Spatio-Spectral Filter Optimization in a Bayesian Framework for EEG-based Brain-Computer Interfaces

Department of Brain and Cognitive Engineering
Seong-Whan Lee
September 18, 2012
A BCI is the translation of human intentions into a technical control signal without using activity of muscles or normal peripheral nerves.

Decoding the neuronal signals of brain activities by means of machine learning techniques
Background

● Spectral filtering
  – Advisable to reduce the frequency content of the EEG signal to some frequency band of interest
    ● ERD/ERS-complex can be found predominantly in the $\mu$ and $\beta$-band
    ● However, highly variable across subjects and across even trials for the same subject

● Spatial filtering
  – To find the signal that originates at a specific scalp location
    ● Every EEG electrode measures a superposition of signals derived from various sources in the brain
    ● To take a linear combination of signals recorded over EEG channels and extract only the component that we are interested in
Prevalent Steps in Motor Imagery Classification

Spectral filtering: $Z = h \otimes X$  \hspace{1cm} (X: single trial EEGs, $h$: spectral filter)
Spatial filtering: $Y = W^T Z$  \hspace{1cm} (W: spatial filter)
Feature extraction: $F = \log(\text{var}(Y))$  \hspace{1cm} (F: feature vector)

- No general method that analytically finds the optimal frequency band
  - Either selected manually based on a visual inspection or unspecifically set to a broadband

- Spatial filter learning (Common Spatial Pattern: CSP [RMGP00]) on bandpass-filtered signals
  - Dependent on the operational frequency band

Learn spatio-spectral filters simultaneously in a unified framework
Related Work

Spatio-spectral filters optimization methods in the literature

<table>
<thead>
<tr>
<th>Authors</th>
<th>Methods and Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemm et al. [LBCM05]</td>
<td>CSP on spatially augmented signals with embedment of the time-delayed signals, empirical determination of the time-delay parameter</td>
</tr>
<tr>
<td>Ang et al. [ACZG08]</td>
<td>Dissection of a broadband into small non-overlapping frequency bands, predefined and fixed-bandwidth</td>
</tr>
<tr>
<td>Thomas et al. [TGL+09]</td>
<td>A subject-dependent filter bank with a coefficient decimation technique, non-overlapping frequency bands and fixed-bandwidth</td>
</tr>
<tr>
<td>Zhang et al. [ZCA+11]</td>
<td>A gradient-based learning for spatial filter optimization in a filter bank, predefined non-overlapping frequency bands and fixed-bandwidth</td>
</tr>
</tbody>
</table>

- Iterative way of selecting the optimal time-delay parameter
- User-defined filter bank: predefined non-overlapping frequency bands and fixed-bandwidth
A Probabilistic Bayesian Approach

- Represent a frequency band as a random vector $\mathbf{B}$
  - **Statistical viewpoint:** optimizing the spatio-spectral filter as probability density estimation function ($pdf$) and exploitation

- Estimate the unknown posterior $pdf \ p(\mathbf{B}|\mathbf{X};\Omega)$
  - Indicates the relative probability of the single trial EEGs $\mathbf{X}$ being correctly classified into $\Omega$ by $\mathbf{B}$-bandpass filtering along with the ensuing computational processes
  - **No functional assumption about the density**

- Learn the spatial filter $\mathbf{W}$ in each of the estimated optimal frequency band individually

**Bayesian Spatio-Spectral Filter Optimization (BSSFO)**
A Novel Bayesian Framework

Discriminative feature extraction by means of the spatio-spectral filters optimization in a probabilistic Bayesian framework

A schematic diagram of the proposed method for class-discriminative feature extraction by means of spatio-spectral filter optimization in a probabilistic Bayesian framework.
Main Contributions

- A novel **probabilistic Bayesian framework** for discriminative feature extraction via spatio-spectral filters optimization
  - To our best knowledge, this is the first time a method of finding the optimal spatio-spectral filters in a probabilistic Bayesian approach has been proposed

- An **iterative particle-based pdf approximation** and an observation model

- A **spectrally-weighted label decision rule** by linearly combining the outputs from multiple classifiers, one for each frequency band, with the weight of the corresponding frequency band
Mathematical Formulation

- Given a set of single trial EEGs $\mathbf{X}$ and the class labels $\Omega$, the posterior pdf can be written as follows (by a Bayes rule)

$$p(B | \mathbf{X}, \Omega) = \frac{p(X, \Omega | B)p(B)}{p(X, \Omega)}$$

- $p(B)$: encodes the uncertainty about the discriminative frequency band
- $p(B | \mathbf{X}; \Omega)$: given a particular frequency band $B$, this likelihood function indicates the probability that the single trial EEGs $\mathbf{X}$ in conjunction with the class labels would have been available to support it
Spectral filtering: $Z = h \otimes X$
Spatial filtering: $Y = W^T Z$
Feature extraction: $F = \log(\text{var}(Y))$

- Once the frequency band is given, the bandpass filtering is a simple linear operation

\[
p(B \mid Z, \Omega) = \frac{p(Z, \Omega \mid B) p(B)}{p(Z, \Omega)}
\]

- Feature extraction needs deterministic matrix operations
- We can estimate the posterior pdf indirectly as follows

\[
p(B \mid Z, \Omega) = p(B \mid F, \Omega)
= \frac{p(F, \Omega \mid B) p(B)}{p(F, \Omega)}
\]

where $p(F, \Omega) = \int_B p(F, \Omega \mid B) p(B) dB$

Our goal is to estimate the posterior pdf $p(B \mid F; \Omega)$
Posterior Estimation

- Factored-sampling algorithm [GCK91]

- Generate particles \( \{b_k\}_{k=1}^K \) from the prior density \( p(B) \)
  - \( b_k \): a particle representing a single frequency band
- Compute the weight of each particle
  
  \[
  \pi_k = \frac{p(F_k, \Omega | b_k)}{\sum_{j=1}^K p(F_j, \Omega | b_j)}
  \]
  - \( F_k \): a feature set extracted in the \( b_k \) frequency band

The weighted particle-set \( B = \{b_k, \pi_k\}_{k=1}^K \) approximates the distribution
Posterior Estimation (cont.)

- Iteratively apply the factored-sampling algorithm until converged with an appropriate criterion
  - To avoid the effect of the biased prior density
    - *i.e.*, an initial particle-set drawn at the beginning of the algorithm
    - *c.f.*, a 'burn-in' process in Markov chain Monte Carlo

- Applying a diffusing process in resampling

\[ b_k' = \begin{cases} 
    b_k + \sigma & \text{if, } \psi(k) > 1 \\
    b_k & \text{otherwise} 
\end{cases} \]

- \( \psi(k) \) denotes the number of times that the \( k \)-th particle in \( B^{(t-1)} \) is chosen while composing the new particle set \( B^{(t)} \)
- \( \sigma \) is a diffusing noise, normally distributed as \( N(0; R) \)
Likelihood Computation

- An information-theoretic observation model defined as

\[ p(F, \Omega | B) \equiv \exp[I(F; \Omega)] \]

- \( I(F; \Omega) \) denotes mutual information (\( \geq 0 \)) between a feature vector set \( F \) and a class label set \( \Omega \)
  - Measures that how likely the feature vectors contribute to the discrimination between classes
  - Clearly reflects our intention of computing a discriminative power between classes based on the features extracted from the \( B \) bandpass-filtered and \( W \) spatially transformed signals
Mutual Information

\[ I(F_k; \Omega) = H(\Omega) - H(\Omega | F_k) = H(F_k) - H(F_k | \Omega) \]

- **Kernel density estimator** [KC02] for estimation of mutual information

\[
H(F_k) \approx -\frac{1}{D} \sum_{i=1}^{D} \log \left[ \frac{1}{D} \sum_{j=1}^{D} \varphi(f^i_k - f^j_k, \nu) \right]
\]

\[
H(F_k | \omega = c) \approx -\frac{1}{D} \sum_{i \text{ s.t. } \omega_i = c} \log \left[ \frac{1}{D_c} \sum_{j \text{ s.t. } \omega_j = c} \varphi(f^i_k - f^j_k, \nu) \right]
\]

- Gaussian kernel: \( \varphi(a, \nu) = \frac{1}{(2\pi)^{d/2} \nu^d |\Sigma|^{1/2}} \exp \left[ -\frac{a^T \Sigma^{-1} a}{2\nu^2} \right] \)
A Novel Bayesian Framework (recap.)

- **Data-driven optimal filter bank:** \( \hat{\mathbf{B}}_{\text{opt}} = \{ \mathbf{b}_k, \pi_k \}_{k=1}^K \)
  - Different bandwidth, possibly overlapped frequency bands
  - Spectral weighting factor \( \pi_k \)
Spectrally Weighted Classification

- Utilize the informative spectral weights $\{\pi_k\}_{k=1}^K$ in constructing a classifier
  - Compose a filter bank with the set of optimal frequency bands $S$ determined by the following rule
    $$S = \bigcup_k (\pi_k > \tau)$$
    - $\tau$: an empirically determined threshold
  - **Weighted linear combination** of the outputs from multiple classifiers

$$\hat{c} = \arg \max_{c \in \{+, -\}} \left\{ \sum_{k=1}^{|S|} \pi_k \cdot \phi_k^c \left( f_k^* \right) \right\}$$

- $f_k^*$: a feature vector from the input EEG $x^*$
- $\phi_k^c \left( f_k^* \right)$: a score of an SVM$_k$, which classifies $x^*$ into the class $c$
Three Public Datasets

- **Technische Universität (TU) Berlin Dataset**
  - Left-hand and right-hand
  - 140 trials per task, 51 electrodes

- **BCI Competition III Dataset-IVa**
  - Right-hand and right-foot, 5 healthy subjects
  - 140 trials per task, 128 electrodes
  - Imbalanced training and test dataset

- **BCI Competition IV Dataset-IIa**
  - Left-hand, right-hand, foot, and tongue
  - 288 trials per task in a session, 2 sessions
  - 22 electrodes, 9 subjects
  - Binary classification: left-hand vs. right-hand, left-hand vs. foot, etc.
Preprocessing and Hyperparameter Setting

- **Preprocessing**
  - Bandpass-filtering: 4Hz and 40Hz
  - **Small Laplacian derivation** for noise and artifacts reduction
  - **Baseline correction** by subtracting the mean of the samples before cue-onset

- **Particle initialization from a mixture of Gaussians**

  \[ p(B) = \frac{1}{2} N(\mu, \Sigma_{\mu}) + \frac{1}{2} N(\beta, \Sigma_{\beta}) \]

  - From neurophysiological knowledge
  - \( \mu \) and \( \beta \): \( \mu \)-rhythm(8-14Hz) and the \( \beta \)-rhythm(14-30Hz)
  - \( \Sigma_{\mu}, \Sigma_{\beta} \): covariance matrices, set to be diagonal
The improvement of classification accuracy slows down after 30 particles and the maximum accuracy was obtained with 30 particles. We used 30 particles for the rest of the experiments on not only this dataset but also the other two datasets.

Classification performance change over the number of particles in the experiments on Technische Universität Berlin Dataset.
Performance on TU Berlin Dataset

Classification accuracy (SD: Standard Deviation)

<table>
<thead>
<tr>
<th></th>
<th>CSP</th>
<th>CSSP</th>
<th>FBCSP</th>
<th>DCSP</th>
<th>OSSFN wFBCSP</th>
<th>OSSFN wDCSP</th>
<th>BSSFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (%)</td>
<td>76.79</td>
<td>93.57</td>
<td>75.07</td>
<td>79.36</td>
<td>84.36</td>
<td>84.37</td>
<td>97.57</td>
</tr>
<tr>
<td>SD (%)</td>
<td>11.75</td>
<td>2.21</td>
<td>4.86</td>
<td>5.52</td>
<td>5.05</td>
<td>5.06</td>
<td>1.22</td>
</tr>
</tbody>
</table>

- Competing state-of-the-art methods in the literature
  - CSP (Common Spatial Pattern) [RMGP00]
  - CSSP (Common Spatio-Spectral Pattern) [LBCM05]
  - FBCSP (Filter Bank CSP) [ACZG08]
  - DCSP (Discriminative FBCSP) [TGL+09]
  - OSSFN (Optimal Spatio-Spectral Filter Network) [ZCA+11]
Visualization of the distribution of the discriminative frequency band for the Technische Universität Berlin dataset by calculating probabilities between 4Hz and 40Hz at an interval of 0.5Hz. The optimized frequency bands are marked with black squares on it.
Performance on TU Berlin Dataset (cont.)

Topography maps of the learned spatial filter for TU Berlin Dataset

The proposed method has well found spatial filters consistent with the neurophysiological knowledge
## Performance on BCI Competition III-IVa

<table>
<thead>
<tr>
<th></th>
<th>aa</th>
<th>al</th>
<th>av</th>
<th>aw</th>
<th>ay</th>
<th>Mean (%)</th>
<th>SD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CSP</strong></td>
<td>66.96</td>
<td>89.29</td>
<td>52.55</td>
<td>47.77</td>
<td>52.38</td>
<td>61.79</td>
<td>16.98</td>
</tr>
<tr>
<td><strong>CSSP</strong></td>
<td>79.46</td>
<td>92.86</td>
<td>52.55</td>
<td>91.52</td>
<td>51.59</td>
<td>73.60</td>
<td>20.33</td>
</tr>
<tr>
<td><strong>FBCSP</strong></td>
<td>69.64</td>
<td>80.36</td>
<td>47.96</td>
<td>55.36</td>
<td>48.41</td>
<td>60.35</td>
<td>14.21</td>
</tr>
<tr>
<td><strong>DCSP</strong></td>
<td>69.64</td>
<td>82.14</td>
<td>54.08</td>
<td>50.89</td>
<td>48.41</td>
<td>61.03</td>
<td>14.41</td>
</tr>
<tr>
<td><strong>OSSFNwFBCSP</strong></td>
<td>75.00</td>
<td>83.93</td>
<td>53.06</td>
<td>74.11</td>
<td>48.41</td>
<td>66.98</td>
<td>15.22</td>
</tr>
<tr>
<td><strong>OSSFNwDCSP</strong></td>
<td>75.00</td>
<td>83.93</td>
<td>52.05</td>
<td>74.11</td>
<td>47.22</td>
<td>66.46</td>
<td>15.93</td>
</tr>
<tr>
<td><strong>BSSFO</strong></td>
<td>79.46</td>
<td>94.64</td>
<td>57.65</td>
<td>91.96</td>
<td>53.57</td>
<td>75.46</td>
<td>19.06</td>
</tr>
</tbody>
</table>
Performance on BCI Competition IV-IIa

The proposed BSSFO clearly outperforms other methods
How statistically significant are the classification performances?

<table>
<thead>
<tr>
<th>Technische Universität Berlin Dataset</th>
<th>Mean (SD)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSP</td>
<td>76.79 (11.75)</td>
<td>0.00034</td>
</tr>
<tr>
<td>CSSP</td>
<td>93.57 (2.21)</td>
<td>0.00082</td>
</tr>
<tr>
<td>FBCSP</td>
<td>75.07 (4.86)</td>
<td>0.000001</td>
</tr>
<tr>
<td>DCSP</td>
<td>79.36 (5.52)</td>
<td>0.00001</td>
</tr>
<tr>
<td>OSSFNwFBCSP</td>
<td>84.36 (5.05)</td>
<td>0.00003</td>
</tr>
<tr>
<td>OSSFNwDCSP</td>
<td>84.37 (5.06)</td>
<td>0.00003</td>
</tr>
<tr>
<td>Proposed method</td>
<td>97.57 (1.22)</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BCI Competition IV Dataset-IIa (Left-Right)</th>
<th>Mean (SD)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSP</td>
<td>73.46 (16.93)</td>
<td>0.00010</td>
</tr>
<tr>
<td>CSSP</td>
<td>79.78 (13.69)</td>
<td>0.02741</td>
</tr>
<tr>
<td>FBCSP</td>
<td>76.31 (17.90)</td>
<td>0.01693</td>
</tr>
<tr>
<td>DCSP</td>
<td>77.55 (18.33)</td>
<td>0.04450</td>
</tr>
<tr>
<td>OSSFNwFBCSP</td>
<td>76.31 (18.59)</td>
<td>0.01294</td>
</tr>
<tr>
<td>OSSFNwDCSP</td>
<td>75.62 (18.84)</td>
<td>0.02071</td>
</tr>
<tr>
<td>Proposed method</td>
<td>83.80 (13.09)</td>
<td>-</td>
</tr>
<tr>
<td>BCI Competition III Dataset-IVa</td>
<td>BCI Competition IV Dataset-IIa (Left-Foot)</td>
<td>BCI Competition IV Dataset-IIa (Left-Tongue)</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td><strong>Mean (SD)</strong></td>
<td><strong>p-value</strong></td>
<td><strong>Mean (SD)</strong></td>
</tr>
<tr>
<td>CSP</td>
<td>61.79 (16.98)</td>
<td>0.15652</td>
</tr>
<tr>
<td>CSSP</td>
<td>73.60 (20.33)</td>
<td>0.10603</td>
</tr>
<tr>
<td>FBCSP</td>
<td>60.35 (14.21)</td>
<td><strong>0.05319</strong></td>
</tr>
<tr>
<td>DCSP</td>
<td>61.03 (14.41)</td>
<td>0.10300</td>
</tr>
<tr>
<td>OSSFNwFBCSP</td>
<td>66.98 (15.22)</td>
<td><strong>0.03206</strong></td>
</tr>
<tr>
<td>OSSFNwDCSP</td>
<td>66.46 (15.93)</td>
<td><strong>0.02145</strong></td>
</tr>
<tr>
<td>Proposed method</td>
<td>75.46 (19.06)</td>
<td>-</td>
</tr>
<tr>
<td>CSP</td>
<td>72.76 (17.93)</td>
<td><strong>0.00134</strong></td>
</tr>
<tr>
<td>OSSFNwFBCSP</td>
<td>80.94 (14.18)</td>
<td><strong>0.00314</strong></td>
</tr>
<tr>
<td>OSSFNwDCSP</td>
<td>84.11 (12.15)</td>
<td><strong>0.00070</strong></td>
</tr>
<tr>
<td>Proposed method</td>
<td>89.89 (10.71)</td>
<td>-</td>
</tr>
<tr>
<td>CSP</td>
<td>73.07 (15.82)</td>
<td><strong>0.00136</strong></td>
</tr>
<tr>
<td>OSSFNwFBCSP</td>
<td>75.54 (16.71)</td>
<td><strong>0.00127</strong></td>
</tr>
<tr>
<td>Proposed method</td>
<td>87.89 (11.29)</td>
<td>-</td>
</tr>
<tr>
<td>CSP</td>
<td>72.38 (17.83)</td>
<td><strong>0.00389</strong></td>
</tr>
<tr>
<td>FBCSP</td>
<td>80.32 (12.89)</td>
<td><strong>0.01354</strong></td>
</tr>
<tr>
<td>OSSFNwFBCSP</td>
<td>79.01 (10.31)</td>
<td><strong>0.02682</strong></td>
</tr>
<tr>
<td>Proposed method</td>
<td>89.04 (11.43)</td>
<td>-</td>
</tr>
<tr>
<td>CSP</td>
<td>67.28 (11.40)</td>
<td><strong>0.00128</strong></td>
</tr>
<tr>
<td>OSSFNwFBCSP</td>
<td>75.08 (12.39)</td>
<td><strong>0.03021</strong></td>
</tr>
<tr>
<td>Proposed method</td>
<td>81.10 (9.73)</td>
<td>-</td>
</tr>
</tbody>
</table>
Effects of # of Particles in Label Decision

Evolution of the classification accuracies according to the number of particles used in the label decision. The number of particles for each subject is considered up to the size of the filter bank that resulted in the performances presented above.

We could achieve the same or better performance with a smaller number of particles than the threshold-determined filter bank. That means we could further reduce the computational cost for those subjects in our BCI system.
Conclusions

- Motor-imagery based brain-computer interfaces
  - Simultaneous optimization of spectral and spatial filter in a unified framework

- Main contributions
  - **Efficient Bayesian integration** of spectral filter optimization and spatial filter learning
  - A **factored-sampling-based approximation method** for posterior $pdf$ estimation and an **information-theoretic observation model** to measure discriminative power of features between classes
  - A **spectrally-weighted label decision rule** by linearly combining the outputs from multiple classifiers, one for each frequency band
Further Issues

• Applicable to other kinds of single trial EEG classification problems, which are based on modulations of brain rhythms, by no means limited to motor imager-based BCIs

• Optimal discriminative time-segment selection

• Incorporating the problem of task-related channel selection into the proposed Bayesian framework
Thank you for your attention