over post-hoc multiple comparisons indicates that the six algorithms differ significantly in kappa value of multi-class classification ($p = 2.7 \times 10^{-5}$). Additional Wilcoxon signed ranks test for each paired method show that the MMISS is significantly better than all of the other methods in terms of mean kappa value at the 5% level.

b) Session-to-session transfer results on unseen evaluation data:

The session-to-session transfer is more challenging since the property of brain signals on training data may change substantially from that of brain signals on the evaluation data which may be recorded on different day. In order to further validate the effectiveness of the MMISS algorithm, we perform the $5 \times 5$-fold cross-validation on the data of training session and then use the derived spatial and spectral filters together with the classifiers to label the data of the unseen evaluation session. This operation can simulate the process of session-to-session transfer. Hence there are $5 \times 5$ classification results on the evaluation data and this operation may stabilize the variation of results and make the session-to-session transfer results statistically comparable. The mean kappa values of the $5 \times 5$-fold cross-validation results on the training data set and unseen evaluation data set for Dataset Iia of the BCI competition IV are shown in Table V. The MMISS achieves the best performance on both the training data session and the unseen evaluation session in terms of mean kappa value. The transfer results decrease by all the methods and this is in accordance with the previous reports [15]. MMISS performs best on the training data session, the statistical analysis (Friedman’s test, $p = 0.002$) for cross-validation on the training data set shows that the comparative methods differ significantly in the classification accuracy. However, the Wilcoxon signed ranks test for each paired method shows that the MMISS is only significantly better than CSP. This result differs from that of Table III, because there are few trials on the training data than trials on the combined data set. Therefore the performance of all the methods decreases due to the decreasing size of data set. The statistical analysis shows that the six filtering algorithms differ significantly in the classification accuracy of session-to-session transfer (Friedman’s test, $p = 0.002$). Additional Wilcoxon signed ranks test for each paired method indicate that the MMISS performs significantly better than two of the comparative methods in terms of classification accuracy of session-to-session transfer. However, the challenge of session-to-session contains the problem of inherent non-stationary across sessions and this might not be the main focus of the proposed method. Whether the proposed method can be used to solve the problem needs further investigation.

V. DISCUSSION

A. Statistical analysis of comparative methods

For statistical analysis, we combine the average classification accuracy of data set Iva (five persons, right hand versus foot) and one of the binary classifications of data set Iia (nine persons, right hand versus foot) together into one group of samples (fourteen persons). The null hypothesis is that all the competing algorithms together with MMISS yield the same mean classification accuracy. Friedman’s test indicates that the six filtering algorithms differ significantly in classification accuracy ($p = 9.8 \times 10^{-5}$). Additionally, Wilcoxon signed ranks test for each paired method and the combination of two data sets are shown in Table IV. The results indicate that the null hypothesis can be rejected at the $\alpha = 0.05$ level of significance. This means the MMISS algorithm is significantly better than a majority of the other methods in terms of mean classification accuracy. The difference between MMISS and SWSCP is marginally insignificant ($p = 0.079$).

Results from Table II, III and Table V suggest that the MMISS algorithm can extract the discriminative features efficiently. Under the same comparative condition, the mean classification accuracy of MMISS outperforms all the other comparative methods. Note that the results in Table II are slightly different from the previous reports [11], [13]. The reason is that the number of best individual features is set to 4 for FBCSP in their computation. Considering a fair comparison, we choose only one pair of spatial filters for each
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2014.2345458, IEEE Transactions on Biomedical Engineering

method, which explains the slight discrepancy. The classification performance of ISSPL is closely related to the time window and electrodes to be selected, which weren’t reported in their research. Therefore, the deterioration for subject ‘av’ is reasonable under the parameter setting of the current study. In our applications, deterioration of performance happens to scenario since the nonparametric estimation of pdf might catch more sophisticated rhythmic activity in self-paced BCI. However, in our implementation, we aim to discover the discriminative information by simple assumption rather than recovering the underlying probability density of rhythmic activity, which is relatively hard to achieve. To elaborate on this point, we draw the scatter plot of the CSP features for one specific subject ‘aw’ in Fig. 5. The marginal probability density function (pdf), which is estimated from the nonparametric method and parametric Gaussian distribution respectively, is plotted alongside the scatter plot for each class. Clearly, the nonparametric method has the flexibility to describe a rather complex distribution. By contrast, the Gaussian distribution, a uni-modal function, seems to sacrifice flexibility. However, it is more robust for cue-based motor imagery if only limited samples are provided since few parameters have to be estimated.

<table>
<thead>
<tr>
<th>Method</th>
<th>Subjects(training session)</th>
<th>mean</th>
<th>p-value(2.7 × 10⁻⁵)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSP</td>
<td>S1 0.602 0.265 0.713 0.339 0.128 0.142 0.253 0.216 0.646</td>
<td>0.454</td>
<td>0.008</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.077 0.095 0.067 0.083 0.085 0.118 0.087 0.083 0.090</td>
<td>0.241</td>
<td>0.734</td>
</tr>
<tr>
<td>CSP</td>
<td>S1 0.661 0.300 0.775 0.398 0.332 0.218 0.679 0.747 0.632</td>
<td>0.527</td>
<td>0.217</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.071 0.092 0.064 0.084 0.069 0.123 0.083 0.066 0.078</td>
<td>0.174</td>
<td>0.055</td>
</tr>
<tr>
<td>SWCSP</td>
<td>S1 0.678 0.273 0.785 0.409 0.209 0.169 0.571 0.740 0.742</td>
<td>0.509</td>
<td>0.245</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.062 0.096 0.058 0.069 0.084 0.091 0.071 0.062 0.077</td>
<td>0.259</td>
<td>0.098</td>
</tr>
<tr>
<td>ISSPL</td>
<td>S1 0.706 0.244 0.817 0.336 0.131 0.204 0.035 0.091 0.713</td>
<td>0.500</td>
<td>0.259</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.074 0.093 0.060 0.080 0.087 0.084 0.078 0.077 0.074</td>
<td>0.500</td>
<td>0.259</td>
</tr>
<tr>
<td>FBCSP</td>
<td>S1 0.672 0.198 0.759 0.200 0.440 0.131 0.724 0.678 0.717</td>
<td>0.502</td>
<td>0.259</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.085 0.087 0.058 0.076 0.094 0.088 0.066 0.073 0.072</td>
<td>0.500</td>
<td>0.259</td>
</tr>
<tr>
<td>MMISS</td>
<td>S1 0.658 0.282 0.826 0.373 0.470 0.226 0.658 0.713 0.748</td>
<td>0.550</td>
<td>0.220</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.065 0.084 0.050 0.078 0.089 0.109 0.074 0.083 0.066</td>
<td>0.550</td>
<td>0.220</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Subjects(unseen evaluation session)</th>
<th>mean</th>
<th>p-value(0.002)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSP</td>
<td>S1 0.596 0.194 0.665 0.432 0.327 0.133 0.577 0.610 0.548</td>
<td>0.425</td>
<td>0.020</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.038 0.034 0.015 0.033 0.036 0.042 0.038 0.031 0.019</td>
<td>0.221</td>
<td>0.203</td>
</tr>
<tr>
<td>CSP</td>
<td>S1 0.628 0.211 0.698 0.451 0.073 0.168 0.620 0.642 0.553</td>
<td>0.449</td>
<td>0.203</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.039 0.033 0.016 0.030 0.047 0.056 0.030 0.025 0.030</td>
<td>0.227</td>
<td>0.164</td>
</tr>
<tr>
<td>SWCSP</td>
<td>S1 0.677 0.196 0.701 0.414 0.136 0.180 0.637 0.599 0.597</td>
<td>0.460</td>
<td>0.222</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.034 0.031 0.022 0.033 0.043 0.051 0.043 0.035 0.043</td>
<td>0.222</td>
<td>0.164</td>
</tr>
<tr>
<td>ISSPL</td>
<td>S1 0.690 0.219 0.744 0.431 0.084 0.177 0.028 0.602 0.599</td>
<td>0.464</td>
<td>0.222</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.015 0.039 0.022 0.023 0.032 0.025 0.055 0.028 0.040</td>
<td>0.225</td>
<td>0.164</td>
</tr>
<tr>
<td>FBCSP</td>
<td>S1 0.578 0.140 0.685 0.308 0.228 0.171 0.509 0.558 0.542</td>
<td>0.413</td>
<td>0.196</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.030 0.049 0.026 0.054 0.024 0.049 0.050 0.037 0.075</td>
<td>0.413</td>
<td>0.196</td>
</tr>
<tr>
<td>MMISS</td>
<td>S1 0.656 0.257 0.781 0.407 0.295 0.195 0.662 0.584 0.608</td>
<td>0.494</td>
<td>0.202</td>
</tr>
<tr>
<td>(S.D.)</td>
<td>S2 0.040 0.025 0.037 0.039 0.038 0.055 0.032 0.040 0.049</td>
<td>0.494</td>
<td>0.202</td>
</tr>
</tbody>
</table>
B. Advantage of the proposed method

To validate whether the Gaussian distribution is reasonable [32], we evaluate the similarity between the distribution of log-power features (see Eq. (3)) and the multivariate normal distribution. The null hypothesis is that the distribution of log-power features conditioned on one class is similar to the normal distribution. Mardia’s test [33], [34] is applied to the data sets of fourteen persons (right hand class) and the results are shown in Table VI. The results show that the distribution of log-power features for nine of fourteen persons is not significantly different from the normal distribution at the \( \alpha = 0.05 \) level of significance. Although the underlying distribution might be more complex than the multi-variate normal distribution especially for some subjects, it’s more difficult to derive the gradient of the object function (6) if we use a more complex distribution. By relying on the Gaussian model assumption, only the mean and variance need to be estimated. Hence we can estimate the parameters more efficiently with lower variance compared to nonparametric estimation [35]. Consequently, the MMISS algorithm outperforms the OSSFN greatly. Similarly, the ISSPL, which outputs the weight coefficients of the spectrum by a classifier, might be overfitting when the number of channels is large and the weight coefficients of the spectrum by a classifier, might be overfitting when the number of channels is large and the separability of the subject’s data is low, e.g. subject ‘av’. The performance of all the methods could be further improved by choosing several pairs of spatial filters and spectral filters [4] or by selecting an optimal channel configuration [36]. We find that the average classification accuracy of subjects ‘aa’ and ‘aw’ achieve to be 91.2 ± 4.9 and 97.5 ± 3.2 by the MMISS algorithm if only part of the channels are used while the constraint of one pair of spatial filters and one spectral filter is maintained. However, how to tune the parameters (such as the number of spatial and/or spectral filters, the length of time window and the channel configuration) with nested cross-validation or other methods to achieve the best results is not the research focus of this paper.

<table>
<thead>
<tr>
<th>Subject</th>
<th>aa</th>
<th>al</th>
<th>av</th>
<th>aw</th>
<th>ay</th>
<th>S1</th>
<th>S2</th>
</tr>
</thead>
<tbody>
<tr>
<td>skewness</td>
<td>0.00</td>
<td>0.15</td>
<td>0.00</td>
<td>0.31</td>
<td>0.00</td>
<td>0.10</td>
<td>0.00</td>
</tr>
<tr>
<td>kurtosis</td>
<td>0.01</td>
<td>0.36</td>
<td>0.33</td>
<td>0.05</td>
<td>0.00</td>
<td>0.41</td>
<td>0.05</td>
</tr>
<tr>
<td>Subject</td>
<td>S3</td>
<td>S4</td>
<td>S5</td>
<td>S6</td>
<td>S7</td>
<td>S8</td>
<td>S9</td>
</tr>
<tr>
<td>skewness</td>
<td>0.10</td>
<td>0.13</td>
<td>0.48</td>
<td>0.08</td>
<td>0.01</td>
<td>0.52</td>
<td>0.87</td>
</tr>
<tr>
<td>kurtosis</td>
<td>0.09</td>
<td>0.64</td>
<td>0.40</td>
<td>0.30</td>
<td>0.85</td>
<td>0.08</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Previous research suggests that CSP is a decomposition technique especially powerful for motor imagery based feature extraction rather than a direct classification technique. A post processing classifier such as linear discriminant analysis (LDA) or SVM is necessary to get the classification results. Usually some spatial patterns derived by CSP are neurophysiologically interpretable. However, these spatial filters derived by CSP are not always optimal for the classification problem [37], [38]. Sometimes the eigenvectors corresponding to the extreme eigenvalue spectrum will concentrate on the strong power which might be contaminated by artifacts (see the spatial filters in Fig. 1(a)). A recent paper discusses the relationship between information theoretic feature extraction and the CSP algorithm [19]. However, the mutual information has only been used as a criterion to select spatial filters in their implementation. We propose to use the mutual information as an unique objective for feature extraction. The optimization of spatial and spectral filters is achieved by maximizing the mutual information, which is a function of spatial and spectral filters. The results obtained by MMISS suggest that the mutual information might be an alternative objective for extracting spatial and spectral features of the ERD/ERS activity. It should be noted that the estimation of mutual information relies on the assumption of Gaussian distribution of spatial spectral features. This assumption, however, is not robust enough to outliers, since the EEG data contains outliers sometimes [4]. Whether the more robust distribution like Student’s t-distribution will perform better in this case is worthy of future investigation.

C. Limitations

In the current study, we propose to optimize the spatial and spectral filters simultaneously in a unified mutual information based framework. The optimizing algorithm is realized by solving the spatial filters and spectral filters separately and iteratively. Due to the mutual information based cost function is usually complex and non-convex, it’s difficult to find the global optimum by the available optimization methods. Our algorithm proceed until \( I(A, \Omega) \) attains its local maximum. We have to point out the following facts. First, it may happen that the convergence of cost function is slow by the gradient search method and this is closely related to the geometric property of designed cost function and the EEG data set. In this situation, advanced gradient algorithms could help to improve the efficiency. The maximum number of iterations can be used alternatively in this situation and this is a common step to guarantee the algorithm stops appropriately. Second, the initial point is important to the optimization algorithm. In the proposed solution, we suggest the CSP-learned spatial filters to be a good starting point. However, this is difficult to prove in theory. Some different initial values can usually be tried to get a better local maximum. Nevertheless, our experiments on the BCI competition datasets show that the proposed solution usually gives good results. The other algorithms proposed in

![Fig. 5. Scatter plot and marginal probability density function (pdf) of CSP features for two classes of motor imagination from one specific subject ‘aw’. Note that the CSP features are log transformed values. The marginal pdf is estimated from the nonparametric (kernel density estimation) method (a) and the parametric Gaussian distribution (b), respectively.](image-url)
the literature may also provide alternative starting point for the MMISS. Note that, the results reported in Table III and V are lower than the winner of data set IIA on BCI competition IV. Because we use a homogenous parameter setting for all the methods such as the same time window, the same number of spatial filters, which are not optimized for each method. Additionally, we only calculated the classification accuracy during the time interval of [0.5,2.5]s rather than picked up the maximum kappa value from the entire continuous output of classification accuracy from the onset of the fixation cross to the end of motor imagery in the competition. These cause the discrepancy between our results and the results published on the website of BCI competition IV. The statistical results have to be interpreted with care in the study due to the discrepancy and the limitations.

VI. CONCLUSION

In this study, we have presented the MMISS algorithm for optimizing spatial and spectral filters simultaneously. In the proposed framework, the learning of spatial and spectral filters are formulated as a maximizing mutual information problem. The alternative and iterative optimization approach is adopted to solve the problem. Additionally, the subspace gradient learning approach, in which spatial and spectral filters are parameterized by lower dimensional vectors, is applied to make the solution robust and efficient. Under the same comparative condition, the MMISS algorithm achieves classification accuracy of 90.7% on average over five subjects in comparative condition, the MMISS algorithm achieves classification accuracy from the onset of the fixation cross to maximum kappa value from the entire continuous output of the data set IVa of BCI competition III and IV. The statistical results have together with the statistical significance tests show that the proposed algorithm perform well in classification for multi-class and session-to-session transfer and these demonstrates the effectiveness of the proposed method as compared with other algorithms.

APPENDIX A

SIMULTANEOUS OPTIMIZATION OF ALL SPATIAL FILTERS

A joint vector, which is formed by concatenating all the spatial filters together, is considered in order to optimize all the spatial filters simultaneously. That is, \( \tilde{w} = [w_1^T, \ldots, w_l^T, \ldots, w_{n_f}^T]^T \)

Due to the assumption of diagonal covariance matrix for each class \( \omega \), each element \( a_{i,l} \) of feature vector \( a_i \) is only a function of \( w_l \). Then the derivative of mutual information \( I(A, \Omega) \) with respect to the joint vector \( \tilde{w} \) can be derived by concatenating the derivatives with respect to each single spatial filter \( w_l \) (i.e., \( \nabla_{w_l} I(A, \Omega) \)). Therefore

\[
\nabla_{\tilde{w}} I(A, \Omega) = [\nabla_{w_1} I(A, \Omega)^T, \ldots, \nabla_{w_l} I(A, \Omega)^T, \ldots, \nabla_{w_{n_f}} I(A, \Omega)^T]^T
\]

The multiple spatial filters for multichannel EEG signals are high-dimensional vectors. However, practical issues arise for gradient-based optimization in the high-dimensional space. To address the problem, we use a similar trick to Zhang et al [16]. The subspace optimization approach is employed to reduce the dimensionality of optimization problem.

Let \( \mathcal{V} \) be a \( n_v \)-dimensional \( (n_v < N) \) subspace, which is linearly spanned by the \( N \)-dimensional column vectors. Denote \( V = [v_1, \ldots, v_k, \ldots, v_{n_v}] \in R^{N \times n_v} \), where \( v_k \in R^{N \times 1} \) is the \( k \)th basis vector of \( \mathcal{V} \). Hence, any spectral filter \( w_l \) in the subspace \( \mathcal{V} \) can be expressed by

\[
w_l = \sum_{k=1}^{n_v} m_{lk} v_k = U m_l
\]

where \( m_l \) is a coefficient vector that uniquely determines \( w_l \) with respect to \( U \).

In the subspace \( \mathcal{V} \), optimization of the spatial filter \( w_l \) is equivalent to optimization of the coefficient vector \( m_l \). Then simultaneous optimization of the concatenated spatial filters \( \tilde{w} \) is reformulated as optimization of the concatenated coefficient vectors

\[
\tilde{m} = [m_1^T, \ldots, m_l^T, \ldots, m_{n_f}^T]^T
\]

Now consider the partial derivatives of \( I(A, \Omega) \) with respect to \( m_l \). Substitute (27) into (8), then the element \( a_{i,l} \) of feature vector \( a_i \) is

\[
a_{i,l} = \log \left( \frac{1}{n_v} \xi_l \right) = \log \left( \frac{1}{n_v} (U m_l)^T \hat{R}_{xi} (U m_l) \right)
\]

Similar to (13), differentiating \( I(A, \Omega) \) with respect to \( m_l \) gives

\[
\nabla_{m_l} I(A, \Omega) = \nabla_{m_l} H(A) - \sum_{\omega \in \Omega} \nabla_{m_l} H(A|\omega) P(\omega)
\]

Here the derivative of \( H(A|\omega) \) with respect to \( m_l \) is given by

\[
\nabla_{m_l} H(A|\omega) = \frac{1}{2}\sum_{i<l} \frac{2}{n_v} \sum_{k=1}^{n_v} (a_{i,k} - \pi_{i,k}) \times \frac{2R_{xi,k} + \frac{1}{N \times n_v} \sum_{j=1}^{n_v} 2R_{xj,k})}{N \times n_v} U m_l
\]

The derivation of \( \nabla_{m_l} H(A) \) is performed similarly and omitted here. Then we have \( \nabla_{m_l} I(A, \Omega) \) by concatenating all the derivatives of \( \nabla_{m_l} I(A, \Omega), l = 1, \ldots, n_l \).

The remaining problem is how to construct the subspace \( \mathcal{V} \). This is an important problem for practical computation, however, we do not study the initialization problem in this paper. We simply use the spatial filters obtained by the CSP algorithm as the subspace basis vectors.

APPENDIX B

SUBSPACE GRADIENT-BASED LEARNING FOR SPECTRAL FILTERS

For a specific dataset, the dimensionality of a spectral filter is determined by the sampling rate and the length of the time window. To reduce the dimension of the spectral filter, a similar subspace optimization technique is used. Let \( \mathcal{V} \) be a \( n_v \)-dimensional \( (n_v < N) \) subspace, which is linearly spanned by the \( N \)-dimensional column vectors. Denote \( V = [v_1, \ldots, v_k, \ldots, v_{n_v}] \in R^{N \times n_v} \), where \( v_k \in R^{N \times 1} \) is the \( k \)th basis vector of \( \mathcal{V} \). Hence, any spectral filter \( b \) in the subspace \( \mathcal{V} \) can be expressed by

\[
b = \sum_{k=1}^{n_v} d_k v_k = V d
\]
where $d$ is a coefficient vector that uniquely determines $b$ with respect to $V$.

Similarly, in the subspace $V$, optimization of the spatial filter $b$ is equivalent to optimization of the coefficient vector $d$. The derivation of $\nabla_d f(A, \Omega)$ is straightforward according to Eq. (21) and is omitted here. In this study, we use a simple method to construct the subspace $V$, i.e., several spectral weights in consecutive frequency bins are set to be the same value, hence, the spectrum in the neighbourhood are flat. This simple operation degrades the dramatic changes of spectral coefficients.

ACKNOWLEDGMENT

The authors would like to thank the Berlin BCI group for sharing their data sets and Pete Shull for help suggestions of writing the manuscript. We would also like to thank Hiroshi Higashi from Tokyo University of Agriculture and Technology for helpful discussion of the DFBCSP method, and Wei Wu from Tsinghua University for helpful discussion of the ISSPL method.

REFERENCES


