Learning optimal EEG features across time, frequency and space.

Jason Farquhar, Jeremy Hill, Bernhard Schölkopf

Max Planck Institute for Biological Cybernetics, Tübingen
Germany

NIPS06 Workshop on Trends in BCI

Outline

Motivation

Source types in EEG based BCI

Automatic Feature Selection

Learning Spatial Features

Feature selection as Model Selection

Spectral/Temporal Filtering

The current approach to learning in BCIs

Current BCI use learning in two distinct phases,

1. Feature Extraction – where we attempt to extract features which lead to good classifier performance,
2. Classification – usually a simple linear classifier, (SVM, LDA, Gaussian), because

> *"Once we have good features the classifier doesn't really matter"*

The current approach to learning in BCIs

Current BCI use learning in two distinct phases,

1. Feature Extraction – where we attempt to extract features which lead to good classifier performance, using,
   - prior-knowledge, the 7-30Hz band for ERDs
   - maximising r-scores,
   - maximising 'independence' (ICA)
   - maximising the ratios of the class variances (CSP)
2. Classification – usually a simple linear classifier, (SVM, LDA, Gaussian), because

   "Once we have good features the classifier doesn't really matter"

The current approach to learning in BCIs

Current BCI use learning in two distinct phases,

1. Feature Extraction – where we attempt to extract features which lead to good classifier performance,
2. Classification – usually a simple linear classifier, (SVM, LDA, Gaussian), because

> *"Once we have good features the classifier doesn't really matter"*

The current approach to learning in BCIs

This seems wrong!

Note

The objectives used in feature extraction are not good predictors of generalisation performance.

Question?

Why, use an objective for the important feature extraction which is a poor predictor of generalisation performance?
When we have provably good predictors (margin, evidence) available in the unimportant classifier?

The current approach to learning in BCIs

This seems wrong!

Note

The objectives used in feature extraction are not good predictors of generalisation performance.

Question?

Why, use an objective for the important feature extraction which is a poor predictor of generalisation performance?
When we have provably good predictors (margin, evidence) available in the unimportant classifier?

The current approach to learning in BCIs

This seems wrong!

Note

The objectives used in feature extraction are not good predictors of generalisation performance.

Question?

Why, use an objective for the important feature extraction which is a poor predictor of generalisation performance?

When we have provably good predictors (margin, evidence) available in the unimportant classifier?

The current approach to learning in BCIs

This seems wrong!

Note

The objectives used in feature extraction are not good predictors of generalisation performance.

Question?

Why, use an objective for the important feature extraction which is a poor predictor of generalisation performance?
When we have provably good predictors (margin, evidence) available in the unimportant classifier?

A Better approach

1. Combine the feature extraction and classifier learning
2. Choose features which optimise the classifier's objective

We show how to learn spatio-spectro-temporal feature extractors for classifying ERDs using the max-margin criterion[1]

[1] We have also successfully applied this approach to LR and GP classifiers and MRP/P300 temporal signals.

A Better approach

1. Combine the feature extraction and classifier learning
2. Choose features which optimise the classifier's objective

We show how to learn spatio-spectro-temporal feature extractors for classifying ERDs using the max-margin criterion[1]

[1] We have also successfully applied this approach to LR and GP classifiers and MRP/P300 temporal signals

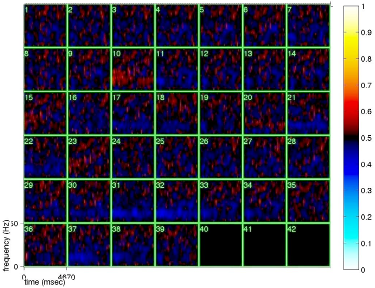
A Better approach

1. Combine the feature extraction and classifier learning
2. Choose features which optimise the classifier's objective

We show how to learn spatio-spectro-temporal feature extractors for classifying ERDs using the max-margin criterion[1]

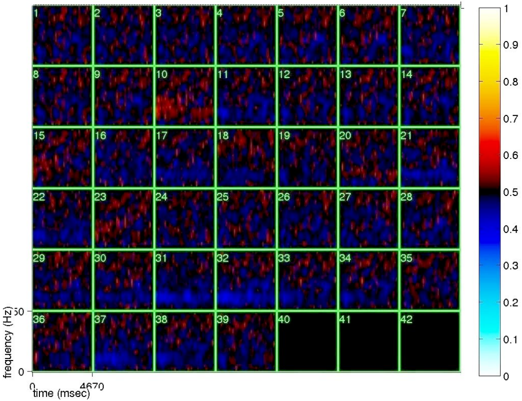
[1] We have also successfully applied this approach to LR and GP classifiers and MRP/P300 temporal signals)

Data-visualisation: the ROC-ogram



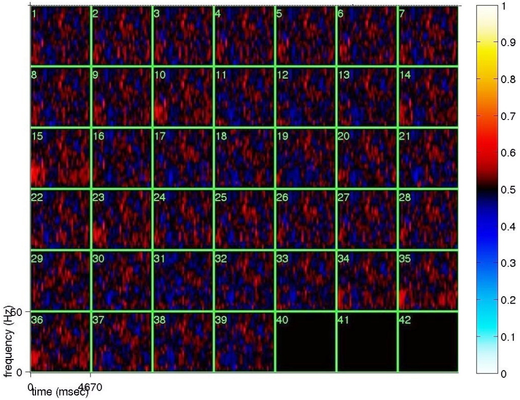
- Raw data is *dxT* time-series for *N* trials
- ROC-ogram : time vs. frequency vs. ROC score for each channel
- allows us to identify where the discriminative information lies

Example raw ROC-ogram: (The good)



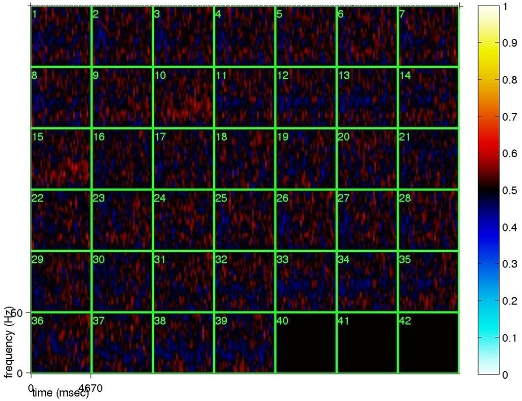
Spatial, Spectral, and Temporal discriminative features are subject specific

Example raw ROC-ogram: (The bad)



Spatial, Spectral, and Temporal discriminative features are subject specific

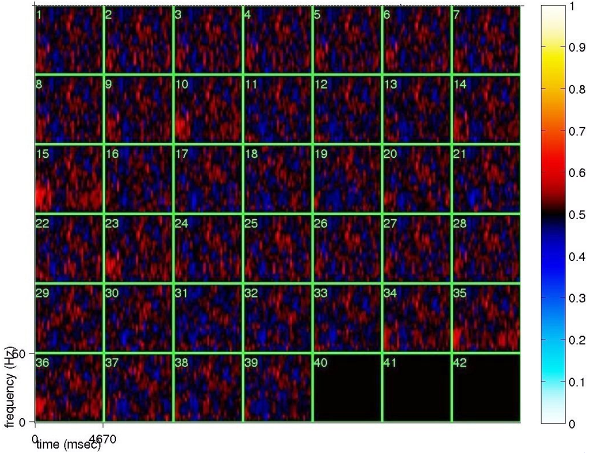
Example raw ROC-ogram: (The ugly)



Spatial, Spectral, and Temporal discriminative features are subject specific

Spatio-Spectro-Temporal feature selection

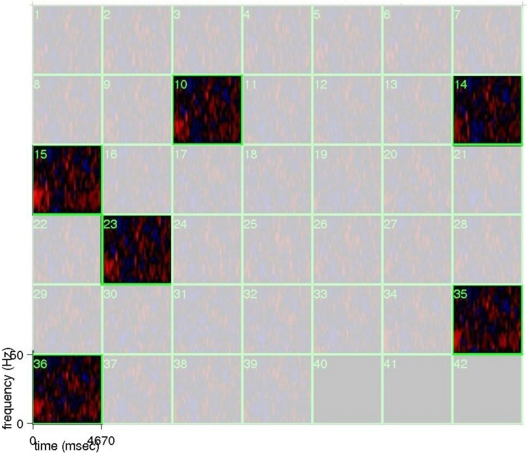
Would like to automatically perform feature selection:



- Spatially,
- Temporally,
- Spectrally

Spatio-Spectro-Temporal feature selection

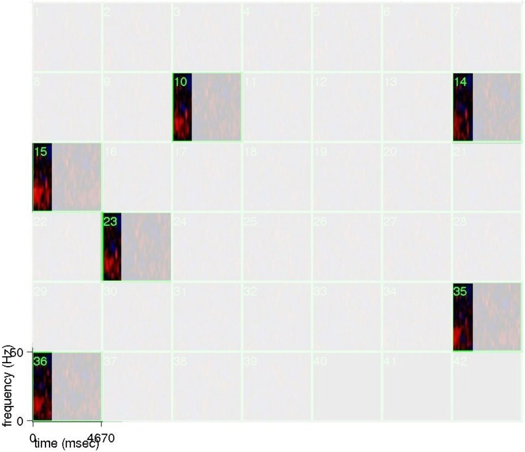
Would like to automatically perform feature selection:



- Spatially,
- Temporally,
- Spectrally

Spatio-Spectro-Temporal feature selection

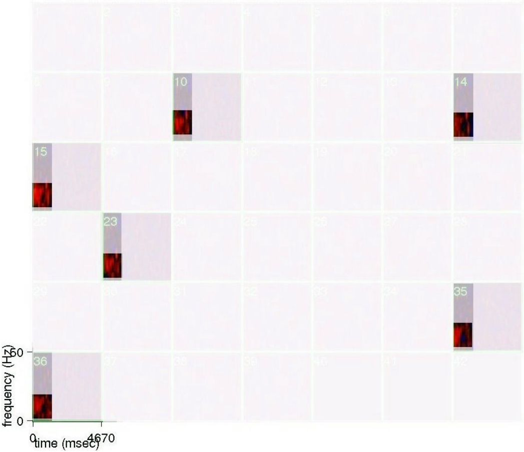
Would like to automatically perform feature selection:



- Spatially,
- Temporally,
- Spectrally

Spatio-Spectro-Temporal feature selection

Would like to automatically perform feature selection:

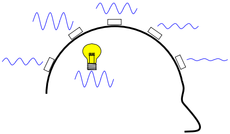


- Spatially,
- Temporally,
- Spectrally

Learning Feature Extractors

- Start by showing how to learning spatial filters with the max-margin criteria,
- Then extend to learning spatial+spectral+temporal

Spatial Filtering



- Volume Conduction – electrodes detect superposition of signals from all over the brain

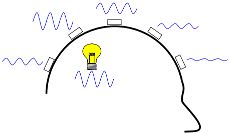
$$X = AS$$

- Spatial filtering undoes this superposition to re-focus on discriminative signals

$$\mathbf{y} = \mathbf{f}_s^\top X$$

- This is a Blind Source Separation (BSS) problem many algorithms available to solve this problem
- In BCI commonly use a fast, supervised method called Common Spatial Patterns [Koles 1990]

Spatial Filtering



- Volume Conduction – electrodes detect superposition of signals from all over the brain

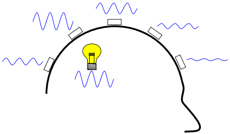
$$X = AS$$

- Spatial filtering undoes this superposition to re-focus on discriminative signals

$$\mathbf{y} = \mathbf{f}_s^\top X$$

- This is a Blind Source Separation (BSS) problem many algorithms available to solve this problem
- In BCI commonly use a fast, supervised method called Common Spatial Patterns [Koles 1990]

Spatial Filtering



- Volume Conduction – electrodes detect superposition of signals from all over the brain

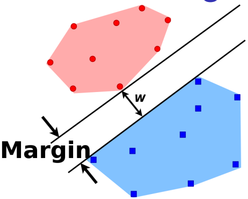
$$X = AS$$

- Spatial filtering undoes this superposition to re-focus on discriminative signals

$$\mathbf{y} = \mathbf{f}_s^\top X$$

- This is a Blind Source Separation (BSS) problem many algorithms available to solve this problem
- In BCI commonly use a fast, supervised method called Common Spatial Patterns [Koles 1990]

The Max-margin Objective



- related to an upper bound on generalisation performance
- the basis for the SVM
- finds w such that the minimal distance between classes is maximised
- in the linear case can be expressed primal objective as,

$$\min_{w,b} \lambda w^\top w + \sum_i max(0, 1 - y_i(x_i^\top w + b))$$

- for non-linear classification we can simply replace x_i with an explicit feature mapping $\psi(x_i)$
- this is how we include the feature extraction into the classification objective

Max-margin optimised spatial filters

1. Define the feature-space mapping, ψ, from time series, X_i to spatially filtered log bandpowers,

$$\psi(X_i, F_s) = \ln(\texttt{diag}({F_s}^\top X_i X_i^\top F_s))$$

where, $F_s = [\mathbf{f}_{s_1}, \mathbf{f}_{s_2}, \dots]$ is the set of spatial filters

2. Include the dependence on ψ explicitly into the classifiers objective, e.g. Linear, Max Margin

$$J_{mm}(X, \mathbf{w}, b, F_s) = \lambda \mathbf{w}^\top \mathbf{w} + \sum_i \max(0, 1 - y_i(\psi(X_i; F_s)^\top \mathbf{w} + b))$$

3. Optimise this objective, treating ψ's parameters as additional optimisation variables

Note

Unconstrained optimisation, solve for $\mathbf{w}, b, F_s$ directly using CG

Max-margin optimised spatial filters

1. Define the feature-space mapping, ψ, from time series, X_i to spatially filtered log bandpowers,

$$\psi(X_i, F_s) = \ln(\mathtt{diag}({F_s}^\top X_i X_i^\top F_s))$$

where, $F_s = [\mathbf{f}_{s_1}, \mathbf{f}_{s_2}, \ldots]$ is the set of spatial filters

2. Include the dependence on ψ explicitly into the classifiers objective, e.g. Linear, Max Margin

$$J_{mm}(X, \mathbf{w}, b, F_s) = \lambda \mathbf{w}^\top \mathbf{w} + \sum_i \max(0, 1 - y_i(\psi(X_i; F_s)^\top \mathbf{w} + b))$$

3. Optimise this objective, treating ψ's parameters as additional optimisation variables

Note

Unconstrained optimisation, solve for $\mathbf{w}$, b, F_s directly using CG

Max-margin optimised spatial filters

1. Define the feature-space mapping, ψ, from time series, X_i to spatially filtered log bandpowers,

$$\psi(X_i, F_s) = \ln(\mathtt{diag}(F_s{}^\top X_i X_i^\top F_s))$$

where, $F_s = [\mathbf{f}_{s_1}, \mathbf{f}_{s_2}, \ldots]$ is the set of spatial filters

2. Include the dependence on ψ explicitly into the classifiers objective, e.g. Linear, Max Margin

$$J_{mm}(X, \mathbf{w}, b, F_s) = \lambda \mathbf{w}^\top \mathbf{w} + \sum_i \max(0, 1 - y_i(\psi(X_i; F_s)^\top \mathbf{w} + b))$$

3. Optimise this objective, treating ψ's parameters as additional optimisation variables

Note

Unconstrained optimisation, solve for $\mathbf{w}$, b, F_s directly using CG

Adding Spectral/Temporal filters

Very simple to include Spectral/Temporal filtering,....

.....just modify the feature-mapping ψ to include them.

Let, $\mathbf{f}_f$ be a spectral filter, and $\mathbf{f}_t$ a temporal filter. Then, the Spatial + Spectral + Temporally filtered band-power is,

$$\begin{aligned} \psi(X; \mathbf{f}_s, \mathbf{f}_f, \mathbf{f}_t) &= \mathcal{F}^{-1}(\mathcal{F}({\mathbf{f}_s}^\top X D_t) D_f)(\mathcal{F}^{-1}(\mathcal{F}({\mathbf{f}_s}^\top X D_t) D_f))^\top \\ &= {\mathbf{f}_s}^\top \mathcal{F}(X D_t) D_f^2 \mathcal{F}(X D_t)^\top {\mathbf{f}_s}^\top / T \end{aligned}$$

where, $\mathcal{F}$ is the Fourier transform, and $D_{(.)} = \mathtt{diag}(\mathbf{f}_{(.)})$

Adding Spectral/Temporal filters

Very simple to include Spectral/Temporal filtering,....
.....just modify the feature-mapping ψ to include them.

Let, $\mathbf{f}_f$ be a spectral filter, and $\mathbf{f}_t$ a temporal filter. Then, the Spatial + Spectral + Temporally filtered band-power is,

$$\begin{aligned}\psi(X; \mathbf{f}_s, \mathbf{f}_f, \mathbf{f}_t) &= \mathcal{F}^{-1}(\mathcal{F}(\mathbf{f}_s{}^\top X D_t) D_f)(\mathcal{F}^{-1}(\mathcal{F}(\mathbf{f}_s{}^\top X D_t) D_f))^\top \\ &= \mathbf{f}_s{}^\top \mathcal{F}(X D_t) D_f^2 \mathcal{F}(X D_t)^\top \mathbf{f}_s{}^\top / T\end{aligned}$$

where, $\mathcal{F}$ is the Fourier transform, and $D_{(.)} = \mathtt{diag}(\mathbf{f}_{(.)})$

Adding Spectral/Temporal filters

Very simple to include Spectral/Temporal filtering,....

.....just modify the feature-mapping ψ to include them.

Let, $\mathbf{f}_f$ be a spectral filter, and $\mathbf{f}_t$ a temporal filter. Then, the Spatial + Spectral + Temporally filtered band-power is,

$$\begin{aligned}\psi(X;\mathbf{f}_s,\mathbf{f}_f,\mathbf{f}_t) &= \mathcal{F}^{-1}(\mathcal{F}({\mathbf{f}_s}^\top X D_t)D_f)(\mathcal{F}^{-1}(\mathcal{F}({\mathbf{f}_s}^\top X D_t)D_f))^\top \\ &= {\mathbf{f}_s}^\top \mathcal{F}(X D_t) D_f^2 \mathcal{F}(X D_t)^\top {\mathbf{f}_s}^\top / T\end{aligned}$$

where, $\mathcal{F}$ is the Fourier transform, and $D_{(.)} = \mathtt{diag}(\mathbf{f}_{(.)})$

Filter regularisation

- The filters, F_s, F_f, F_t are unconstrained so may overfit
- We have prior knowledge about the filters shape, e.g.
  - spatial filters tend to be over the motor regions
  - temporal and spectral filters should be smooth
- Include this prior knowledge with quadratic regularisation on the filters,

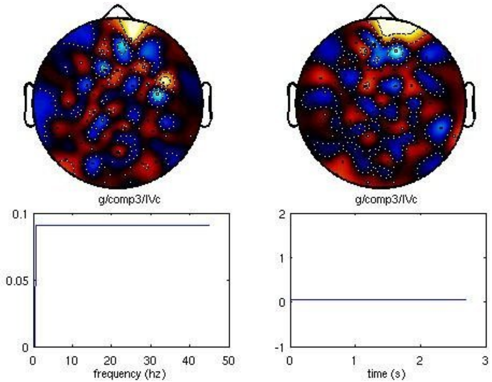
$$J_{mm} = \lambda \mathbf{w}^\top \mathbf{w} + \sum_i \max(0, 1 - y_i(\ln(\psi(X_i; F_s, F_f, F_t))^\top \mathbf{w} + b)) + \lambda_s \mathtt{Tr}(F_s^\top R_s F_s) + \lambda_f \mathtt{Tr}(F_f^\top R_f F_f) + \lambda_t \mathtt{Tr}(F_t^\top R_t F_t)$$

where, $R_{(.)}$ is a positive definite matrix encoding the prior knowledge

Implementation issues

- Optimising J_{mm} for all the filters directly, results in a "stiff" problem and very slow convergence
- Further, evaluating $\psi(X; \mathbf{f}_s, \mathbf{f}_f, \mathbf{f}_t)$ requires a costly FFT
- Coordinate descent on the filter types solves both these problems,
  1. Spatial optimisation, where, $\psi_s(X, \mathbf{f}_s) = \mathbf{f}_s^\top X_{f,t} X_{f,t}^\top \mathbf{f}_s$
  2. Spectral optimisation, where $\psi(X; \mathbf{f}_f) = \widetilde{X}_{s,t} D_f^2 \widetilde{X}_{s,t}^\top$
  3. Temporal optimisation, where $\psi_t(X, \mathbf{f}_t) = X_{s,f} D_t^2 X_{s,f}^\top$
  4. Repeat until convergence
- Non-convex problem – seed with good solution found by another method, e.g. CSP or prior knowledge.

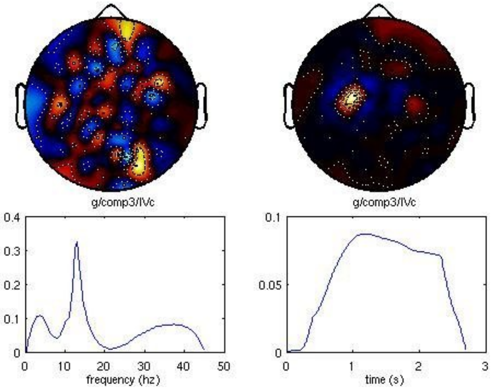
Example – Optimisation trajectory for CompIII,Vc



Iteration 0: 45% Error

- Noisy CSP seed
- Finds:
  - motor region,
  - 15Hz band,
  - >.5s temporal band
- Finds foot region

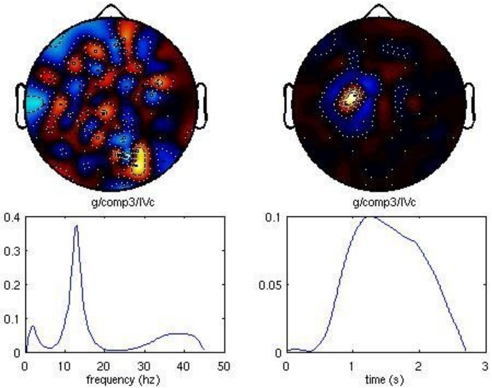
Example – Optimisation trajectory for CompIII,Vc



Iteration 1: 16% Error

- Noisy CSP seed
- Finds:
  - motor region,
  - 15Hz band,
  - $>$.5s temporal band
- Finds foot region

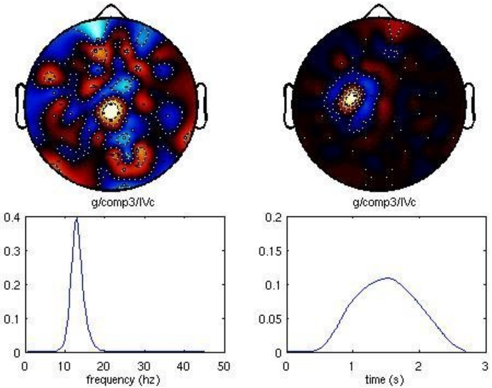
Example – Optimisation trajectory for CompIII,Vc



Iteration 2: 3.5% Error

- Noisy CSP seed
- Finds:
  - motor region,
  - 15Hz band,
  - >.5s temporal band
- Finds foot region

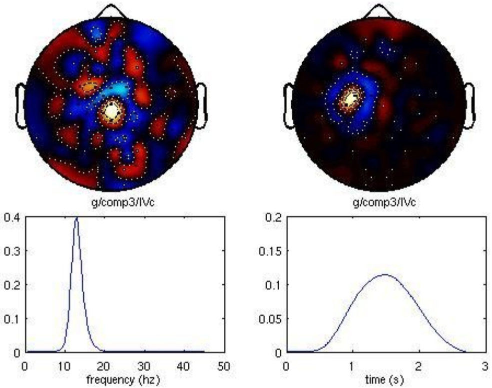
Example – Optimisation trajectory for CompIII,Vc



Iteration 3: 3.5% Error

- Noisy CSP seed
- Finds:
  - motor region,
  - 15Hz band,
  - >.5s temporal band
- Finds foot region

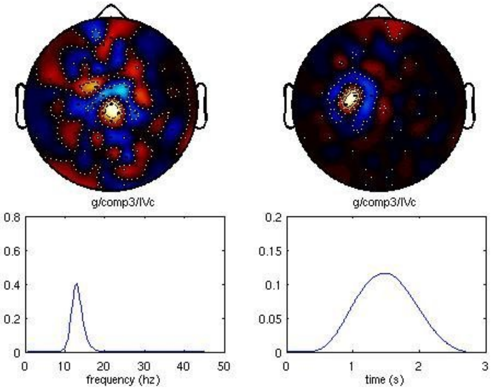
Example – Optimisation trajectory for CompIII,Vc



Iteration 4: 2.6% Error

- Noisy CSP seed
- Finds:
  - motor region,
  - 15Hz band,
  - >.5s temporal band
- Finds foot region

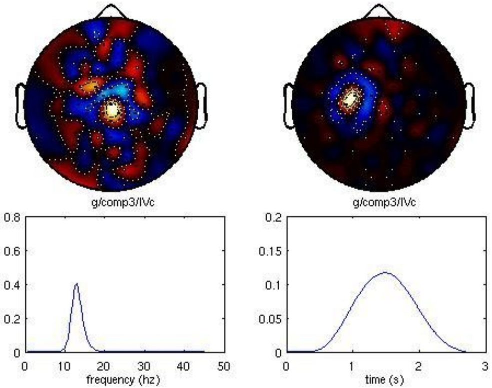
Example – Optimisation trajectory for CompIII,Vc



Iteration 5: 2.6% Error

- Noisy CSP seed
- Finds:
  - motor region,
  - 15Hz band,
  - >.5s temporal band
- Finds foot region

Example – Optimisation trajectory for CompIII,Vc



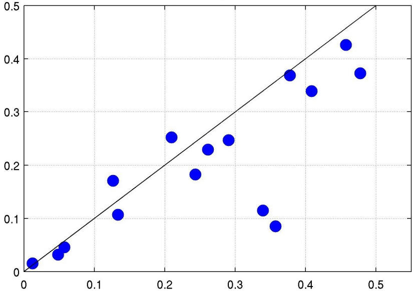
Iteration 6: 2.6% Error

- Noisy CSP seed
- Finds:
  - motor region,
  - 15Hz band,
  - $>$.5s temporal band
- Finds foot region

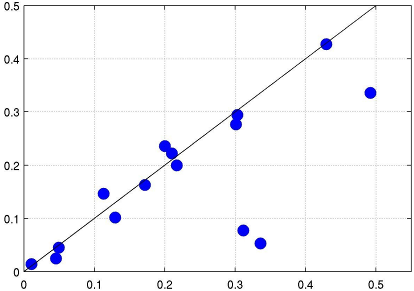
Experimental analysis

- We show binary classification error from 15 imagined movement subjects:
  - 9 from BCI competitions (Comp 2:IIa, Comp 3:IVa,IVc) and
  - 6 from an internal MPI dataset.
- pre-processed by band-pass filtering to .5–45Hz
- Baseline performance is from CSP with 2 filters computed on the signal filtered to 7-27Hz.
- CSP solution used as the spatial filter seed,
- flat seeds used for spectral and temporal filters

Results – Spatial Optimization



100 Training Points

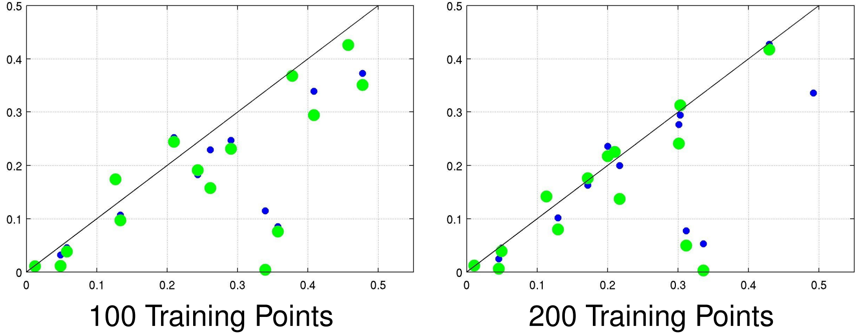


200 Training Points

Spatial

- General Improvement in performance; particularly for low numbers of data-points (when overfitting is an issue)
- Huge improvement in a few cases

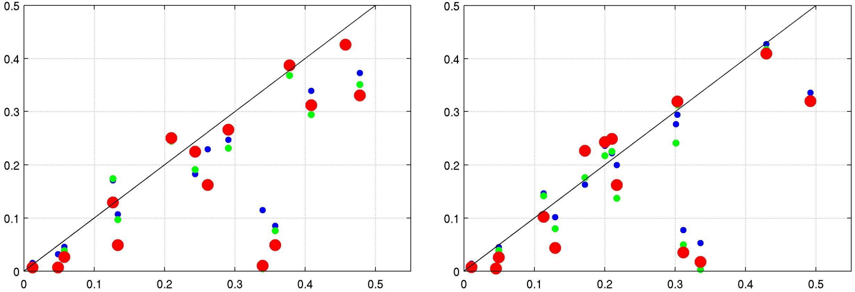
Results – Spatial + Spectral Optimization



Spatial + Spectral

- Further improvement, for the subjects helped before
- large benefit in a few cases

Results – Spatial + Spectral + Temporal Optimization



100 Training Points | 200 Training Points

Spatial + Spectral + Temporal

- Further improvements for some subjects
- slight decrease for others

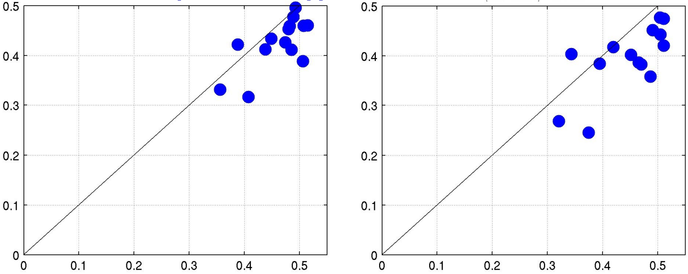
Summary

- EEG BCI performance depends mainly on learning subject-specific feature extractors
- These can be learnt by direct optimisation of the classification objective (Max-margin)
- Results show significant improvement over independent feature-extractor/classifier learning (better in 12/15 cases)

Future work

- Alternative objective functions – SVM, LR and Gaussian Process objectives implementated already.
- Better priors – particularly for the spatial filters, found by cross-subject learning?
- Other feature/signal types – wavelets, MRPs, P300, etc.
- On-line feature learning

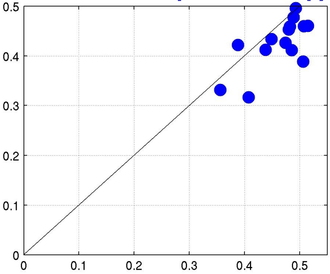
Results – Temporal Signal Extraction



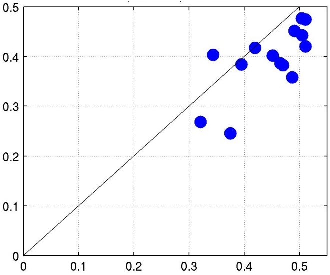
100 Training Points   200 Training Points

- learn a rank-1, i.e. 1-spatial + 1-temporal, approximation to the full svm weight vector
- this regularisation significantly improves classification performance
- and produces readily interpertable results..

Results – Temporal Signal Extraction



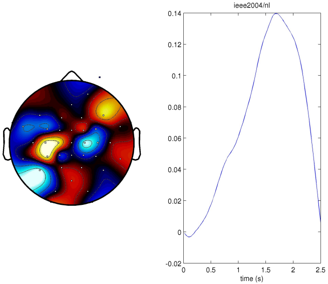
100 Training Points

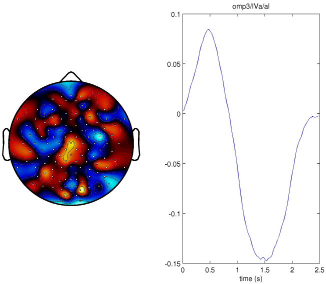


200 Training Points

- learn a rank-1, i.e. 1-spatial + 1-temporal, approximation to the full svm weight vector
- this regularisation significantly improves classification performance
- and produces readily interpertable results..

Results – Example solutions





- spatially – differential filter between left/right motor regions
- temporally?

- spatially – differential filter between foot and motor regions
- temporally?