Sparse Common Spatial Patterns with Recursive Weight Elimination

Fikri Goksu*, Firat Ince*† and Ibrahim Onaran†
*Department of Electrical and Computer Engineering
†Department of Neurosurgery
University of Minnesota, Twin Cities, MN 55455 USA

***Abstract*—The past decade has shown the importance of adapting spatial patterns of neural activity while decoding it in a Brain Machine Interface (BMI) framework. The common spatial patterns (CSP) algorithm tackles this problem as feature extractor in binary BMI setups in which a number of spatial projections are computed while maximizing the variance of one class and minimizing of the other. Recent advances in data acquisition systems and sensor design now make recording the neural activity of the brain with dense electrode grids a possibility. However, high density recordings also pose new challenges such as overfitting to data as the number of recording channels increases dramatically compared to the number of training trials. In this study, we tackle this problem by constructing a sparse CSP algorithm through recursive weight elimination (CSP RWE), in which the spatial projections are computed using a subset of the recording channels. The sparse projections are expected to yield increased robustness and eliminate overfitting. We show promising results towards the classification of multichannel Electrocorticogram (ECoG) and Electroencephalogram (EEG) datasets with CSP RWE for a BMI.**

I. Introduction

In a BMI the neural activity of impaired subjects are translated into communication and control commands, e.g. in the form of binary labels as used in [1] to select a letter on a virtual keyboard. Recent advances in data acquisition systems and sensor design now make it possible to record the neural activity of the brain with dense electrode grids. The CSP method first proposed in [2] is a powerful signal processing technique for feature extraction from multichannel neural recordings in a BMI framework. The main task is to obtain a set of spatial filters where each filter has the same length that is equal to the number of recording channels available. Using each of these filters the multi channel neural data is linearly projected into a one-dimensional signal such that the variance obtained in this dimension is maximized for one class and minimized for the other. This is achieved by using the correlation among a number of recording channels whose spatial extend is related to the executed task. The CSP method has become one of the commonly used methods in BMI applications due to its success, [3].

One of the major problems of CSP especially arises in setups where the number of recording channels is higher than the number of training trials. This results in overfitting hence poor generalization performance. Another problem is lack of robustness over time. The robustness in this case can be described as the sensitivity of the BMI system to the variation in data. Multi channel neural recordings obtained at different times cause variation in data. The extracted CSP features are quite sensitive to such variations due to extraction of features by linear combination of all channels. Minimization of these handicaps requires the placements of electrodes on the scalp with the same setup from one session to another which is difficult to accomplish. Moreover, the chance that CSP uses a noisy or corrupted channel is linearly increased with increasing number of recording channels. These problems were previously reported in [4]–[6].

The handicaps summarized above can be reduced by seeking sparse solutions when extracting the spatial filters. Regularization by sparse solutions in the CSP formulation has been previously explored in [5], [7]. Basically, with slightly different formulations, they attempted to obtain sparse CSP (sCSP) filters by adding an L1 norm constraint on the size of the spatial filter. In both studies it was reported that although the number of channels are reduced considerably, this is obtained with slight decrease in classification performance. Recently, the authors of this paper attempted to obtain sCSP filters by employing greedy search methods which resulted a decrease in the number of channels and an improvement in the classification performance [8]. When extending the greedy search based methods to obtain sCSP filters in [8], we only considered two search approaches; Forward Search (FS) and Backward Elimination (BE). The former builds the sparse solution starting with the empty set and adds variables into it as going forward. On the other hand, the latter starts with full solution and removes variables iteratively. In the experiments performed it was observed that both FS and BE methods outperformed the standard CSP where the BE method provided better classification performance with low number of channels than the FS method. However, this better performance of BE method came with the cost of considerably higher computational complexity when extracting the sparse solutions. In this paper we propose a sparse CSP algorithm through recursive weight elimination (CSP RWE) that is expected to have lower computational complexity than the BE method while providing a comparable classification performance. Basically, instead of searching all possibilities while implementing the BE search, we rank the weights of the CSP vector and remove the least important one. This reduces the computational complexity of original BE search method considerably where the reduction is especially appreciated when the number of channels is

978-1-4673-0323-1/11/$26.00 ©2011 IEEE

large. We explored the performance of the RWE based sCSP method on publicly available EEG and ECoG datasets of BCI competitions III and IV. Its performance is evaluated against the greedy FS and BE methods as well as the traditional CSP. In the following sections we first introduce the CSP based feature extraction. Then we shortly review the greedy search based sparse CSP methods and provide the details of the proposed RWE method. The paper continues in Section IV with the experimental setup and obtained results. Finally, we discuss our results and future work in Section V.

II. Common Spatial Patterns

Let us here shortly describe the traditional CSP and its optimization formulation. The reader is referred to [4] and references therein for a detailed review of the CSP method.

The CSP is a supervised signal processing technique that solves the following optimization problem to find the spatial filters (w),

$$\arg\max_{w} \frac{w^T A w}{w^T B w} \tag{1}$$

where A and B denote the sample *spatial* covariance matrices of two different classes. The objective function is known as the Rayleigh quotient (RQ). Deriving the Lagrangian of the optimization problem in (1) and taking the derivative w.r.t. variable w gives us;

$$Aw = \lambda B w \tag{2}$$

Equation (2), with positive definite matrices is known as the generalized eigenvalue decomposition (GED) problem, [9], which has a closed form solution that diagonalizes both of the covariance matrices when multiplied from left and right. Let C be the number of recording channels. There are C generalized eigenvector and eigenvalue pairs. Each eigenvector of this solution can be used as a spatial filter in a CSP application. Furthermore, corresponding eigenvalues are variances; hence, indicate which spatial filters to select for feature extraction. In practice equal number eigenvectors are selected from both end of the spectrum. The spatial filters in one half maximize the variance for one class and in the other half maximize for the other class.

III. Sparse Csp

To obtain sparse solutions of the GED problem we follow the approach developed in [10] which is based on the observation that finding a sparse solution to a GED problem with L0 norm penalty on the eigenvector is combinatorial as it is discussed in the following. We assume that the solution of a GED problem without cardinality constraint on the solution vector is available.

Assume the covariance matrices A and B and a sparse eigenvector w with k nonzero elements are given. That means w is the solution of the following optimization problem;

$$\arg\max_{w} \frac{w^T A w}{w^T B w} \text{ s.t.} \|w\|_0 = k \tag{3}$$

We observe that multiplication of A (B) from left by w^T selects rows of A (B) corresponding to non zero indices of w. Similarly, multiplication of A (B) from right by w selects columns of A (B) corresponding to the same indices. Therefore, the objective function in (3) is equivalent to;

$$\arg\max_{w} \frac{w^T A w}{w^T B w} \text{ s.t.} \|w\|_0 = k \tag{4}$$

Note that the full solution vector s using A_k and B_k is the sparse solution vector w with appropriate indices such that A_k, B_k are $k \times k$ dimensional submatrices obtained by keeping the rows and columns of (A, B) corresponding to nonzero indices of w. As a result, since we know how to solve for s which maximizes the right hand side of the equality in (4), then we have a sparse vector w with cardinality k that maximizes the objective in (3). The catch here is how to decide which $k \times k$ submatrices to keep that is the list of indices of k rows and columns. Searching all possible $k \times k$ submatrices of (A, B) will be infeasible for covariance matrices of large sizes. The alternative is to employ suboptimal greedy search algorithms such as FS and BE. We obtained promising results in our previous work in [8] using FS and BE.

A. Forward Search

This search starts with empty index set and adds variables into the set one by one going forward. At the m^{th} step, it searches for all $m \times m$ possible submatrices by adding one variable at a time and adds the index to the set whose inclusion increases the variance the most. This sequential search continues until the desired cardinality is reached.

B. Backward Elimination

The BE search starts with the full set and removes variables one by one going backward until the desired cardinality is reached. At any given it searches for all possible submatrices with the size of interest by removing one variable at a time and removes the variable (row and column index) from the set whose exclusion decreases the variance the most.

C. Recursive Weight Elimination

The search method described in detail below is our proposed solution for decreasing the computational complexity of the BE method. Assume that the full size of the covariance matrices in the CSP method that we want to obtain a sparse solution from is C. The BE method in the first step searches $C-1$ separate submatrices and solves GED problem for each of them to find a sparse solution whose cardinality is $C-1$. Hence, a GED is solved $C-1$ times on $C-1 \times C-1$ matrices. In each step the size of the submatrices becomes one less and that is also equal to the number of separate GED solutions that is performed at each step. As a result, until the desired cardinality is reached the total number of separate GED solutions dominates the computational complexity. The computational complexity is even higher when C is large and the desired cardinality is small. Based on this observation we propose employing only one GED solution per step and recursively eliminate variables based on their weights. This recursive weight elimination approach is motivated by the work of [11] which employed recursive feature elimination in an SVM framework. The authors of [11] assumed that the coefficients of the weight vector are related to their contribution to the maximum margin of the SVM. In a recursive fashion they eliminated small weights and corresponding features and recomputed the SVM margin until desired number of features remained. Here, with the same spirit, we assume that the values of the spatial filter coefficients represent their contribution to the maximized RQ objective function. Our proposed RWE method proceeds such that in any step instead of searching all possible submatrices, we get the full solution from the previous step, rank the coefficients based on the absolute values and remove the coefficient (index) that has the minimum value. In the next step we obtain the full solution with this new channel subset by solving the corresponding GED problem and repeat the ranking and removing procedure until the desired cardinality is reached. For instance, in the first step the BE method solves the GED problem in $C-1$ separate submatrices whereas the RWE method uses a single GED solution of C channels. Consequently, the RWE approach will decrease the computational complexity of the search dramatically.

D. Multiple Sparse CSP Filters

At this point, we have search methods that provide CSP filters with cardinalities ranging from 1 to C. After computing a filter with certain cardinality, we can calculate the next sparse CSP filter by employing the same search methods following a deflation procedure [12]. The next spatial filter is computed using the following equation

$$[I - TD^T(DTB^{-1}TD^T)^{-1}]Aw_m = \lambda Bw_m \qquad (5)$$

where $D = [w_1 \ \ldots \ wm-1]^T$. When T is set to equal to B the solution vectors not only diagonalize the matrix A but also the matrix B. We can think of the multiplication of the matrix A from left with D as removing the effect of previous eigenvectors from it. The problem in (5) is still a GED with different but known left hand side and any of the greedy search methods, FS, BE, or RWE can be employed to find the next sparse eigenvector.

IV. Performance Evaluation

A. Dataset and Preprocessing

We evaluated the performance of the proposed approach on two class ECoG (dataset I) and EEG (dataset IVa) datasets of BCI competition 2005, [13] and multiclass ECoG (dataset 4) dataset of BCI competition 2008 [14]. The ECoG dataset of BCI competition 2005 gives the opportunity for evaluating the robustness of the proposed approach over time since the training and test sets are recorded in two different sessions with one week apart. On the other hand, the EEG data is provided with small training data that gives the opportunity to evaluate the classification capacity of the proposed solutions against overfitting. Finally, we tested our algorithm with a multiclass finger movement BMI problem where small amount of training data is available,

TABLE I
Classification performance for ECoG data

	Sparse CSP			*Standard CSP*
Search Method	***RWE***	***BE***	***FS***	
Sparseness Level	16	7	5	64
Test Error (%)	11	10	12	13

The ECoG data BCI competition 2005 is recorded with 64 channels with total of 278 trials available for training and 100 trials for testing where the number of trials is evenly distributed between the two classes. During the experiment, the subject imagined either tongue or small left finger movements. The ECoG data is filtered in 8 to 16 Hz (-band). The EEG dataset of BCI competition 2005 was collected from five subjects with 118 channels. The subjects were asked to imagine either foot or right index finger movements. For each subject there are 280 trials in total. The number of trials for each subject is 80%, 60%, 30%, 20%, and 10%. The EEG data is filtered in 8-30Hz band prior to feature extraction. The ECoG dataset of BCI competition 2008 was recorded from three subjects during finger flexions and extensions [14] with a sampling rate of 1 kHz. Each electrode array contained 48 (8x6) or 64 (8x8) platinum electrodes. The finger index to be moved was shown with a cue on a computer monitor. The subjects moved one of their five fingers 3-5 times during the cue period. The ECoG data of each subject was subband filtered in the gamma frequency band (65-200Hz). The dataset contains around 146 trials for each subject. In all these dataset, we used 1 sec data following the cue onset for feature extraction. In order to evaluate the classification performance of all methods we computed four spatial filters. We used an LDA classifier on this four dimensional feature space for final decision.

B. Results

In both of the two class ECoG and EEG datasets of BCI competition 2005, a 10-fold cross validation method is employed to select the optimum cardinality (sparseness level) on the training data only. We investigated cardinality values of (1, 2, 3, 4, 5, 7, 9, 11, 16, 32, and 64) channels. For multiclass ECoG data, a 10-fold cross validation was not feasible due to the high computational complexity of the BE method. Therefore, we studied the RQ as a function of cardinality in the training data. With decreasing cardinality we observed a decrease in RQ value We selected the cardinality that corresponds to the elbow of the RQ curve, which indicates loss of informative channels. The optimum cardinality was found to be 2 in the training data.

The summary of the results for two-class ECoG data is provided in Table I. All sparse CSP based methods outperformed the standard CSP where the classification errors were 11%, 10% and 12% for the RWE, BE and FS, respectively. The total numbers of channels used by the sCSP methods are 42, 21,

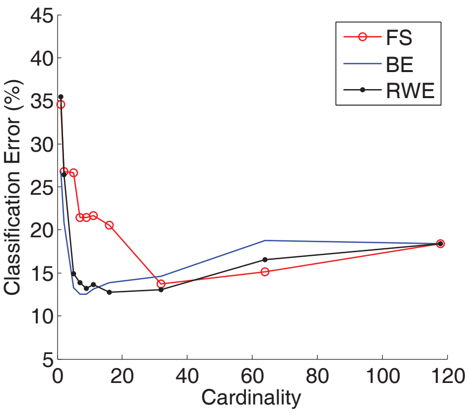
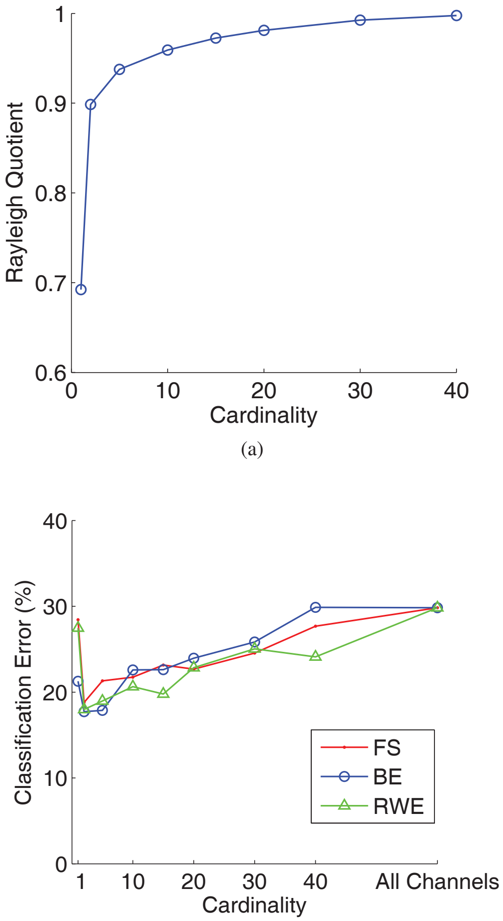


Fig. 1. Classification Error Comparison for EEG dataset.



(a)

(b)

Fig. 2. (a) The RQ values for the ECoG finger movement dataset of the RWE method. (b) The Classification Error rates (%) for FS, BE and RWE.

and 18. Not only the number of channels is reduced but also sCSP methods resulted in a better generalization performance.

The test errors curves averaged over five subjects for the two-class EEG data are provided in Fig. 1. The results obtained by the RWE method is similar to those of the BE method and they are consistently better than the FS method. They all outperform the traditional CSP, last point on the x-axis. When averaged over five subjects the total numbers of channels used by the sCSP methods (on the test data) are 26.2, 22.2, and 61.2 for RWE, BE, and FS methods, respectively. It is clear that the sCSP methods are using considerably lower numbers of channels than available 118 channels. The RWE method performs similarly to BE method in terms of classification error and cardinality.

Finally, we provided the RQ curve as a function of cardinality and the classification error curve of the multiclass ECoG data in Fig. 2. We observed that the RQ curve had an elbow at two channels on the training data. This cardinality level also provided the best classification error for all sparse methods. The RWE method resulted to a 17.9% error rate on the test data. We observed that the standard CSP method achieved a dramatically higher error rate of 29.8%.

C. Computational Complexity

As explained in Section III-D, the proposed RWE method should have considerably lower computational complexity than the BE method. Fig. 3 shows the elapsed time to calculate one filter on simulated data with sparseness level of 2 using either BE or RWE methods with different size of number of channels on the x-axis. The experiment was performed on a desktop computer with 4GB of RAM and equipped with a CPU running at 2.66 Ghz. The computational complexity of the RWE is dramatically lower than the BE method. For instance from 128 channels the RWE method computes a sparse filter with cardinality 2 in less than one second. The same process takes around one and a half minute for the BE method. In case of using cross validation on the training data as a parameter or sparseness estimator, the experiments exploiting the BE method can take several hours. This large execution time makes BE method unfeasible to be used in BMI applications. On the other hand, RWE can be executed in a few minutes. The RWE method performs comparable to the BE method, the RWE method can be used easily where the number of channels is large.

V. Conclusion

The CSP method is widely used for feature extraction in BMI applications with large number of recording channels. However, use of all available channels during spatial filtering results in several problems such as overfitting and lack of robustness over sessions. To overcome the drawbacks of the traditional CSP method, we proposed a sparse CSP method based on recursive weight elimination. The RWE based CSP is an extension of previously proposed greedy search methods to find sparse CSP filters which are associated with higher complexity. In this study we have shown that the proposed RWE based CSP method selects a few numbers of channels for feature extraction. Compared to standard CSP, estimated spatial filters provide better classification performance in a binary and multi class BMI setup. The RWE method performed

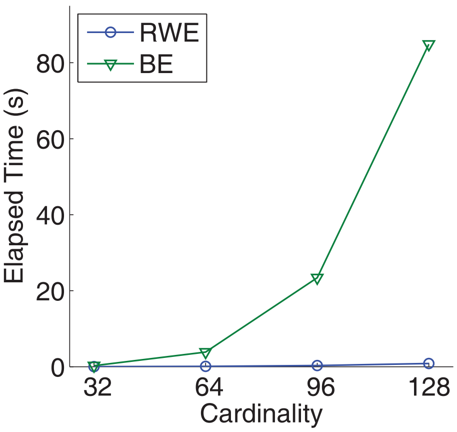


Fig. 3. Computational complexity of BE and RWE methods.

similarly to the BE method with a dramatic decrease in computational complexity. As a result the proposed RWE method is applicable to other domains to select sparse solutions where the problem of interest involves solving a GED and the size of the covariance matrices is large. In particular, it is possible to extend this this method to select taps of the common spatio-spectral pattern algorithm of [15] in which the neural data is filtered in space and frequency simultaneously.

References

[1] J. R. Wolpaw and D. J. McFarland, "Multichannel EEG-based brain-computer communication," *Electroencephalography and Clinical Neurophysiology*, vol. 90, no. 6, pp. 444 – 449, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/001346949490135X

[2] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial Patterns Underlying Population Differences in the Background EEG," *Brain Topography*, vol. 2, pp. 275–284, 1990.

[3] C. Guger, H. Ramoser, and G. Pfurtscheller, "Real-time eeg analysis with subject-specific spatial patterns for a brain-computer interface (bci)," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, no. 4, pp. 447 –456, dec 2000.

[4] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing Spatial filters for Robust EEG Single-Trial Analysis," *Signal Processing Magazine, IEEE*, vol. 25, no. 1, pp. 41 –56, 2008.

[5] J. Farquhar, N. J. Hill, T. N. Lal, and B. Schlkopf, "Regularised CSP for sensor selection in BCI," in *In Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course*, 2006.

[6] B. Reuderink and M. Poel, "Robustness of the Common Spatial Patterns algorithm in the BCI-pipeline," Univ. of Twente, Enschede, July 2008.

[7] X. Yong, R. Ward, and G. Birch, "Sparse spatial filter optimization for EEG channel reduction in brain-computer interface," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, Apr. 2008, pp. 417 –420.

[8] F. Goksu, N. Ince, and A. Tewfik, "Sparse common spatial patterns in brain computer interface applications," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, may 2011, pp. 533 –536.

[9] J. Demmel, *Applied Numerical Linear Algebra*. SIAM, 1997.

[10] B. Moghaddam, Y. Weiss, and S. Avidan, "Generalized spectral bounds for sparse lda," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 641–648. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143925

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002. [Online]. Available: http://citeseer.ist.psu.edu/guyon02gene.html

[12] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 10, no. 6, pp. 978 –983, nov 1988.

[13] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Bci Competion III, Dataset IVa," 2005. [Online]. Available: http://www.bbci.de/competition/iii/desc_IVa.html

[14] K. J. Miller and G. Schalk, "Prediction of Finger Flexion 4th Brain-Computer Interface Data Competition," 2008. [Online]. Available: http://www.bbci.de/competition/iv/desc_4.pdf

[15] 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*, vol. 52, no. 9, pp. 1541 –1548, sept. 2005.