ROBUSTNESS OF THE COMMON SPATIAL PATTERNS ALGORITHM IN THE BCI-PIPELINE

Boris Reuderink, Mannes Poel

HMI, University of Twente
Drienerlolaan 5, 7522NB, Enschede, The Netherlands
phone: + (31) 534893718, email: {reuderin, mpoel}@cs.utwente.nl

ABSTRACT

When we want to use brain-computer interfaces (BCI) as an input modality for gaming, a short setup procedure is necessary. Therefore a user model has to be learned using small training sets. The common spatial patterns (CSP) algorithm is often used in BCI. In this work we investigate how the CSP algorithm generalizes when using small training sets, how the performance changes over time, and how well CSP generalizes over persons. Our results indicate that the CSP algorithm severely overfits on small training sets. The CSP algorithm often selects a small number of spatial filters that generalize poorly, which can have in impact on the classification performance. The generalization performance does not degrade over time, which is promising, but the signal does not seem to be stationary. In its current form, the CSP generalizes poorly over persons.

1. INTRODUCTION

A brain-computer interface (BCI) provides a direct communication channel between the brain of a subject and a computer, such that mental activities can be used to influence the computer. BCIs are used to enable patients in a late stage of amyotropic lateral sclerosis (ALS) or locked-in syndrome to communicate. Other applications for BCI include research in neuroscience, and applications for healthy users such as training and gaming.

In the context of gaming, a short setup time is required for a pleasant gaming experience. This implicates that the BCI has to learn user models from a small training set, while the model needs to stay reliable during the whole gaming session. For games, a multi-electrode electroencephalogram cap (EEG cap) is suitable to extract the signal of mental activities. Using spatial filters, mental activities such as imagined movement can be extracted from the EEG signals. A technique to compute these spatial filters is the common spatial patterns (CSP) algorithm, which has been used successfully for the analysis of imagined movement. The CSP algorithm uses labeled trials to produce a transformation that maximizes the variance for one class while minimizing the variance for the other class. The difference in variance can be used to classify a fragment of EEG signals into one of two classes. While the CSP algorithm is often used in the BCI pipeline, it tends to overfit [6, 8], resulting in a suboptimal performance. Several improvements for the CSP have been suggested, but it remains unclear under what circumstances the CSP is more susceptible to overfitting.

To find out if the CSP algorithm is usable for the short setup-time of games, we investigate the influence of the amount of trials in the training set, the time between training and use for classification, and the generalization performance over subjects for BCI based on EEG measurements.

In the next section we describe previous work on the CSP algorithm. Section 3 will explain the CSP algorithm in more detail and outline the experiments we performed. Subsequently we will present the results, and conclude with a discussion and recommendations for future work.

2. PREVIOUS WORK

The CSP algorithm was first presented by Koles [7] as a method to extract the abnormal components from EEG, using a set of patterns that are common to both the normal and the abnormal recordings and have a maximally different proportion of the combined variances (see Section 3.3). The EEG is decomposed into orthogonal signals that can be judged by experts. Afterwards, the signal can be recomposed using only the abnormal signals so that the spatial topography of the abnormal components can be inspected.

Later CSP was used by Ramoser et al. [10] to create features for classification of the event-related desynchronization (ERD) in EEG caused by imagined movements. The first and last few CSP components (the spatial filters that maximize the difference in variance) are used to classify the trials with a high accuracy.

Since then the CSP has often been used [1] for classification of imagined movement. Hill et al. noticed that the CSP algorithm has a tendency to overfit [6], and a few extensions to the CSP algorithm have been developed in order to improve the generalization performance. A few of these extensions incorporate spatio-spectral filters [8, 3], that include frequency filtering in the CSP algorithm to make it more robust against artifacts. Another extension focuses on sparsifying the CSP to prevent overfitting [4]. While these approaches seem to increase the performance it is still unclear what influences the overfitting observed with the CSP algorithm, and therefore it is not known if these changes are adequate.

3. METHODS

3.1 Dataset

For our experiments we selected Dataset IVa of the BCI Competition III [1]. This dataset was selected because it contained multiple subjects, and two classes of cued motor imagery. The dataset contains trials recorded on five healthy subjects, ranging from 28 to 224 labeled trials per subject. The CSP algorithm has been used before on this dataset and produced good results.

Visual cues indicated for 3.5 seconds which motor imagery the subject should perform. No feedback was given during the experiment. The recording was made using BrainAmp amplifiers and a 128 channel Ag/AgCl electrode cap from ECI and were subsequently down-sampled from 1000Hz to 100Hz.

3.2 Preprocessing

Before we applied the CSP algorithm, the EEG recording was re-referenced, bandpass filtered and separated into trials of 3.5 seconds. First the data was re-referenced using the Common Average Reference (CAR). After re-referencing the data, we used a zero-phase forward-backward FIR-filter to bandpass filter the data to the alpha and beta frequency ranges. The filter we used was a 400 taps Blackman-window FIR-filter with a bandpass range from 8 to 30 Hz and an approximate transition band of 1 Hz. After filtering the data was split into trials according to the markers provided with the dataset, and unlabeled trials were removed.

3.3 CSP

The CSP algorithm calculates a matrix W with spatial filters with a high variance for the first class and a low variance for the second, and vice versa. It is a $M \times N$ transformation matrix W with the following property:

$$Cov(WX_1) = D \quad \text{and} \quad Cov(WX_1) + Cov(WX_2) = I \tag{1}$$

where D is a diagonal matrix with elements monotonely descending, I is the identity matrix and $Cov(X)$ is the covariance matrix of X. N is the number of channels and M is the rank of $Cov(X)$, and X_i is a matrix with observations in the rows, and EEG channels in the columns for class i. In other words: the transformed channels with a low variance for one class will have a high variance for the other class. This variance can be used for classification. Eq. 1 is equivalent to the CSP equations found in [9, 8]:

$$WCov(X_1)W^T = D \quad \text{and} \quad WCov(X_2)W^T = I - D \tag{2}$$

We can calculate a matrix P using Singular Value Decomposition (SVD) [1] that will transform the data to have an identity covariance matrix:

$$Cov(PX) = I \Rightarrow Cov(X) = (P^T P)^{-1} = U\lambda U^T \tag{3}$$

$$P = \sqrt{\lambda^{-1}}U \tag{4}$$

Where U is an orthogonal matrix, and λ is a diagonal matrix. This transfomation is equivalent to performing a principal component analysis (PCA) and normalizing the variance to one. When $rank(Cov(X)) < M$, only the significant eigenvalues and eigenvectors are used to compute P.

While $Cov(PX) = I$, $Cov(PX_1)$ and $Cov(PX_2)$ will generally have a covariance matrix that is not completely diagonal. To create a diagonal covariance matrix for $Cov(WX_1)$ and $Cov(WX_2)$, we perform an additional SVD after the whitening transform P:

$$Cov(PX_1) = BDB^T \Rightarrow W = B^T P \tag{5}$$

This definition of W satifies Eq. 1.

3.4 Generalization

If the results on the test set differ significantly from the results on the training set, overfitting occurs, and we cannot expect a good generalization performance of a classifier trained on the output of the CSP. As the CSP algorithm is defined in terms of component variance, we have chosen to measure the generalization performance in terms of variance. More specifically, we measure if the distribution of the component variances for each of the two classes is statistically different for the training set and the test set. As we want to compare non-normal values, we test for differences with the Mann-Whitney U test [5]. If the distribution for a specific CSP component is significantly ($p < 0.05$) different for the test set, we regard this as a generalization error. For each of the following experiments we display these generalization errors graphically.

3.4.1 Generalization for different amounts of training data

In the first experiment we performed, we measured the influence of the amount of training data on the generalization performance as described above.

For each subject, we selected n trials using stratified random sampling to form the training set. The test set was formed by stratified sampling of 10 trials from the remaining trials. The CSP is calculated on the training set, the generalization error for each CSP component was computed on the test set and stored. In order to get more stable results this was repeated 10 times for each n; the mean error used used in the evaluation.

We expect to find less generalization error for the first and last CSP components, as these components are used for classification in previous work, and should therefore have good generalization properties. The generalization error should decrease when more trials in the training set are used.

3.4.2 Generalization over time

The first experiment gives some insight on the influence of the number of trials used to construct the CSP. For a real BCI system, not only the number but also the performance over time is of importance.

In this second experiment we measure the generalization error as a function of the delay t between training and testing. For training the first 30 trials were used, for testing we used the first 20 trials found at $t > 0$ seconds after the last trial in the training set.

Because the EEG patterns are known to differ from session to session, we expect that the overfitting increases over time.

3.4.3 Generalization over subjects

While generalization over time is necessary to produce a usable BCI system, it would be convenient to train a BCI system once on a number of subjects, and use this system for new subjects. In a final experiment we measure how well the properties of the CSP algorithm hold when it is trained on 4 of the 5 subjects, and tested on the other subject. Koles [7] mentions that magnitude variations in EEG exist between subjects, therefore we do not expect a good generalization performance in this experiment.

[1] Note that for a symmetric matrix – such as a covariance matrix – the SVD is equivalent to diagonalization, or the solution to the eigenvalue problem.

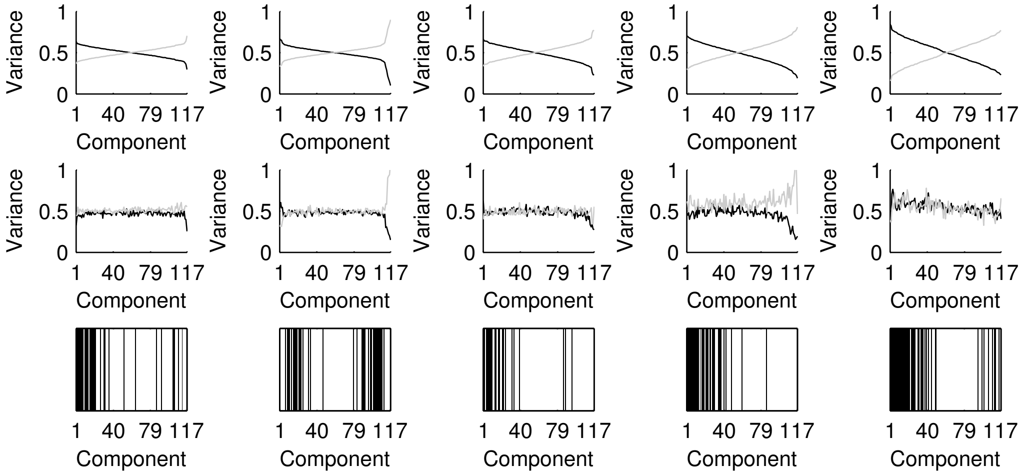

Figure 1: Variances on the training set (first row), the test set (second row), and the generalization error (last row) for each subject. Black represents trials of the class "right hand", gray represents trials of the class "foot". On the training set the complementary variance for the two classes is clearly visible, on the test set this property is less clear. When there is a significant difference between the component variances on the training and test set, a black bar is drawn for that component in the last row.

4. RESULTS

We will begin with the results for the generalization performance depending on the number of trials in the training set.

Figure 1 shows the variance of the CSP components on the training set (80% of the trials) and test set (20% of the trials). The CSP algorithm does indeed overfit; the complementary, ordered variances are more chaotic on the test set. Usually the first and last *m* components are used for classification. Only for subject "al" these most discriminative components seem to generalize well, and for this subject the best scores were obtained during BCI Competition III as well. The generalization error is displayed on the last row. Black bars represent components that have a statistically different distribution on the training and test set.

In Figure 2 the mean generalization error is plotted for different numbers of trials in the training set. As we expected, the generalization error decreases for all subjects when more trials are used for training. The low-numbered components (these have a high variance for trials of imagined movement of the right hand) show remarkably more generalization error than the components associated with the imagined movement of the feet for all subjects. The good generalization performance of the lowest and highest components of subject "al" is visible as clear white bands on the top and the bottom of the generalization-error plot. All plots contain less generalization error for the middle components. These components are often not discriminatory, and are of little use for classification.

Now we have some information about the influence of the number of trials on the generalization performance, we can take a look at the generalization performance over time. Figure 3 shows the generalization error of the CSP trained on the first 30 trails of a subject, and a test set delayed by *t* seconds. Experiment 1 showed that overfitting is to be expected when using 30 trials. Unfortunately this dataset did not contain enough trials to use a larger training set and still measure over a large range of delay. Figure 3 clearly shows that the generalization performance is not stationary, most likely the source signal changes over time. Periods of about 200 seconds seem to be more or less stationary. While the signal changes over time, there is no evidence that the generalization performance drops as the delay is increased. The changes over time that are visible in this experiment show that the analysis of BCI signals should incorporate tempo-

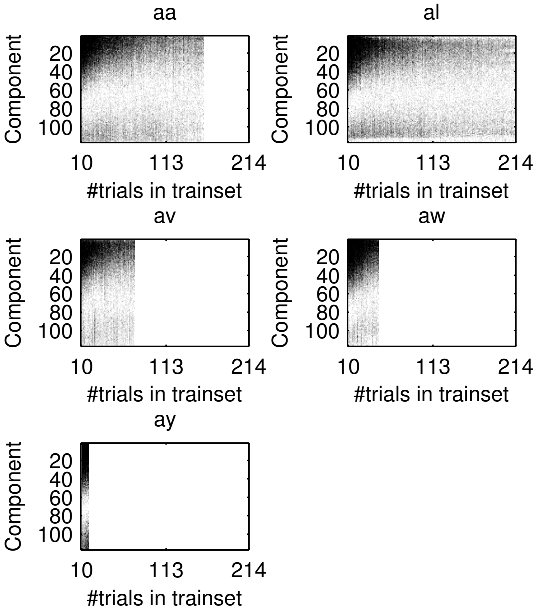

Figure 2: Mean generalization error depending on the number of trials in the training set. In order to compare the performance of different subjects, all errors are plotted using the same time scale, which explains the white areas for all subjects except "al".

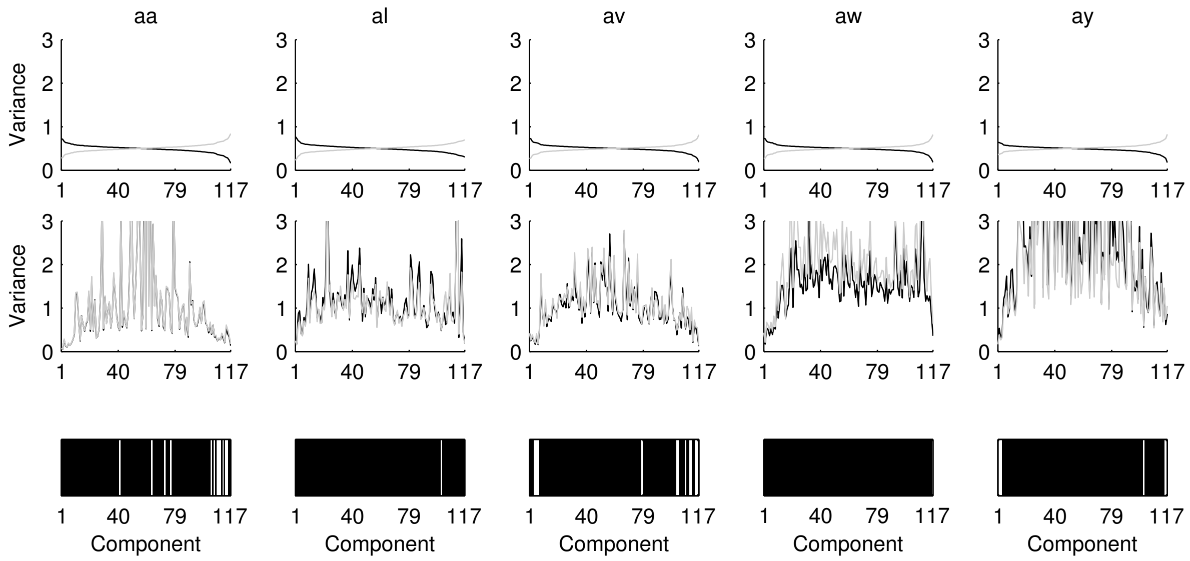

Figure 4: Variances on the training set consisting of 4 subjects, the test set consisting of the 5th person and the generalization error. The black columns in the last row indicate statistically significant generalization errors of the individual components.

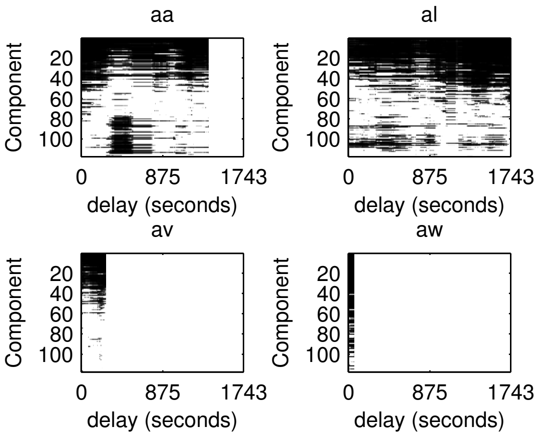

Figure 3: Generalization error as the result of time between the training and test set. Subject "ay" was removed because not enough trials were available to perform the experiment. In the images, time is represented from left to right, components are displayed from top to bottom. Statistically significant generalization errors are displayed from light gray (p = 0.05) to black (p = 0).

rally distinct train and test sets.

In our last experiment we tested the CSP on subjects that were not present in our training set. All subjects were instructed to produce the same imagined movements, and the classes corresponded which allowed us to combine the trials of different subjects. In Figure 4 we can see that the CSP does not generalize over persons; nearly all components produce variances that are statistically different on the training and test set. Koles [7] mentions that magnitude variations in the EEG data exist between subjects. These variations can be the cause of the bad generalization performance over subjects. To prevent these magnitude varations from influencing the results, Koles and later Ramoser et al. [10] used normalized covariance matrices to train the CSP. However, for testing a different normalization procedure was applied, which makes it hard to replicate their normalization procedure in a meaningful way.

5. CONCLUSIONS AND FUTURE WORK

Our first experiment showed that the CSP algorithm does overfit severely on small training sets. As expected, the generalization performance of the CSP algorithm improved when more trials were used in the training sets. However, the first and last CSP components that are often used do not generalize very well, except for for subject "al", on which also the highest scores are obtained in the BCI Competition III using the CSP algorithm.

Hill et al. obtained the best generalization performance using components with high eigenvalues [6]. This contradicts our results. A possible explanation is that the support vector machine (SVM) classifier they used on the CSP-filtered data is robust to the overfitting we observed by comparing variances. Another cause for this contradiction could be the presence of artifacts in the data. Artifacts are most likely captured by a component with a high variance [2], which would impact the generalization performance.

The results of the second experiment show us that the variance of the CSP components are not stationary. There is no clear evidence of drifting; the performance does not constantly degrade over time, which we did expect.

Our third experiment shows that current generalization over subjects does not work very well. The most likely cause is the magnitude variations in the EEG that exist between individuals. Normalization could be used to improve the generalization over subjects.

The CSP does severely overfit, which makes sparse CSP-like algorithms an interesting development. Evaluation using chronologically separated training sets and test sets is needed to measure real-world performance. A number of studies already use chronologically split training and test sets [3, 8], others do not [4, 6]. For a convenient BCI, short training times are required. From our second experiment we know that variations in the variance of the CSP do occur. It is therefore unlikely that a short training period contains enough information to generalize over these variances. For future work we would recommend investigating the source of this drifting over time, perhaps the users change their imagery of the motor task, or the pattern changes in a way for which the CSP is not invariant. As the first and last components often seem to overfit on small training sets, we would recommend the use of feature selection to select discriminatory, generalizing features for classification.

References

[1] B. Blankertz and K. R. Müller. The BCI Competition III. 2005.

[2] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Müller. Optimizing Spatial Filters for Robust EEG Single-Trial Analysis. *IEEE Signal Proc. Magazine*, 2008.

[3] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K. R. Muller. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE Trans. Biomed. Eng*, 53:2274–81, 2006.

[4] J. Farquhar, J. Hill, T. N. Lal, and B. Schölkopf. Regularised CSP for sensor selection in BCI. 2006.

[5] J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. CRC Press, 2003.

[6] N. J. Hill, T. N. Lal, B. Scholkopf, M. P. I. B. Cybernetics, G. Tubingen, M. Tangermann, T. Hinterberger, B. Neurobiology, G. Widman, C. E. Elger, and Others. Classifying Event-Related Desynchronization in EEC, ECoG, and MEG Signals. *Toward Brain-Computer Interfacing*, 2007.

[7] Z. J. Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical eeg. *Electroencephalography and Clinical Neurophysiology*, 79(6):440–447, December 1991.

[8] S. Lemm, B. Blankertz, G. Curio, and K. R. Muller. Spatio-spectral filters for improving the classification of single trial EEG. *Biomedical Engineering, IEEE Transactions on*, 52(9):1541–1548, 2005.

[9] F. Popescu, S. Fazli, Y. Badower, B. Blankertz, and K. Müller. Single trial classification of motor imagination using 6 dry EEG electrodes. *PLoS ONE*, 2(7), 2007.

[10] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *Rehabilitation Engineering, IEEE Transactions on [see also IEEE Trans. on Neural Systems and Rehabilitation]*, 8(4):441–446, 2000.