BASELINE REGULARIZED SPARSE SPATIAL FILTERS

Ibrahim Onaran[1,2] *N. Firat Ince*[1,3] *A. Enis Cetin*[2]

[1] Department of Neurosurgery, University of Minnesota, Minneapolis, MN 55455 USA
[2] Department of Electrical Engineering, Bilkent University, Ankara, Turkey
[3] Department of Biomedical Engineering, University of Houston, Houston, TX 77204 USA

ABSTRACT

The common spatial pattern (CSP) method has large number of applications in brain machine interfaces (BMI) to extract features from the multichannel neural activity through a set of linear spatial projections. These spatial projections minimize the Rayleigh quotient (RQ) as the objective function, which is the variance ratio of the classes. 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 to some adequate number. We introduce a spatially sparse projection (SSP) method that renders unconstrained minimization possible via a new objective function with an approximated ℓ_1 penalty. We apply our new algorithm with a baseline regularization to the ECoG data involving finger movements to gain stability with respect to the number of sparse channels.

Index Terms— Baseline regularization, Brain machine interfaces, Common spatial patterns, Sparse spatial projections, Unconstrained optimization

1. INTRODUCTION

The aim of the BMI technology is to help disabled people by establishing a communication channel with their environment using only their brain signals. The recent advances in electrode design technology allow BMI applications to use increasing number of electrodes. In this scheme, the common spatial pattern (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].

The benefits of the CSP method come with some drawbacks. One major drawback of the CSP is that it generally overfits the data when it is recorded from a large number of electrodes and there is limited number of train trials. Furthermore, the chance that CSP uses a noisy or corrupted channel linearly increases with increasing number of recording channels. Another major problem is the robustness over time in CSP applications [3, 4]. Using all channels in spatial projections of CSP may reduce the classification accuracy in case the electrode locations slightly change in different sessions. In this case, CSP method 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 increases or decreases the variance ratio of two classes. The variance ratio of two classes can be represented in terms of Rayleigh Quotient of the spatial covariance matrices. 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. The solution of the CSP is the generalized eigenvalue decomposition of matrices A and B. This problem can also be solved in an unconstrained 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. Since RQ does not depend on the magnitude of the filter w, we observed that the solution to this optimization problem is essentially scaled version of the generalized eigenvalue decomposition (GED) solution and does not depend on λ. Therefore, we introduced a novel objective function which has dependency on its magnitude and rise the same solution as GED when λ is equal to zero [6].

A number of studies investigated putting the CSP into alternative optimization forms to obtain a sparse solution for it. 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 classifi-

This research was supported in part by The National Science Foundation, award CBET-1067488, by a grant from the University of Minnesota Interdisciplinary Informatics (UMII) and by a grant from The Science and Technological Research Council of Turkey (TUBITAK), project no: 111E057.

cation accuracy while decreasing the number of channels significantly. Recently, quasi ℓ_0 norm based methods 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), backward elimination (BE) [9] and recursive weight elimination (RWE) [10] to decrease the computational complexity. It has been shown that BE was better than RWE and FS (less myopic) in terms of classification error and sparseness level but associated with very high complexity making it difficult to use in rapid prototyping scenarios.

Selecting the sparsity level that produces high accuracy is crucial for the sparse spatial filters. We observed that the small variations in sparsity may lead to large change in the classification accuracy [6]. So a more representative sparse spatial filters needs to be constructed to eliminate large deviations on the classification accuracy.

In this paper, we develop a baseline regularization algorithm to improve the classification accuracy and eliminate instability over the sparsity levels. The baseline regularization make the sparse spatial patterns to represent the fingers, instead of discriminating them from each other. The SSP which is computationally efficient sparse spatial projection based on a novel objective function and RWE are used to demonstrate the efficiency of the baseline regularization. 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 IV ECoG dataset involving individuated movements of five fingers [11]. We also compare our method to standard CSP. Finally, we investigate the contribution of the baseline regularization to the classification accuracies by constructing a mixed generative/discriminative sparse filters.

2. MATERIAL AND METHODS

The CSP filters are weighted linear combination of recording channels, which are specialized 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 ECoG data.

2.1. Sparse Spatial Filter

We sparsify the spatial filters to overcome the drawbacks of the CSP method that are described earlier and to increase the classification accuracy and the generalization capability of the method. We assume that a few channel of the data has the discriminatory information and the number of these channels is much smaller than the actual number of all recording 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.

$$G(w) = w^T A w + \frac{1}{w^T B w} \tag{4}$$

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

$$G(w) = a + \frac{1}{b} = \frac{ab}{b} + \frac{1}{b} = Rb + \frac{1}{b} \tag{5}$$

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 solving Equation 6,

$$\frac{\partial G(w)}{\partial b} \quad = \quad R - \frac{1}{b^2} = 0 \tag{6}$$

we note that b is equal to $\sqrt{R^{-1}}$. By inserting b value into the Equation 5 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 put $G(w)$ into unconstrained optimization formulation in (2) as the objective function. We placed a twice differentiable smooth version of ℓ_1 (epsL1) which is sufficiently close to minimizing ℓ_1 [12] as a regularization parameter. 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 $L(w)$ [13]. The epsL1 is defined as

$$\|w\| = \sum_{i=1}^{K} \sqrt{w_i^2 + \epsilon} \tag{7}$$

where ϵ is a sufficiently small parameter and K is the dimension of w. The epsL1 approximates the ℓ_1 norm and they are identical when ϵ is equal to zero. Twice differentiability of the epsL1 norm allows us to use it when w_i is equal to zero unlike the regular ℓ_1 norm which is not differentiable at zero.

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 computed the desired cardinality which is the number of channels to be selected for the spatial projection by implementing a bisection search [14] 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.

2.2. Recursive Weight Elimination

Recursive weight elimination (RWE) is an ℓ_0 norm based greedy search algorithm to obtain sparse filters in very a efficient and effective way [10]. The algorithm starts with a full size covariance matrices of the traditional CSP method. Assume that the size of these covariance matrices is $K \times K$. In the very first step, RWE solves general CSP problem and finds the weight vector w. The contribution of the smallest magnitude coefficient can be ignored compared to the other coefficients, since we have a high number of channels. Assume that the index of this small coefficient is k. We can remove this coefficient by removing k^{th} row and column of the full size covariance matrices and solving the CSP on these new $K-1 \times K-1$ matrices. We can decrease the number of channels to the desired cardinality level by recursively applying this algorithm to the smaller matrices. Each cardinality reduction involves solving a traditional CSP, therefore this method is faster than other ℓ_0 norm based greedy search algorithms such as BE or FS [9].

2.3. Baseline Regularized Sparse Spatial Filters

The data set consists of finger movement and baseline regions. We used the baseline data to regularize the finger to finger contrast. In other words, each multichannel finger data is contrasted with a mixture of baseline and another finger. Let's assume A is the spatial covariance matrix of the first finger, and C is spatial covariance matrix of one of the other four fingers and D is the covariance matrix of the baseline, we find a solution to the following optimization problem,

$$L(w) = w^T(\alpha C + (1-\alpha)D)w + \frac{1}{w^T A w} + \lambda\|w\| \quad (8)$$

where α is the mixing coefficient ranging from 0 to 1. We contrast a finger to another finger when α is 1 to obtain discriminative spatial filters. On the other hand when α is equal to zero, we contrast each finger with baseline which yields representative spatial filters. Therefore, α determines level discrimination or representation characteristic of the constructed sparse filter. Similarly, we also apply this approach to RWE method.

In this scheme, we computed the first spatial filter w that minimizes the $L(w)$ to obtain the sparse filter that maximize the variance of the first finger. Then we interchanged the matrices A and C to find the spatial filter that maximizes the variance of the other finger. In order to find multiple sparse filters we deflated the covariance matrices with these initial sparse vectors using the Schur complement deflation method described in [15]. Using this new deflated matrices, we find the second set of spatial filters and obtain a total of 4 spatial filters.

2.4. ECoG Dataset

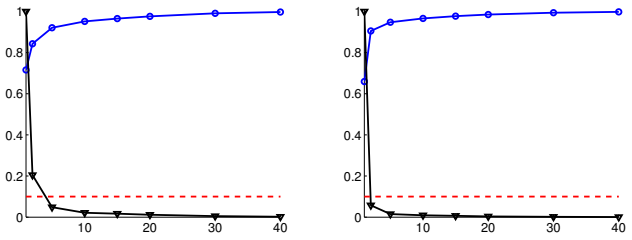
The ECoG data was recorded from three subjects during finger flexions and extensions [11] with a sampling rate of 1 kHz. The electrode grid was placed on the surface of the brain. Each electrode array contained 48 (8x6) or 64 (8x8) platinum electrodes. The finger index to be moved was shown with a cue 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 (65-200 Hz) as in [16]. We used one second data following the movement onset and 500 ms data before the movement onset in the analysis. The dataset contains around 146 trials for each subject.

The multichannel signal was transformed into four channel signal using the spatial filters are derived using 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 [17].

Since we are tackling a multiclass problem for the ECoG dataset, we used the pairwise discrimination strategy of [2] to apply the CSP to the five-class finger movement data. In other words, we constructed sparse spatial filters tuned to contrast pairs of finger movements such as 1 vs. 2; 1 vs. 3; 2 vs. 4 etc.

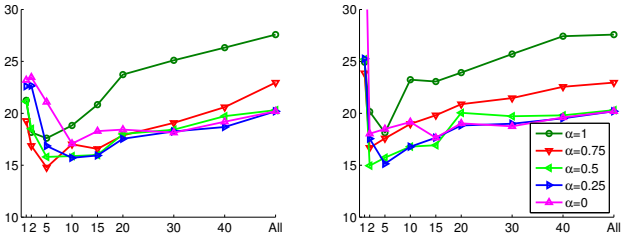
We studied the classification accuracy as a function of cardinality and the mixing parameter α. On the training data with the purpose of finding optimum *sparseness* level for the classification, we computed several sparse solutions, with decreasing cardinality. The sparse CSP methods were employed with $k \in \{40, 30, 20, 15, 10, 5, 2, 1\}$. For each cardinality, 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.

Two times two fold cross validation were run on the entire data set and the results were averaged over the folds and iterations. In average, we used 15 ± 2 train trials per finger. The value of the ϵ in epsL1 regularization term was chosen to be 10^{-6}. We used $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ for the baseline regularization experiments.



(a) (b)

Fig. 1. The average IRQ of all subjects versus cardinality for SSP method (a) and RWE method(b) for the α values 0.75 and 0.5 respectively. The red line is the 10 percent threshold that determines the optimum cardinality to be used in the test data. The optimum cardinality levels are five and two respectively. The line with circle markers is IRQ curve and the line with triangle markers is derivative of the IRQ curve.



(a) (b)

Fig. 2. The classification error curve versus the cardinality for SSP method (a) and RWE method(b). The last data point corresponds to the results obtained from standard CSP which uses all channels.

3. RESULTS

We depicted the change in IRQ values for each cardinality as shown in Fig. 1a and 1b. 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. 1. The cardinality value was found to be 5 for SSP method. For the RWE method the cardinality value was 2. These indices perfectly corresponded to the elbow of the IRQ curve, which indicates loss of informative channels. In Table 1, we provide the classification results and selected cardinalities using SSP, CSP and ℓ_0 based greedy solution, RWE with a mixing parameter α that provides minimum accuracy error. In order to give a flavor about the change in error rate versus the cardinality, we provided the related classification error curves in Fig. 2.

On all subjects we studied, we observed that the SSP method consistently outperformed the CSP method. We noted that the minimum error rate was obtained with SSP method. SSP and RWE methods used cardinality of 5 and 2 to achieve the minimum error rate respectively. As expected the full CSP solution did not perform as good as the other sparse methods and likely overfitted the training data.

We also note that the baseline regularization removes overfitting of the classifier and provides robustness to the sparsity level. In Fig. 2 it is shown that the increase in cardinality did not affect the regularized ($\alpha \neq 1$) sparse filters as much as unregularized ($\alpha = 1$) sparse filters. On the other, hand pure generative ($\alpha = 0$) sparse filters accuracy error tends to increase with decreasing cardinality below the cardinality level 10.

Table 1. Classification error rates (%) for each subject using SVM classifier

	Cardinality	α	Subject 1	Subject 2	Subject 3	Avg
RWE	2	0.5	17.7	14.3	12.9	14.95
SSP	5	0.75	19.6	12	12.8	14.79
CSP	All	0.25	25.7	19.6	15.3	20.19

4. CONCLUSION

In general the dimensionality of the BMI data is larger than the number of training data. This imbalance between the amount of training data and the number of channels results overfitting on the training data. To minimize overfitting and eliminate noisy channels, we introduced a spatially sparse projection technique (SSP) based on a novel objective function. 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 ECoG dataset and compared its classification capacity to standard CSP and to an ℓ_0 norm based greedy technique. The sparse methods outperformed the standard CSP method. We observed that the sparse methods are sensitive to the cardinality, therefore we regularized the sparse spatial filters using the baseline data. We study the effect of regularization on classification accuracy by implementing a baseline/movement mixing method. Our results indicate that baseline regularization improves the classification accuracies as well as it provides stability with respect to the cardinality level.

5. 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] Nuri F. Ince, Rahul Gupta, Sami Arica, Ahmed H. Tewfik, James Ashe, and Giuseppe 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, pp. 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," 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] Ibrahim Onaran, N. Firat Ince, and A. Enis Cetin, "Sparse spatial filter via a novel objective function minimization with smooth 1 regularization," *Biomedical Signal Processing and Control*, , no. 0, pp. –, 2012.

[7] Xinyi Yong, R.K. Ward, and G.E. 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, Cuntai Guan, Kai Keng Ang, and Chai 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] F. Goksu, N.F. Ince, and A.H. 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.

[10] Fikri Goksu, Firat Ince, and Ibrahim Onaran, "Sparse common spatial patterns with recursive weight elimination," in *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, November 2011, pp. 117 –121.

[11] Kai J. Miller and G. Schalk, "Prediction of Finger Flexion 4th Brain-Computer Interface Data Competition," 2008.

[12] Su-in Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng, "Efficient L1 Regularized Logistic Regression," in *In AAAI*, 2006.

[13] Mark Schmidt, Glenn Fung, and Rmer Rosales, "Fast Optimization Methods for L1 regularization: A Comparative Study and Two New Approaches," 2009.

[14] R.L. Burden and J.D. Faires, *Numerical Analysis*, Thomson Brooks/Cole, 8 edition, 2005.

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

[16] Ibrahim Onaran, N. Firat Ince, and A. Enis Cetin, "Classification of Multichannel ECoG Related to Individual Finger Movements with Redundant Spatial Projections," in *International IEEE EMBS Conference*, August 2011.

[17] Chih-Chung Chang and Chih-Jen Lin, "A library for Support Vector Machines.," 2001.