A Sparse Common Spatial Pattern Algorithm for Brain-Computer Interface

Li-Chen Shi[1], Yang Li[1], Rui-Hua Sun[1], and Bao-Liang Lu[1,2,⋆]

[1]Center for Brain-Like Computing and Machine Intelligence
Department of Computer Science and Engineering
[2]MOE-Microsoft Key Lab. for Intelligent Computing and Intelligent Systems
Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai 200240, China
bllu@sjtu.edu.cn

Abstract. Common spatial pattern (CSP) algorithm and principal component analysis (PCA) are two commonly used key techniques for EEG component selection and EEG feature extraction for EEG-based brain-computer interfaces (BCIs). However, both the ordinary CSP and PCA algorithms face a loading problem, i.e., their weights in linear combinations are non-zero. This problem makes a BCI system easy to be over-fitted during training process, because not all of the information from EEG data are relevant to the given tasks. To deal with the loading problem, this paper proposes a spare CSP algorithm and introduces a sparse PCA algorithm to BCIs. The performance of BCIs using the proposed sparse CSP and sparse PCA techniques is evaluated on a motor imagery classification task and a vigilance estimation task. Experimental results demonstrate that the BCI system with sparse PCA and sparse CSP techniques are superior to that using the ordinary PCA and CSP algorithms.

Keywords: sparse common spatial pattern, sparse principal component analysis, EEG, brain-computer interface.

1 Introduction

Brain-computer interface (BCI) is usually defined as a direct communication pathway between the brain and a computer or a device. And electroencephalogram (EEG) is the most commonly used brain signals for BCIs. Over the last twenty years, with the advances of signal processing, pattern recognition, and machine learning techniques, the field of BCI research has made great progress [1,2]. Through BCIs, people can directly control an external device just by using EEG signals generated from motor imagery, visual evoked potentials, or people's mental states. However EEG signals are very noisy and unstable. Therefore, relevant EEG components selection and feature extraction are very important for BCIs. For traditional BCIs, spatial filters based on common spatial

⋆ Corresponding author.

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part I, LNCS 7062, pp. 725–733, 2011.

pattern (CSP) are usually used for selecting the relevant EEG components from the linear combination of the original EEG signals of different channels [3], and principal components analysis (PCA) technique is usually used for extracting features from the linear combination of the original EEG features.

However, both the ordinary CSP and PCA algorithms face a loading problem, i.e., their weights in the linear combinations for PCA and CSP are non-zero. That problem makes a BCI system easy to be over-fitted during training process, because not all of the EEG channels or the EEG features are relevant to the given tasks. As a result, to develop efficient algorithms for EEG channel selection and EEG feature selection is highly desirable.

In this paper, we introduce sparse loading representations for both CSP and PCA algorithms. Our proposed sparse technique can accomplish EEG channel selection, relevant EEG component selection, and EEG feature selection. For sparse PCA, Zou's method is adopted [4], where PCA is considered as a regression-type problem and elastic net is used to calculate the sparse loading of PCA. The performance of a BCI system using sparse PCA is evaluated on an EEG-based vigilance estimation task. For sparse CSP, we propose a novel sparse CSP algorithm and consider CSP as a Rayleigh quotient problem. We use sparse PCA and elastic net to calculate the sparse loadings of CSP. The performance of a BCI system with our proposed sparse CSP algorithm is evaluated on a motor imagery task from the BCI Competition III, Data sets IIIa [5]. Experimental results demonstrate that both BCI systems using sparse representation techniques have outperformed the traditional BCI systems.

This paper is organized as follows. In Section 2, the sparse PCA and sparse CSP algorithms are presented. In Section 3, the experimental setups and the EEG data processing of vigilance task and motor imagery task are described. In Section 4, experimental results are presented and discussed. Finally, some conclusions are given in Section 5.

2 Sparse PCA and CSP Algorithms

As both sparse PCA and sparse CSP algorithms are based on elastic net, the elastic net algorithm is briefly introduced first, and then sparse PCA and our proposed sparse CSP algorithms are described.

2.1 Elastic Net

Consider a data set $\{X, Y\}$, here $X = (x_1, ..., x_m)$ is the input set, $x_i = (x_{i,1}, ..., x_{i,n})^T$, $i = 1, ..., m$, is the *i-th* feature of input set, n is the number of data, m is the feature dimension, and $Y = (y_1, ..., y_n)^T$ is the response set. For linear regression model, a criterion is usually formed as

$$\hat{\beta} = \arg\min_{\beta} |Y - X\beta|^2, \tag{1}$$

where β is the linear coefficients to be estimated. However, the elements of β are typically nonzero, even some features $\{x_i\}$ are almost not correlated with the response set. This makes the linear regression model easy to be overfitted.

To solve this problem, various kinds of methods have been proposed. Lasso is one of the famous methods, which adds a L_1 norm penalty to the ordinary criterion. The Lasso criterion is formed as

$$\hat{\beta} = \arg\min_{\beta} |Y - X\beta|^2 + \lambda|\beta|^1, \tag{2}$$

where λ is the penalty factor and $|\cdot|^1$ stands for L_1 norm.

By tuning λ, Lasso can continuously shrink the linear coefficients toward zero and accomplish feature selection, and then improve the prediction accuracy via the bias-variance tradeoff. Lasso can be efficiently solved by the LARS algorithm [6]. However, LARS has a drawback: the number of selected features is limited by the number of training data or the number of linear unrelated features in the training data. To overcome this promlem, naive elastic net and elastic net have been proposed [7], which add a L_2 norm penalty to the Lasso criterion. The naive elastic net criterion is formed as

$$\hat{\beta} = \arg\min_{\beta} |Y - X\beta|^2 + \lambda_1|\beta|^1 + \lambda_2|\beta|^2, \tag{3}$$

where λ_1 and λ_2 are the penalty factors.

The naive elastic net usually makes too much coefficients shrinkage, and causes more bias to the ELM. But it only reduces a little variances. To correct the bias, elastic net is proposed, whose solution is a rescaled naive elastic net solution with a factor $(1 + \lambda_2)$. The elastic net criterion is formed as

$$\hat{\beta} = (1 + \lambda_2)\arg\min_{\beta} |Y - X\beta|^2 + \lambda_1|\beta|^1 + \lambda_2|\beta|^2. \tag{4}$$

Both naive elastic net and elastic net can be efficiently solved by the LARS-EN algorithm [7]. The elastic net can simultaneously produce an accurate and sparse model without the limitation of LARS.

2.2 Sparse PCA

Sparse PCA used in this paper was proposed by Zou et al. [4]. They reformulate the PCA problem as a regression model and solve it by using the following four theorems.

In theorem 1, let Z_i denote the *i-th* principal component of X. The corresponding PCA loadings V_i can be calculated from the following regression model,

$$\hat{\beta} = \arg\min_{\beta} |Z_i - X\beta|^2 + \lambda|\beta|^2, \tag{5}$$

where λ can be assigned with any positive value, and $V_i = \frac{\hat{\beta}}{|\hat{\beta}|}$.

In theorem 2, another connection between PCA and a regression model is formed as

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{j=1}^{n} |X_{.,j} - \alpha\beta^T X_{.,j}|^2 + \lambda|\beta|^2 \quad \text{subject to } |\alpha|^2 = 1, \tag{6}$$

where $X_{.,j}$ is the row vector of X, α and β are $m \times 1$ vectors, and $V_1 = \frac{\hat{\beta}}{|\hat{\beta}|}$.

In theorem 3, let α and β be $m \times k$ matrices. The connection between PCA and a regression model is formed as

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{j=1}^{n} |X_{.,j} - \alpha\beta^T X_{.,j}|^2 + \lambda \sum_{i=1}^{k} |\beta_i|^2 \quad \text{subject to } \alpha^T\alpha = I_k, \tag{7}$$

where $V_i = \frac{\hat{\beta}_i}{|\hat{\beta}_i|}$, for $i = 1, ..., k$.

To achieve sparse loadings, a L_1 penalty is added into (7)

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{j=1}^{n} |X_{.,j} - \alpha\beta^T X_{.,j}|^2 + \lambda \sum_{i=1}^{k} |\beta_i|^2 + \sum_{i=1}^{k} \lambda_{1,i}|\beta_i|^1 \quad \text{subject to } \alpha^T\alpha = I_k, \tag{8}$$

where $\lambda_{1,i}$ is the penalty factor. This is a naive elastic net problem, and can be efficiently solved after fixing α.

In theorem 4, suppose the SVD of $X^T X\beta$ is $X^T X\beta = P\Sigma Q^T$. It is proved that the solution of α in (8) should be

$$\hat{\alpha} = PQ^T. \tag{9}$$

Then Eq. (8) can be solved by alternated updating $\hat{\alpha}$ and $\hat{\beta}$ until they converge. When solving Eq. (8), only the covariance matrix of X is need. For more details, please refer [4].

2.3 The Proposed Sparse CSP Algorithm

Let X denote the original EEG signals, where X is a $p(\text{channel}) \times l(time)$ matrix. The CSP-based spatial filter is to determine some linear projections, $y = v^T X$, that can maximize the variance (yy^T or $v^T XX^T v$) of signals of one condition and at the same time minimize the variance of signals of another condition in a specific frequency band. The variance of a specific frequency band is equal to the band-power. Then, CSP can be formulated as a maximum power-ratio problem or a Rayleigh quotient problem as follows:

$$\hat{V} = \{v | \max \frac{v^T R_1 v}{v^T R_2 v} \text{ or } \max \frac{v^T R_2 v}{v^T R_1 v}\} \tag{10}$$

where R_i is the covariance matrix of original EEG signals on condition i, and $\hat{V}$ are the projection vectors or loadings of CSP.

Equation (10) can be solved as follows. Let

$$v^T R_2 v = u^T u, \tag{11}$$

and then,

$$v = P\Sigma^{-1/2} u, \tag{12}$$

where P and Σ are the PCA decomposition of R_2, $R_2 = P\Sigma P^T$.

By applying Eqs. (11) and (12), $\frac{v^T R_1 v}{v^T R_2 v}$ can be reformed as

$$\frac{u^T D u}{u^T u}, \tag{13}$$

where $D = \Sigma^{-1/2} P^T R_1 P \Sigma^{-1/2}$.

It is easy to show that the *i-th* largest value of Eq. (13) is the *i-th* largest eigenvalue of D, and u is the corresponding eigenvector. The *i-th* smallest value of Eq. (13) corresponds to the *i-th* largest value of $\frac{v^T R_2 v}{v^T R_1 v}$. Usually, not two projections but several projections corresponding to the large values of $\frac{v^T R_1 v}{v^T R_2 v}$ and $\frac{v^T R_2 v}{v^T R_1 v}$ are used for EEG spatial filtering. The loadings, v, of CSP can be calculated by using Eq. (12) together with the eigenvectors corresponding to some large eigenvalues or small eigenvalues of D.

To achieve sparse loadings of CSP, we can reformulate Eq. (12) as an elastic net problem as follows:

$$\hat{v} = \arg\min_v |u - \Sigma^{1/2} P^T v|^2 + \lambda_1 |v|^1 + \lambda_2 |v|^2, \tag{14}$$

and solve it by the LARS-EN algorithm.

2.4 Complexity Analysis of the Proposed Sparse CSP Algorithm

In EEG data analysis, the number of features, m, is usually less than the number of data, n. Therefore, the complexities of the proposed spare CSP algorithm can be analyzed only on $m < n$ condition.

For elastic net, the time cost is $O(m^3 + nm^2)$ [7], which is equivalent to the cost of least square problem. For sparse PCA, the time cost is $nm^2 + pO(m^3)$ [4], where p is the number of iterations when solving the sparse PCA. As a result, the cost of sparse PCA is comparable with the cost of the ordinary PCA, $O(m^3)$.

For our proposed sparse CSP algorithm, the extra time cost is $kO(nm^2 + m^3)$ in comparison with the ordinary CSP algorithm, where k is the number of components extracted by CSP. The cost of ordinary CSP algorithm is $O(m^3 + nm^2)$. Therefore, the total cost of the proposed sparse CSP algorithm is $(k+1)O(m^3 + nm^2)$, which is comparable with the cost of the ordinary CSP algorithm.

3 Experiment

3.1 Experimental Setup

Vigilance Task. This is a monotonous visual task [8,9,10]. The subjects are asked to sit in a comfortable chair, two feet away from the LCD. There are four colors of traffic signs being presented in the LCD randomly by the NeuroScan $Stim^2$ software. Each trial is 5.5∼7.5 seconds long, including 5∼7 seconds black screen and 500 millisecond traffic signs presented. The subjects are asked to recognize the sign color, and press the correct button on the response pad. A total of 11 healthy subjects have participated in this experiment. After training, each subject has finished at least two sessions (one for train, and others for test). For each session, a total of 62 EEG channels are recorded by the NeuroScan system sampled at 500Hz. Each session continues for more than one hour, during 13:00∼15:00 after lunch. The local error rate of the subject's performance is used as the reference vigilance level, which is derived by computing the target false recognition rate within a 2-minute time window at 2-second step.

Motor Imagery Task. This data set comes from BCI Competition III, data sets IIIa, provided by the Laboratory of Brain-Computer Interfaces (BCI-Lab), Graz University of Technology [5]. It is a 4 classes (left hand, right hand, foot, and tongue) cued motor imagery experiment from 3 subjects. After trial begin, the first 2s were quite, at t=2s an acoustic stimulus indicated the beginning of the trial, and a cross + is displayed; then from t=3s an arrow to the left, right, up or down was displayed for 1 s; at the same time the subject was asked to imagine a left hand, right hand, tongue or foot movement, respectively, until the cross disappeared at t=7s. There are 60 trials per class for each subject. A total of 60 EEG channels are recorded by the NeuroScan system sampled at 250Hz.

3.2 Data Processing

Vigilance Task. Six EEG channels (P1, Pz, P2, Po3, Poz, Po4) are used for the vigilance estimation task, which are measured from the posterior regions of the scalp. The vigilance estimation process consists of the following five main components: a) a bandpass filter (1Hz-50Hz) is used to remove the low-frequency noise and the high frequency noise; b) the power spectral density (PSD) of each channel is calculated by every 2 seconds with a 2 Hz frequency resolution as the original features; c) the features are smoothed with a 2 min moving-average filter; d) the top 10 principal components of the PSD are calculated by the sparse PCA algorithm as features; and e) a least square regression model is adopted for vigilance estimation by every 2 seconds.

Each subject has an individual vigilance estimation model. For each vigilance estimation model, one session of a subject is used for training, while other sessions of this subject are used for test.

Motor Imagery Task. All 60 EEG channels are used for motor imagery classification. The 4-class motor imagery data sets are paired into 6 groups of 2-class motor imagery data sets for classification. The classification process consists of the following four main components: 1) a bandpass filter (8Hz-32Hz) is used to remove the noises and EEG signals which are unrelated to motor imagery; 2) the top 10 motor imagery related EEG components are extracted by the proposed sparse CSP algorithm; 3) the variance of each component in each single motor imagery trial is calculated as the feature; and 4) SVMs with RBF kernel is adopted as the motor imagery classifiers.

The classification model is trained for each subject and each pair of 2-class motor imagery separately. For each classification model, half of each 2-class motor imagery data set is used for training, while the other half is used for test. The parameters used in SVMs are fine tuned by 5-fold cross validation.

4 Experimental Results

The performance of BCI system using sparse PCA is evaluated on the vigilance estimation task. The parameter λ_1 in sparse PCA is used to control the sparseness of loadings. Instead of tuning λ_1, we directly set the number of nonzero coefficients in the loadings of sparse PCA. An early stopping strategy is used for the LARS-EN algorithm. When the number of nonzero coefficients of β_i meets the predefined number, the LARS-EN algorithm used for solving the naive elastic net in sparse PCA is stopped. In this study, without fine-tuning, λ is assigned to 10^{-5}, and the number of nonzero coefficients in each principal component loading is set to 20.

For comparison, another BCI system with using the ordinary PCA is used for vigilance estimation. There are totally 30 pairs of training and test data set from the 11 subjects. The linear correlation coefficient and mean square error between the estimated vigilance level and the reference vigilance level are used for performance evaluation. The experimental results of vigilance estimation is

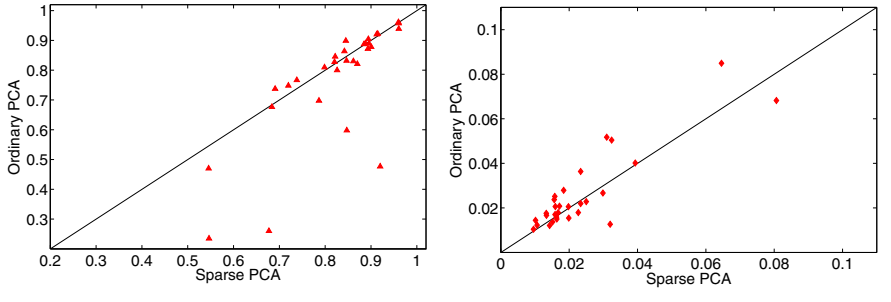

Fig. 1. The result of linear correlation coefficient between the estimated vigilance level and the reference vigilance level (left), and the result of mean square error between the estimated vigilance level and the reference vigilance level (right)

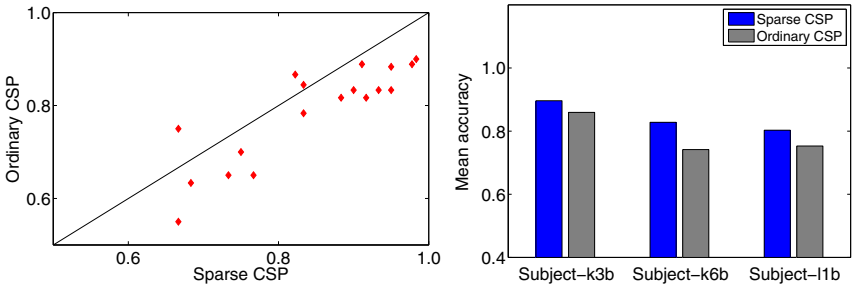

Fig. 2. Comparison of classification accuracies of all 2-class motor imagery data set from 3 subjects (left), and the means of two-class classification accuracies for each subject (right)

shown in Fig. 1. From this figure it can be seen that the average performance of the BCI with sparse PCA is better than that of the BCI system with the ordinary PCA. For those data set the BCI with the ordinary PCA performed well, and the BCI with sparse PCA also performed well. But for those data set the BCI with the ordinary PCA didn't perform well, and the BCI with sparse PCA still performed well, or at least performed much better than the BCI with ordinary PCA.

The performance of the BCI system with the proposed sparse CSP algorithm is evaluated on the motor imagery task. There are totally 6 pairs of training and test data set for each subject. The LARS-EN algorithm used in the sparse CSP algorithm also adopts an early stopping strategy. The number of nonzero coefficients in each CSP loading is set to 30, and λ_2 is assigned to 0.01.

For comparison, a BCI system with the ordinary CSP algorithm is also applied to the motor imagery classification. The experimental results are shown in Fig. 2. From this figure it can be seen that, for most two-class data sets, the BCI system with the proposed sparse CSP algorithm performed better than the BCI system with the ordinary CSP algorithm; and for each subject, the average performance of the BCI system with the proposed sparse CSP algorithm is better than that of the BCI system with the ordinary CSP algorithm.

5 Conclusions

In this paper, sparse PCA and sparse CSP techniques are introduced to EEG-based BCIs. And a novel sparse CSP algorithm has been proposed. The performance of BCI systems with sparse PCA and sparse CSP algorithms have been evaluated on a vigilance estimation task and a motor imagery classification task. Experimental results demonstrate that the BCI systems with sparse PCA and CSP techniques have outperformed the ordinary BCI systems. This result indicates that sparse subspace learning technique is very useful for EEG data processing, which can improve the robustness of EEG-based BCI systems.

In addition, as the solution of LARS-EN is global optimal in comparison with other spare subspace learning techniques such as non-negative matrix factorization [11], the solutions of sparse PCA and sparse CSP can be much more stable.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (Grant No. 90820018), the National Basic Research Program of China (Grant No. 2009CB320901), and the European Union Seventh Framework Programme (Grant No. 247619).

References

1. Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., Arnaldi, F.: A review of classification algorithms for EEG-based brain-computer interfaces. Journal of Neural Engineering 4, R1–R13 (2007)
2. Brunner, P., Bianchi, L., Guger, C., Schalk, G.: Current trends in hardware and software for brain-computer interfaces. Journal of Neural Engineering 8(2) (in press, 2011)
3. Koles, Z.: The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. Electroencephalogr. Clin. Neurophysiol. 79(6), 440–447 (1997)
4. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics 15(2), 265–286 (2006)
5. BCI Competition III, `http://www.bbci.de/competition/iii/`
6. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angel regression. Annals of Statistics 32(2), 407–499 (2004)
7. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67(2), 301–320 (2005)
8. Shi, L.C., Lu, B.L.: Dynamic clustering for vigilance analysis based on EEG. In: Proce. of the 30th International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, Canada, pp. 54–57 (2008)
9. Shi, L.C., Lu, B.L.: Off-Line and On-Line Vigilance Estimation Based on Linear Dynamical System and Manifold Learning. In: Proce. of the 32nd International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, pp. 6587–6590 (2010)
10. Ma, J.X., Shi, L.C., Lu, B.L.: Vigilance estimation by using electrooculographic features. In: Proce. of the 32nd International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, pp. 6591–6594 (2010)
11. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)