MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Spectrally weighted Common Spatial Pattern algorithm for single trial EEG classification

Ryota TOMIOKA, Guido DORNHEGE,
Guido NOLTE, Benjamin BLANKERTZ,
Kazuyuki AIHARA, and Klaus-Robert MÜLLER

METR 2006–40 July 2006

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

Spectrally weighted Common Spatial Pattern algorithm for single trial EEG classification

Ryota TOMIOKA[1,2], Guido DORNHEGE[2],
Guido NOLTE[2], Benjamin BLANKERTZ[2],
Kazuyuki AIHARA[1], and Klaus-Robert MÜLLER[2,3]

[1] Dept. Mathematical Informatics, IST,
The University of Tokyo, Japan
[2] Fraunhofer FIRST.IDA, Berlin, Germany
[3] Technical University Berlin, Germany

ryotat@sat.t.u-tokyo.ac.jp

July 5, 2006

Abstract

We propose a simultaneous spatio-temporal filter optimization algorithm for the single trial ElectroEncephaloGraphy (EEG) classification problem. The algorithm is a generalization of the Common Spatial Pattern (CSP) algorithm, which incorporates non-homogeneous weighting of the cross-spectrum matrices. We alternately update the spectral weighting coefficients and the spatial projection. The cross validation results of our SPECtrally-weighted CSP (SPEC-CSP) algorithm on 162 EEG datasets show significant improvements when compared to wide-band filtered CSP and similar accuracy as Common Sparse Spectral Spatial Pattern (CSSSP), however, with far less computational cost. The proposed method is at the same time highly interpretable and modular because the temporal filter is parameterized in the frequency domain. The possibility of incorporating any prior filter opens up the applicability of the method far beyond brain signals.

1 Introduction

The goal of the Brain-Computer Interface (BCI) research is to provide a direct control pathway from human intentions reflected in their brain signals to computers. Recently, a considerable amount of effort has been done in the development of BCI systems [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. We will be focusing on non-invasive, electroencephalogram (EEG) based BCIs. Such a system not only provides disabled people more direct and natural control over a neuro-prosthesis or over a computer application (e.g. [3, 4]) but also opens up an opportunity for healthy people to communicate solely by their intentions.

The study of BCIs consists of two parts; the first part is the techniques that development of such a system requires and the second part is the relationship between the system and the user. The design of an experimental paradigm or a study on subject-to-subject or session-to-session variability belongs to the second category, whereas a feature extraction technique (e.g., this study) or a classification algorithm belongs to the first category. Although novel techniques are often more general in the sense that they could be applied to other types of problems, yet the importance of interpretability and transparency of the method as the basis of second type of BCI studies is often neglected.

Recently machine learning approaches to BCI have proven to be effective by making the subject training required in a classical subject "conditioning" framework unnecessary and allow to compensate for the high inter-subject variability.

The task is to extract subject-specific discriminative patterns from high-dimensional spatio-temporal EEG signals. We study a BCI based on the motor imagination paradigm. Motor imagination can be captured through spatially localized band-power modulation in μ- (7-15Hz) and β- (15-30Hz) rhythms; underlying neuro-physiology is well known as Event Related Desynchronization (ERD) [14]. With respect to the topographic patterns of brain rhythm modulations, the Common Spatial Pattern (CSP) (see [15, 16, 17]) algorithm, which was first introduced as a decomposition method that finds projections common to two states of brain activity (e.g., abnormal or normal) and afterwards successfully applied to the classification problem of the two states, has also proven to be very useful for motor-imagery BCI. On the other hand, the frequency band on which the CSP algorithm operates, has been either selected manually or unspecifically set to a broad band filter [17, 5]. Recently, Lemm *et al.* [18] proposed a method called Common Spatio Spectral Pattern (CSSP) which applies the CSP algorithm to a time-delay embedded signal; they doubled the number of electrodes with the addition of τ delayed channels. The challenge in their approach was that the selection of the time-lag parameter τ, which embodies the whole problem of choosing characteristic temporal structure, was only possible through cross validation on the training set. Dornhege *et al.* [19] proposed a method called Common Sparse Spectral Spatial Pattern (CSSSP) which solves the CSP problem not only for the spatial projection but also for the Finite Impulse Response (FIR) temporal filter coefficients. The difficulty in this work was the computational inefficiency of the optimization procedure, in which solving a generalized eigenvalue problem was required for the evaluation of the objective function at every point.

In this paper, we present a method for simultaneous spatio-temporal filter optimization, which is an iterative procedure of solving a spectrally weighted CSP problem and updating the spectral weighting coefficients. The proposed method is highly interpretable and modular at the same time because the temporal filter is parameterized in the frequency domain. Moreover it is capable of handling arbitrary prior filters based on neurophysiological insights. The proposed method outperforms wide-band filtered CSP in most datasets. Moreover, a detailed validation shows how much of the gain is obtained by the theoretically obtained filter and how much is obtained by imposing a suitable prior filter.

This paper is organized as follows: in Sec. 2 the method is proposed; in Sec. 3 our novel SPEC-CSP is compared against all three recent filter methods, namely, CSP, CSSP, and CSSSP; in Sec. 4 the optimal combination of a spectral filter obtained from the statistical criterion and prior filters is investigated in detail; finally in Sec. 5 concluding remarks are given.

2 Method

Let us denote by $X \in \mathbb{R}^{d\times T}$ the EEG signal of a single trial of an imaginary motor movement[1], where d is the number of electrodes and T is the number of sampled time-points in a trial. We consider a binary classification problem where each class, e.g. right or left hand imaginary movement, is called the positive $(+)$ or negative $(-)$ class. The task is to predict the class label for a single trial X.

In this study, we use a feature vector, namely *log-power features* defined as follows:

$$\phi_j(X; \boldsymbol{w}_j, B_j) = \log \boldsymbol{w}_j^\dagger X B_j B_j^\dagger X^\dagger \boldsymbol{w}_j \qquad (j = 1, \ldots, J), \tag{1}$$

where the upper-script $\dagger$ denotes a conjugate transpose or a transpose for a real matrix, $\boldsymbol{w}_j \in \mathbb{R}^d$ is a spatial projection that projects the signal into a single dimension, and $B_j \in \mathbb{R}^{T\times T}$ denotes the linear time-invariant temporal filter. A *log-power feature* ϕ_j captures a brain-rhythm modulation, which is not only spatially localized (captured by $\boldsymbol{w}_j$) but also localized in the frequency domain (captured by B_j). The training of a classifier is composed of two steps. In the first step, the coefficients $\boldsymbol{w}_j$ and B_j are optimized. In the second step, the Linear Discriminant Analysis (LDA) classifier is trained on the feature vector. Note that the whole procedure is equivalent to the conventional CSP based classifiers (see [17, 5, 20]) except that in this study the temporal filter B_j is also optimized, which was manually chosen and fixed in the previous work.

[1] For simplicity, we assume that the trial mean is already subtracted and the signal is scaled by the inverse square root of the number of time-points. This can be achieved by a linear transformation $X = \frac{1}{\sqrt{T}} X_{\text{original}} \left(I_T - \frac{1}{T}\mathbf{1}\mathbf{1}^\dagger\right)$.

We adopt the Common Spatial Pattern (CSP) [15, 16, 17] algorithm for the optimization of the spatial projection $\boldsymbol{w}$. The idea is to simultaneously diagonalize the sensor covariance matrices corresponding to two motor-imagery classes. Here, a sensor covariance refers to the covariance between channels averaged over time as well as trials[2]:

$$\Sigma^{(c)}(B) := \left\langle XBB^{\dagger}X^{\dagger} \right\rangle^{c},$$

where angled brackets $\langle \cdot \rangle^{c}$ denote expectation within a class $c \in \{+, -\}$. Using the fact that a linear time-invariant temporal filter B is diagonal in the frequency domain $U^{\dagger}BB^{\dagger}U = \mathrm{diag}(\alpha_1, \ldots, \alpha_T)$ [3], we can rewrite the sensor covariance matrix as a weighted sum of cross-spectrum matrices as follows:

$$\Sigma^{(c)}(\boldsymbol{\alpha}) := \sum_{k=1}^{T'} \alpha_k \left\langle V_k \right\rangle^{c} := \sum_{k=1}^{T} \alpha_k \left\langle \hat{\boldsymbol{x}}_k \hat{\boldsymbol{x}}_k^{\dagger} \right\rangle^{c} = \left\langle XUU^{\dagger}BB^{\dagger}UU^{\dagger}X^{\dagger} \right\rangle^{c} = \Sigma^{(c)}(B),$$

where $U := \{\frac{1}{\sqrt{T}} e^{-2\pi ikl/T}\}_{kl} \in \mathbb{C}^{T \times T}$ is the Fourier transformation (thus $U^{\dagger}U = I_T$), $\hat{\boldsymbol{x}}_k \in \mathbb{C}^{d}$ is the k-th frequency component, $V_k := 2 \cdot \mathrm{Re}\left[\hat{\boldsymbol{x}}_k \hat{\boldsymbol{x}}_k^{\dagger}\right]$ and $\langle V_k \rangle^{c} \in \mathbb{R}^{d \times d}$ is the real-part of the k-th frequency component in the cross spectrum. Here, without loss of generality we only take the real-part of the cross spectrum because (a) the imaginary part cancels out since the spectrum of the filter $\{\alpha_k\}_{k=1}^{T}$ is symmetric around the Nyquist frequency and (b) the phase of the signal is irrelevant to the log-power feature (Eq. (1)). Note that only the $T' = \lceil \frac{n_{\mathrm{FFT}}+1}{2} \rceil$ independent frequency components below the Nyquist frequency are taken into the sum. Furthermore, the complexity of the spectrum estimation is controlled by using Welch method, which averages estimations on overlapping windows with shorter time-points $n_{\mathrm{FFT}} \leq T$. Now, a decomposition that is common to two brain states, or a set of bases that simultaneously diagonalize the two sensor covariance matrices, can be found by solving the following generalized eigenvalue problem [15]:

$$\Sigma^{(+)}(\boldsymbol{\alpha})\boldsymbol{w} = \lambda \Sigma^{(-)}(\boldsymbol{\alpha})\boldsymbol{w}. \tag{2}$$

Note that since for each pair of eigenvector and eigenvalue $(\boldsymbol{w}_j, \lambda_j)$ the equality $\lambda_j = \boldsymbol{w}_j^{\dagger}\Sigma^{(+)}(\boldsymbol{\alpha})\boldsymbol{w}_j \Big/ \boldsymbol{w}_j^{\dagger}\Sigma^{(-)}(\boldsymbol{\alpha})\boldsymbol{w}_j$ holds,

$$\boldsymbol{w}_1 = \underset{\boldsymbol{w}}{\mathrm{argmin}} \frac{\boldsymbol{w}^{\dagger}\Sigma^{(+)}(\boldsymbol{\alpha})\boldsymbol{w}}{\boldsymbol{w}^{\dagger}\Sigma^{(-)}(\boldsymbol{\alpha})\boldsymbol{w}},$$

$$\boldsymbol{w}_d = \underset{\boldsymbol{w}}{\mathrm{argmax}} \frac{\boldsymbol{w}^{\dagger}\Sigma^{(+)}(\boldsymbol{\alpha})\boldsymbol{w}}{\boldsymbol{w}^{\dagger}\Sigma^{(-)}(\boldsymbol{\alpha})\boldsymbol{w}},$$

[2] More precisely, it should be called the *cross-power matrix*, because a projection $\boldsymbol{w}^{\dagger}\Sigma\boldsymbol{w}$ gives the power of the projected signal.

[3] We assume that T is sufficiently large compared to the tap-length of the filter so that the problem due to the boundary is negligible.

where the eigenvectors are sorted in the ascending order of the eigenvalues. Note that the maximization of the ratio of the powers corresponds to the maximization of the separation of two classes in the *log-power feature* space if the class centers are well approximated by the logarithm of the class-averaged powers. It is common practice that only the first n_{of} largest eigenvectors and the last n_{of} smallest eigenvectors are used to construct a low dimensional feature representation.

The next question is how to optimize the coefficients $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^{T'}$. In order to achieve the trade-off between the good discrimination and the reliability of the band-power estimation, we formulate this problem as follows:

$$\max_{\boldsymbol{\alpha}} \frac{\langle s(\boldsymbol{w},\boldsymbol{\alpha})\rangle^{+} - \langle s(\boldsymbol{w},\boldsymbol{\alpha})\rangle^{-}}{\sqrt{\mathrm{Var}\left[s(\boldsymbol{w},\boldsymbol{\alpha})\right]^{+} + \mathrm{Var}\left[s(\boldsymbol{w},\boldsymbol{\alpha})\right]^{-}}}, \qquad (3)$$
$$\text{s.t. } \alpha_k \geq 0 \quad (\forall k = 1, \ldots, T'),$$

where we define

$$s(\boldsymbol{w},\boldsymbol{\alpha}) := \sum_{k=1}^{T'} \alpha_k s_k(\boldsymbol{w}) := \sum_{k=1}^{T'} \alpha_k \boldsymbol{w}^{\dagger} V_k \boldsymbol{w}.$$

Note that Eq. (3) can be viewed as the signed square root of the Rayleigh quotient used in Fisher discriminant analysis with an additional constraint that all coefficients must be positive; therefore, if we exchange the labels, Eq. (3) yields a different solution; thus we take the maximum of Eq. (3) for the "+" class and the minimum for the "−" class just like choosing CSP projections from the both ends of the eigenvalue spectrum of Eq.(2).

The optimal filter coefficient is explicitly written as follows:

$$\alpha_k^{(+)\mathrm{opt}} \propto \begin{cases} \dfrac{\langle s_k(\boldsymbol{w})\rangle_{+} - \langle s_k(\boldsymbol{w})\rangle_{-}}{\mathrm{Var}\left[s_k(\boldsymbol{w})\right]_{+} + \mathrm{Var}\left[s_k(\boldsymbol{w})\right]_{-}} & \langle s_k(\boldsymbol{w})\rangle_{+} - \langle s_k(\boldsymbol{w})\rangle_{-} \geq 0, \\ 0 & \text{otherwise}, \end{cases} \qquad (4)$$

because the spatio-temporally filtered signal $s(\boldsymbol{w},\boldsymbol{\alpha})$ is linear with respect to the spectral filter coefficients $\{\alpha_k\}_{k=1}^{T}$ and we additionally assume that the signal is a stationary Gaussian process, where the frequency components are independent to each other for a given class label; thus $\mathrm{Var}\left[s(\boldsymbol{w},\boldsymbol{\alpha})\right]_c = \sum_{k=1}^{T} \alpha_k^2 \mathrm{Var}\left[s_k(\boldsymbol{w})\right]_c$. Note that the labels (+ and −) are exchanged for the filter for the "−" class $\{\alpha_k^{(-)\mathrm{opt}}\}_{k=1}^{T'}$. The norm of the vector $\boldsymbol{\alpha}$ cannot be determined from the problem (3). Therefore, in practice we normalize the coefficients so that they sum to one.

Furthermore, we can incorporate our prior knowledge on the spectrum of the signal during the task. This can be achieved by generalizing Eq. (4) as follows:

$$\alpha_k^{(c)} = \left(\alpha_k^{(c)\mathrm{opt}}\right)^{q} \cdot (\beta_k)^{p} \qquad (c \in \{+,-\}), \qquad (5)$$

where $\{\beta_k\}_{k=1}^{T'}$ denotes the prior information, which we define specific to a problem (see Sec. 3.1.4). The optimal values for p and q should depend on the data, preprocessing, and the prior information $\{\beta_k\}_{k=1}^{T'}$. Therefore one can choose them by cross validation.

To summarize, the optimal spatial projection $\boldsymbol{w}$ is the eigenvector with the largest eigenvalue of the generalized eigenvalue problem (2) and the optimal spectral filter $\boldsymbol{\alpha}$ is the solution to the problem (3) and is explicitly written as Eq. (4). Moreover, one can incorporate any prior filter $\{\beta_k\}_{k=1}^{T'}$ as Eq. (5). Since both of the optimal spatial and spectral filter depend on each other, we use an iterative method that starts from conventional CSP (solving Eq. (2) with $\forall k, \alpha_k = 1$) and updates one fixing the other alternately. We found that using a fixed number of iterations results in comparable performance with using cross validations to select the number of iterations (see Sec. 3); alternatively one can use the eigenvalues of Eq. (2) to decide when the iteration should stop. The details are summarized in Table 1.

```
1.  Initialize $\alpha_k^{(1)} = 1\,(k = 1, \ldots, T')$ and $J = 1$.
2.  **for** each step
3.      **for** each set of spectral coefficients $\boldsymbol{\alpha}^{(j)}$ $(j = 1, \ldots, J)$
4.          Calculate the sensor covariance matrices $\Sigma^{(c)}(\boldsymbol{\alpha}^{(j)})$ $(c \in \{+, -\})$.
5.          Solve the generalized eigenvalue problem (2) and let
                $W_j^{(c)} \in \mathbb{R}^{d \times n_{\mathrm{of}}}$   : the set of $n_{\mathrm{of}}$ top eigenvectors, and
                $\lambda_j^{(c)}$                                        : the top eigenvalue $(c \in \{+, -\})$.
6.      **end** (for each set of spectral coefficients)
7.      set  $W^{(-)} := W_{j^*}^{(-)}$ with $j^* = \mathrm{argmin}_{j=1,\ldots,J} \lambda_j^{(-)}$ and
             $W^{(+)} := W_{j^*}^{(+)}$ with $j^* = \mathrm{argmax}_{j=1,\ldots,J} \lambda_j^{(+)}$
8.      **for** each projection $\boldsymbol{w}_j \in \{W^{(-)}, W^{(+)}\}$ $(j = 1, \ldots, J = 2n_{\mathrm{of}})$
9.          Calculate $\langle s_k(\boldsymbol{w}_j) \rangle^c$ and $\mathrm{Var}\,[s_k(\boldsymbol{w}_j)]^c$ for $c \in \{+, -\}$.
10.         Update $\alpha_k^{(j)} := \left(\alpha_k^{\mathrm{opt}}\right)^q \cdot (\beta_k)^p$ according to Eqs. (4) and (5)
            and normalize $\boldsymbol{\alpha}^{(j)}$ so that it sums to unity.
11.     **end** (for each projection)
12. **end** (for each step)

Note:   The top eigenvectors in step 5 are the eigenvectors
        corresponding to the largest and the smallest eigenvalues
        for the positive and the negative classes, respectively.
```

Table 1: The implementation of the proposed method.

3 Results

3.1 Experimental setup

3.1.1 Validation

We test four different preprocessing techniques, namely, the proposed method, wide-band filtered CSP [15, 17], CSSP [18], and CSSSP [19] on 162 datasets of BCI experiments from 29 subjects by cross validation. We use the log-power feature (Eq. (1)) with $n_{\text{of}} = 3$ features for each class (thus $J = 2n_{\text{of}} = 6$) and LDA as a classifier.

3.1.2 Data acquisition

A Berlin Brain-Computer Interface (BBCI) experiment consists of two parts, the calibration block and the feedback block [10]. In the calibration block, subjects performed 3-3.5 seconds of one of the imaginary movement tasks, namely, right hand (R), left hand (L) or foot (F), instructed by the corresponding letter displayed on the screen during this period. These trials were repeated every 4.5-6 seconds. Brain activity was recorded at the sampling frequency 100Hz from the scalp with multi-channel EEG amplifiers using 32, 64 or 128 channels. Besides EEG channels, we recorded the electromyogram (EMG) from both forearms and the leg as well as horizontal and vertical electrooculogram (EOG) from the eyes. The EMG and EOG channels were used exclusively to make sure that the subjects performed no real limb or eye movements correlated with the mental tasks that could directly (artifacts) or indirectly (afferent signals from muscles and joint receptors) be reflected in the EEG channels and thus be utilized by the classifier, which operates on the EEG signals only. Varying from a dataset to another, from 70 to 600 (median 280) trials were recorded. No feedback or response to the subject's motor imagination was presented in the calibration block. On the other hand, in the feedback block, the subject could steer a cursor or play a computer game like *brain-pong* by BCI control. The data from the feedback block is not used in this study because they depend on intricate interactions of the subject with the original classification algorithm in use with the feedback. In this study, since we investigate only binary classifications, different combinations of imaginary movements produced several binary problems in the datasets from a single experiment. The multi-class CSP proposed by [20] can also be generalized to incorporate spectral weighting.

3.1.3 Preprocessing of the signals

After removing the EOG, EMG, and a few outermost channels of the cap, we band-pass filter the signal from 7-30Hz and cut out the interval of 500-3500ms after the appearance of the visual cue on the screen from the continuous EEG signal for each trial of imaginary movement. Only in Sec. 4, we also use the signal without the band-pass filter step, in order to investigate the effect of assuming the above mentioned band on the design of a filter; except the band-pass filtering, the signal was equally processed as described above.

3.1.4 Prior information

The prior filter $\{\beta_k\}_{k=1}^{T'}$ is defined as follows:

$$\beta_k = I_k^{[7,30]} \cdot \left(\langle s_k(\boldsymbol{w}) \rangle^{+} + \langle s_k(\boldsymbol{w}) \rangle^{-} \right) / \, 2, \tag{6}$$

where $\{I_k^{[7,30]}\}_{k=1}^{T'}$ is an indicator function that takes value one only in the band 7-30Hz, and otherwise zero. Since we have already band-pass filtered the signal, it is reasonable to restrict the resulting filter to take values only within this band. The second term, which is the average activity of two classes, expresses our understanding that in the motor imagery task that involves ERD [14], good discrimination is most likely found at frequency bands that correspond to strong oscillations, i.e., μ- and β-rhythms. Since the optimal filter (Eq. (4)) and the prior filter (Eq. (6)) scale with powers -1 and 1 of the spectrum, respectively, we reparameterize the hyperparameters as $p = p' + q'$ and $q = q'$. Here, if $p' = c$ the filter scales with the power c regardless of which q' one chooses. Thus, the contributions of the scale and the discriminability are separated in the new parameterization. Now, using p', the scaling exponent of the filter and q', the intensity of the label information, Eq. (5) can be written as follows:

$$\alpha_k^{(+)} \propto I_k^{[7,30]} \cdot \begin{cases} \left(\frac{\left(s_k^{(+)} - s_k^{(-)}\right)\left(s_k^{(+)} + s_k^{(-)}\right)}{v_k^{(+)} + v_k^{(-)}} \right)^{q'} \cdot \left(s_k^{(+)} + s_k^{(-)} \right)^{p'} & s_k^{(+)} - s_k^{(-)} \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where the following short hands are used: $s_k^{(c)} := \langle s_k(\boldsymbol{w}) \rangle^{c}$ and $v_k^{(c)} := \mathrm{Var}\,[s_k(\boldsymbol{w})]^{c}$. Note that for $\alpha_k^{(-)}$, the labels (+ and $-$) should be exchanged.

3.2 Visual examples

Here we present some figures which are not directly relevant to the comparison but are helpful in understanding the proposed method better.

Figure 1 shows the construction of a spectral filter. The averaged spectrum of spatially filtered signals is shown for each class in Fig. 1(b). The spatial projection coefficients are also topographically shown in Fig. 1(c). The conventional CSP is purely an operation in the spatial domain. Therefore, as a spectral filter it has a flat spectrum as shown in Fig. 1(a). The proposed method (Fig. 1(d)), on the other hand, is a combination of the theoretically obtained filter (Eq. (4), Fig. 1(e)) and the prior filter (Eq. (6), Fig. 1(f)). The theoretically obtained filter scales with the power -1 of the spectrum (see Sec. 4 for the detailed discussion), therefore not only the discrimination around 12Hz but also that around 24Hz, which can be hardly seen in the original scale (Fig. 1(b)), are detected. Although the increase in the amplitude just before 30Hz might seem problematic, the effect is limited because the power of the wide-band filtered signal (Fig. 1(b)) goes to zero at 30Hz and the theoretical filter is only counterbalancing the decrease in the power of the original signal. The prior filter (Fig. 1(f)) peaks at frequency components corresponding to a strong rhythmic activity (μ-rhythm, 12Hz) regardless of whether it has discriminative information or not. Since the signal is already band-pass filtered from 7-30Hz, taking the average spectrum (Eq. (6)) is sufficient to tell where the rhythmic activity is. The resulting filter (Fig. 1(d)), which is the elementwise product of the two filters in this case (because $(p', q') = (0, 1)$), has two peaks, one larger peak at 12Hz and a smaller peak at 24Hz, reflecting the compromise between the theoretical and the prior filters. The optimal combination of the two filters is discussed in Sec. 3.4

Figure 2 shows the process of iteratively updating the spatial projection and the spectral filter. The iteration starts from uniform spectral coefficients (step 0). In an odd step the spatial projection is updated, whereas in an even step the spectral coefficients are updated. Note that “step 1” is the CSP with the wide-band filter. At “step 5”, we obtain a clear pattern corresponding to synchronized brain activity in the foot area of the motor cortex, which could not be found by the wide-band filtered CSP.

Figure 3 shows the improvement in the cross-validation error by iterative updates. An odd step and an even step corresponds to a spatial projection update and a spectral filter update similarly to Fig. 2. The 10×10 cross-validation errors are shown for six subjects. In addition, median over 162 datasets are also shown (gray dashed line). For some subjects (e.g., in subject F) no improvement were observed, most likely due to artifacts whose effects are not localized in the frequency spectrum (e.g. blinking, chewing or other muscle movements).

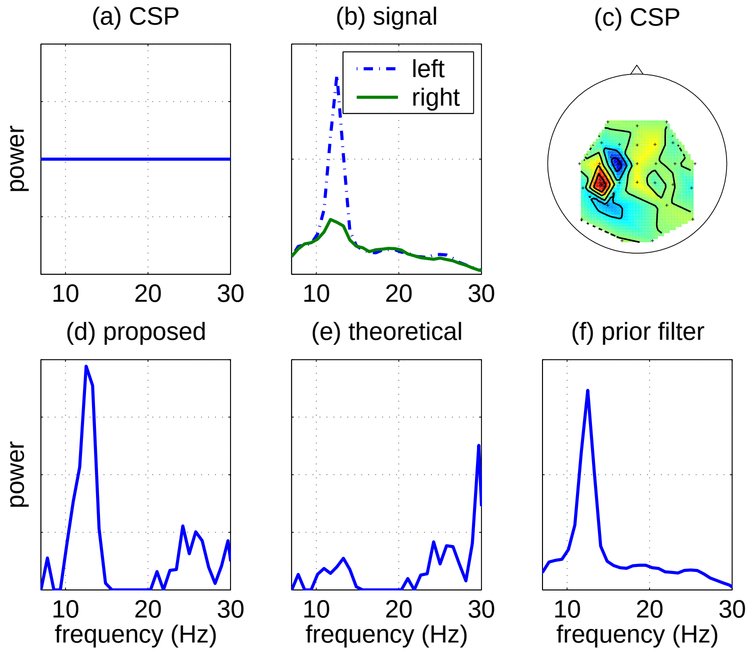


Figure 1: (a) The conventional CSP in the frequency domain. (b) The class-averaged spectrum of the original signal projected with a CSP projection. (c) The CSP projection topographically mapped on a head viewed from above. The head is facing the top of the paper. (d) The filter spectrum obtained by the proposed method. (e) The theoretically obtained filter (Eq. (4)). (f) The prior filter (Eq. (6)). All vertical axes show the powers in arbitrary units.

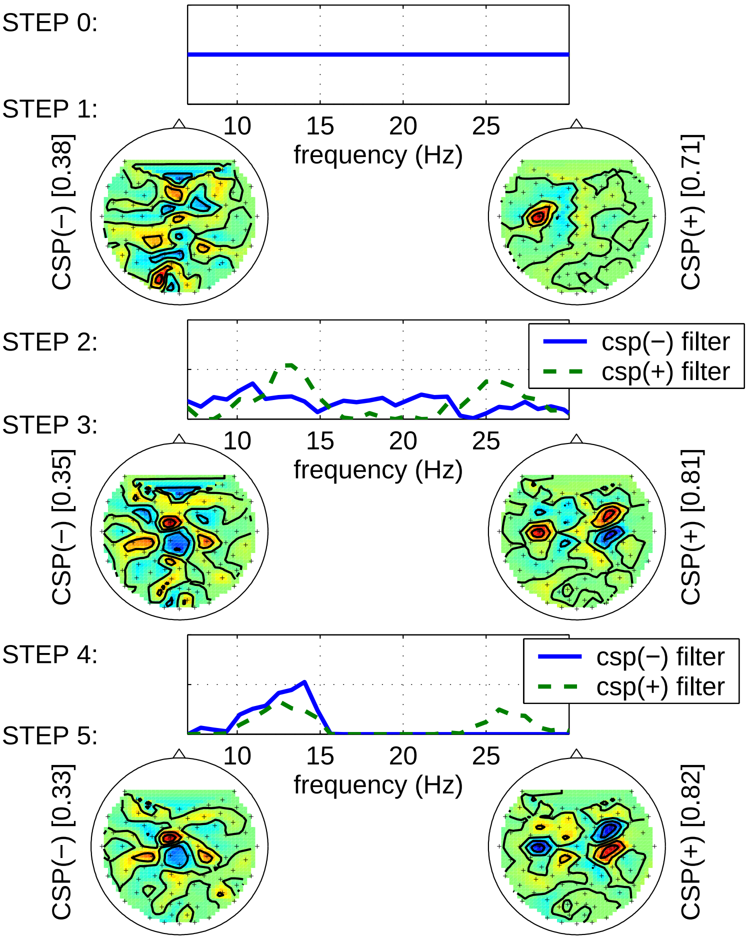


Figure 2: The topographical patterns of the CSP projections and the spectra of the filters are shown for each step of iteration for a Foot (−) vs. Left (+) dataset. The iteration starts from a homogeneous spectral filter (step 0) and the spatial projection and spectral filter are updated alternately (step 1-5). Note that although we use n_{of} features for each class, only the top patterns are shown here for the visualization purpose.

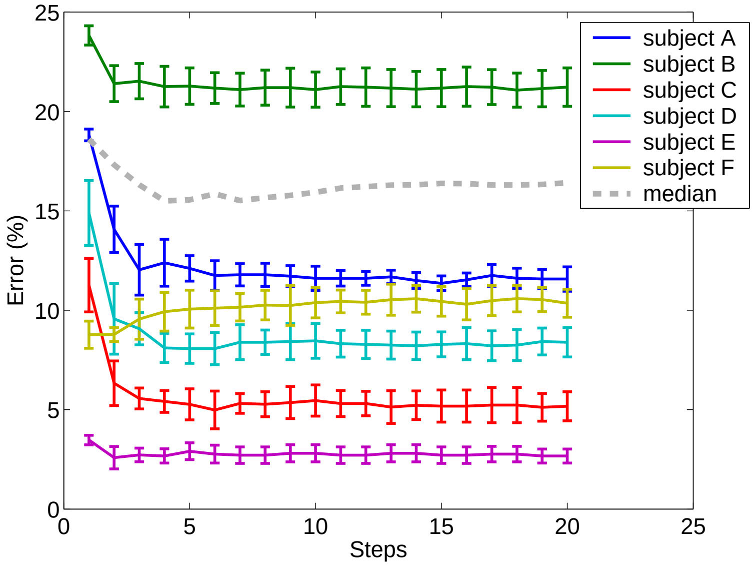


Figure 3: The 10×10 cross-validation errors of the proposed method for each step are shown for six subjects from very good classification accuracy (subject E) to moderate accuracy (subject B). The median over 162 datasets is also shown (dashed line). The hyperparameters were fixed at $p' = 0$ and $q' = 1$ (the elementwise product of Eqs. (4) and (6)). The odd steps correspond to spatial projection updates and the even steps are spectral filter updates. Note that the first step is the wide-band filtered CSP.

3.3 Comparison with conventional algorithms

Figure 4 shows the 10×10 cross validation errors of CSP and the proposed method for each dataset as a single data point. The hyperparameters are fixed at $p' = 0$ and $q' = 1$, which corresponds to the elementwise product of the theoretical optimum Eq. (4) and the prior filter Eq. (6). The iteration was performed 10 times, which contain 10 times of spatial projection updates and 10 times of spectral filter updates; therefore the total number of steps $n_{\text{step}} = 20$.

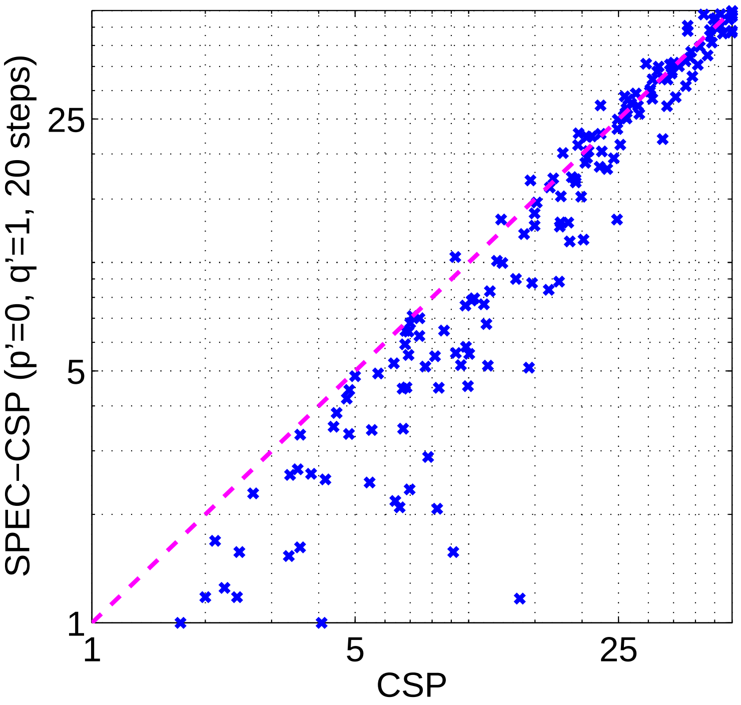


Figure 4: The 10×10 cross-validation errors of CSP and the proposed method on 162 datasets. Points lower than the diagonal correspond to datasets where the proposed method outperforms CSP. The hyperparameters for the proposed method were fixed at $p' = 0$ and $q' = 1$ (the elementwise product of Eqs. (4) and (6)). The iteration was performed 10 times, i.e., the number of steps $n_{\text{step}} = 20$. For a better visualization, data points outside of 1-50% intervals are shown on the edge of the figure box.

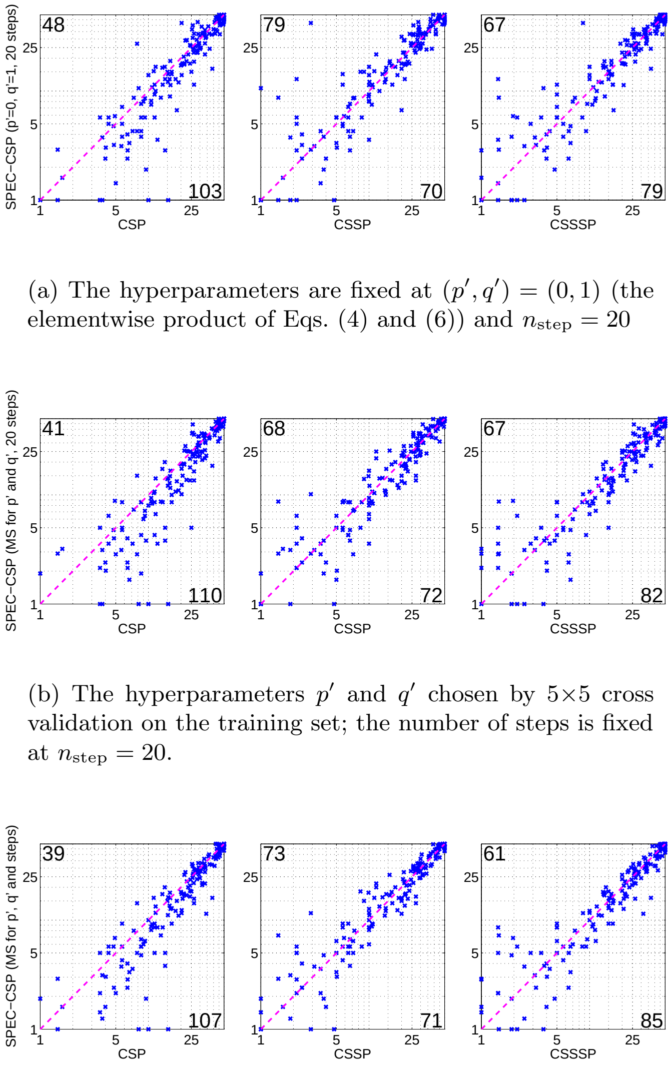
Figure 5 shows the test error of the proposed method against three conventional methods, namely CSP, CSSP, and CSSSP. All methods were trained on the first half of the dataset and applied to the remaining half (chronological validation). The time-lag parameter τ for CSSP was chosen out of all values from 0ms to 150 ms at 10ms interval by leave-one-out cross validation on the training set [18]; the regularization constant C for CSSSP was chosen out of $\{0, 0.01, 0.1, 0.2, 0.5, 1, 2, 5\}$ by 2×5-fold cross validation on the training set [19]. The hyperparameters for the proposed method were (a) all fixed at $p' = 0$, $q' = 1$ (the elementwise product of Eqs. (4) and (6)) and $n_{\mathrm{step}} = 20$; (b) p' and q' were chosen by 5×5 cross validation on the training set out of $p' \in \mathcal{P}' := \{-1, -0.5, 0, 0.5, 1\}$ and $q' \in \mathcal{Q}' := \{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$ and the number of steps was fixed at $n_{\mathrm{step}} = 20$; (c) all parameters were chosen out of $p' \in \mathcal{P}'$, $q' \in \mathcal{Q}'$ and $n_{\mathrm{step}} \in \{1, 2, \ldots, 20\}$. The improvement by choosing p' and q' by cross validation for each dataset (from Fig. 5(a) to Fig. 5(b)) is notable, however further choosing the number of steps n_{step} by cross validation yielded no significant improvement (from Fig. 5(b) to Fig. 5(c)). The fact that the proposed method with the fixed number of steps at $n_{\mathrm{step}} = 20$ performs as good as that with n_{step} chosen by cross validation implies the iterative procedure in the proposed method suffers no serious over-fitting problem.

3.4 Hyperparameter dependency

Figure 6 shows the contour plot of the average bitrate (per decision) over 162 datasets. The bitrate is defined as follows:

$$1 - \left(p_{\mathrm{err}} \log_2 \frac{1}{p_{\mathrm{err}}} + (1 - p_{\mathrm{err}}) \log_2 \frac{1}{1 - p_{\mathrm{err}}} \right), \qquad (8)$$

i.e., the capacity of a binary symmetric channel with the error probability p_{err} obtained by 4×4 cross validation for each combination of $(p', q') \in \mathcal{P}' \times \mathcal{Q}'$. For instance, a bitrate 0.38 corresponds to 15.4% error and also corresponds to 7.6 bits per minute provided that the subject can make a decision every 3 seconds. The maximum average bitrate is achieved in the area including $(p', q') = (0, 1)$ (the hyperparameters used in Sec. 3.3) The hyperparameters corresponding to CSP $((p', q') = (0, 0))$ or the prior filter outperforms the theoretically obtained filter $((p', q') = (-1, 1))$. We further elucidate the underlying rationale for incorporating both the theoretical optimum and the prior filter in Sec. 4



(a) The hyperparameters are fixed at $(p', q') = (0, 1)$ (the elementwise product of Eqs. (4) and (6)) and $n_{\text{step}} = 20$

(b) The hyperparameters p' and q' chosen by 5×5 cross validation on the training set; the number of steps is fixed at $n_{\text{step}} = 20$.

(c) The hyperparameters n_{step} as well as p' and q' are chosen by 5×5 cross validation on the training set.

Figure 5: The chronological test error of the proposed method compared to three conventional methods, namely CSP, CSSP, and CSSSP on 162 datasets. The time-lag parameter τ for CSSP and the regularization constant C for CSSSP were chosen by cross validation on the training set. The data points outside of 1-50% interval are shown on the edge of the figure box for a better visualization. The number of datasets lying above/below the diagonal is shown at top-left/bottom-right corners of each plot, respectively.

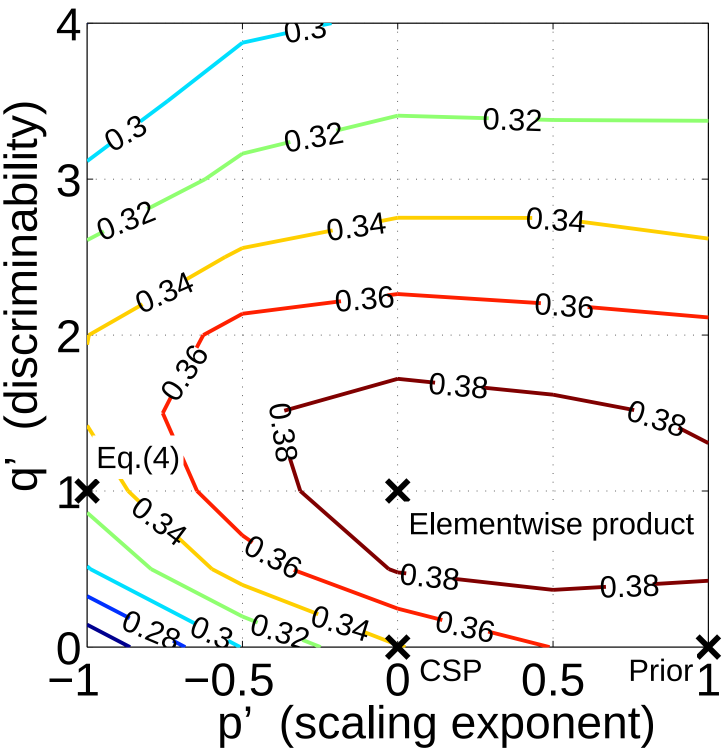


Figure 6: The contour plot of the average bitrate over 162 datasets in the two-dimensional hyperparameter space. The bitrate is defined as Eq. (8) with the error probability p_{err} obtained by 4×4 cross validation for each (p', q'). The number of steps $n_{\mathrm{step}} = 20$. The filter is defined as Eq. (7) with two hyperparameters p', the scaling exponent, and q', the discriminability. $(p', q') = (-1, 1)$ is the theoretical optimum (Eq. (4)). $(p', q') = (0, 0)$ corresponds to the wide-band filtered CSP. $(p', q') = (1, 0)$ is the prior filter itself (Eq. (6)). $(p', q') = (0, 1)$ corresponds to the elementwise product of Eqs. (4) and (6), which is used in Sec. 3.3.

4 Discussion: the effect of prior information

Since the solution (Eq. (4)) of the problem (3) has the form of "mean over variance", it scales with the power -1 with respect to the spectrum. Theoretically speaking, this is favorable because the filter compares all frequency components in a fair manner regardless of the power at each frequency component. In other words, it whitens the spectrum before the comparison. The scaling exponent $p' = -1$ is also favorable from another point of view, namely invariance; one can apply an arbitrary (non-zero) spectral filter to the signal before calculating Eq. (4) yet the effect is canceled out by Eq. (4). However, the cross-validation result in Sec. 3.4 shows that the filter having the scaling exponent $p' = 0$ is better than $p' = -1$ compared at any q'. This is analogous to the fact that the wide-band filtered CSP ($\forall \alpha_k = 1$) works quite well in general, because the scaling exponent $p' = 0$ implies that the power of the filtered signal is dominated by rhythmic activities, e.g., μ- and β-rhythms, which have overwhelmingly strong power.

In order to fill the gap between the theoretical scaling exponent $p' = -1$ and the empirically obtained scaling exponent $p' = 0$, here we carry out an additional validation. The validation consists of two steps. In the first step we optimize the spatial projection. Each dataset is band-pass filtered from 7-30Hz and the CSP projection with $n_{\mathrm{of}} = 3$ patterns for each class is calculated on the whole dataset. In the second step, in order to investigate the optimal design of a temporal filter, we conduct a cross-validation on the signal without pre-filtering. Note that this validation differs from that in section 3.4 in two folds: first, the optimization of the spatial projection was done on the whole dataset in the first step and fixed during the validation, second, the spatial projection was calculated on the pre-filtered signal but applied to the signal without pre-filtering. Furthermore, in the cross-validation we test two prior filters $\boldsymbol{\beta}$ namely,

- with the wide-band 7-30Hz assumption:

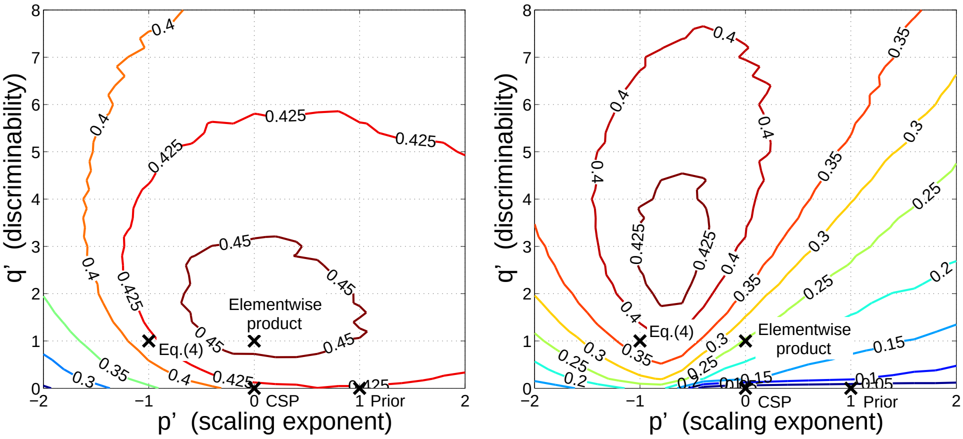
$$\beta_k = I_k^{[7,30]} \cdot \left(\langle s_k(\boldsymbol{w}) \rangle^+ + \langle s_k(\boldsymbol{w}) \rangle^- \right) / \, 2, \tag{9}$$

- without the assumption:

$$\beta_k = \left(\langle s_k(\boldsymbol{w}) \rangle^+ + \langle s_k(\boldsymbol{w}) \rangle^- \right) / \, 2. \tag{10}$$

Note that for the signal already band-pass filtered as in Sec. 3 the wide-band assumption is only useful in avoiding numerical instability and improving interpretability, whereas here with the signal without pre-filtering, the assumption imposes a real constraint on the design of a spectral filter. We test the filter (5) with the two prior filters for $p = p' + q'$ and $q = q'$ with $p' \in [-2, 2]$ and $q' \in [0, 8]$.

Figures 7(a) and 7(b) show the contour plot of the average bitrate for all combinations of $p' \in [-2, 2]$ and $q' \in [0, 8]$ on a 0.2 interval grid for the prior filters Eqs. (9) and (10), respectively. Figure 7(a) is similar to Fig. 6 where the maximum bitrate is achieved approximately at $(p', q') = (0, 1)$. However, it is clearer here, since the spatial projection is not recalculated, that the weighting of cross-spectrum matrices according to Eq. (4) improves the classification accuracy ($(p', q') = (0, 1)$ is better than $(p', q') = (0, 0)$) and incorporating the prior filter is also effective ($(p', q') = (0, 1)$ is better than $(p', q') = (-1, 1)$). On the other hand, Fig. 7(b) shows a completely different picture. Since the wide-band assumption is not adopted in the prior filter (Eq. (10)), it weights not only μ- and β-bands but also the brain activity lower than 7Hz, which has nothing to do with motor imagery task or even which cannot be considered a rhythmic activity. Thus the prior information is not so much useful anymore. The highest bitrate is now obtained in the area $p' < 0$ where the filter scales inversely to the spectrum. The theoretical optimum (Eq. (4)) is now the best performer among the wide-band filtered CSP, the prior filter (Eq. (10)), and the elementwise product of Eqs. (4) and (10). Note that however the overall bitrate is higher in Fig. 7(a) compared to that in Fig. 7(b). Therefore, in practice the wide-band assumption appears to help though the aim of this section was to show that in general it is necessary that the filter scales inversely to the power of the signal (Eq. (4)). Also note that since the all the trials in a dataset are used to calculate the spatial projection for each dataset, the bitrate does not reflect the real generalization performance in Fig. 7; thus one cannot directly compare Fig. 7 to Fig. 6.



(a) with the wide-band 7-30Hz assumption (see Eq. (9)).

(b) without the assumption (see Eq. (10))

Figure 7: The contour plot of the average bitrate over 162 datasets in the two-dimensional hyperparameter space. The bitrate is defined as Eq. (8) with the error probability p_{err} obtained by 4×4 cross validation. Unlike in section 3.4 the cross-validation was carried out on the signal without pre-filtering with pre-computed spatial projections. Points corresponding to the wide-band filtered CSP, the theoretically derived filter (Eq. (4)), the prior filter, and the elementwise product of the two filters $((p', q') = (0, 1))$ are marked.

5 Conclusion

In this paper, we have proposed a novel technique for spatio-temporal filter optimization in the context of single-trial EEG classification. The method works in the spatial domain and in the frequency domain alternately. The spatial projection optimization is a generalized version of CSP [17], in which a weighted sum of cross-spectrum matrices in the frequency domain is calculated for each class and then simultaneously diagonalized. The spectral filter, which is the weighting coefficients of the cross-spectrum matrices, is optimized through a novel optimization criterion. Since both the spatial and spectral filter optimization depends on each other, the two steps are iterated alternately.

The cross validation on 162 BCI datasets show improved classification accuracy compared to CSP [17] and comparable accuracy with CSSP [18] and CSSSP [19]. In comparison to CSP [17], we have shown that the non-homogenous weighting of the spectrum improves the classification accuracy. In comparison to CSSP [18], the problem of temporal filter optimization is directly addressed through a statistical criterion (3) and the new criterion has proven to be capable of handling more flexible and interpretable representation of a temporal filter without a serious over-fitting problem. In comparison to CSSSP [19], the proposed method is far more computationally efficient with comparable classification accuracy as well as being easily interpretable because the temporal filter is parameterized not as a FIR filter but in the frequency domain. Note that in online application it is straightforward to realize the obtained spectral filter as an AR filter or ARMA filter by various existing methods, e.g., the Yule-Walker method and its extensions.

Furthermore, we have investigated the best combination of the theoretical optimum (4) and the prior filter (6) by cross validation. We have found that the best combination is approximately obtained by the elementwise product of the theoretical optimum and the prior filter ($(p', q') = (0, 1)$, or $(p, q) = (1, 1)$ in the original parameterization). Moreover, we have found that CSP ($(p', q') = (0, 0)$) or the prior filter itself ($(p', q') = (1, 0)$) gives better classification accuracy than the theoretical optimum ($(p', q') = (-1, 1)$).

The fact that the models with a larger scaling exponent p' perform better than the theoretical optimum, motivated us to conduct an additional validation in order to investigate the effect of the wide-band 7-30Hz assumption on the optimal filter design. The validation was done on the signal without pre-filtering. Moreover, the recalculation of the spatial projection was not performed in order to ensure that the difference only arises from the temporal filter.

We have found that without the wide-band assumption, the prior filter, which assumes the discrimination to be found at frequency regions that are strongly active, fails because the activity below 7Hz will tend to dominate without contributing to discriminability. On the other hand, the theoretically optimal scale exponent $p' = -1$, which whitens the signal, has proven to be favorable than $p' = 0$ or $p' = 1$ in this situation. Thus, the "strong activity implies good discrimination" assumption that is behind the prior filter is only valid together with the wide-band assumption (7-30Hz). Note that either CSP or the elementwise product of the theoretical optimum and the prior filter ($(p', q') = (0, 1)$, or $(p, q) = (1, 1)$ in the original parameterization), which we have used in Sec. 3.3, already incorporates this prior knowledge. In fact, after band-pass filtering from 7-30Hz, a homogeneously weighted spectrum is dominated by some *a priori* important activities (e.g. μ- and β-rhythms).

The proposed method gives a highly interpretable spatial projection naturally because we solve the generalized CSP problem. In addition, the spectral representation of the temporal filter is favorable not only from the interpretability but also from providing possibility to incorporate any prior information about the spectral structure of the signal as we have demonstrated in section 3.

The applicability of the proposed method is not limited to brain signals because we use a very simple statistical criterion and we have clearly separated the effect of the statistical criterion and the prior knowledge specific to EEG signals in the implementation.

Acknowledgment

This research was partially supported by MEXT, Grant-in-Aid for JSPS fellows 17-11866 and Grant-in-Aid for Scientific Research on Priority Areas, 17022012, by BMBF-grant FKZ 01IBE01A/B, and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[1] Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. Clin. Neurophysiol. **113** (2002) 767–791

[2] Curran, E.A., Stokes, M.J.: Learning to control brain activity: A review of the production and control of EEG components for driving brain-computer interface (BCI) systems. Brain Cogn. **51** (2003) 326–336

[3] Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J., Birbaumer, N.: Brain-computer communication: Unlocking the locked in. Psychol. Bull. **127**(3) (2001) 358–375

[4] Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., Flor, H.: A spelling device for the paralysed. Nature **398** (1999) 297–298

[5] Pfurtscheller, G., Neuper, C., Guger, C., Harkam, W., Ramoser, R., Schlögl, A., Obermaier, B., Pregenzer, M.: Current trends in Graz brain-computer interface (BCI). IEEE Trans. Rehab. Eng. **8**(2) (2000) 216–219

[6] Pfurtscheller, G., Neuper, C., Müller, G., Obermaier, B., Krausz, G., Schlögl, A., Scherer, R., Graimann, B., Keinrath, C., Skliris, D., Woertz, M., Supp, G., Schrank, C.: Graz-BCI: state of the art and clinical applications. IEEE Trans. Neural Sys. Rehab. Eng. **11**(2) (2003) 177–180

[7] Pfurtscheller, G., Neuper, C., Birbaumer, N.: Human Brain-Computer Interface. In Riehle, A., Vaadia, E., eds.: Motor Cortex in Voluntary Movements. CRC Press, New York (2005) 367–401

[8] Pfurtscheller, G., Müller-Putz, G.R., Schlögl, A., Graimann, B., Scherer, R., Leeb, R., Brunner, C., Keinrath, C., Lee, F., Townsend, G., Vidaurre, C., Neuper, C.: 15 years of BCI research at Graz University of Technology: current projects. IEEE Trans. Neural Sys. Rehab. Eng. **14**(2) (2006) 205–210

[9] Blankertz, B., Dornhege, G., Schäfer, C., Krepki, R., Kohlmorgen, J., Müller, K.R., Kunzmann, V., Losch, F., Curio, G.: Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. IEEE Trans. Neural Sys. Rehab. Eng. **11**(2) (2003) 127–131

[10] Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.R., Kunzmann, V., Losch, F., Curio, G.: The Berlin Brain-Computer Interface: EEG-based communication without subject training. IEEE Trans. Neural Sys. Rehab. Eng. **14**(2) (2006) in press.

[11] Trejo, L., Wheeler, K., Jorgensen, C., Rosipal, R., Clanton, S., Matthews, B., Hibbs, A., Matthews, R., Krupka, M.: Multimodal neuroelectric interface development. IEEE Trans. Neural Sys. Rehab. Eng. (11) (2003) 199–204

[12] Parra, L., Alvino, C., Tang, A.C., Pearlmutter, B.A., Yeung, N., Osman, A., Sajda, P.: Linear spatial integration for single trial detection in encephalography. NeuroImage **7**(1) (2002) 223–230

[13] Penny, W.D., Roberts, S.J., Curran, E.A., Stokes, M.J.: EEG-based communication: A pattern recognition approach. IEEE Trans. Rehab. Eng. **8**(2) (2000) 214–215

[14] Pfurtscheller, G., da Silva, F.H.L.: Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin. Neurophysiol. **110**(11) (1999) 1842–1857

[15] Koles, Z.J.: The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. Electroencephalogr. Clin. Neurophysiol. **79** (1991) 440–447

[16] Koles, Z.J., Soong, A.C.K.: EEG source localization: implementing the spatio-temporal decomposition approach. Electroencephalogr. Clin. Neurophysiol. **107** (1998) 343–352

[17] Ramoser, H., Müller-Gerking, J., Pfurtscheller, G.: Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. Rehab. Eng. **8**(4) (2000) 441–446

[18] Lemm, S., Blankertz, B., Curio, G., Müller, K.R.: Spatio-spectral filters for improved classification of single trial EEG. IEEE Trans. Biomed. Eng. **52**(9) (2005) 1541–1548

[19] Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., Müller, K.R.: Combined optimization of spatial and temporal filters for improving brain-computer interfacing. IEEE Trans. Biomed. Eng. (2006) accepted.

[20] Dornhege, G., Blankertz, B., Curio, G., Müller, K.R.: Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. IEEE Trans. Biomed. Eng. **51**(6) (2004) 993–1002