CSP patches: an ensemble of optimized spatial filters. An evaluation study

Claudia Sannelli¹, Carmen Vidaurre¹, Klaus-Robert Müller¹,2 and Benjamin Blankertz¹,2,3

¹ Machine Learning Laboratory, Berlin Institute of Technology, Berlin, Germany
² Bernstein Focus: Neurotechnology, Berlin, Germany
³ Intelligent Data Analysis Group, Fraunhofer FIRST, Berlin, Germany

E-mail: claudia.sannelli@tu-berlin.de

Received 31 July 2010
Accepted for publication 25 November 2010
Published 24 March 2011
Online at stacks.iop.org/JNE/8/025012

Abstract
Laplacian filters are widely used in neuroscience. In the context of brain–computer interfacing, they might be preferred to data-driven approaches such as common spatial patterns (CSP) in a variety of scenarios such as, e.g., when no or few user data are available or a calibration session with a multi-channel recording is not possible, which is the case in various applications. In this paper we propose the use of an ensemble of local CSP patches (CSPP) which can be considered as a compromise between Laplacian filters and CSP. Our CSPP only needs a very small number of trials to be optimized and significantly outperforms Laplacian filters in all settings studied. Additionally, CSPP also outperforms multi-channel CSP and a regularized version of CSP even when only very few calibration data are available, acting as a CSP regularizer without the need of additional hyperparameters and at a very low cost: 2–5 min of data recording, i.e. ten times less than CSP.

1. Introduction

Brain–computer interfaces (BCIs) decode brain activity to establish a direct communication channel from a person to a computer system, e.g., for rehabilitation after stroke or to support daily activities of patients suffering from amyotrophic lateral sclerosis (ALS).

BCIs based on sensory motor rhythms (SMR) use the modulation of the spontaneous rhythm of the electroencephalogram (EEG) in the α and/ or β frequency band over the somatosensory cortex which is caused by the imagination of a limb movement (left hand, right hand, one or both feet), cf [23]. The voluntary SMR modulation evoked by motor imagery can be detected by the BCI and thus allows control of applications.

The machine learning (ML) approach to BCI utilizes newly developed algorithms which learn subject-specific parameters and adapt automatically to the user’s brain signals, see [3, 22]. The experimental design for ML-based BCIs typically includes a calibration session in which the user does not obtain feedback. Thus collected calibration data are used to train the ML methods, after which the true BCI feedback application may start, cf [29].

Spatial filters which are optimized specifically for each user by a common spatial pattern (CSP) analysis [25] are particularly successful for this application (see [4] for a review). The drawback of this method is that it requires a considerable number of training data, resulting in long and exhausting calibration runs difficult to perform outside the laboratory. In fact, the number of parameters to be estimated increases with the number of electrodes used, bearing the risk of overfitting when using few training samples or in the presence of outliers. Additionally, reference [26] recently showed that without subject-specific prior knowledge at least 32 channels and 120 trials are required for CSP analysis to work properly. For these reasons, Laplacian channels are often used as spatial filters of the EEG data in the context of BCI [2, 10, 24, 30], while reference [9] suggests the construction of unsupervised spatial filters which are more robust against artifactual data. Furthermore, due to practical constraints, long multi-channel recordings might be difficult to realize and researchers have investigated ways to allow good performance with just a few calibration data [11–13, 18, 19] or in the absence...
of calibration data, i.e. with subject-independent spatial filters and classifiers previously trained on a data set of other users [17].

In this work we propose a novel spatial filter that can be considered as a compromise between Laplacian and CSP filters. Since it consists of an ensemble of local ‘CSP patches’ the method is called CSPP. The performance of CSPP is validated for several channel configurations (with size ranging from 13 to 90 electrodes) and several training sample sizes (from 10 to 75 trials). Moreover, CSPP is compared in all settings with Laplacian derivations, classical CSP and regularized CSP (R-CSP) [20]. Among several recent CSP versions [1, 7, 11, 16, 31] and the unsupervised spatial filters of [9], we chose R-CSP with an analytical estimation of the shrinking parameter [15, 20]. Following this procedure, the shrinking hyperparameter can be estimated without the need of shrinking parameter [15, 20].

The results presented show that CSPP outperforms Laplacian filters in all settings studied. It also outperforms CSP and R-CSP in several multi-channel settings, using as few as ten trials for calibration, i.e. 2 min recording.

2. Experimental setup

In this study, 80 BCI motor imagery data sets were used. All experiments belong to the same study described in [5], each data set was acquired in one single session and all users were BCI novices. In order to obtain a fair comparison among the four investigated spatial filters, just the calibration data of the experiment are used in this study. During the calibration session, an arrow directed to the left, to the right or to the bottom indicates to the user which movement, respectively left hand, right hand, one or both feet, she/he should imagine in each trial. Each calibration data set contains 75 trials per class. After the calibration session a semi-automatic procedure was used to select the class combination to use in the feedback session and a subject-specific frequency band and time interval in which the two classes are better discriminable. Apart from the class combination, this subject-specific information (band and time interval) is not used for the validation in the following, since it is assumed that at the beginning of the experiment no subject-specific parameters can be calculated. Instead, a broad frequency band (8–35 Hz) and a fixed time interval (750–3750 ms after stimulus presentation) were used for the calculation of the spatial filters and the features. The stimulus duration was 3 s, while the inter-stimulus interval (ISI) was between 4.5 and 6 s. EEG was recorded using 128 channels and a sample frequency of 1000 Hz.

3. Methods

3.1. Laplacian filter

A Laplacian derivation (see [21]) of one channel is calculated subtracting the activity of n surrounding channels weighted by 1/n from the activity of the channel itself. The Laplacian derivation then weights all involved channels always in the same way, without considering the class labels. It is mainly used to eliminate the background noise, which is supposed to be present in all involved channels. For the configurations with Laplacian filters on C3, Cz and C4, and in order to obtain the same number of features as for the other methods (six), the log-band power was separately calculated in two frequency bands (8–15 Hz and 16–35 Hz) instead of in the broadband. For the configuration with five channels, the broadband was used, obtaining then five features instead of six. For all other configurations, where more than six Laplacian derivations are obtained, six features are automatically selected, as described in section 3.5.

3.2. CSP

CSP is a discriminative algorithm (see [4]) which determines spatial filters W from band-pass filtered EEG data X such that the difference between the variances of the filtered data $X_{CSP} = X \cdot W$ for the two classes is maximized. This is done by solving the following generalized eigenvalue problem:

$$\Sigma_2 W = (\Sigma_1 + \Sigma_2) W \Lambda$$

where $\Sigma_1$ and $\Sigma_2$ are the covariance matrices of data belonging to classes 1 and 2, respectively. Each column of W is a spatial filter $w_i$ corresponding to the eigenvalue $\lambda_i$, the $i$th element of $\Lambda$, with $i = 1, 2, \ldots, Nc$, where $Nc$ is the number of channels in X. Choosing N filters corresponding to extreme eigenvalues (either close to 1 or close to 0) the filtered data $X_{CSP'} = X \cdot W'$ will have smaller dimensionality $N < Nc$ and the two classes will be maximally separated by their variance. For each configuration, from the resulting spatial filters, the six most informative ones, three per class, are chosen as described in section 3.5.

3.3. R-CSP

R-CSP is the application of CSP on regularized covariance matrices. It has been proven to be very effective for small sample sizes [20], where the sample class covariance matrices $\Sigma_1$ and $\Sigma_2$ in equation (1) tend to be wrongly estimated [8]. The regularization is carried out by shrinkage, i.e. by biasing the covariance matrix toward a multiple of an identity matrix [15]. Here, the R-CSP filters were calculated following [19] and [20]. Since in the present setting no data from other users are employed, a simplified version of the R-CSP is applied, which involves just one regularization parameter instead of two. In practice, $\Sigma_c$ with $c = 1, 2$ in equation (1) is replaced by

$$\tilde{\Sigma}_c(\gamma) = (1 - \gamma) \Sigma_c + \gamma \nu I$$

where $\gamma \in [0, 1]$ is the regularization parameter and $\nu$ is the average eigenvalue of $\Sigma_c$. An optimal $\gamma$ is automatically calculated as in [15]. For each configuration, from the resulting spatial filters, the six most informative ones, three per class, are chosen as described in section 3.5.
3.4. CSPP

CSPP is the application of the CSP analysis to small sets of channels (patches) and the combination of the resulting features. The hypothesis is that, applying CSP analysis on just a few channels, the risk of overfitting is reduced in comparison to usual CSP, which is calculated on all channels available. This is because the number of parameters to fit for each patch is less than that for CSP.

The patches can include a different number of surrounding channels. Also the position of the centers of the patches can be chosen, depending on the number of channels available and on the task. The patches analyzed in this study are shown on the left of figure 1, while the centers of the patches are shown on the right of figure 1. For each patch, a number of filters equal to the number of involved channels is a result of the CSP analysis. From those filters, one per class is selected as described in section 3.5, i.e. two filters per patch are obtained. From the resulting ensemble of filters, the six most informative ones, three per class, are chosen as described in section 3.5.

In the context of co-adaptive calibration (see [35]), CSPP can be used in combination with a subject-independent classifier and online adaptation, to allow local spatial flexibility to adjust the system to possibly changing SMR modulations, as shown in [27].

3.5. Feature selection

In order to obtain a uniform comparison, the same feature selection has been applied for all four methods. In particular, the variance $v_{i,j} = w_j^T X_i X_i^T w_j$ of the $j$th feature is calculated within each trial $i$ and the corresponding ‘ratio of medians’ is taken as the score $s_j$ of that feature:

$$s_j = \frac{m_{j,2}}{m_{j,2} + m_{j,1}} \quad (3)$$

where $m_{j,1}$ and $m_{j,2}$ are the medians of $v_{i,j}$ across all trials $i$ belonging to classes 1 and 2, respectively. A score $s_j$ close to 1 indicates that the corresponding feature maximizes the variance for class 2 while a score close to 0 indicates that the corresponding feature maximizes the variance for class 1. Choosing the features with an extreme score implies that the log variance features of the two classes will be maximally separated. This ‘ratio of medians’ score has been suggested in the CSP review [4] as being more robust with respect to outliers than the classical eigenvalue score. Moreover, for this study it was revealed as particularly useful because, differently from the eigenvalue score, it can also be used for the Laplacian filter and for CSPP. In fact, using the eigenvalues of each patch directly as the score for CSPP features would not allow a fair and global comparison among all available features.

4. Results

In the following, always the same number of channels are used to compare the four methods: Laplacian, CSP, R-CSP and CSPP filters. In particular, a Laplacian filter is calculated using the same channels as the corresponding CSP patch, and CSP and R-CSP are calculated using all the channels of the different patches together. Also six features are always used and calculated in the same frequency band (8–35 Hz) and time interval (750–3750 ms after stimulus presentation), except for the configuration with five Laplacian filters, as explained in section 3.1. For classification a shrinkage linear discriminant analysis (LDA) is used [6, 28, 34], i.e. a LDA regularized by shrinkage of the covariance matrix of the features, as done for the covariance matrix of the data in section 3.3.

4.1. Evaluation of configurations and training set size

Laplacian filters, CSP, R-CSP and CSPP were evaluated using as training set the first $n$ trials of the calibration set, with $n = 10, 15, 20, \ldots, 75$. The second half of the data (75 trials) was used as a test set. All channel configurations and patch sizes in figure 1 were tested.

In general, with small channel sets (up to 30 electrodes) and small patches (up to eight electrodes), all CSP-based methods outperform the Laplacian filters but they are not significantly different from each other.

As the number of electrodes increases, the error rate decreases for all methods, but especially for Laplacian and CSPP filters. This results in a significant difference between CSPP and the rest of the filters studied. In particular, with the patch ‘twelve’ CSPP starts to be superior to CSP and R-CSP with as few as 33 channels and 15 training trials, while with larger patches like ‘eighteen’ and ‘twenty two’, i.e. with at least 39 channels, the improvement obtained with CSPP with respect to the other three methods is significantly better already with 10 training trials.

In figures 2 and 3, the results with patches ‘twelve’ respectively ‘eighteen’ are shown. In particular, the median
performance was calculated as the area under the receiver operating characteristic curve (AUC) across all 80 users. The test error (100 AUC) versus the number of training samples is shown (for the four methods and for all six channel sets). Also the results of a Wilcoxon signed rank test for equality of medians (see, e.g., [14]) are shown by stars, in the case of a not significant difference, and by asterisks in the case of a significant improvement ($p < 0.05$).

Thus, CSPP is always better than Laplacian filters at the low cost of recording 2 min of data (20 trials). Additionally, CSPP outperforms CSP and R-CSP in all configurations studied where the number of channels exceeds 39, using any amount of training data (at least up to 75 trials), with the great advantage that it needs very few training data to be tuned (10 trials).

4.2. Improving the co-adaptive calibration setting

The results presented in [35] showed that adaptive ML techniques can effectively help a variety of users to gain BCI control. In that work, Laplacian derivations were used at the beginning of the experiment to provide immediate feedback to BCI users. Although this approach was successful for good performing people, other users could not benefit from it. In this section we show that CSPP can be used instead of Laplacian filters even when no data from the user is available.
The data recorded in [35] are used for the analysis. Eleven volunteers participated in the study. A subject-independent classifier was trained using the good BCI performing users from the 80 data sets and the basic parameter setting with a broad frequency band and ‘small’ patches centered in C3, Cz and C4. This setting is parallel to that used in the original study, but with CSP patches instead of Laplacian filters. Then, the performance of CSPP on the first three runs of the co-adaptive study was evaluated and compared with the real performance (obtained using Laplacian filters). In both settings the classifier (LDA) is adapted as in [32, 35]. The results of this analysis are shown in figure 4, where each point represents the classification performance of one user in one run (with three runs per user). It can be observed that CSPP performs significantly better ($p = 0.027$) than the Laplacian filters, especially for poor BCI performing users (accuracy $\leq 70\%$).

5. Conclusion and discussion
The classical BCI machine learning approach requires a calibration session where data are acquired in order to train
complex algorithms that can then decode the brain activity during the feedback application. Here we have shown that CSPP is a data-driven approach that, unlike CSP, only needs very few data to be tuned.

In comparison with previous approaches to the problem of small training sets [1, 11–13, 19], which use 128 channels, this study also evaluates the performance of the methods depending on the number of channels and trials, giving a hint on the minimum number of channels necessary to obtain a robust classification result. Moreover, a large database of 80 users with feedback performances ranging from the chance level to 100% was used for the validation. Finally, in this study, no information of other users is used for the calculation of the spatial filters. While the suitability of CSPP to online adaptation in combination with a subject-independent classifier has been shown in section 4.2, a comparison with previous similar approaches for subject-independent spatial filters [11, 17, 18] is not performed here but is part of our future work.

Results show that CSPP always outperforms Laplacian filters, at the small expense of recording about 2 min of data. In larger multi-channel recordings, it also outperforms CSP and R-CSP which can be considered as the state-of-the-art methods. In particular, CSPP can be seen as a CSP regularizer without the need of estimating hyperparameters. It employs covariance matrices of lower dimension, improving the estimate of the parameters and subsequently, significantly improving the user’s performance using just a few minutes of training data.

It can be concluded that the CSPP approach is useful with few channels or with multi-channel recordings and highly efficient, as the number of calibration data needed is significantly reduced in comparison to the state-of-the-art.

Acknowledgments

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG), MU 987/3-1, by Bundesministerium für Bildung und Forschung (BMBF), FKZ 01IB001A/B and 01GQ0850, and by TOBI-FP7-224631.

References

[18] Lotto F and Guan C 2010 Learning from other subjects helps reducing brain–computer interface calibration time Int.
[19] Lu H, Plataniotis K and Venetsanopoulos A N 2009
Regularized common spatial patterns with generic learning for EEG signal classification 31st Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC 2009) pp 6599–602


[27] Sannelli C, Vidaurre C, Müller K-R and Blankertz B 2010


[34] Vidaurre C, Krämer N, Blankertz B and Schlögl A 2009 Time domain parameters as a feature for EEG-based brain computer interfaces Neural Netw. 22 1313–9