

One Step Feature Extraction and Classification For Brain Computer Interface

Arun K Bharathan

NSS College of Engineering, Palakkad

arunkbharathan@gmail.com

April 4, 2015

Overview

Introduction

- Project Overview

- BCI Overview

- Brain Rhythms

- Motor Imagery Events

Steps in BCI along with work

- Signal Acquisition

- Pre-processing

- Feature Extraction Techniques

- Combined Feature Extraction and Classification

 - Signal Analysis Framework

Modification

Results

Conclusion

References

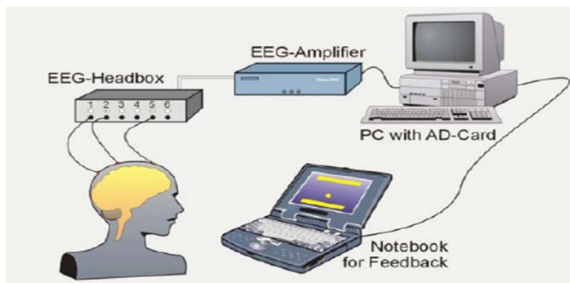
Last Section

Project Overview

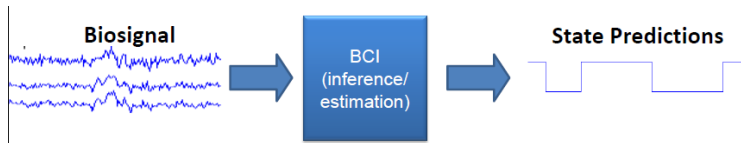
Common Spatial Patterns (CSP) is a popular algorithm in BCI field for learning spatial filters for oscillatory processes. But the CSP method is not optimal. A few spatial filters are chosen arbitrarily from the learned spatial filter matrix W of CSP. Here a method based on optimization approach is used to select optimum weight matrix, which combines the feature extraction and classification stages in normal BCI. Thus use of an independent classification stage like LDA is avoided.

Brain Computer Interface

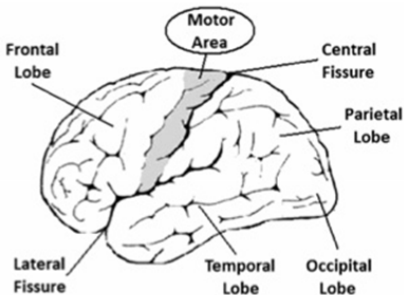
- Brain Computer Interface (BCI) is a hardware and software communications system that permits cerebral activity alone to control computers or external devices.
- The immediate goal of BCI research is to provide communications capabilities to severely disabled people who are totally paralysed or 'locked in' by neurological neuromuscular disorders.



As a Signal Processing Scheme



Brain Rhythms



- Delta (less than 4 Hz)
- Theta (4 to 7 Hz)
- Alpha (8 to 12 Hz) and Mu (7 to 13 Hz)
- Beta (12 to 30 Hz)
- Gamma (30 to 100 Hz)

Events

- Different parts of brain generate oscillations when idle.
- The amplitude of oscillations starts to decrease when we think of moving a limb.
- It reaches a minimum just after the onset of motion called Event Related De-synchronization (ERD), then it reverts back called Event Related Synchronization (ERS)

Steps that form a standard BCI

1. **Signal Acquisition**
2. **Pre-processing**
3. **Feature Extraction**
4. **Classification**
5. **Control Interface**

Signal Acquisition

EEG

- ▶ Poor quality weak signals.
- ▶ Low spatial resolution and high spectral resolution.
- ▶ severely affected by background noise.
- ▶ Non-invasive technique, so widely used.

Used 4 class dataset, 2a from BCI Competition 4, has 22 EEG and 3 EOG electrodes.

ECoG

- ▶ Good quality strong signals
- ▶ High spatial and spectral resolution.
- ▶ Low in artifacts.
- ▶ Requires Craniotomy.

Used 2 class dataset, dataset 1 from BCI Competition 3, has 8×8 ECoG electrodes.

Pre-processing

EEG

- ▶ $4C_2$ binary combination was extracted each with 144 testing and training trials.
- ▶ RLS filtered, BPF at 7-30 Hz.
- ▶ 0.5 to 3.5 sec epoch was cut out after visual cue.
- ▶ Down-sampled to 100 Hz and whitened the covariance matrices.

ECoG

- ▶ Available as 3 sec epoch cut out after visual cue at $F_s=1000$ Hz.
- ▶ Filtered at 7-30 Hz with *filtfilt* Matlab command.
- ▶ Down-sampled to 100 Hz and whitened the covariance matrices.

Dataset Description

BCI Competition IV Dataset 4a

- ▶ 4 class EEG dataset with 22 electrodes and 3 EOG electrodes.
- ▶ Left hand [Class 1], Right hand [Class 2], Feet [Class 3], Tongue [Class 4].
- ▶ 9 subjects, each with 144 testing and 144 training trials.
- ▶ Sampled at 250Hz , BPF at 0.5Hz and 100Hz and a notch filter at 50Hz .

Cont.

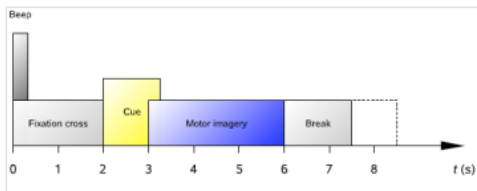


Figure : Timing Scheme

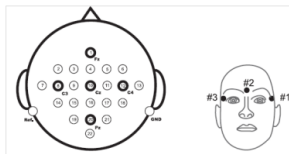


Figure : Electrode Positions

Dataset Description

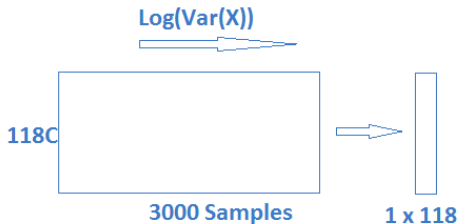
BCI Competition III Dataset 1

- ▶ 8x8 ECoG electrode grid placed on right motor cortex.
- ▶ Imagined figure and tongue movements recorded for 3 seconds duration.
- ▶ Sampled at 1000 Hz.
- ▶ Recording started after 0.5 second to avoid VEP.
- ▶ Available as data epoch of $64 \times 3000 \times 278$ for training and $64 \times 3000 \times 100$.

Feature Extraction Techniques

Logarithmic Band-power

- Band pass filter trials at (7 - 30) Hz.
- Only 118 weights and 1 bias term to learn.



- Suppose $X \in \mathbb{R}^{118 \times 3000}$.
- But
 - Does not capture time variation in oscillations.
 - Combine multiple bands.
 - No data adaptive feature.

Common Spatial Patterns

- A data dependent spatial filtering.
- Can exploit ERS and ERD localized in the motor cortex.
- Simple, fast and relatively robust.
- Projects multichannel EEG signals into a subspace, where differences are maximized and similarities are minimized.
- Works on normalized spatial covariance matrix.

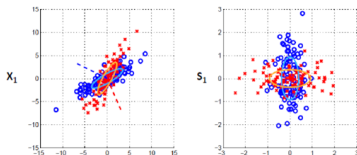
$$C = \frac{XX^T}{\text{trace}(XX^T)}$$

- $C \in \mathbb{R}^{118 \times 118}$.

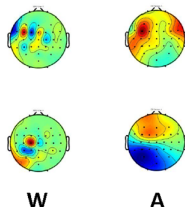
-

$$f_{\theta}(X) = \sum_{j=1}^J \beta_j \log \left(w_j^T S^T B_j B_j^T S w_j \right) + \beta_0$$

Common Spatial Patterns



- $X = AS$, $X, S \in \mathbb{R}^{118 \times 3000}$, $A \in \mathbb{R}^{118 \times 118}$
- $S = WX$, $X, S \in \mathbb{R}^{118 \times 3000}$, $W \in \mathbb{R}^{118 \times 118}$
 - A is forward mapping matrix.
 - W is inverse mapping matrix.



Common Spatial Patterns

CSP spatial filters

- W is the generalized eigenvectors of covariance matrices C_1 and $(C_1 + C_2)$
- By optimization approach

$$w_c = \max_w \frac{w^T C_1 w}{w^T C_2 w} \quad s.t. \quad w^T C_2 w = 1$$

- But we take only 1-3 pair of filters and patterns from either side of W and A , respectively.
- If we take only 1 pair

$$W_{2 \times 118} \times X_{118 \times 3000} = S_{2 \times 3000}$$

- Taking log variance of S , LDA has to learn only 3 parameters (including bias term).

Combined Feature Extraction and Classification

Combined Feature Extraction and Classification

- ▶ The method learns in a single step both the spatial filters and the relative weights.
- ▶ A unified, globally optimal solution to spatial filter estimation (an alternative to CSP+LDA).
- ▶ This is an optimization-based approach.
- ▶ Offers a lot of flexibility, a number of parameters are available for fine tuning.
- ▶ Used to investigate neuroscientific questions about the underlying process.
- ▶ Can impose various regularizers and loss terms while optimizing.
- ▶ Using various regularizers different assumptions on the structure of data can be made.

Signal Analysis Framework

The framework consists of three components.

1. Probabilistic predictor model.
2. Detector function.
3. Regularization.

Probabilistic predictor model

Learns a probability model from input data to given label.

- ▶ Probabilistic predictor are facing two tasks.
 - ▶ How to learn the predictor from a collection of labelled examples.
 - ▶ How to decode the intention of a user given the brain signal and the predictor.

Let

- $X \in \mathcal{X}$, input brain signal
- $q(Y|X)$ is the predictor which assigns probabilities to the user's command $y \in Y$ given the brain signal X .

The task of decoding is to find the most likely command \hat{y} given the input X and the predictor q as follows:

$$\tilde{y} = \arg \max_{y \in Y} q(Y = y | X)$$

Probabilistic predictor model

- The task of learning is to find a predictor from a suitably chosen collection of candidate models.
- Assume that a model is parametrized by a parameter $\theta \in \Theta$.
- The loss function is defined as the negative logarithmic pay off (or the Shannon information content in information theory) as follows:

$$\ell_L((X, y), \theta) = -\log q_\theta(Y = y \mid X)$$

- Loss is smaller if the predictor predicts the actual intention of the user with high confidence.

Probabilistic predictor model

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell_L((X_i, y_i), \theta)$$

- Minimization of $L_n(\theta)$ leads to over-fitting due to small sample size.
- The parameter θ is determined by solving the following constrained minimization problem:

$$\min_{\theta \in \Theta} L_n(\theta) \text{ s.t. } \Omega(\theta) \leq C$$

- The second term is called the regularizer.
- C is a hyper-parameter that controls the complexity of the model.

Probabilistic predictor model

- ▶ If we suppose that the training examples $\{X_i, y_i\}_{i=1}^n$ are sampled i.i.d from some probability distribution $p(X, Y)$, the above function $L_n(\theta)$ can be considered as the empirical version of the following function.

$$L(\theta) = D(p(Y | X) \| q_\theta(Y | X)) + H(p(Y | X))$$

- ▶ $D(p \| q)$ is Kullback–Leibler divergence between two probability distributions p and q .
- ▶ Second term is the conditional entropy of Y given X and is a constant that does not depend on the model parameter .

Probabilistic predictor model

Logistic Model

- ▶ The logistic model assumes the user command Y to be either one of the two possibilities; e.g., $Y = -1$ and $Y = +1$ for left and right-hand movement, respectively.
- ▶ The logistic predictor q_θ is defined through a latent function f_θ (Detector function).
- ▶ The detector function outputs a positive number if $Y = +1$ is more likely than $Y = -1$ and vice versa.
- ▶ The logistic model converts it into the probability of $Y = +1$ given X

$$q_\theta(Y = y | X) = \frac{1}{1 + \exp(-yf_\theta(X))} \quad (y \in \{+1, -1\})$$

Probabilistic predictor model

Logistic loss function

- ▶ The loss function for the logistic model.

$$\ell_L((X, y), \theta) = \log(1 + \exp(-yf_\theta(X)))$$

- ▶ The function f_θ is called a detector because in the BCI context it captures some characteristic spatio-temporal activity in the brain.

Detector Function

- ▶ The commonly used CSP based detector model can be written as follows:

$$f_{\theta}(X) = \sum_{j=1}^J \beta_j \log \left(w_j^T S^T B_j B_j^T S w_j \right) + \beta_0$$

- ▶ We use the following linear detector function:

$$f_{\theta}(X) = \langle W, S^T S \rangle + b$$

Detector Function

- ▶ We can set X as a block diagonal concatenation of the covariance as follows:

$$X = \begin{pmatrix} \frac{1}{\eta_{(1)}} \Xi^{(1)} & & \\ & \frac{1}{\eta_{(2)}} \Xi^{(2)} & \\ & & \frac{1}{\eta_{(k)}} \Xi^{(k)} \end{pmatrix}$$

- ▶ Where,
 - ▶ $\Xi^{(k)}$ is the covariance matrix of a short segment of band-pass filtered EEG signal.
 - ▶ $\eta_{(k)}$ is the normalization factor to prevent biasing the selection of terms with large power, it is the square root of the total variance of each block element.

Regularization

- ▶ Regularizer used is the linear sum of singular-values of the weight matrix W , which is called the dual spectral (DS) norm.

$$\Omega_{DS}(\theta) = \sum_{j=1}^r \sigma_j(W)$$

- $\sigma_j(W)$ is the j^{th} singular value of the weight matrix W .
- r is the rank of W .
- ▶ The DS regularization can be considered as a case of the ℓ_1 -regularization; it induces sparsity in the singular-value spectrum of the weight matrix W .
- ▶ That is, it induces low-rank matrix W .
- ▶ The DS regularization automatically tunes the feature detectors as well as the rank of W .

CSP with Tikhonov Regularization

$$w_c = \max_w \frac{w^T C_1 w}{w^T C_2 w + \alpha w^T K w} \quad s.t. \quad w^T C_2 w = 1$$

- ▶ ' C_x ' is the average covariance matrix of class 1 and 2.
- ▶ Where K is identity matrix or any diagonal matrix that encode channel prior.
- ▶ The above regularization is equal to minimizing the squared euclidean norm of each channel.
- ▶ α is the regularization parameter, that is to be fine tuned by cross validation on training data.
- ▶ α fixes how much it should believe Identity matrix than covariance matrix.

Cont.

- ▶ To prevent over-fitting in CSP due to short data set or noisy data set.
- ▶ Restricts w to have small norms.
- ▶ Performs better than CSP for noisy and short data set.

Modification

- ▶ In our detector function W is a symmetric matrix and $W = w \times w^T$.

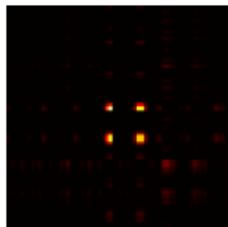
$$f_{\theta}(X) = \langle W, X \rangle + b$$

- ▶ The diagonal element is equal to squared euclidean norm of channel.
- ▶ Change regularization term to $\text{trace}(W)$ with our predictor model.
- ▶ Here W is constraint to diagonal matrix.

Results

Result for ECoG data set.

- ▶ 91% of accuracy is achieved with single second order model (7-30 Hz).



- ▶ Figure shows learned low rank weight matrix $W \in R^{64 \times 64}$.

Results

Results for EEG data set.

Table : Result for Dataset 2a

	[1 2]	[1 3]	[1 4]	[2 3]	[2 4]	[3 4]
A1	86.81	85.42	99.31	91.67	100	67.36
A2	73.61	77.78	68.75	78.47	82.64	72.92
A3	98.61	93.75	84.03	98.61	97.22	86.11
A4	78.47	90.97	86.11	95.14	87.50	75.00
A5	70.18	77.78	81.25	75.00	79.86	80.56
A6	69.44	79.86	72.22	74.30	75.00	75.69
A7	84.72	97.92	97.92	97.92	97.92	84.03
A8	99.31	94.44	97.22	95.14	97.22	95.83
A9	92.36	95.14	100	90.97	96.53	97.92

Results (Cont.)

Results for EEG data set.

Table : Comparison with CSP

	DS	CSP
A1	86.81	88.89
A2	73.61	51.39
A3	98.61	96.53
A4	78.47	70.14
A5	70.18	54.86
A6	69.44	71.53
A7	84.72	81.25
A8	99.31	93.75
A9	92.36	93.75

Results (Cont.)

Results for EEG data set.

Table : Performance Comparison of TRCSP Using One Step Process and by eigen decomposition

	One Step TRCSP	TRCSP
A1	85.68	88.89
A2	60.42	54.17
A3	90.72	96.53
A4	73.61	70.83
A5	58.33	62.50
A6	70.56	67.36
A7	81.94	81.25
A8	89.56	95.87
A9	90.00	91.67

Results (Cont.)

Results for EEG data set.

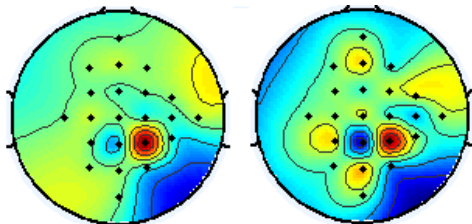
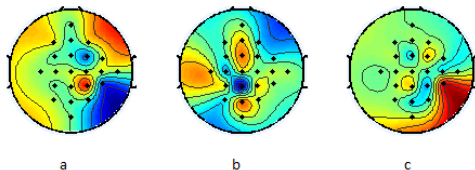


Figure : Simulation Results using topoplot for TRCSP and one step method side by side

Topoplot plots a topographic map of a scalp data field in a 2-D circular view.

Results (Cont.)

Results for EEG data set.



- (a) Filter captured by combined alpha and beta band.
- (b) Filter captured by alpha band alone.
- (c) Filter captured by beta band alone.

Control Interface

- ▶ 2-dimensional cursor control.
- ▶ To gaming devices
- ▶ Orthoses and prostheses control.
- ▶ Robotic arms.
- ▶ Mobile Robots.

Conclusion

- The issues of feature learning, feature selection, and feature combination are addressed through regularization.
- The key idea of the approach is to focus on directly predicting the intention of a user.
- The method can be employed in the multi-class environment.
- Can be applied to other multiple sensor recordings fMRI, Computer vision.
- New Detector model, Predictor model, Regularization term can be applied.
- Signal analysis Framework can be applied to regression problems.

List of Publication

Presented

- "Review of Feature Extraction Techniques for motor imagery signals in BCI", Arun K Bharathan, Nandakumar P, PIC 2013, IETE Palakkad.






Accepted

- "One Step Feature Extraction and Classification with Tikhonov Regularization for BCI", Arun K Bharathan, Arun Ashok, Soujya V R, Nandakumar P, ICGCE 2013, Kavarapettai.
- "Tikhonov Regularized Spectrally Weighted Common Spatial Patterns", Arun Ashok, Arun K Bharathan, Soujya V R, Nandakumar P, ICCCE 2013, Thiruvananthapuram.

Communicated

- "Optimizing Spatial Filters By Minimizing Within Class Dissimilarities Using Different Metric Measures In EEG Based Brain Computer Interfaces", Soujya V R, Arun Ashok, Arun K Bharathan, Nandakumar P, SPINCON 2014, Noida.

References

-  Tomioka, R., and Mller, K. R. (2010). "A regularized discriminative framework for EEG analysis with application to brain-computer interface". Neuroimage, 49(1), 415-432.
-  Lotte, Fabien, and Cuntai Guan. (2011), "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms." Biomedical Engineering, IEEE Transactions on 58, no. 2 : 355-362.
-  Lotte, Fabien, and Cuntai Guan. (2010), "Spatially regularized common spatial patterns for EEG classification." In Pattern Recognition (ICPR), 20th International Conference on, pp. 3712-3715. IEEE, 2010.
-  Tomioka, Ryota, and Kazuyuki Aihara.(2007) "Classifying matrices with a spectral regularization." In Proceedings of the 24th international conference on Machine learning, pp. 895-902. ACM.
-  Ramoser, H., Muller-Gerking, J., and Pfurtscheller, G. (2000). "Optimal spatial filtering of single trial EEG during imagined hand movement". Rehabilitation Engineering, IEEE Transactions on, 8(4), 441-446.

References



Boyd, Stephen Poythress, and Lieven Vandenberghe. (2004), Convex optimization. Cambridge university press.



Vandenberghe, Lieven, and Stephen Boyd. (1996) "Semidefinite programming." SIAM review 38, no. 1: 49-95.



Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Muller, K. R. (2008). "Optimizing spatial filters for robust EEG single-trial analysis". Signal Processing Magazine, IEEE, 25(1), 41-56.



Tomioka, Ryota, Kazuyuki Aihara, and Klaus-Robert Müller.(2007) "Logistic regression for single trial EEG classification." Advances in neural information processing systems 19: 1377-1384.



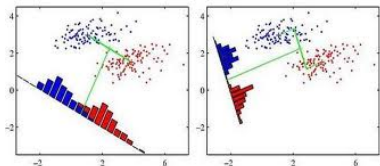
Grant, Michael, Stephen Boyd, and Yinyu Ye. (2008) "CVX: Matlab software for disciplined convex programming."



Delorme, Arnaud, and Scott Makeig.(2004) "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." Journal of neuroscience methods 134.1: 9-21.

The End

Linear Discriminant Analysis



```
Mu1=mean(C1feature)';  
Mu2=mean(C2feature)';  
S1=cov(C1feature);  
S2=cov(C2feature);  
Sw=S1+S2;  
SB= (Mu1-Mu2)*(Mu1-Mu2)';  
[V,D]=eig(inv(Sw)*SB);  
FDV=V(:,1);  
a=C1feature*FDV;  
b=C2feature*FDV;
```

Implementation DS norm

$$\min_{W \in R^{R \times C}, b \in R, z \in R^n, Q_1 \in S_+^C, Q_2 \in S_+^R} \sum_{i=1}^n \ell_{LR}(z_i) + \lambda (Tr[Q_1] + Tr[Q_2]), \quad i = 1, \dots, n \quad (1)$$

$$y_i \left(Tr[W^T X_i] + b \right) = z_i, \quad i = 1, \dots, n$$

$$\begin{pmatrix} Q_1 & -\frac{1}{2}W \\ -\frac{1}{2}W^T & Q_2 \end{pmatrix} \succeq 0$$

CVX DS Code

```
function [W, bias, z]=lrl1(X, Y, lmd)
C = size(X,1); n = length(Y);
cvx_begin sdp
variable W(C,C) symmetric;
variable U(C,C) symmetric;
variable bias;
variable z(n);
minimize sum(log(1+exp(-z)))+lmd*trace(U);
subject to
for i=1:n
Y(i)*(trace(W*X(:, :, i))+bias)==z(i);
end
U >= W; U >= -W;
cvx_end
end
```


CVX Tikhonov Code

```
function [W, bias, z]=lrl1(X, Y, lmd)
C = size(X,1); n = length(Y);
cvx_begin sdp
variable W(C,C) diagonal;
variable bias;
variable z(n);
minimize sum(log(1+exp(-z)))+lmd*trace(W);
subject to
for i=1:n
Y(i)*(trace(W*X(:, :, i))+bias)==z(i);
end
cvx_end
end
```

RLS Code

```
N=7;
W=eps*ones(1,N);
P=eye(7)*1000;
lambda= 0.95;
tic
for i=N:length(x)
    u=x(i:-1:i-(N-1));
    y(i)=W*u;
    e(i)=d(i)-y(i);
    k=(lambda+u'*P*u)\P*u;
    W=W+k'*e(i);
    P=(1/lambda)*(P-k*u'*P);
    MSE(i-6)=e(i)^2;
end
```