Neurocomputing 108 (2013) 69–78

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Greedy solutions for the construction of sparse spatial and spatio-spectral filters in brain computer interface applications

Fikri Goksu [a,*], Nuri F. Ince [a], Ahmed H. Tewfik [b]

[a] University of Minnesota Twin Cities, Electrical & Computer Engineering, 4–174 Keller Hall, 200 Union St SE, Minneapolis, MN 55455, USA
[b] The University of Texas at Austin, Electrical & Computer Engineering, 2501 Speedway, Austin, TX 78712-0240, USA

ARTICLE INFO

Article history:
Received 24 December 2011
Received in revised form
14 November 2012
Accepted 6 December 2012
Communicated by Sanqing Hu
Available online 5 January 2013

Keywords:
Brain computer interface
Common spatial pattern
Sparse
Greedy search

ABSTRACT

In the original formulation of common spatial pattern (CSP), all recording channels are combined when extracting the variance as input features for a brain computer interface (BCI). This results in overfitting and robustness problems of the constructed system. Here, we introduce a sparse CSP method in which only a subset of all available channels is linearly combined when extracting the features, resulting in improved generalization in classification. We propose a greedy search based generalized eigenvalue decomposition approach for identifying multiple sparse eigenvectors to compute the spatial projections. We evaluate the performance of the proposed sparse CSP method in binary classification problems using electrocorticogram (ECoG) and electroencephalogram (EEG) datasets of brain computer interface competition 2005. We show that the results obtained by sparse CSP outperform those obtained by traditional (non-sparse) CSP. When averaged over five subjects in the EEG dataset, the classification error is 12.3% with average sparseness level of 11.6 compared to 18.4% error obtained by the traditional CSP with 118 channels. The classification error is 10% with sparseness level of 7 compared to that of 13% obtained by the traditional CSP using 64 channels in the ECoG dataset. Furthermore, we explored the effectiveness of the proposed sparse methods for extracting sparse common spatio-spectral patterns (CSSP).

1. Introduction

The use of neural activity as a source of information has enabled the subjects with motor impairments to communicate with their environment using brain computer interfaces. A brain computer interface (BCI) extracts critical patterns from neural activity which are induced by purely mental tasks and processes to identify the mental state of the subject [1,2].

Recent advances in microprocessor and microcircuit design have enabled recording of neural data over large number of channels. One of the most crucial steps in designing a BCI system is to extract parsimonious features from the multi-channel neural recordings. The CSP method is a signal processing technique that extracts features by combining signals from all available recording channels. The method was first proposed in [3] to analyze abnormal EEG patterns. Since then, it has been one of the most effective feature extraction tools of current BCI technology in binary and multi task classification problems [4–7]. The CSP method [3,8] finds spatial filters which correspond to linear weighting of each channel in a multi-channel setup. Namely, the relationship between the input multi-channel brain signal, $x(t)$, and the output, $x_{CSP}(t)$, after CSP filtering is given by $x_{CSP}(t)=W^T x(t)$. Here, each column of W is a distinct spatial filter that captures different spatial localizations of the underlying brain activity. In a binary BCI application, the solution of the spatial filters is achieved by solving a generalized eigenvalue decomposition (GED) problem in which the variance of one class is maximized while minimizing the variance of the other [2]. The CSP filters achieve this task by using the spatial correlation patterns which are sensed from a number of recording channels. Consequently, dense neural recordings have higher likelihood in capturing the discriminative spatial patterns as they cover most of the surface available to assess brain activity. However, this results in redundancy of information and makes the current BCI systems more prone to artifacts since it is difficult to obtain robustness over sessions.

Recent studies [2,9,10] have shown that the CSP method suffers from a number of problems which pose new challenges when using this method in practice. Let us shortly summarize these challenges. Generally, the multi-channel neural recordings are obtained at different times or sessions. This means that the parameters necessary to extract features and the classifier are obtained using the data collected in one session. These parameters

* Corresponding author. Tel.: +1 202 5093671.
E-mail addresses: goks0002@umn.edu (F. Goksu), firat@umn.edu (N.F. Ince), tewfik@austin.utexas.edu (A.H. Tewfik).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.12.003

are used to classify the neural data in another session. The time difference introduces variation in the neural patterns. Furthermore, in the case of EEG the electrodes are removed and reattached between sessions. Variation in the data creates even larger variation in the extracted features due to linear combinations. Even the failure of a single channel or an outlier in the data might cause significant changes in the features. This increases the sensitivity of the algorithm to intersession variability; therefore, degrades overall performance of the CSP method.

Another problem of the CSP method arises when the size of the training data is smaller than the number of recording channels [10]. The extraction of the spatial filters in CSP formulation relies on the estimated correlation matrices for the two tasks we want to discriminate. In the presence of large number of recording channels, the estimation of the correlation matrices are poor. As a result, the algorithm overfits the data and deteriorates the generalization performance. Consequently, a regularization step is necessary to overcome the robustness and overfitting problems of the standard CSP algorithm.

In the standard CSP method there is no constraint on the number of nonzero components in each spatial filter. Recently, in [9,11], sparse spatial filters are extracted by adding L1 norm constraint in to the CSP formulation. In both of the studies, it has been shown that the number of channels can be reduced significantly but with a decrease in the classification accuracy. The solution employed in these approaches most likely gets trapped in a local minimum and is not the best possible sparse spatial filter for a given cardinality. This is due to the non-convexity of the optimization problem to be solved in the CSP formulation [12, Chapter 3]. At the same time, finding a solution with a predefined number of nonzero weights in the spatial filter is not a straightforward task with an L1 norm based approach.

Here, we reformulate the traditional CSP problem to obtain sparse spatial filters by employing greedy search methods used in [13,14]. The results given by [13] in the case of principal component analysis (PCA) problem using L1 norm regularization compared to that of using greedy search are our motivation to employ a greedy search based method to find the sparse CSP filters. In particular, our goal is to extract spatial filters by solving the CSP formulation such that the resulting filters have only a few nonzero components. With regularization via sparseness in the spatial filters, we expect to decrease the sensitivity of the CSP algorithm to the variations in the multi-channel input data. The essential idea of the regularization via sparsifying the solution is presented in a regression framework in [12, Chapter 6]. Basically, since in CSP framework the features are obtained by multiplying the spatial filter with the input signal, variation in the input signal will result in even more variation in the features. Therefore, forcing the spatial filter to be sparse will diminish the variation in the features. Clearly, the likelihood of a failure of a single channel is small in a sparse projection compared to projection using all the channels. Hence, an improvement in robustness is expected. Moreover, the covariance matrices can be better estimated using a small sensor suite where a limited number of training trials is available. Therefore, our approach will enhance the generalization performance of the CSP method.

With these motivations we constructed an L_0-norm based sparse CSP approach with multiple eigenvectors where the initial idea was presented in [15]. Moreover, we extended the sparse CSP framework to spatio-spectral filtering. Our contribution in this study is twofold; (a) extending the greedy search based methods of [14] to find multiple sparse solutions to a GED problem and (b) studying multiple sparse solutions in CSP and CSSP frameworks in noninvasive and invasive BCI applications. The goal in (a) is achieved by formulating the optimization problem in Lagrange form as employed in [16]. We tested our approach on two different modalities, the EEG and ECoG datasets which have distinct characteristics. While EEG is a noninvasive and robust recording technique, it suffers from spatial specificity and is prone to artifacts. On the other hand, the ECoG is a highly invasive technique with superior spatial resolution and high SNR. Consistently, on all subjects, we observed that our method outperformed the standard CSP and CSSP method on both EEG and ECoG datasets. Moreover, the selected cardinality in the sparse spatial projections for the ECoG dataset was lower than the EEG dataset. This perfectly correlates with the nature of these modalities where the ECoG is more spatially localized. Our results show that the method we proposed can be effectively used on both non-invasive and invasive modalities.

The rest of the paper is organized as in the following. In the next section we first provide the details of extracting the standard CSP filters via the GED formulation. Then, we introduce the details of our approach to find sparse CSP filters. We explored the performance of the greedy search based sparse CSP method in a binary classification problem both on EEG and ECoG datasets used in BCI competition 2005 [17]. We describe the experimental setups in Section 3. Then in Section 4, we provide the performance evaluation results obtained by the sparse CSP and those obtained by standard CSP. We provide an exploratory analysis of the proposed methods in CSSP framework at the end of this part. Finally, in Section 5 we discuss our results and future work.

2. Methods

2.1. Traditional CSP

Let us shortly describe the traditional CSP method and its optimization formulation. By traditional CSP we mean the CSP in its original formulation to distinguish it from our approach as sparse CSP (sCSP). We refer the reader to [2] and references therein for a recent review on CSP and its applications.

Let us consider a binary BCI problem with two classes, Y_l, $l=1,2$. Let C be the number of channels and N be the number of time samples in each channel. Then, $X_i \in R^{C\times N}$ represents the multi-channel neural data in the ith trial. The labels of each trial are known during the training. The CSP solves the following optimization problem to find the spatial filters (w);

$$\underset{w}{\operatorname{argmax}}\ \frac{w^T\Sigma_1 w}{w^T\Sigma_2 w} \tag{1}$$

in this equation, Σ_l, $l=1,2$ denote estimated covariance matrices for each class and are $C\times C$ dimensional. A simple interpretation of Eq. (1) can be stated as finding a vector w such that the variance of class Y_1 is maximized while that of class Y_2 is minimized. This formulation in Eq. (1) is equivalent to the following:

$$\underset{w}{\operatorname{argmax}}\, w^T\Sigma_1 w \quad \text{s.t. } w^T\Sigma_2 w = 1 \tag{2}$$

After writing the latter formulation in Lagrange form and taking the derivative with respect to the variable w gives us the following equality:

$$\Sigma_1 w = a\Sigma_2 w \tag{3}$$

Eq. (3) is known as the GED problem, and its closed form solution is available via joint diagonalization of both of the numerator and denominator that can be found in [18]. There are C eigenvector and eigenvalue pairs (w_i, α_i), all eigenvalues are positive. Since, from (3) $(1/\alpha)\Sigma_1 w = \Sigma_2 w$, the eigenvector which maximizes the variance in (2) for one class also minimizes the variance for the other class. Hence, variance is used as feature in CSP framework. It is a common practice to select an equal number eigenvectors from both end of the spectrum as the spatial filters. Finally, given $W=[w_1, \ldots, w_C]$ is the collection of column vectors of spatial filters; the output of spatially filtered signal is given as $Z=W^TX$. Therefore, each row of Z is a linear combination of all channels. Assuming W is invertible, then $(W^{-1})^TZ=X$. This means each sample of multi-channel X can be written as a linear combination of columns of $(W^{-1})^T$. In this case, each column can be considered as an invariant source distribution which is called a pattern; hence, the name for CSP. This provides an easy visualization tool for interpretation.

We conclude this part by noting the steps of the CSP based feature extraction and classification.

i. Each channel of the input signals is filtered in time to extract information in the relevant frequency band (s).
ii. The CSP filters are computed using the covariance matrices which are estimated from a manually selected time window with a typical length of 0.5–1 s.
iii. For each spatial filter, w_i, corresponding feature is calculated as the logarithm of the energy via $\log(||w_i^TX||_2^2)$.
iv. A supervised learning method is applied on these extracted features.

2.2. Sparse CSP (sCSP)

Our goal here is to find sparse CSP filters. Mathematically speaking, this is equivalent to solving the GED problem such that some coefficients of each eigenvector are constrained to be zero. Before going any further, we would like to mention potential approaches for solving the sparse GED problem. At this point it is worthwhile to make a connection between CSP and Principal Component Analysis, (PCA). Consider the following CSP formulation:

$$\max w^TAw \quad \text{s.t.} w^TBw=1 \tag{4}$$

Define another variable $z=B^{1/2}w$ and $M=B^{-1/2}AB^{-1/2}$, then Eq. (4) will be equivalent to

$$\max z^TMz \quad \text{s.t. } z^Tz=1 \tag{5}$$

The formulation in (5) is known as the PCA problem when the matrix M is the sample covariance matrix. Hence, by a simple variable transformation, one can solve the PCA eigenvectors, z, then using $w=B^{-1/2}z$ one can find the CSP filters. This potentially gives us the opportunity to employ the existing sparse PCA algorithms for finding the sparse CSP filters. For a number of sparse PCA algorithms that enforce sparseness in the solution via L1 norm regularization, we refer the reader to the studies of [13,19] and references therein. Note however that these approaches are not directly applicable; a sparse z may not necessarily give a sparse w since $w=B^{-1/2}z$ and the columns of $B^{-1/2}$ are not necessarily sparse. Furthermore, in the case of solving sparse PCA problem, a number of experiments were run in [13] that compare solutions obtained via L1 norm regularization and greedy search. It has been shown that the method based on a greedy search does provide better results.

2.2.1. Sparse CSP via greedy search

We follow the search approach developed in [14] to obtain sparse solution to CSP problem. In this scheme, we first explain how to compute only one sparse eigenvector given the two covariance matrices. Later, we show how to iteratively compute multiple sparse eigenvectors using the formulation employed in [16].

Let us assume that we are given two covariance matrices A and B and a sparse eigenvector w with cardinality k (number of nonzero coefficients) such that w is a solution of the following optimization problem:

$$\underset{w}{\text{argmax}} \frac{w^TAw}{w^TBw} \quad \text{s.t.} ||w||_0=k \tag{6}$$

Considering the quadratic term in the numerator, multiplication of A from left by w^T selects rows of A corresponding to nonzero indices of w. Similarly, multiplication of A from right by w selects columns of A corresponding to the same indices. Therefore, maximizing of (6) is equivalent to

$$\text{argmax} \frac{w^TAw}{w^TBw} = \text{argmax} \frac{s^TA_ks}{s^TB_ks} \tag{7}$$

Here, s is the vector obtained by only keeping the nonzero values of w. Similarly, A_k and B_k are $k\times k$ dimensional submatrices obtained by keeping the rows and columns of (A, B) corresponding to nonzero indices of w. Hence, if we know how to find a vector s which maximizes the right hand side of the equality in (7), then that is a sparse vector w, with k nonzero elements, which also maximizes the left hand side of the equality in (6). This approach requires searching all $k\times k$ sub-matrices of (A, B) and selecting the one with the largest eigenvalue. Setting k to any number provides the opportunity to reach the desired cardinality. However, such a search is computationally complex. At this point it is worth noting that when extracting a $k\times k$ submatrix of (A, B), the row and column indices are the same. Although the symmetric structure of (A, B) reduces the computational complexity, searching all possibilities is infeasible for covariance matrices of large sizes. Therefore, suboptimal search algorithms are employed to find good solutions without searching all combinations, [20]. To find a sparse eigenvector with a given cardinality $k \in\{1,\ldots,C\}$, in the following we explain the greedy search methods with reduced complexity.

2.2.1.1. Forward selection (FS). FS starts with an empty index set. In the first step, $k=1$, it searches for all 1×1 sub-matrices of (A, B) which are the diagonal elements, and then picks the index which has the largest value. In the second step, $k=2$, the search is done now for all 2×2 sub-matrices of (A, B) where one of the indices is the one selected in the previous step. The GED solution is computed for each sub-matrix and the eigenvector with the largest eigenvalue is selected. This sequential forward search continues until the desired cardinality of the eigenvector is reached. The main advantage of using FS to find a sparse eigenvector with a desired cardinality is its low computational complexity when the cardinality is small.

2.2.1.2. Backward elimination (BE). In BE search, the initial search index set consists of all the indices. In the first step, $k=C$, the GED solution is obtained by the covariance matrices and this provides the maximum eigenvalue possible. In the second step, $k=C-1$, all $(C-1)\times(C-1)$ sub-matrices are searched and corresponding eigenvalues are calculated. Then, the index (row and column) which yields the least eigenvalue difference from the first step is deleted from the index set. This deletion of row and columns continues until the desired cardinality is reached. Obviously, the BE method has higher computational complexity than the FS method when the desired cardinality is small.

2.2.2. Multiple sparse CSP filters

So far, we have explained and shown search methods which provide a single eigenvector with a sparseness level of interest. However, in the CSP applications multiple spatial filters are desired. In order to extract multiple sparse CSP filters, one can employ the same search methods explained above. However, the contribution of the previously selected eigenvector(s) should be removed from the sample covariance matrices. For this particular purpose we follow the work of [16]. Clearly, we are interested in finding the eigenvectors of a GED problem that is obtained by joint diagonalization of the covariance matrices in an iterative manner.

We assume now that we have found a number of sparse eigenvectors, w_i, $i=1,\ldots,m-1$; and we are interested in finding the next one, w_m. The idea is to rewrite the GED objective (2), such that the relationships among the eigenvectors are stated explicitly in the optimization formulation as constraints. The problem of interest we want to solve becomes

$$\underset{w_m}{\operatorname{argmax}} w_m^T A w_m \quad \text{s.t. } w_m^T B w_m = 1 \cdot w_m^T T w_i = 0, \quad i=1,\ldots,m-1. \tag{8}$$

In (8), (A, B) are sample covariance matrices. The objective function is the same as before but the eigenvector sought for is also constrained to be T-orthogonal with the existing eigenvectors. The matrix T is selected such that it is useful for the problem. For example, in the special case when T is the identity matrix, the additional constraint involving T means that the new eigenvector is desired to be orthogonal to the previous eigenvectors. Since we are interested in the solution which diagonalizes A and B; hence, T is set to B. This choice makes the projected outputs from each class uncorrelated. Eq. (8) is the original GED problem with an additional constraint that the eigenvectors diagonalize B. To see how this setup also diagonalizes A, we proceed as in the following. Since any solution to (8) satisfies$Aw_m = \lambda B w_m$, multiplication of both sides with any of the previously found eigenvectorw_k^T, $w_k^T A w_m = \lambda w_k^T B w_m$, makes the left hand side of the equality zero. Therefore, the solution of (8) is equivalent to the solution of (2) which diagonalizes A and B. Now, we can proceed to extract the vectors w_m iteratively. The Lagrangian formulation of the problem with constraints is

$$L(w_m,\lambda,\mu) = w_m^T A w_m + \lambda(1 - w_m^T B w_m) - \mu_{m\text{-}1} w_m^T T w_{m\text{-}1} - \cdots - \mu_1 w_m^T T w_1. \tag{9}$$

At the optimum solution the partial derivative of L with respect to w_m equals to zero

$$\frac{dL}{dw_m} = 0. \to 2Aw_m - 2\lambda B w_m - \mu_{m-1} T w_{m-1} - \cdots - \mu_1 T w_1 = 0 \tag{10}$$

Multiplying (10) by $w_i^T T B^{-1}$from left for $i=1,\ \ldots,\ m-1$ gives us $m-1$ equations of the from

$$2w_i^T T B^{-1} A w_m - 2\lambda w_i^T T B^{-1} B w_m - \mu_{m-1} w_i^T T B^{-1} T w_{m-1} - \cdots - \mu_1 w_i^T T B^{-1} T w_1 = 0 \tag{11}$$

we note here that the second term in (11) disappears due to second constraint in (8). By defining a matrix $D = [w_1 \cdots w_{m-1}]^T$and a vector $\mu = [\mu_1 \cdots \mu_{m-1}]^T$we can write (11) as

$$2DTB^{-1}Aw_m = DTB^{-1}TD^T\mu \to \mu = 2\left(DTB^{-1}TD^T\right)^{-1} DTB^{-1}Aw_m \tag{12}$$

Similarly, (10) can be re-written as $2Aw_m\text{-}TD^T\mu = 2\lambda B w_m$, in which we substitute μ that is obtained in (12) and end up with the following GED problem:

$$[I - TD^T\left(DTB^{-1}TD^T\right)^{-1}DTB^{-1}]Aw_m = \lambda B w_m \tag{13}$$

If we define a new matrix, $A_n = [I - TD^T(DTB^{-1}TD^T)^{-1}DTB^{-1}]A$, then Eq. (10) becomes

$$A_n w_m = \lambda B w_m \tag{14}$$

This means that the problem to be solved is still a GED problem but with a different matrix in the numerator. Now, we can employ any of the greedy search based sparse eigenvector selection method to solve the new GED problem. This is our approach to find multiple sparse eigenvectors in an iterative way. We conclude this section by providing our proposed approach of finding multiple sparse eigenvectors in an algorithmic form. It is worth noting that when the matrices A and B in the above discussion are replaced with appropriate covariance matrices, then we can find all sparse spatial CSP filters.

In the following, the notation (A, B) refers to solving the CSP problem via maximizing variance for class A while minimizing that for class B.

Algorithm 1. Extraction of m CSP filters per class with given sparseness levels:

Input: Estimated covariance matrices for both classes; A and B
Sparseness levels for each filter, s_i, $i=1,\ \ldots,\ 2m$.

i. Find an initial eigenvector for (A, B) with the sparseness level s_i using one of the methods explained in Section 2.2.1.
ii. Remove the effect of the selected eigenvector(s) to get A_n via equation defined in Section 2.2.2.
iii. Let $A=A_n$, repeat steps 1 and 2 until all m filters are extracted
iv. Repeat the first three steps with order of covariance matrices reversed, (B, A).

3. Datasets and preprocessing

In order to evaluate the performance of our proposed approach, we implemented a number of experiments using two publicly available datasets of BCI competition 2005 [17]. In particular we tackle the intersession variability problem in a multi-channel ECoG data set and limited training data challenge on an EEG database of five subjects. In all cases the problem was a binary motor imagery classification task.

3.1. Raw data

3.1.1. ECoG data

The ECoG data are recorded from subdural electrodes and consists of 64 channels. During the experiment, the subject performed either tongue or small left finger movement imaginations. Each 768 samples long trial starts 0.5 s after the stimulus and lasts 3 s. There are 139 trials available for training, and 50 trials for test from each class. Training dataset was recorded one week earlier than test dataset. Hence, this dataset poses a great challenge in generalization over sessions. As a preprocessing step, we filtered the ECoG data in 8–16 Hz (α-band) using a dB6 wavelet filter. The selection of this frequency band is based on our previous study using the same dataset [21]. The length of analysis windows was selected as 1 s.

3.1.2. EEG data

The multi-channel EEG data was recorded from five subjects (*aa, al, av, aw, ay*) during either foot or right index finger movement imagination. Signals were collected from 118 electrodes on the scalp. The number of training and test trials for each subject is different which provides the opportunity to evaluate the effect of training size on the overall performance. Namely, starting from the subject *aa* to subject *ay* in order with equal number of trials for each class; there are total 168, 224, 84, 56, and 28 trials for training and 112, 56, 196, 224, and 252 trials for test, respectively. Here the challenge is to estimate robust filters with limited amount of training samples. Prior to estimating the spatial filters, the signals are filtered in 8–30 Hz frequency bands which cover both α and β components. The length of analysis windows was also selected as 1 s.

3.2. Feature generation

We are interested in generating features via CSP framework both using traditional and sparse CSP by following our proposed approach. We used a 1 s analysis window for both ECoG and EEG datasets. In each dataset, using the data in the analysis window the covariance matrices are estimated and 2+2 spatial filters are extracted via the Algorithm 1 provided in Section 2.2.2. The features are generated both via traditional and sparse CSP as provided at the end of Section 2.1. An LDA classifier is considered as the decision function.

4. Performance evaluation

4.1. Experimental procedure

We executed several experiments to compare the performance of sCSP to the standard solution. Moreover, we investigated the efficacy of the greedy forward selection and backward elimination search techniques.

As mentioned earlier, in both EEG and ECoG cases the training and test datasets are provided separately. In order to evaluate the generalization performances, the following procedure is employed. In the case of EEG data, 10-fold cross validation method is used to select the optimum sparseness level for the filters except the subject where only 14 trials are provided for training. For that subject, a leave one trial out cross validation method is employed instead. During cross validation, in each fold, four CSP filters are computed via sparse CSP method where all filters have the same sparseness that is allowed to vary among (1, 2, 5, 7, 9, 11, 16, 32, 64). This way the optimum sparseness level is selected in the training data. After deciding the sparseness level, filters and other classifier parameters using all available training data and we applied learned model to the test data in order to obtain the test error.

Similarly, in the ECoG data, a 10-fold cross validation method is employed using the training data to select the optimum sparseness level. After this point, the same procedure explained above is implemented to get the test error.

4.2. Results

4.2.1. EEG

4.2.1.1. General trend. In Table 1, we provide the summary of the experiments performed. The parameters used to achieve the test errors are also provided. In all subjects the classification errors obtained via sCSP based method is substantially lower than those obtained via traditional CSP in all subjects. The average test error (12.3%) obtained by proposed sparse methods over five subjects is significantly lower than the average error obtained by traditional CSP (18.4%) (p-value=0.013, paired t-test). Another observation is in the number of channels used to reach the test errors. We observed that by employing a sCSP based feature extraction, we not only reduce the number of channels but also improve the generalization capability of the constructed system.

4.2.1.2. Search method comparison. We provide the results obtained by the two search methods to compute the sCSP features in Table 2. Among the five subjects we studied, the FS method achieved the minimum classification error only once, for subject *aw*, and achieved the same error with

BE for subject *al*. For the rest of the subjects, the BE method achieved better classification results. The test errors obtained by sparse methods, the BE method (p-val=0.019) and the FS method (p-val=0.013, paired t-test), are significantly lower than those obtained by traditional CSP.

In addition, the BE method also achieved lower error rates using fewer number of channels. The reason can be explained the way the individual search methods work. The full covariance information is available in the beginning of the BE method whereas that is reached only at the end in the FS method. We observe that the FS method starts generally selecting those electrodes which have higher energy or outliers. Therefore, the available information is better utilized in the BE method. On the other hand the better performance of BE comes with a trade off in computational complexity. This is especially apparent when the desired number of channels is very few and the total number of channels is high. We observed that the BE method reached to a better classification error of 12.6% with only an average sparseness level of 6.6 whereas this level was 21 with the FS procedure. The spatial filters with that constraint will be extracted only in a couple of steps via the FS method but will take far more steps with the BE method. The spatial filters have the same number of nonzero coefficients but not necessarily the filters share the same channels. In order to better evaluate this issue, we counted the number of distinct channels used by all four filters at each cross validation fold and calculated the mean using all available 10 folds. This average is shown under the "Avg. # of Channels" columns in Table 2. In particular, the total number of channels used in test by four filters for subjects *aa*, *al*, *av*, *aw*, *ay* are 17, 16, 35, 25, and 18 for the BE method and 47, 54, 81, 89, and 35 for the FS method, respectively. The number of channels used by all four filters is far less when the method is BE compared to that when the method is FS. But this is because the optimum sparseness levels are lower with the BE method. In other words, both of the methods select almost distinct channels per filter.

In Fig. 1, we provide the average of classification errors using all five subjects at each cardinality as another comparison between the BE and FS. When the cardinality is low, the error level obtained via the BE method is clearly lower than that obtained via the FS method. This behavior changes after the

Table 1
Classification performances, sparseness levels and search methods for the EEG data.

Subject	*Sparse CSP*			*Traditional CSP*	
	Method	Cardinality	Test error (%)	Number of channels	Test Error (%)
aa	BE	5	15.8	118	28.3
al	BE	5	1.8	118	3.6
av	BE	11	25	118	30.1
aw	FS	32	12.5	118	17
ay	BE	5	6.3	118	13.1
Avg:		**11.6**	**12.3**	**118**	**18.4**

Table 2
Classification performances for EEG data, BE compared to FS.

Subject	*BE*			*FS*		
	Avg. # of channels	Cardinality	Test error (%)	Avg. # of channels	Cardinality	Test error (%)
aa	17.5	5	15.8	47.1	16	21.05
al	14.9	5	1.8	49.5	16	1.8
av	37.8	11	25	77.1	32	26.5
aw	24.7	7	14.3	84.6	32	12.5
ay	16.9	5	6.3	32.2	9	11.9
Avg.	**22.4**	**6.6**	**12.6**	**58.1**	**21**	**14.8**

cardinality level is greater than 32. Given the last error point is the one obtained via the traditional CSP; it is also clear that sCSP based features if the sparseness level is correctly selected do provide better results when compared to those obtained via the traditional CSP.

In order to give a flavor about the distribution of spatial patterns and sparse spatial filters, in Fig. 2 we provided head plots of two representative subjects (*aa*, *ay*) with which the sparse solution obtained noticeable improvements with respect the standard CSP. The patterns and filters were ordered according to their topographic distribution. The spatial patterns were located on the left hemisphere and central region which are in correlation with the cortical areas responsible for the right hand and foot control. Since the subjects executed foot and hand motor imageries, the locations of estimated patterns and filters were in accordance with the responsible brain regions. We note that in both subjects although more emphasis was given to the left motor and central area, several other channels in the frontal, temporal and occipital areas were used by the standard CSP method. As expected the sCSP filters were generally distributed in the center of spatial patterns or in the neighboring regions. In both subjects the best sparseness level is five. When compared to the standard CSP solution, the use of sCSP filters resulted in a drop in classification error around 50% in these subjects.

In [22], Arvaneh et al. used the L1/L2 ratio as a penalty term and they applied their algorithm to the BCI competition III EEG dataset IVa [11] which we used in this paper. Here, we present the results that were obtained by employing one filter from each end

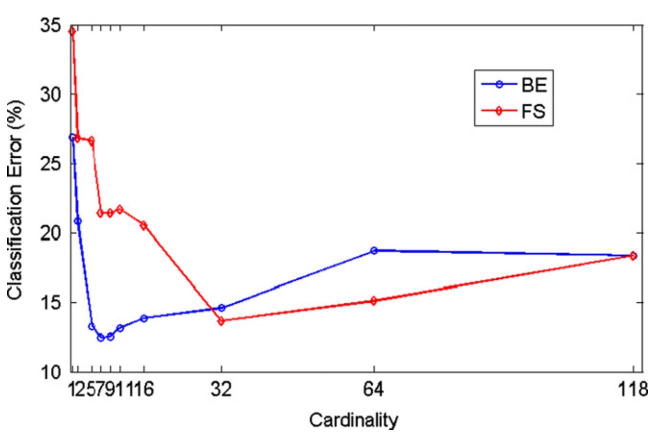


Fig. 1. Average classification errors for EEG data, BE compared to FS.

of the sparse solutions for comparison purposes. We achieved a mean error rate of $10.6 \pm 10.7\%$ for the BE method using 17.6 ± 14 channels. They achieved mean error rate of $17.7 \pm 15.4\%$ using 22.6 ± 11 channels. The one tail t-test between their results and our BE results showed that there was a significant improvement (p-value$=0.018$) without a significant change in the number of channels (p-value$=0.167$).

4.2.2. ECoG

In this part we provide the results obtained from the ECoG data. Summary of the results are provided in Table 3.

4.2.2.1. General trend. Recall that the optimum sparseness level is selected via cross validation in the training data. We observed that the FS method and the traditional CSP achieved the same test error of 13%. For the BE method the test error was 10%. Both sparse methods achieved the best classification error with a sparseness level of 7. We note that the sCSP based methods provide similar, as in the case of FS, or lower test errors as in the case of BE when compared to the traditional method. It should be noted that the FS method achieved the same error of the traditional method with less number of channels. As we verified in the previous part for EEG setup, constraining the number of channels used by the spatial filters does result in a better generalization performance in a CSP framework also on the experiments with ECoG signals.

4.2.2.2. Search method comparison. In order to provide a comparison between the sCSP methods, the classification error curves versus the sparseness level are shown in Fig. 3. First thing to note is that although both methods use the same sparseness level, the test errors are 10% and 13% for the BE and the FS method, respectively. Note also that the results obtained at different sparseness levels with the BE method are consistently better than those obtained by the sparse FS method. With the BE

Table 3
Classification performances for ECoG data.

Search method	Sparse CSP		Traditional CSP
	BE	FS	
Cardinality	7	7	64
Test error (%)	10	13	13
Avg. # of channels	21.2	22	64

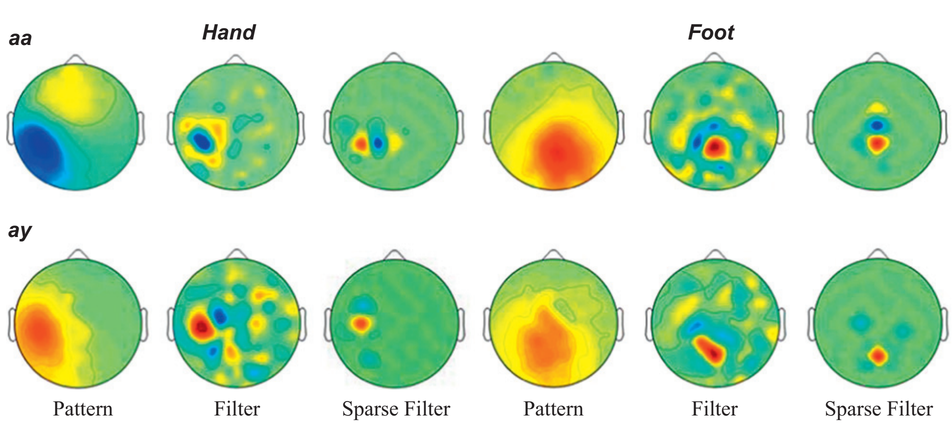


Fig. 2. For two representative subjects, the first and last CSP patterns, related filters and sparse filter solutions are presented from left to right. The sCSP filters were estimated with BE search. The spatial patterns were located on the left hemisphere and central region representing the hand and foot imageries respectively. Note that the standard CSP method estimated filters using all channels where the filter weights were spread all over the head. For both subjects the sCSP filters used only five channels in each filter.

method, the test error of 10% has been achieved with three more sparseness levels in addition to optimum selected level of 7.

It can be claimed that the results obtained by the BE method are more stable than those of by the FS method. Namely, change in the sparseness does not result in big difference in the classification error. Finally, we evaluate the number of channels for each filter. We counted the number of channels for each cross validation fold and take the average to find the average number of channels per filter. The total number of distinct channels used by four spatial filters during test are 23 and 20 for the FS method and the BE method, respectively. This means that the sparse CSP methods provide comparable and better classification results than those obtained by the traditional CSP method by using only 1/3rd of the available channels.

In Fig. 4 we provided the distribution of first spatial patterns and corresponding full and sparse spatial filters from each class. The sparse filter is obtained using the BE method with sparseness level of 7. One observation to note is that the spatial patterns obtained via the ECoG data is well localized as opposed to those of obtained via the EEG data. This is due to the higher spatial resolution available with the former. The spatial pattern shown in the first row was located in the part of the brain responsible for left hand imagery and the spatial pattern shown in the second row was located in the part of brain region responsible for speech imagery. These results confirm our expectation with respect to imagined movements. As they are clearly seen in the figure, the sparse filters select the channels whose locations overlap well with the spatial patterns. On the contrary, the full spatial filters utilize channels over the entire area.

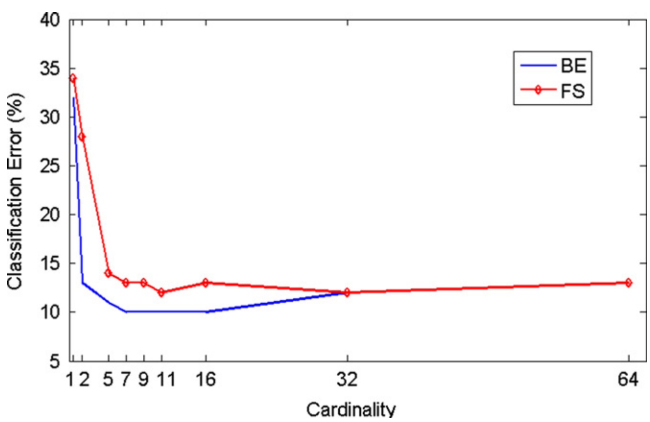


Fig. 3. Classification errors for ECoG data, BE compared to FS.

4.3. Computational complexity

It is worth mentioning the scalability of the proposed methods in comparison to the traditional approach. The worst case computational complexities of finding one generalized eigenvector for the FS, BE, are reported as $O(n^3)$, $O(n^4)$, [14] and the computational complexity of the traditional method is $O(n^3)$, [18], where n denotes the size of the covariance matrices. The cases where the desired cardinality is large and where the desired cardinality is small correspond to the worst case situations for the FS and BE methods, respectively. Therefore, for the case where the goal is to extract sparse enough CSP filters, we can expect the FS and BE methods to be implemented considerably faster and slower than the traditional method, respectively.

Since the BE method has high computational complexity, it will not be the desired method to obtain sparse solutions when the dimension of the problem is high and this maybe the case for some other application areas. But with around 100 channels in a typical BCI application, computational complexity should not be a problem when using the greedy search based methods.

4.4. Sparse spatio-spectral filtering

We have shown the superior performance of our proposed sCSP methods in the previous sections over the traditional technique. Here, we argue that a method which is derived from traditional CSP may benefit from employing proposed sparse methods. In this scheme, we conducted an exploratory analysis of using proposed sCSP methods in spatio-spectral filtering. The common spatio-spectral pattern (CSSP) method of [23] is an

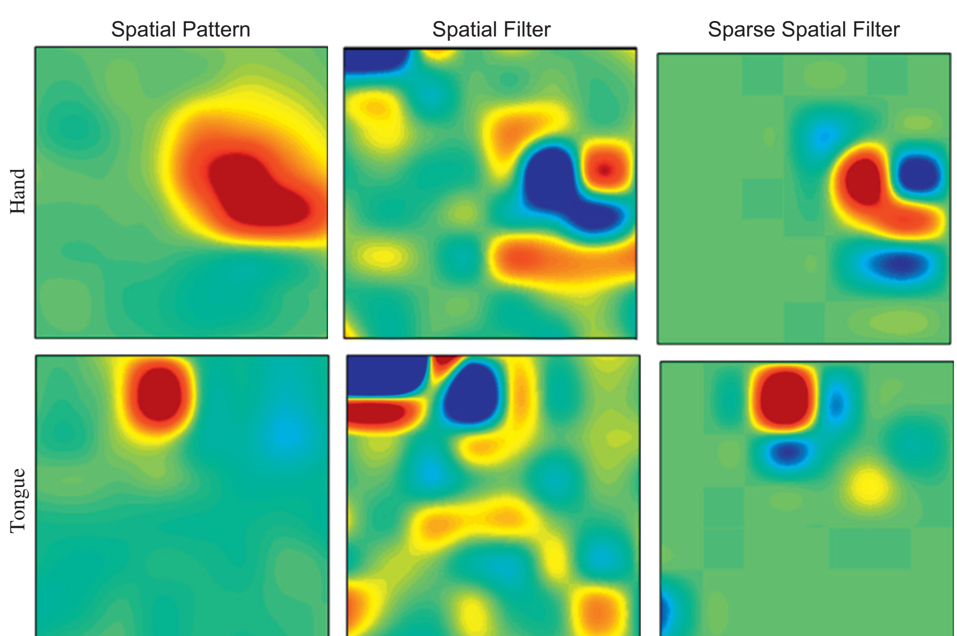


Fig. 4. Spatial pattern and spatial filters: full versus sparse. The filters in the first row minimize the variance of the finger movement whereas the filters in the second row minimize the variance of the tongue imaginations.

extension of the traditional CSP that combines spectral filtering with the original spatial filtering. Their approach was simple and efficient as we described next. In Section 2, $X_i \in R^{C \times N}$ is defined as the ith trial of the multi-channel neural data where C is the number of channels and N is the number of time samples in each channel. First by considering only one coordinate in time a delayed trial is obtained, X_i^{τ}, where τ denotes the amount of delay. Then the traditional CSP algorithm is employed on the enlarged space that is obtained by merging the original and delayed channels to extract the spatial filters. Therefore, in this enlarged space the size of the covariance matrices are $2C$, hence, the length of the spatiao-spectro filters. In each filter half of the coefficients apply to X_i and the other half to X_i^{τ} to derive the CSP filtered multi-channel neural data as in $Z_i = W^T X_i + W_{\tau}^T X_i^{\tau}$. Here, W denotes the collection of spatial filters that applies to the original signals and W_{τ} the same that applies to the delayed signals. Since each time sample of the spatially filtered signal z_i is expressed as the linear combination of the signal and its delayed version, the whole process described so far can be interpreted as an application of the CSP algorithm to the finite impulse response (FIR) filtered multi-channel neural data, hence, it is named as CSSP.

The advantage of this approach is that it extracts the spectral (frequency) information from the data that is specific to the subject. This advantage on the other hand comes with an additional (delay) parameter, τ, to tune for during the calibration phase. Finally, it is clear that when the delay parameter is set to zero we get the original CSP algorithm. For further information about the details of the CSSP algorithm and its performance in a BCI framework we refer the reader to [23]. Since the CSSP algorithm is an extension of the original CSP algorithm as described earlier, the proposed sparse methods are also applicable to CSSP as well, hence, we would like to obtain sparse CSSP. It is worth noting that employing sparseness in CSSP framework was attempted in [24]. We have to emphasize that in [24] sparseness is sought in the length of the spectral FIR filter which was allowed to be arbitrary long. On the contrary, we are seeking sparseness in the spatial domain.

In our exploration the goal is to analyze the effect of sparseness in CSSP as a function of the number of channels. In the first step the best delay parameter is decided for each subject in both of the EEG and ECoG datasets by the standard CSSP algorithm.

In the second step, sparse CSSP filters are extracted with the proposed BE and FS methods on the twice enlarged space that is obtained by the original and delayed data. In the above analysis in order to keep the spectral information that is obtained by processing the original signal and its delayed version, we modify the greedy search methods such that a channel and its delayed version are considered at each step. In more detail, in each iteration, a channel and its delayed version is removed in the BE method or added in the FS method. The spatial sparseness levels are allowed to be one of (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 20, 24, 28, 32, 40, 48, 64). The results as classification errors are provided in Fig. 5. In the case of EEG dataset the errors are averaged over five subjects. In addition to classification errors obtained by each of the proposed sparse methods in CSSP algorithm as a function of sparseness level, the errors obtained by the traditional CSP algorithm are provided in the plots as well.

In both of the plots in Fig. 5, the last error points of FS and BE correspond to the errors that are obtained by the original CSSP method working on all channels with the best delay possible. We observed that the standard CSSP method resulted in 18% error in the ECoG dataset and 13.8% error in the EEG dataset. The spectral filtering in addition to spatial filtering results in worse performance in ECoG dataset and better performance in the EEG dataset when all the channels are used. Employing sparseness in the spatial domain improved the CSSP performance in ECoG dataset considerably by both FS and BE methods especially when the spatial filters are highly to moderately sparse. The better performance is only valid in the EEG dataset when the spatial filters are highly sparse. In both of the datasets the best performances are obtained by spatial filters that are considerably sparse and moderately sparse with the BE method and the FS method, respectively. Nevertheless, the sparse CSSP method provided lower classification error rates than the standard CSSP and sparse CSP methods in both datasets.

5. Conclusion

In this study, we show that by employing L0 norm based greedy search methods, sparse spatial filters are easily extracted in a CSP framework. These sparse spatial filters provide better performances than those obtained by the traditional non-sparse approaches.

Our motivation when seeking for a sparse CSP solution is simple. By making each spatial filter sparse in terms of the number of channels used, there will be a decrease in the amount of variance explained by linear combination of channels but this decrease will be small since only a subset of channels are expected to contribute to real variance difference. Therefore,

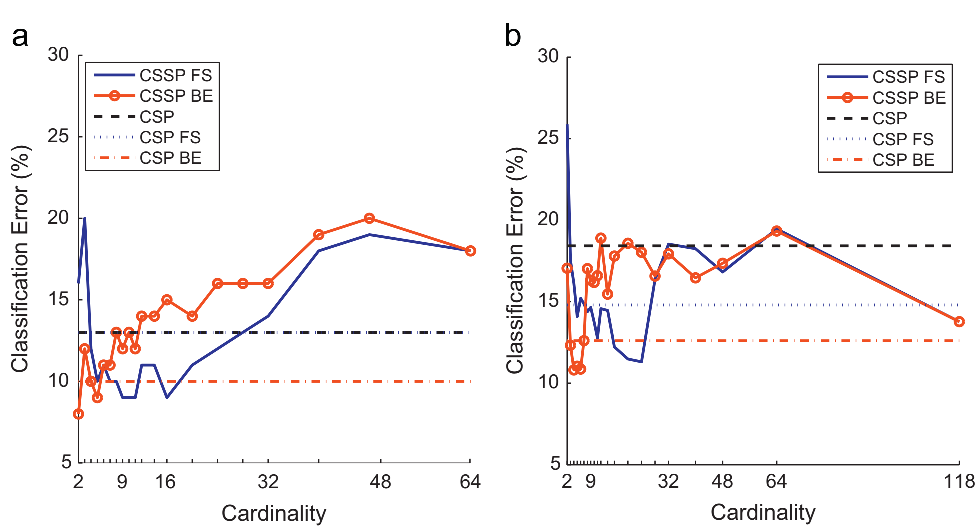


Fig. 5. Classification errors for sparse CSSP using different search strategies. The last data points using all channels correspond to standard CSSP.

constructing spatial filters from a subset of all available channels yields improved generalization performance. First, we observe the CSP optimization formulation as a GED problem. After showing the combinatorial nature of the sparse GED problem solution, we employed greedy search based techniques to find multiple sparse eigenvectors. For this particular purpose, we utilized two different greedy search techniques, (i) forward selection and (ii) backward elimination with low computational complexity. Finally, we evaluated the performance of our sparse CSP based approach on two binary BCI problems.

The sparse CSP method was applied to the multi-channel EEG and ECoG datasets of BCI competition 2005. These two different setups gave us the opportunity to examine the performance of our algorithm on two different challenges; (i) availability of limited amount of training data and (ii) intersession variability, respectively. In more detail, the number of subjects and the varying training and test data sizes in the EEG based data setup did provide us to investigate the generalization capacity of our sparse CSP method in a unique way. The test errors obtained by the sparse CSP method in EEG data over five subjects was 12.3% whereas the test error rate of the traditional CSP was 18.4%. We observed that the BE method reached to a better classification error of 12.6% with only an average sparseness level of 6.6 over five subjects. This is a remarkable decrease in the number of channels used compared to the standard CSP. Our results show that the generalization capabilities of the sparse CSP methods compared to those obtained by the traditional CSP method do improve not only when there is large amount of recording channels but also when the size of the training data is small.

Using the ECoG data we are able to evaluate our method's performance when there is intersession variability between the training and test sets. The sparse CSP methods, based on the *BE* and *FS* search approaches achieved lower or comparable test errors when compared to those of traditional CSP method, respectively. This result verifies our claim that by reducing the number of channels in a CSP problem using greedy search, the extracted features are made more robust and less prone to the variations in the data between training and test phases.

The classification performances obtained by the sparse CSP based methods are superior to those obtained by the traditional CSP method. To the best of our knowledge our study is the first showing that sparse spatial filters can provide better classification accuracy of neural activity than the traditional CSP in a BCI framework. When individual search methods are investigated in detail, the results obtained by the BE method are better than those obtained by the FS method especially when the sparseness level is low. However, the computational complexity of BE was much higher than the FS. This in turn gives an opportunity to trade off between computational complexity and classification performance. One would prefer the BE method if time required to extract the spatial filters is not a concern. On the other hand, the FS method is preferable for fast extraction of the spatial filters especially when the desired sparseness level for each filter is low.

We observed that in both EEG and ECoG datasets, on average the channel overlap among the extracted sparse spatial filters is rather low. Furthermore, the optimum sparseness levels on average are very low when compared to the number of available channels for BE in EEG and for BE and FS in ECoG datasets. On average, the total number of channels used by the BE method does not exceed 19% and 33% of all available channels in EEG and ECoG, respectively. On the other hand, the FS method used 49% and 34% of all available channels in EEG and ECoG, respectively. The reason we believe is peculiar to the source signal. ECoG signals do have high signal to noise ratios and are spatially more localized since the electrodes are directly in contact with the brain whereas the EEG signals contain a considerable amount of noise and have lower spatial resolution as it is measured on the scalp. Due to these facts the FS method could not find a good solution to the CSP problem with lower sparseness level on the EEG data as the channels with outliers or those ones with high variance at the temporal and occipital regions were generally selected at initial iterations. However, the FS method did find reasonable sparse filters with the ECoG data.

In the studies [9 and 11], where sparse CSP solutions are achieved via L1 norm regularization, the number of channels used by each spatial filter is reduced considerably but with a loss in the classification accuracy. This is in contrast with our results obtained via greedy search based sparse CSP. We believe the reason can be explained that for the same sparseness level, search based methods are able to find sparse eigenvectors with higher variance difference between classes [14]. Similar results were observed in [13] in the case of PCA problem where using L1 norm regularization compared to that of using greedy search explained less variance of the data. Finally, with respect to greedy search based methods for finding sparse CSP filters, it should be noted that this approach is flexible in the sense that it provides eigenvectors (solutions) with all sparseness levels. Using this flexibility, it is possible to force a certain pattern when searching for the sparse solution which is not available by L1 norm based solution when extracting the solution.

We extended our sparse solution to the CSSP method which is an extension of CSP method where a spatio-spectral filter is constructed on the data in a FIR filter framework by embedding a delayed version of the original. This gives the CSSP the opportunity to select subject specific bands in the 8–30 Hz prefiltered EEG/ECoG data. While providing a spectral selectivity in the selected 8–30 Hz band, it should be noted that even a single delay doubles the amount of the data in the CSSP framework. For instance, a neural recording obtained from 64 channels and a single tap delay results a covariance matrix of 128×128 to be used in the estimation of spatio-spectral filters. Consequently the curse of dimensionality becomes a more important issue in CSSP than the CSP. As evidenced with the results we obtained from multi-channel EEG and ECoG data, selection of a sparse spatio-spectral filter will likely increase the generalization capacity of the CSSP method as in a very high dimensional space the algorithm might easily find spurious projections and overfit the data. Consequently, finding sparse solutions might have higher impact in CSSP applications.

In this scheme, we note that the sparseness in spectral domain was introduced in [24] by so called "Common Sparse Spectral Spatial Pattern" (CSSSP) algorithm. The approach in [23] only explored the best choice among single tap filters where the delay was investigated between 0 and 15. While employing the sparse CSSP method, we focused on the sparseness in the spatial domain. We need to note that, the L0 norm based greedy search method that we propose can seek sparseness in spatial, spectral and jointly in spatio-spectral domain depending on the implementation of the search technique. However, as we describe below, implementation of the latter two are infeasible due to the computational complexity. Assume that we are seeking for sparseness in spectral domain using the same amount of channels in our EEG data. In this scheme, as in [24], a 16 tap filter will yield a covariance matrix size of 1856×1856 (118 channels $\times$ 16 filter taps). After incorporating all delays of the FIR filter in the CSSP framework, the covariance matrix which we will use in greedy search becomes quite large. On this large matrix, a column (simultaneously a row) can be eliminated at a time. This will allow one to select channel subsets with arbitrary length filters for each. Consequently, using our technique, it is possible to obtain sparse solutions in spatio-spectral domain jointly. However, working with such a big matrix will increase the computational complexity dramatically. Implementing a search in a high dimensional space will cause severely extended training times which would not be feasible for practical BCI scenarios.

It should be noted that, as in [24], it is also possible to seek sparseness only in spectral domain using our approach. For this particular purpose, one can simply restrict the elimination to delayed samples of all channels in the covariance matrix. However, such an approach is still far too complex as the dimension of the search space is quite large. Therefore, in this manuscript, we elected to set a single fixed delay for all channels as in [23] and eliminated a channel and its delayed version at a time. Consequently, the greedy search is implemented in the spatial domain (118×118 for EEG and 64×64 for ECoG), which enabled us to run/train the algorithm on a standard computer architecture within a reasonable time limit.

Compared to CSP and CSSP, we observed noticeable better results with our spatially sparse CSSP method. Interestingly, on similar motor imagery tasks, the authors of [24] reported the median classification error rate for the CSSSP, CSSP and CSP algorithms as 20.7%, 21.0% and for 23.3%, respectively. Although they concluded that the CSSSP is superior to CSSP and CSP, to our understanding, these error rates indicated that the sparseness in spectral domain did not provide any noticeable improvements over standard CSSP. In contrary, our results indicate that there is a clear difference between standard CSSP and its sparse derivation (18% vs. 8% for ECoG and 13.8% vs. 10.8% for EEG datasets using CSSP and sparse CSSP respectively). This could be due to the amount of sparseness achieved in the final setup. Specifically, the difference between the number of electrodes (118 for EEG and 64 for ECoG) and temporal delays (16 filter taps) is large. Consequently, compressing the spatial domain might be more advantageous than the spectral due to the difference in their dimensions.

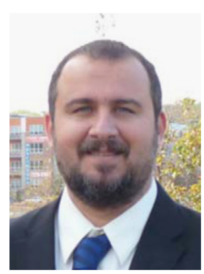
Acknowledgments

This research was supported in part by the National Science Foundation, award CBET-1067488. The authors express their appreciation for the constructive comments provided by the referees, which served to considerably improve the manuscript.

References

[1] G. Pfurtscheller, C. Neuper, Motor ımagery and direct brain-computer ı nterface, Proc. IEEE 89 (2001) 1123–1134.
[2] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K.R. Muller, Optimizing spatial filters for robust EEG single trial analysis, IEEE Signal Proc. Mag. (2008) 41–56.
[3] Z.J. Koles, M.S. Lazar, S.Z. Zhou, Spatial patterns underlying population differences in the background EEG, Brain Topogr. 2 (4) (1990) 275–284.
[4] C. Guger, H. Ramoser, G. Pfurtscheller, Real-time EEG analysis with subject-specific spatial patterns for a brain computer ınterface (BCI), IEEE. Trans. Neural Syst. Rehab. Eng. 8 (4) (2000) 447–456.
[5] H. Ramoser, J. Muller-Gerking, G. Pfurtscheller, Optimal spatial filtering of single trial EEG during imagined hand movement, IEEE. Trans. Rehab. Eng. 8 (4) (2000) 441–446.
[6] N.F. Ince, R. Gupta, S. Arica, A.H. Tewfik, J. Ashe, and G. Pellizzer, Movement direction decoding with spatial patterns of local field potentials, in: Proceedings of the IEEE Neural Engineering Congress, Antalya, Turkey, 2009.
[7] N.F. Ince, R. Gupta, S. Arica, A.H. Tewfik, J. Ashe, G. Pellizzer, High accuracy decoding of movement target direction in non-human primates based on common spatial patterns of local field potentials, PloS ONE 5 (12) (2010) e14384, http://dx.doi.org/10.1371/journal.pone.0014384.
[8] Z.J. Koles, The quantitative extraction and topographic mapping of the abnormal components in the clinica EEG, Electrcoencephalogr. Clin. Neurophysiol. 79 (6) (1991) 440–447.
[9] J. Farquhar, N.J. Hill, T.N. Lal, and B. Schölkopf, Regularised CSP for sensor selection in BCI, in: Proceedings of the Third International BCI Workshop Training Course, Graz, Austria, 2006, pp. 14–15.
[10] B. Reuderink, M. Poel, Robustness of the Common Spatial Patterns Algorithm in the BCI Pipeline, Internal Report, University of Twente, 2008.
[11] X. Yong, R.K. Ward, G.E. Birch, Sparse spatial filter optimization for EEG channel reduction in brain computer ınterface, IEEE. ICAASP (2008) 417–420.
[12] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, New York, 2004.
[13] B. Moghaddam, Y. Weiss, S. Avidan, Spectral Bounds for Sparse PCA: Exact and Greedy Algorithms, NIPS 18, Vancouver, BC, Canada, 2005.
[14] B. Moghaddam, Y. Weiss, S. Avidan, Generalized spectral bounds for sparse LDA, in: Proceedings of the 23rd ICML, Pittsburgh, PA, 2006.
[15] F. Goksu, N.F. Ince, A.H. Tewfik, Sparse common spatial patterns in brain computer ınterface applications, in: 36th IEEE – ICAASP, Prague, Czech Republic, 2011 533–536 .
[16] J. Duchene, S. Leclercq, An optimal transformation for discriminant and principal component analysis, IEEE Trans. Pattern Anal. Mach. Intell. 10 (6) (1988) 978–983.
[17] B. Blankertz and K.R. Muller, The BCI Competition III, 2005 ⟨http://www.bbci.de/competition/iv/⟩.
[18] J. Demmel, Applied Numerical Linear Algebra, SIAM, Philadelphia, 1997, pp 178–179.
[19] A. D'Aspremont, F. Bach, L. El-Ghaui, Optimal solutions for sparse principal component analysis, J. Mach. Learn. Res. 9 (2008) 1269–1294.
[20] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd Edition, Springer-Verlag, 2009. (Chapter 3).
[21] F. Ince, F. Goksu, A.H. Tewfik, An ECoG Based Brain Computer Interface with Spatially Adapted Time Frequency Patterns, International Conference (BIOSTEC), Portugal, January 2008.
[22] M. Arvaneh, C. Guan, K.K. Ang, C. Quek, Optimizing the channel selection and classification accuracy in EEG-based BCI, IEEE Trans. Biomed. Eng. 58 (6) (2011) 1865–1873.
[23] S. Lemm, B. Blankertz, G. Curio, K.R. Muller, Spatio-spectral filters for improving the classification of single trail EEG, IEEE Trans. Biomed. Eng. 52 (9) (2005) 1541–1548.
[24] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, K.R. Muller, Combined optimization of spatial and temporal filters for improving brain-computer interfacing, IEEE Trans. Biomed. Eng. 53 (11) (2006) 2274–2281.

Fikri Goksu received the Ph.D. degree from University of Minnesota, Twin Cities, MN, USA, in 2011. Fikri is currently a Senior Analyst at Thomson Reuters, Washington, DC. His areas of interest included pattern recognition and machine learning methods and their applications to real world problems.



Nuri Firat.Ince received his BSc., M.Sc. and Ph.D. degrees all in Electrical and Electronics Engineering from Cukurova University, Adana, Turkey in 1999, 2002 and 2005. respectively. During his Ph.D. he worked as a visiting scholar at the University of Minnesota. His thesis was supported by the National Scientific Research Council of Turkey (TUBITAK) with an International Joint Ph.D. scholarship. Between 2005 and 2008, he completed dual post doctoral training in the fields of biomedical signal processing and computational neuroscience in the department of Neuroscience at the University of Minnesota and the Brain Sciences Center VA Medical Center Minneapolis. Currently, Dr. Ince is a research assistant professor holding joint appointments with the Departments of Electrical Engineering and Neurosurgery at the University of Minnesota. He is serving as a reviewer to several journals in signal processing and biomedical engineering. His research interests include neural engineering, wearable medical sensors, in-home assistive systems, programmable wireless sensor networks, adaptive time frequency analysis and classification of biomedical and acoustic signals. In the past he worked on cardivascular signal processing and biomedical instrumentation.

Ahmed H. Tewfik received the B.Sc. degree from Cairo University,Cairo, Egypt, in 1982 and the M.Sc., E.E., and Sc.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1984, 1985, and 1987, respectively. He was a Distinguished Lecturer of the IEEE Signal Processing Society in 1997–1999 and received the IEEE Third Millennium award in 2000. He is currently both the Chair of the Department of Electrical and Computer Engineering and a Cockrell Family Regents Chair in Engineering at the University of Texas at Austin.