2006 G/>pJsO@E*3X=,M]O@%o!<%/%7%9%C%W
2006 Workshop on Information-Based Induction Sciences (IBIS2006)
Osaka, Japan, October 31 - November 2, 2006.

Adapting Spatial Filtering Methods for Nonstationary BCIs

Ryota Tomioka *†‡ Jeremy Hill § Benjamin Blankertz† Kazuyuki Aihara*

Abstract: A major challenge in applying machine learning methods to Brain-Computer Interfaces (BCIs) is to overcome the possible nonstationarity in the data from the datablock the method is trained on and that the method is applied to. Assuming the joint distributions of the *whitened* signal and the class label to be identical in two blocks, where the whitening is done in each block independently, we propose a simple adaptation formula that is applicable to a broad class of spatial filtering methods including ICA, CSP, and logistic regression classifiers. We characterize the class of linear transformations for which the above assumption holds. Experimental results on 60 BCI datasets show improved classification accuracy compared to (a) fixed spatial filter approach (no adaptation) and (b) fixed spatial pattern approach (proposed by Hill et al., 2006 [1]).

1 Introduction

Brain-Computer Interfaces (BCIs) are devices that translate the intent of a subject measured from brain signals directly into control commands, e.g. for a computer application or a neuroprosthesis [2]. We focus on EEG based imaginary movement BCIs. From the machine learning point of view, the task is to predict the class of imagination y from the EEG signal X of a single trial based on training examples $\{X_i, y_i\}_{i=1}^n$. Consider a situation when a classifier is trained on one block of recording and then applied to another block. A potential drawback is that the characteristic of the signal can be considerably affected by altered mental states with respect to, e.g., concentration or excitedness, variable demands in visual processing, or changes in the impedance of the electrodes. The challenge is to adapt the classifier which is optimized on the first block to the second block. If one could access the labels in the second block, the problem becomes considerably easier. However, in practice, especially in online BCI experiments, the real intention of the controller is unknown to the system. Therefore, we restrict ourselves to accessing only the marginal distribution of X, in particular to the estimation of the covariance matrix Σ_2 on the new datablock. Such an adaptation is a highly important step, since the calibration measurement is typically a repetitive task without feedback, in which the richness of available stimuli and the subject's level of arousal are relatively low, in contrast to the more dynamic phase of the actual BCI operation.

This study was conducted within the Berlin Brain-Computer Interface (BBCI) project which develops an EEG-based system operating on the spatio-spectral changes during different kinds of motor imagery. The BBCI uses machine learning techniques to adapt to the specific brain signatures of each user ([3]).

2 Materials

For the demonstration of the proposed method we will use 60 EEG datasets from 20 experiments recorded from 16 subjects (some subjects took part in more than one experiment). In each experiment subjects performed trials of cued imaginary movements with an inter-cue interval of 6-7s. There was no feedback given in these recordings. The subjects performed one of the three imaginary movements namely, left(L), right(R), or foot(F) for 3-3.5s in each trial. There were two different blocks of measurements; in the block called "lett", subjects were fixating at the center of the screen and the instruction of the imaginary movement he/she should perform was given as the corresponding letter fixed at the center of the screen. On the other hand, in the "move" block, subjects were fixating at a cursor randomly bouncing inside the screen (the movement being uncorrelated with the required task). The instruction was given as the change in the shape of the cursor. For most subjects, both blocks contained equally 70 trials for each class, but for some subjects it was 35 trials in the first block and 105 trials in the second block (or the other way around). Since we concentrate on the binary classification problem, 20 experiments produced 60 datasets by taking all the binary combination of three classes.

A first inspection of the datasets reveals a systematic difference of brain activity during the "lett" and the "move" blocks: the log band-power (7-30Hz) shows

*Dept. Mathematical Informatics, IST, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656, Japan.
†Fraunhofer FIRST (IDA), Kekuléstr. 7, 12 489 Berlin, Germany.
‡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.
§MPI for Biological Cybernetics, Spemannstrase 38, 72 076 Tübingen, Germany.

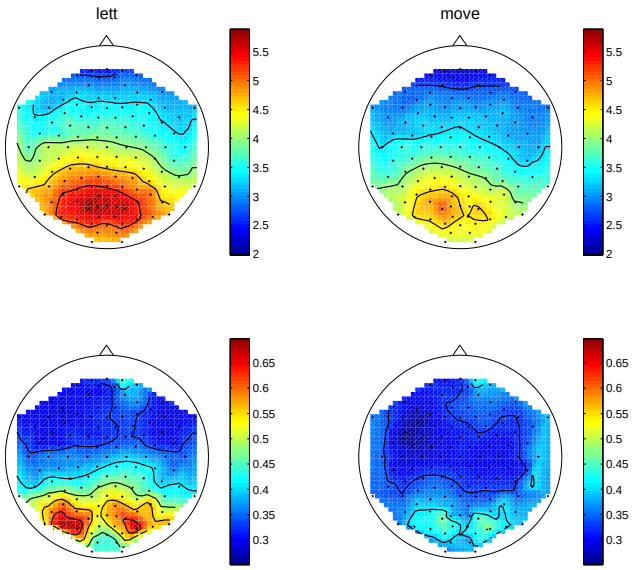


Figure 1: The upper scalp maps show the average log band-power (7-30Hz) in the datasets "lett" and "move" calculated by taking the log-variance of the band-pass filtered signals. The maps in the lower row show the trial-to-trial standard deviation of log band-power. The maps are calculated for one representative subject with good BCI performance.

a much stronger parietal activity during "lett" compared to "move", see upper row of Figure 1. Similarily the variation of log band-power is larger in that area, as can be seen in the lower rows of Fig. 1.

3 Spatial filtering methods

Let $X \in \mathbb{R}^{d \times T}$ be the EEG signal of a single trial with d channels and T sampled time points[1]. We consider a binary classification problem where each class, e.g. right or left hand imaginary movement, is called positive $(+)$ or negative $(-)$ class. Let $y \in \{+1, -1\}$ be the class label. Given a set of trials and labels $\{X_i, y_i\}_{i=1}^n$, the task is to predict the class label y for an unobserved trial X.

In this section, we show two examples of single trial EEG classification methods that the proposed adaptation method can be applied, namely, Common Spatial Pattern (CSP) [4] and the logistic regression classifier with rank=2 approximation [5].

3.1 Common spatial pattern

Common Spatial Pattern (CSP) [4] is a spatial filtering method widely used in motor imagination based BCIs [6, 3], where the task is to classify two different state of brain activity, e.g., imagining the movement of the left or the right hand. In this context, the event related (de-)synchronization (ERD/ERS; [7]) of rhythmic brain activity is a widely used and well studied physiology. The EEG signal is commonly band-pass filtered around μ- (7-15Hz) and/or β- (15-30Hz) rhythms and two covariance matrices $\Sigma^{(+)}$ and $\Sigma^{(-)}$ are calculated for the two classes. CSP tries to find a spatial filter $\boldsymbol{w} \in \mathbb{R}^d$ that maximizes the difference in the average band power of the filtered signal while keeping the sum constant.

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \boldsymbol{w}^{\mathrm{T}}(\Sigma^{(+)} - \Sigma^{(-)})\boldsymbol{w}, \\ \text{s.t.} \quad & \boldsymbol{w}^{\mathrm{T}}(\Sigma^{(+)} + \Sigma^{(-)})\boldsymbol{w} = 1. \end{aligned} \tag{1}$$

The problem can be solved through a generalized eigenvalue problem:

$$\left(\Sigma^{(+)} - \Sigma^{(-)}\right) W = \left(\Sigma^{(+)} + \Sigma^{(-)}\right) W \Lambda. \tag{2}$$

The matrix of generalized eigenvectors $W \in \mathbb{R}^{d \times d}$ are written as $W = PR$ where $P = (\Sigma^{(+)} + \Sigma^{(-)})^{-1/2}$ and R is the eigenvector of $P^{\mathrm{T}}(\Sigma^{(+)} - \Sigma^{(-)})P$. It is easy to see that with the eigenvector $\boldsymbol{r}_1$ corresponding to the largest eigenvalue, the optimum of the above problem is obtained as $\boldsymbol{w}^* = P\boldsymbol{r}_1$.

The CSP based classifier has the following form,

$$f_{\mathrm{CSP}}(X; W, \alpha) = \sum_{j=1}^{p} \alpha_j \log \boldsymbol{w}_j^{\mathrm{T}} X X^{\mathrm{T}} \boldsymbol{w}_j + \alpha_0.$$

Here the classifier is trained in two stages. First, we take $p = 2n_{\mathrm{of}}$ columns of W from the generalized eigenvalue decomposition (Eq. (2)), namely eigenvectors corresponding to n_{of} largest and smallest eigenvalues. Second, the coefficients $\{\alpha_j\}_{j=0}^{2n_{\mathrm{of}}}$ are learned through linear discriminant analysis (LDA).

3.2 Logistic regression classifier with rank=2 approximation

The logistic regression classifier [5] is also a method developed for ERD based motor imaginary BCI. The key idea in CSP, which is to *classify two classes of zero mean time series with different covariance matrices*, is formulated in a more principled manner in a logistic regression framework. In [5] the rank=2 approximation of the symmetric logit transform of the posterior class probability $\log P(y = +1|X)/P(y = -1|X)$ is proposed.

$$f(X; \boldsymbol{\theta}) = \frac{1}{2}\mathrm{tr}\left[\left(-\boldsymbol{w}_1\boldsymbol{w}_1^{\mathrm{T}} + \boldsymbol{w}_2\boldsymbol{w}_2^{\mathrm{T}}\right) X X^{\mathrm{T}}\right] + b, \tag{3}$$

where $\boldsymbol{\theta} := (\boldsymbol{w}_1, \boldsymbol{w}_2, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$. The function $f(X; \boldsymbol{\theta})$ is optimized in a regularized maximum likelihood problem,

$$\begin{aligned} \min_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d, b \in \mathbb{R}} & \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i f(X_i; \boldsymbol{\theta})}\right) \\ & + \frac{C}{2}\left(\boldsymbol{w}_1^{\mathrm{T}}\Sigma_1\boldsymbol{w}_1 + \boldsymbol{w}_2^{\mathrm{T}}\Sigma_1\boldsymbol{w}_2\right), \end{aligned} \tag{4}$$

[1] For simplicity, we assume that the DC component 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 $X = \frac{1}{\sqrt{T}} X_{\mathrm{original}} \left(I_T - \frac{1}{T}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)$.

where Σ_1 is the covariance matrix of X.

Note that for the above two methods the invariance to linear transformation holds, which forms the basis of the proposed *normalizing* approach for spatial filter adaptation (see Sec. 4.1).

Remark 1 *The above two methods satisfies the following property: when the data X is transformed as $\tilde{X} = AX$, the optimal spatial filter W is transformed as $\tilde{W} = A^{-\mathrm{T}}W$, where $A^{-\mathrm{T}} = \left(A^{-1}\right)^{\mathrm{T}}$.*

The above invariance holds for broad class of well behaving linear spatial filtering methods, for example Independent Component Analysis (ICA).

Furthermore, the following decomposition is possible for the above two methods, which forms the basis of *fixed spatial pattern* approach (see Sec. 4.2 and [1]).

Remark 2 *The filter coefficient matrix $W \in \mathbb{R}^{d\times p}$ in the above two methods can be decomposed as follows*

$$W = PRD. \tag{5}$$

Here, $P = \Sigma_1^{-1/2} \in \mathbb{R}^{d\times d}$ is the whitening part*, where $\Sigma_1 = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^{\mathrm{T}}$ is the covariance matrix of the signal X on the datablock the method is trained on; $R \in \mathbb{R}^{d\times p}$ is a set of p orthonormal vectors in $\mathbb{R}^d$ (*orthogonal part*); $D \in \mathbb{R}^{p\times p}$ is a diagonal matrix (*scaling*).*

Proof: For CSP, the proof is trivial. For the logistic regression classifier, one can easily see that the problem (4) can be considerably simplified by transforming the variables as $W = [\boldsymbol{w}_1, \boldsymbol{w}_2] = P\tilde{W}$ with a whitening matrix $P = \Sigma_1^{-1/2}$ and optimizing on $\tilde{W} = [\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2] \in \mathbb{R}^{d\times 2}$.

$$\min_{\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2 \in \mathbb{R}^d, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i f(\tilde{X}_i; \tilde{\boldsymbol{\theta}})}\right) + \frac{C}{2}\left(\|\tilde{\boldsymbol{w}}_1\|^2 + \|\tilde{\boldsymbol{w}}_2\|^2\right), \tag{6}$$

where $\tilde{X} = PX$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2, b)$. Furthermore when the regularization constant $C > 0$, it is shown in the following lemma that the optimal $\tilde{W}$ can be decomposed as $\tilde{W} = RD$, which completes the proof.

□

Lemma 3 *In the whitened version of logistic regression with rank=2 approximation (Eq. (6)), for every $\tilde{W}$, one can find a "orthogonalization" $\tilde{W}_o = \tilde{W}B = RD$ with $R = [\boldsymbol{r}_1, \boldsymbol{r}_2] \in \mathbb{R}^{d\times 2}$ that satisfies $R^{\mathrm{T}}R = I_2$ and a diagonal matrix $D \in \mathbb{R}^{2\times 2}$ without changing the function (3), i.e., $\forall X$, $f(X; \tilde{W}, b) = f(X; \tilde{W}_o, b)$. Furthermore, the orthogonalization always decreases the regularization term in Eq. (6).*

Proof: See Appendix.

4 Adaptation methods

4.1 Normalizing approach

We assume the following (the assumption is justified *empirically* in Sec. 5),

Assumption 1 *The joint distributions of the whitened signal $Z = \Sigma^{-1/2}X$ and the label y, namely $P_{Z,y}(Z, y)$ are identical in both datablocks, i.e.,*

$$P^{(1)}_{X,y}\left(\Sigma_1^{1/2}Z, y\right)\left|\Sigma_1^{1/2}\right| = P^{(2)}_{X,y}\left(\Sigma_2^{1/2}Z, y\right)\left|\Sigma_2^{1/2}\right|,$$

where $P^{(1)}_{X,y}$ and $P^{(2)}_{X,y}$ are the probability densities of the datablock the method is trained on and that it is applied to, respectively and Σ_1 and Σ_2 are covariance matrices similarly defined.

Thus, in order to predict the label for an unobserved X from the second block, the classifier should be trained on $\{\Sigma_2^{1/2}\Sigma_1^{-1/2}X_i, y_i\}_{i=1}^n$ instead of $\{X_i, y_i\}_{i=1}^n$. In the case of the spatial filtering methods described in Sec. 3, because they are invariant to linear transformation (see Remark 1), one only needs to transform the spatial filter W according to the following adaptation formula, which we call the *normalizing approach*:

$$W_{\text{adapt}} = \Sigma_2^{-1/2}\Sigma_1^{1/2}W. \tag{7}$$

Note that Eq. (7) corresponds to keeping the orthogonal part R and replacing the whitening part P with the new whitening $\Sigma_2^{-1/2}$ in Eq. (5), which is the direct consequence of Assumption 1. In fact, the following statement holds,

Remark 4 *For a spatial filtering method that satisfies Remarks 1 and 2, the orthogonal part R in Eq. (5) is kept constant to a linear spatial transformation if and only if the transformation is written as follows $\tilde{X} = C\Sigma_1^{-1/2}X$, where C is an arbitrary symmetric positive definite matrix.*

The proof is trivial. However, it must be noted that the invariance to linear transformation (Remark 1) does not imply that the classifier trained on the first block can be applied to the second datablock that is transformed with an arbitrary linear transformation; in general the transformation A in Remark 1 *cannot* be assessed without the labels on the second block; Remark 4 identifies the special class of linear transformations that the transformation can be assessed only from the estimation of covariance matrix Σ_2 on the second block.

In this paper, we use the batch estimation of Σ_2 on the whole test block; one can also estimate Σ_2 in an online manner. We call our method the normalizing approach because it normalizes the covariance of each measurement block independently to identity.

4.2 Fixed spatial pattern approach

The fixed spatial pattern (FSP) approach proposed in [1] assumes that the task relevant (or discriminative) columns of $A = W^{-\mathrm{T}}$, namely $A^{[r]} \in \mathbb{R}^{d\times p}$ are kept constant while allowing irrelevant columns to change. Throughout this paper, we refer to each column of A as *spatial pattern* corresponding to the same column of W, which we call *spatial filter*, because they correspond one-to-one by a transformation,

$$A = \Sigma_1 W D^{-2}. \tag{8}$$

The above principle gives the following adaptation rule for the filter (derivation see below),

$$W^{[r]}_{\mathrm{adapt}} = \Sigma_2^{-1}\Sigma_1 W^{[r]} \left(W^{[r]\mathrm{T}}\Sigma_1\Sigma_2^{-1}\Sigma_1 W^{[r]}\right)^{-1} \times \left(W^{[r]\mathrm{T}}\Sigma_1 W^{[r]}\right), \tag{9}$$

where $W^{[r]} = \Sigma_1^{-1}A^{[r]}D^2$ are the filters corresponding to the task relevant patterns $A^{[r]}$. In CSP, the task relevant components are chosen according to the magnitude of the eigenvalues of the generalized eigenvalue problem (Eq. (2)) as described in Sec. 3.1. In the logistic regression classifier, setting $W = [\boldsymbol{w}_1, \boldsymbol{w}_2]$ in the right hand side of Eq. (8) gives the pattern $A \in \mathbb{R}^{d\times 2}$ to be preserved. Equation (9) is derived according to the method described in [1] as follows:

Derivation of Eq. (9): We first split patterns into task-relevant part and irrelevant part $A = \left[A^{[r]}, A^{[i]}\right]$, which has the following form under the new whitening $P_2 = \Sigma_2^{-1/2}$,

$$A_2 = P_2^{-1}\left[C, U\right],$$

where C is called the constrained part, which should satisfy the following

$$A^{[r]} = P_2^{-1}C, \tag{10}$$

and U is the unconstrained part. We assume that C and U are orthogonal subspaces. Then, the filter can be written as follows:

$$\begin{aligned} W_{\mathrm{adapt}} = A_2^{-\mathrm{T}} &= P_2\left[C, U\right]^{-\mathrm{T}} \\ &= P_2\left[C(C^{\mathrm{T}}C)^{-1}, U(U^{\mathrm{T}}U)^{-1}\right]. \end{aligned}$$

The task-relevant part of the adapted filter can be written without the unconstrained part. Using Eq. (10) we have,

$$W^{[r]}_{\mathrm{adapt}} = P_2^2 A^{[r]}(A^{[r]\mathrm{T}}P_2^2A^{[r]})^{-1}.$$

Finally, substituting $P_2^2 = \Sigma_2^{-1}$, Eq. (8) and $D^2 = W^{[r]\mathrm{T}}\Sigma_1 W^{[r]}$, we obtain Eq. (9).

□

5 Results

We evaluate the following three approaches for spatial filter adaptation, namely (a) fixed spatial filter (FSF) approach, in which no adaptation is performed as baseline, (b) fixed spatial pattern (FSP) approach (Eq. (9); see also [1]) and (c) the normalizing approach (Eq. (7)). As the spatial filtering methods to be adapted, we choose CSP (Eq. (1)) and logistic regression with rank=2 approximation (Eq. (4)).

We band-pass filter the EEG signals from the 60 datasets described in Sec. 2 at the μ- and β-range, i.e. 7-30Hz and cut out the 500-3500ms interval after the appearance of visual stimuli as a trial X. Signals were sampled at 1000Hz but down-sampled to 100Hz before processing.

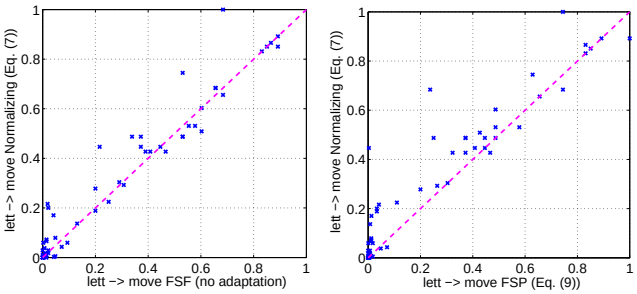
In each dataset, we trained a classifier on one of the two datablocks and try to adapt the classifier based only on additionally estimating the covariance matrix of the other datablock. No label information is used. We evaluated the adaptation methods on both directions, namely, "lett"→"move" and "move"→"lett". The parameter $n_{\mathrm{of}} = 3$ is used for CSP. The regularization constant C for the logistic regression classifier is chosen by 2×10 cross validation on the training block.

Figure 2 shows the results of adapting CSP. The upper row shows the results of "lett"→"move" and the lower row shows the results of "move"→"lett". The bitrate of a pair of methods are plotted for each dataset. Here, bitrate (per decision) is defined based on the classification test error rate p_{err} as the capacity of a binary symmetric channel with the same error probability: $1 - \left(p_{\mathrm{err}}\log_2\frac{1}{p_{\mathrm{err}}} + (1-p_{\mathrm{err}})\log_2\frac{1}{1-p_{\mathrm{err}}}\right)$. The first column in Fig. 2 compares the normalizing approach to FSF approach. The second column compares the normalizing approach to FSP approach. The normalizing approach clearly outperforms FSF in the case of "move"→"lett", however the improvement is not clear in the case of "lett"→"move". The normalizing approach appears favorable to FSP in both cases.

Figure 3 shows the same comparison for the logistic regression classifier. A similar characteristics can also be seen here: the improvement of the normalizing approach is much clearer in the "move"→"lett" situation. In addition, FSP also shows much better performance in the "move"→"lett" situation.

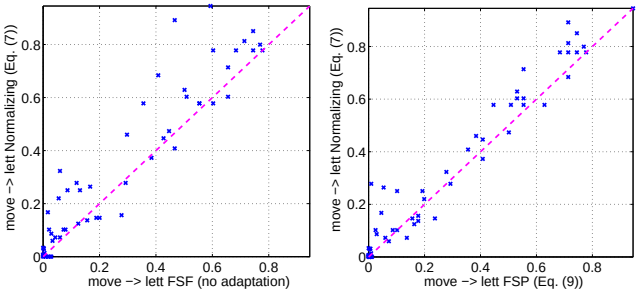
6 Discussion

Why do we see more improvements for "move" → "lett" than "lett"→"move"? The "lett" block contains increased parietal α-activity and also more variability in that area, see Fig. 1. Since this variability is uncorrelated to the task, a classifier trained on "lett" becomes invariant to those variations. On the other hand, a classifier trained on "move" might fail to become invariant in this respect due to the lack of variation in the training data and then fail on block "lett".



(a) Comparison of the normalizing approach to FSF for "lett"→"move".

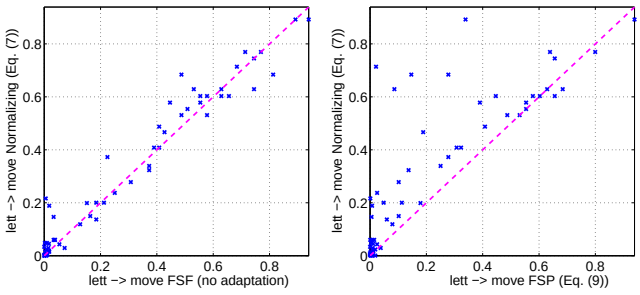
(b) Comparison of the normalizing approach to FSP for "lett"→"move".



(c) Comparison of the normalizing approach to FSF for "move"→"lett".

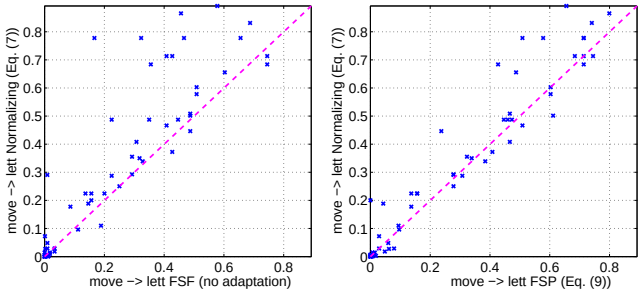
(d) Comparison of the normalizing approach to FSP for "move"→"lett".

Figure 2: Adaptation of CSP (see Sec. 3.1).



(a) Comparison of the normalizing approach to FSF for "lett"→"move".

(b) Comparison of the normalizing approach to FSP for "lett"→"move".



(c) Comparison of the normalizing approach to FSF for "move"→"lett".

(d) Comparison of the normalizing approach to FSP for "move"→"lett".

Figure 3: Adaptation of the logistic regression classifier (see Sec. 3.2).

Why isn't the effect of increased α-activity localized in a CSP component? There are two possible reasons. First, if the occipital α-activity, which mainly arises from the visual cortex is statistically independent from the task relevant sources in the motor cortex, the task relevant components of W from *ICA* will not depend on the the α-activity. However, this may not be case for *CSP*; because the labels are taken into account, CSP components are classification biased. Second, as considered in [1] if the change in the α-activity involves a linear transformation of a task-irrelevant subset of columns of A, the effect will not be localized in the corresponding subset of $W = A^{-\mathrm{T}}$. **Why is the proposed normalization approach more stable than FSP?** Comparing Eq. (7) to Eq. (9), one can see that the major correction term is $\Sigma_2^{-1/2}\Sigma_1^{1/2}$ for the proposed approach and $\Sigma_2^{-1}\Sigma_1$ for FSP. Because the correction of scale is of the order $1/2$ instead of 1, the risk of overly correcting the shift when the actual shift is small, is smaller in the proposed method. **Why is keeping the filter or the pattern not appropriate?** It can be conjectured that because both depends on the whitening part $P = \Sigma^{-1/2}$, which is influenced by difference in the scale of the signal from a datablock to another.

Fig. 4 shows logistic regression (Eq. (3)) coefficients topographically mapped on a head viewed from top (nose pointing up). In this dataset, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are called "left" filter and "right" filter because they have large response in the "left hand" and "right hand" motor imagination, respectively. Both the coefficients obtained from "lett" block and "move" block are shown. Three rows corresponds from top to bottom (1) the filter ($W = \Sigma^{-1/2}R$) (2) the orthogonal part (R) and (3) the pattern ($A = \Sigma^{1/2}R$). The scaling coefficient D is omitted. One can see that although the orthogonal parts (the middle row) are mainly focusing on the motor area, the patterns (the bottom row) have broad spread in the occipital direction, which is substantially different from "lett" block to "move" block, reflecting the variability in the activity in this area. The filters, which have to compensate for this activity, also seem to be different from "lett" block to "move" block.

7 Summary

We have proposed a novel formula for adapting a spatial filtering method which is trained on a block of recording and being applied to another block which possibly has a different distribution. The formula can be applied to broad class of spatial filtering methods

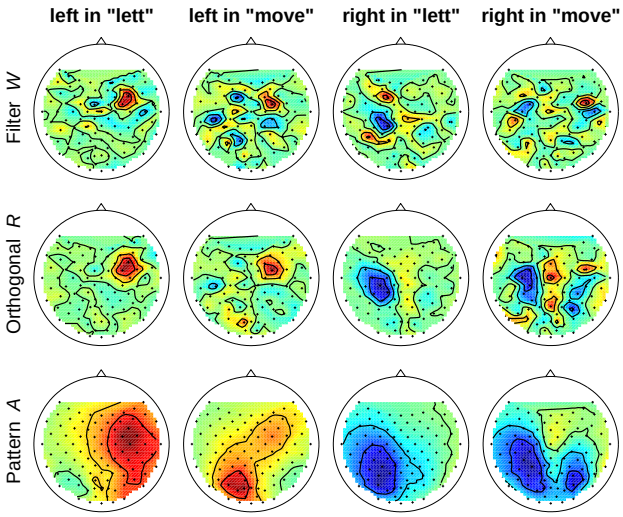


Figure 4: The filters, the orthogonal part, and the patterns from two different blocks of recording "lett" and "move" for the same subject in the same day using logistic regression with rank=2 approximation (see Sec. 3.2 and [5]).

for BCI, e.g., ICA, CSP or the logistic regression classifier and requires no label information on the test set; only the covariance matrix of the new data is required. The formula is equivalent to datablock-wise normalization by whitening transformation. The underlying assumption that the joint distribution of the whitened signal and the label is kept unchanged was tested on 60 BCI datasets. The result shows a improved classification compared to FSF approach (no adaptation) and FSP adaptation, which preserves the task-relevant "patterns", i.e., a subset of columns of the inverse filter matrix corresponding to task-related components. In fact, the orthogonal part seems to be relatively preserved even under a large change in the filter and the pattern (see Fig. 4). At the moment, the method uses batch estimation of covariance matrix on the whole test block. In order to apply the method in the BCI feedback experiment, the estimation has to be done in an online manner. Theoretical justification of Assumption 1 is also necessary.

Appendix: Proof of Lemma 1

There exists $A \in \mathbb{R}^{2\times 2}$ that satisfies the following two conditions,

$$A^{\mathrm{T}}\tilde{W}^{\mathrm{T}}\tilde{W}A = I_2,$$
$$A^{-1}\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix}A^{-\mathrm{T}} = \begin{pmatrix}-d_1^2 & 0\\ 0 & d_2^2\end{pmatrix},$$

because $\tilde{W}^{\mathrm{T}}\tilde{W}$ is a positive definite matrix and the left hand side of the second equation is symmetric. Then, the following holds,

$$-\tilde{\boldsymbol{w}}_1\tilde{\boldsymbol{w}}_1^{\mathrm{T}} + \tilde{\boldsymbol{w}}_2\tilde{\boldsymbol{w}}_2^{\mathrm{T}} = \tilde{W}\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix}\tilde{W}^{\mathrm{T}}$$
$$= \tilde{W}AA^{-1}\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix}A^{-\mathrm{T}}A^{\mathrm{T}}\tilde{W}^{\mathrm{T}}$$
$$= R\begin{pmatrix}d_1 & 0\\ 0 & d_2\end{pmatrix}\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}d_1 & 0\\ 0 & d_2\end{pmatrix}R^{\mathrm{T}}.$$

Here $R = \tilde{W}A$ satisfies $R^{\mathrm{T}}R = I_2$ from the first condition. Therefore, letting $B = A\left(\begin{smallmatrix}d_1 & 0\\ 0 & d_2\end{smallmatrix}\right)$ we have the first part of the lemma. In order to prove the second part, let us write $A^{-1} = \left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right)$. From a simple calculation,

$$A^{-1}\left(\begin{smallmatrix}-1 & 0\\ 0 & 1\end{smallmatrix}\right)A^{-\mathrm{T}} = \begin{pmatrix}-a^2+b^2 & -ac+bd\\ -ac+bd & -c^2+d^2\end{pmatrix} = \begin{pmatrix}-d_1^2 & 0\\ 0 & d_2^2\end{pmatrix}.$$

Furthermore, since R is orthogonal,

$$\|\tilde{\boldsymbol{w}}_{1(o)}\|^2 + \|\tilde{\boldsymbol{w}}_{2(o)}\|^2 = d_1^2 + d_2^2$$
$$= a^2 - b^2 - c^2 + d^2 \le a^2 + b^2 + c^2 + d^2$$
$$= \mathrm{tr}\left[A^{-\mathrm{T}}A^{-1}\right] = \mathrm{tr}\left[\tilde{W}^{\mathrm{T}}\tilde{W}\right]$$
$$= \|\tilde{\boldsymbol{w}}_1\|^2 + \|\tilde{\boldsymbol{w}}_2\|^2. \qquad \square$$

References

[1] Hill, N.J., Farquhar, J., Lal, T.N., Schölkopf, B.: Time-dependent demixing of task-relevant EEG sources. In: Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006, Verlag der Technischen Universität Graz (2006) accepted.

[2] 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

[3] 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.

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

[5] Tomioka, R., Aihara, K., Müller, K.R.: Logistic Regression for Single Trial EEG Classification. In Schölkopf, B., Platt, J., Hofmann, T., eds.: Advances in Neural Information Processing Systems 19, Cambridge, MA, MIT Press (2007) accepted.

[6] 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

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