A NOVEL OBJECTIVE FUNCTION MINIMIZATION FOR SPARSE SPATIAL FILTERS

Ibrahim Onaran, N. Firat Ince
Biomedical Engineering Department
University of Houston
{ionaran, nfince}@uh.edu

A. Enis Cetin
Department of Electronics and Electrical Engineering
Bilkent University
cetin@bilkent.edu.tr

***Abstract*—Common spatial pattern (CSP) method is widely used in brain machine interface (BMI) applications to extract features from the multichannel neural activity through a set of spatial projections. The CSP method easily overfits the data when the number of training trials is not sufficiently large and it is sensitive to daily variation of multichannel electrode placement, which limits its applicability for everyday use in BMI systems. To overcome these problems, the amount of channels that is used in projections, should be limited. We introduce a spatially sparse projection (SSP) method that exploits the unconstrained minimization of a new objective function with approximated ℓ_1 penalty. The SSP method is employed to classify the two class EEG data set. Our method outperforms the standard CSP method and provides comparable results to ℓ_0 norm based solution and it is associated with less computational complexity.**

***Keywords*—Brain Machine Interfaces, Common Spatial Patterns, Sparse Spatial Projections, Rayleigh Quotient, Unconstrained Optimization**

I. INTRODUCTION

The BMI technology aims to help disabled people to establish communication with their environment solely by their brain signals. With the recent advances in electrode design and recording technology, the number of recording channels used in BMI applications is increasing to capture signals from a larger area of the brain or to get more information from smaller regions using dense electrode grids. Therefore, a dimension reduction algorithm needs to be employed to decrease the correlation between channels and improve the signal to noise ratio (SNR). In this scheme, the CSP algorithm is widely used due to its simplicity and lower computational complexity to extract features from high-density recordings both using noninvasive and invasive modalities [1], [2].

Despite the benefits of the CSP method, it also has a number of drawbacks. One major problem of the CSP is that it generally overfits the data when it is recorded from a large number of electrodes and when there is limited number of train trials. Moreover, the chance that CSP uses a noisy or corrupted channel is linearly increased with increasing number of recording channels. Robustness over time is also a major drawback in CSP applications [3], [4]. Since all channels are used in spatial projections of CSP, the classification accuracy may reduce in case the electrode locations slightly change in different sessions. This requires almost identical electrode positions over time, which is difficult to realize [5]. The sparseness of the spatial filter might have an important role to increase the robustness and generalization capacity of the BMI system.

The CSP method minimizes the Rayleigh Quotient (RQ) of the spatial covariance matrices to achieve the variance imbalance between the classes of interest. The RQ is defined as

$$R(w) = \frac{w^T A w}{w^T B w} \tag{1}$$

where A and B are the spatial covariance matrices of two different classes and w is the spatial filter that we want to find. One way to reduce the number of channels used in the projection w, is to transform the CSP algorithm into a regularized optimization problem in the form of

$$L(w) = R(w) + \lambda \|w\| \tag{2}$$

where $R(w)$ is the objective function, $\|w\|$ is the ℓ_1 norm based penalty and λ is a constant that controls the sparsity of the solution.

A number of studies investigated putting the CSP into alternative optimization forms to obtain a sparse solution for it. Recently, in [6] quasi ℓ_0 norm based criterion was used for obtaining the sparse solution which resulted an improved classification accuracy. Since ℓ_0 norm is non-convex, combinatorial and NP-hard, they implemented greedy solutions such as Forward Selection (FS) and Backward Elimination (BE) to decrease the computational complexity. It has been shown that BE was better than FS in terms of classification error but associated with very high complexity making it difficult to use in rapid prototyping scenarios. In [7] the authors converted CSP into a quadratically constrained quadratic optimization problem with ℓ_1 penalty; others used an ℓ_1/ℓ_2 [3], [8] norm based solution. These studies have reported a slight decrease or no change in the classification accuracy while decreasing the number of channels significantly.

In this paper, we construct a computationally efficient spatially sparse projection (SSP) based on a novel objective function with similar characteristics to RQ. This new objective function can be minimized in the form of (2) to address the drawbacks of regular CSP method. We show that our new objective function has the same minimization solution as RQ and it depends on the magnitude of the spatial filter. The magnitude dependency of our new objective function allows us to use a continuous and differentiable function approximating ℓ_1 norm [9] as regularization term in an unconstrained

978-1-4673-5563-6/13/$31.00

optimization framework and can be solved using standard algorithms with low complexity. The rest of the paper is organized as follows. In the following section, we describe our novel objective function and its relation to RQ. Then we explain its use in an unconstrained optimization problem. Next, we apply our method on the BCI competition III EEG dataset IVa [10] involving imaginary foot and hand movements. We also compare our method to standard CSP and the ℓ_0 norm based BE solution given in [6]. Finally, we discuss our results and provide future directions.

II. Material and methods

A. Standard CSP and a New Objective Function

In the CSP framework, the spatial filters are a weighted linear combination of recording channels, which are tuned to produce spatial projections maximizing the variance of one class and minimizing the other. The spatial projection is computed using

$$X_{CSP} = W^T X \tag{3}$$

where the columns of W are the vectors representing each spatial projection and X is the multichannel EEG data.

Maximizing the RQ (1) is identical to the following optimization problem.

$$\begin{aligned} \underset{w}{\text{maximize}} \quad & w^T A w \\ \text{subject to} \quad & w^T B w = 1. \end{aligned} \tag{4}$$

After writing this optimization problem in the Lagrange form and taking the derivative with respect to w, we obtain the identical problem in the form of $Aw = \mu Bw$ which is the Generalized Eigenvalue Decomposition (GED). The solutions of this equation are the joint eigenvectors of A and B and μ is the associated eigenvalue of a particular eigenvector.

We assume that the discriminatory information is embedded in a few channels where the number of these channels is much smaller than the actual number of all recording channels. So the discrimination can be obtained with a sparse spatial projection, which uses only informative channels. In this scheme assume that the data was recorded from K channels. We are interested in obtaining a sparse spatial projection using an unconstrained minimization problem in the form of (2), where w has only k nonzero entries, $card(w) = k$ and $k \ll K$.

Since $R(w)$ does not depend on the gain of w, the optimizer arbitrarily reduces the gain of w to minimize regularization term $\lambda\|w\|$ after finding the direction that minimizes $R(w)$. Thus, the solution of the optimization problem that uses $R(w)$ as an objective function is essentially the same as the GED solution.

To find a sparse solution we need to have an objective function that depends on the gain of w. In this scheme, we replaced $R(w)$ with the following objective function.

$$G(w) = w^T A w + \frac{1}{w^T B w} \tag{5}$$

This function is bounded from below and has interesting properties. Let us define $a = w^T A w$ and $b = w^T B w$. If we define RQ in terms of a and b such that $R = a/b$ then our new objective function can be expressed as

$$G(w) = a + \frac{1}{b} = \frac{ab}{b} + \frac{1}{b} = Rb + \frac{1}{b} \tag{6}$$

The derivative of $G(w)$ with respect to R is equal to b which is always positive. This indicates that our objective function $G(w)$ decreases with a decrease in R value. After taking the derivative of $G(w)$ with respect to b and finding the roots of the derivative, we note that b is equal to $\sqrt{R^{-1}}$. By inserting b value into the Equation 6 we obtain the minimum value of $G(w)$ as $2\sqrt{R}$. This result shows that the direction that minimizes R also minimizes $G(w)$.

We plug $G(w)$ into unconstrained optimization formulation in (2) as the objective function. Rather than working to solve (2) with a non-differentiable ℓ_1 penalty, we replaced it with a twice differentiable smooth version of ℓ_1 (epsL1) which is sufficiently close to minimizing ℓ_1 [9]. The main advantage of this approach is that, since epsL1 and $G(w)$ are both twice differentiable, we can directly apply an unconstrained optimization method to minimize $L(w)$ [11].

The solution w that minimizes the function $L(w)$ tends to become sparse as λ gets bigger. The entries of w generally were not exactly equal to zero, so we normalized w to its maximum absolute value and eliminated the weights consequently corresponding channels that do not exceed a predefined threshold (=10^{-2}). We used "fminunc" function of Matlab to find the solution of our unconstrained minimization problem. We computed the desired cardinality by implementing a bisection search [12] on the λ. The upper border of λ was determined initially using the $G(w_c)/\|w_c\|$ ratio where w_c is the full CSP solution. In case the initial upper border results a cardinality larger than the desired value, we kept doubling the λ parameter until we obtained a λ that results a cardinality which is less than or equal to the target value.

Following the above procedure, we computed the first spatial filter w that minimizes the $G(w)$ which also minimizes the $R(w)$. The solution that maximizes $R(w)$ is also a useful spatial filter. Therefore, we interchanged the matrix A and B to find a solution that maximizes $R(w)$. In order to find multiple sparse filters we deflated the covariance matrices with sparse vectors using the Schur complement deflation method described in [13].

B. EEG Dataset

We applied the SSP method on two class EEG of the BCI competition III dataset IVa [10]. The dataset is recorded from five subjects (aa, al, av, aw, ay) who were asked to imagine either right foot or right index finger movements. The sampling rate of the data was 1 kHz and data was recorded from 118 channels. The EEG signal was filtered in the range of 8-30 Hz. There were 140 trials available for each class. Once again, one second data following the cue was used in the analysis.

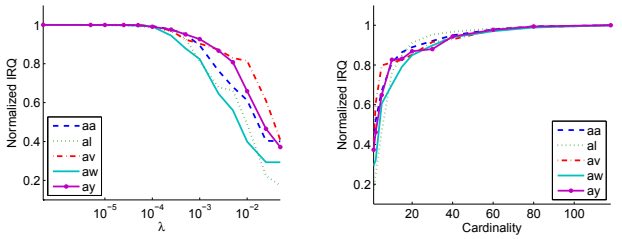
The signal was transformed into four spatial filters by taking first and last two eigenvectors for each CSP methods. After computing the spatial filter outputs, we calculated the energy of the signal and converted it to log scale for each sparse filter and we used them as input features to lib-SVM classifier with an RBF kernel [14].

We compared the SSP to the standard CSP and to the ℓ_0 norm based BE method of [6] as it provided superior results in terms of classification accuracy and reduced cardinality. We studied the classification accuracy as a function of cardinality. With the purpose of finding optimum *sparsity* level for the classification, we computed several sparse solutions, with decreasing number of cardinality on the training data. The sparse CSP methods were employed with $k \in \{80, 60, 40, 30, 20, 15, 10, 5, 2, 1\}$ levels. For each level we computed the corresponding RQ value. We studied the inverse of the RQ (IRQ) curve and determined the optimal cardinality where its value suddenly dropped indicating we started to lose informative channels.

The dataset contains 140 trials per class and subject. We used 70 trials in training to estimate the sparse filters, and 70 trials for testing. The value of the ϵ in epsL1 regularization term was chosen to be 10^{-6}.

III. Results

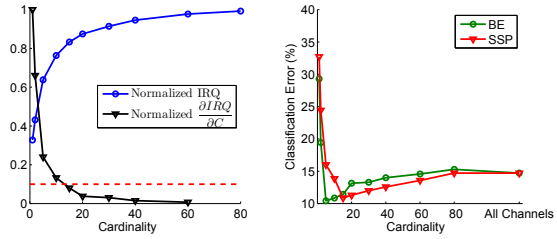
We observed that for the SSP method, any particular λ value can lead to different cardinality and normalized IRQ values for different subjects as shown in Fig. 1. In particular, this inter subject variability of IRQ did not allow us to use the same λ value for all subjects (See Fig. 1a). However, the variability of IRQ values of different subjects was lower when we fixed the cardinality as shown in Fig. 1b. Consequently, due to this reduced variability and to compare our method to the BE technique, we studied the classification error as a function of cardinality. In order to decide on the optimal cardinality level to be used on the test data, the IRQ values were computed on the training data, scaled to their maximum value and averaged over subjects. In the following step, we computed the slope of the IRQ curve and normalized it to its maximum value to get an idea about the relative change in the IRQ.



(a) (b)

Figure 1: Normalized IRQ values are shown in (a). The Normalized IRQ values vs cardinality for each subject is shown in (b).

We depicted the change in IRQ values for each cardinality as shown in Fig. 2a. As expected, decreasing the cardinality of the spatial projection resulted to a decrease in the IRQ value. To determine the optimum cardinality to be used in classification on the test data, we selected the cardinality that is below 10 % of the maximum relative change (See the dashed lines in Fig. 2a). The cardinality value was found to be 15 for the SSP method. For the BE method this value was 10. These indices perfectly corresponded to the elbow of the IRQ curve, which indicates loss of informative channels. In Table I, we provide the classification results and selected cardinalities using different methods including SSP, CSP and ℓ_0 based greedy solution, BE. In order to give a flavor about the change in error rate versus the cardinality, we provided the related classification error curves in Fig 2b. Although the minimum classification error was obtained at cardinality 5 for the BE method, we noticed that we identified the optimum cardinality as 10 on the training data.



(a) (b)

Figure 2: The average IRQ of all subjects versus cardinality (a). The red line is the 10 percent threshold that determines the optimum cardinality to be used in the test data. The optimum cardinality level is 15. The classification error curves of SSP and BE methods versus the cardinality are given in (b). The last data point corresponds to the results obtained from standard CSP which uses all channels.

On all subjects we studied, we observed that the SSP method consistently outperformed the CSP method. As expected the full CSP solution did not perform as good as the other sparse methods and likely overfitted the training data. We obtained comparable results on EEG data using the SSP and BE methods (p-value $= 0.5$, paired t-test). The error difference between regular CSP and SSP is 3.8%. We studied the effect of the amount of training data on the classification accuracy and presented the results in Fig. 4a. When a small number of training trials, as low as 15 are used in the EEG dataset, the difference between the sparse and standard CSP technique was more than 6%. Interestingly, with increasing number of training trials the SSP method consistently provided better results and the difference remained between 3-4%. There was no noticeable difference between SSP and BE.

Fig. 3 illustrates the distribution of the spatial filters obtained using SSP and CSP algorithms for all subjects. We observed that the SSP filter coefficients are localized on the left hemisphere and the central area, which is in accordance with the cortical regions related to right hand and the foot movement generation.

Arvaneh *et al.* [8] used the ℓ_1/ℓ_2 ratio as a penalty term and they applied their algorithm to the BCI competition III EEG dataset IVa [10] which we used in this paper as well. They achieved a mean error rate of $17.7 \pm 15.4\%$ using

Table I: Classification error rates (%) for each subject

	Cardinality	aa	al	av	aw	ay	Avg
BE	10	13.6	2.9	30.7	2.1	5.0	10.9
SSP	15	19.3	1.4	23.6	4.3	5.7	10.9
CSP	118	23.6	3.6	32.1	2.9	11.4	14.7

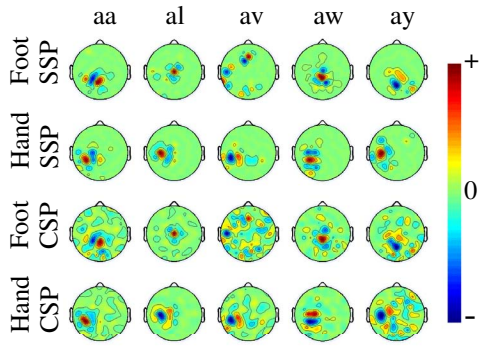
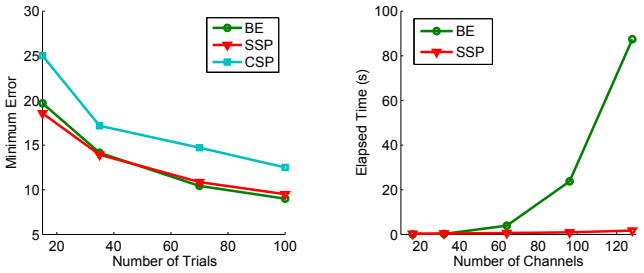


Figure 3: The CSP and SSP filters for hand and foot movement imagination.

22.6 ± 11 channels. Here, we compared our method with the study of Arvaneh *et al.* by extracting one filter from each end of the sparse solutions. The SSP method achieved a mean error rate of $12 \pm 11.3\%$ with an average number of channels 25.6 ± 2.3. The obtained results indicated that the SSP method provided a significant improvement (p-value$= 0.024$, paired t-test) over the ℓ_1/ℓ_2 based algorithm on the classification accuracy without any significant difference between number of channels used (p-value=0.28).



(a) (b)

Figure 4: (a) The minimum error vs. the number of trials. (b) The average elapsed time to estimate a spatial filter with a cardinality of two vs. the number of total recording channels.

In order to compare the computational complexity of SSP method to the BE, we computed sparse filters with a cardinality of two from an increasing number of recording channels on simulated data. The training was performed on a regular desktop computer with 4 GB of RAM and equipped with a CPU running at 2.66 GHz. The elapsed time per filter computation increased exponentially for the BE method and linearly for the SSP method as shown in Fig. 4b. With 128 channels, the BE algorithm computed a single spatial filter with two nonzero entries in 90 seconds. For the SSP method with the same setup above, the elapsed time was less than a second. Although, we used the relative change in the IRQ to identify the optimum sparsity level, one can also run a typical k-fold cross validation procedure to identify the optimum level. However, in such a case training the system with BE method will take several hours which may not be feasible for BMI applications. On the other hand with the SSP method training through cross validation can be executed in a few minutes.

IV. Conclusion

The need for the sparse filters is apparent when there is large number of recording electrodes and insufficient amount of training data. To minimize overfitting on the training data and eliminate noisy channels, we introduced a spatially sparse projection technique (SSP) based on a novel objective function. Unlike the RQ, this new objective function has a dependency on the filter magnitude. By using an approximated ℓ_1 norm, we computed the sparse spatial filters through an unconstrained minimization formulation with standard optimization algorithm. We applied our method to EEG dataset and compared its efficiency to standard CSP, and to a ℓ_0 norm based greedy technique. The SSP method outperformed the standard CSP and provided comparable results to ℓ_0 norm based method, which is associated with higher computational complexity. The SSP method provided 26% decrease in the error rate. The SSP algorithm was able to reach a minimum error rate with only 15 channels. Our results indicate that SSP method can be effectively used to extract features from EEG dataset with large number of recording channels.

References

[1] 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.

[2] N. F. Ince, R. Gupta, S. Arica, A. H. Tewfik, J. Ashe, and G. Pellizzer, "High Accuracy Decoding of Movement Target Direction in Non-Human Primates Based on Common Spatial Patterns of Local Field Potentials," *PLoS ONE*, vol. 5, no. 12, p. e14384, 12 2010.

[3] 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.

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

[5] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, Dec 2000.

[6] 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.

[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] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the Channel Selection and Classification Accuracy in EEG-based BCI," *Biomedical Engineering, IEEE Transactions on*, vol. 58, no. 6, pp. 1865 –1873, june 2011.

[9] S.-i. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient L1 Regularized Logistic Regression," in *In AAAI*, 2006.

[10] 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

[11] M. Schmidt, G. Fung, and R. Rosales, "Fast Optimization Methods for L1 regularization: A Comparative Study and Two New Approaches," 2009.

[12] R. Burden and J. Faires, *Numerical Analysis*, 8th ed. Thomson Brooks/Cole, 2005.

[13] L. Mackey, "Deflation Methods for Sparse pca," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2009, pp. 1017–1024.

[14] C.-C. Chang and C.-J. Lin, "A library for Support Vector Machines." 2001. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libs