Apprentissage statistique appliquée aux interfaces cerveau-machine

A. Rakotomamonjy

LITIS EA 4108
Université de ROUEN

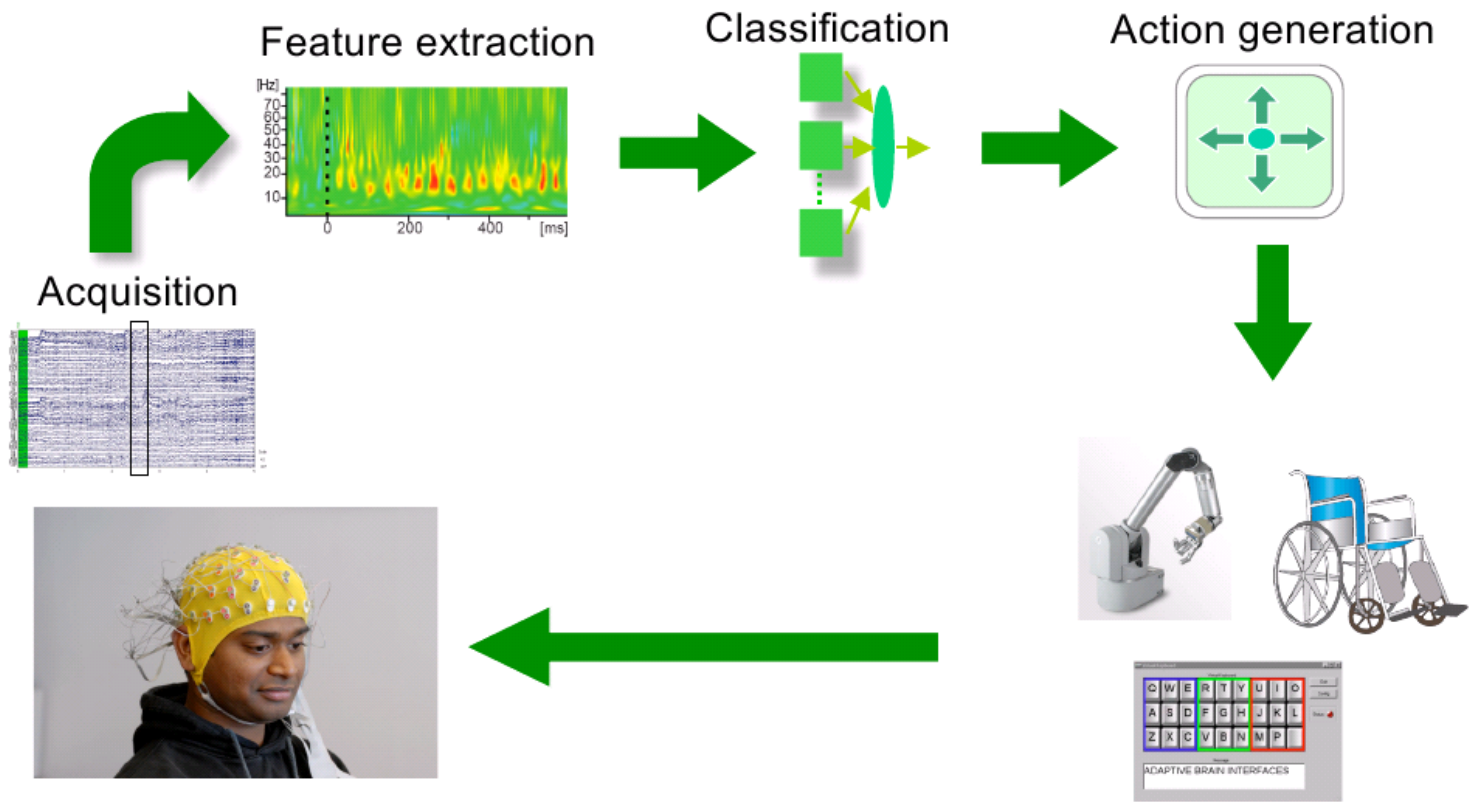
Peyresq 2010

Roadmap

1. Introducing BCI
2. On the role of Machine Learning for BCI
3. Subject-specific approaches
   - Addressing the P300 Speller
   - Learning Spatial Filter for BCI motor imagery
4. Subject-independent : zero training approaches
5. Conclusions and challenges

What is a Brain-Computer Interface ?

BCI : A communication and control system that does not consider usual brain's normal output pathways like muscles or peripherical nerves. It translates human intentions or response to a stimulus into a command.



1

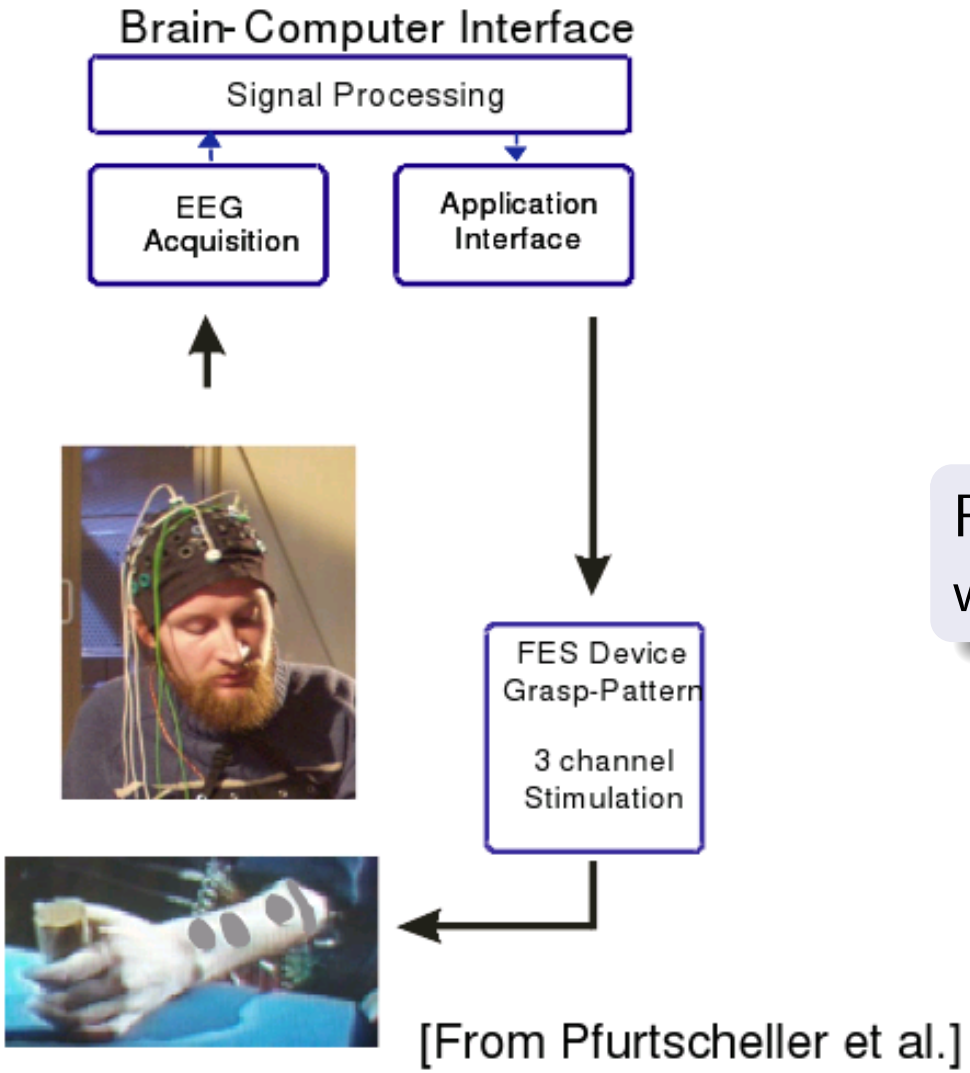
1. from Chavarriaga, 2009

Clinical applications

Help patient with Lock-In Syndrome to communicate with their environments through BCI Speller

[From Nijolt et al.]

Clinical applications



Robotic arm command for patients with amputed arm

Gaming applications

game control without keyboard, joystick nor wiimote

from Nijholt et al.

Different kinds of BCI

- invasive
  - sensors implanted
- asynchronous
  - BCI outputs are generated at user's will
- unstimulated
  - user voluntarily produces the required signals

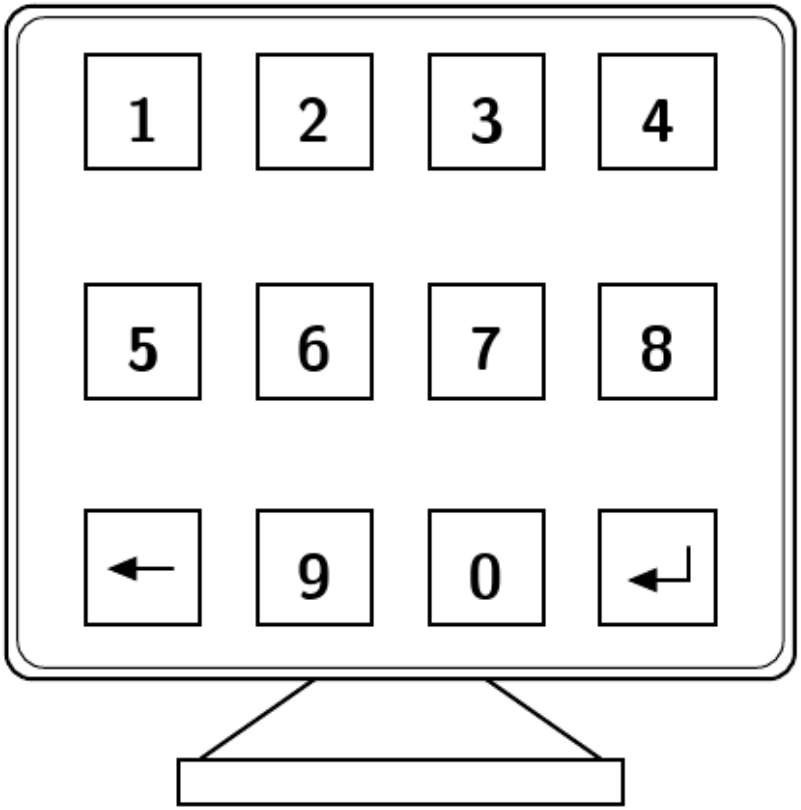
- non-invasive
  - surface sensors : EEG
- synchronous
  - BCI outputs are synchronyzed with a stimulus generator .
- evoked potential
  - requires the user to focus on stimuli to produce change in brain responses

Example : BCI P300 Speller

- Experimental set-up proposed by Donchin et al.
- 6×6 matrix of symbols
- a sequence : each row and column flashes in a random order
- subject concentrates on a symbol and counts as it flashes
- visual P300 elicited when the symbol to spell flashes
- several sequences needed for spelling a single symbol

synchronous, evoked potential

Example : BCI SSEP



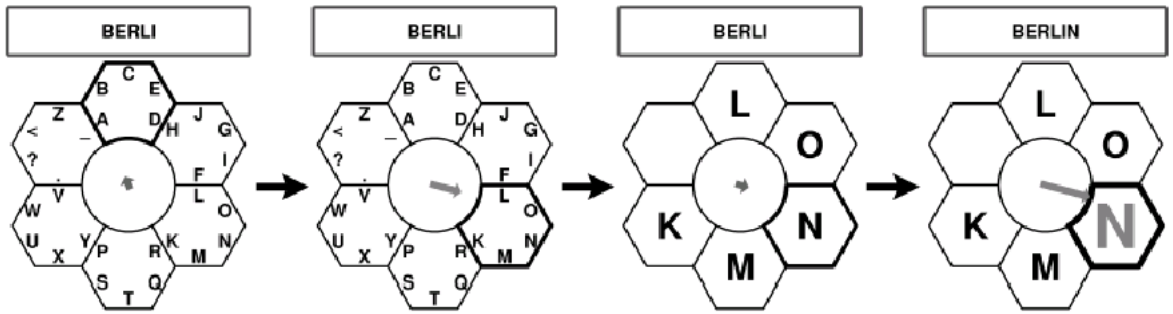
- stimulus presented repetitively at high rate : each symbol flashes at individual frequency
- task : patient looks at the desired symbol
- looking at the stimulus blinking evokes a rythm of same frequency in the visual cortex
- spectral analysis of EEG

[Cheng et al. 2002]

Independent, Asynchronous, Evoked potential

Example : BCI Motor Imagery Hex-O-Speller

- Speller based on imagined movement of hands and feets. (developed by Blankertz et al.)
- μ-rythm amplitudes vary with movements, movements preparation, or movement imagery
- turn the hex-o-speller with imagined left and right hand movements and selection with imagined feet movement



Independent, Asynchronous, Unstimulated

BCI approaches

Operant conditioning

- user learns to voluntary change features of EEG
- need user's training based on some feedback related to EEG

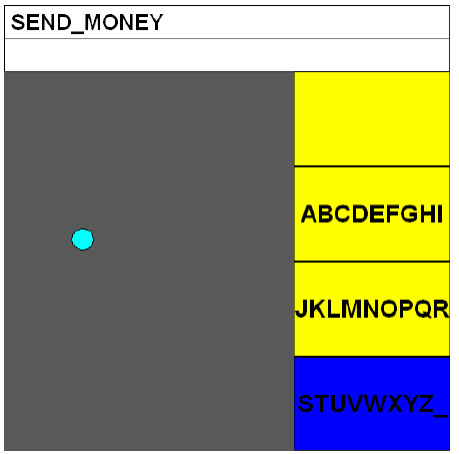
Machine-based

- machine-based detection of user's mental state
- need to produce brain signals according to requested mental states
- train a pattern recognition classifier to recognize mental states

Examples of two μ-rythm based spellers

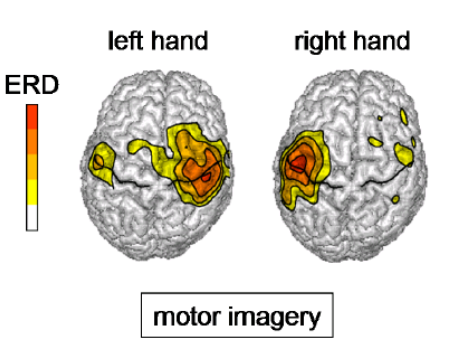
Albany's one. Wolpaw et al.

- user controls his $\mu-$ rhythm
- ball travels horizontally at constant speed
- vertical movement controlled by power of μ-rythm



Hex-O-Speller

- $\mu-$ rhythm controlled speller
- mental states are characterized by modulation in the $\mu-$ rhythm.



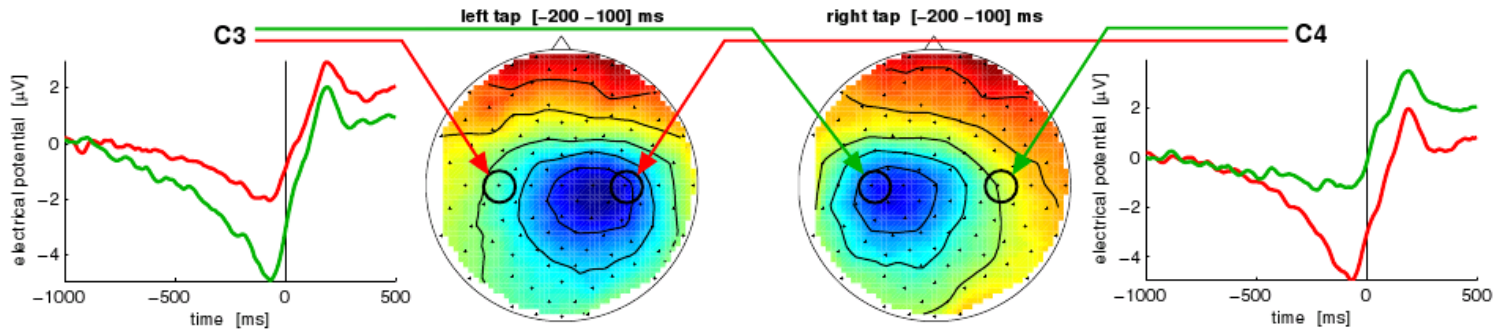
Detecting mental states using machine learning

But do we need a machine learning approach ?

- neurophysiology of BCI paradigms are well-known (P300, motor imagery, $\cdots$).
- Possible to extract features from EEG that discriminate mental states

Example in motor imagery : Lateralized Readiness Potential

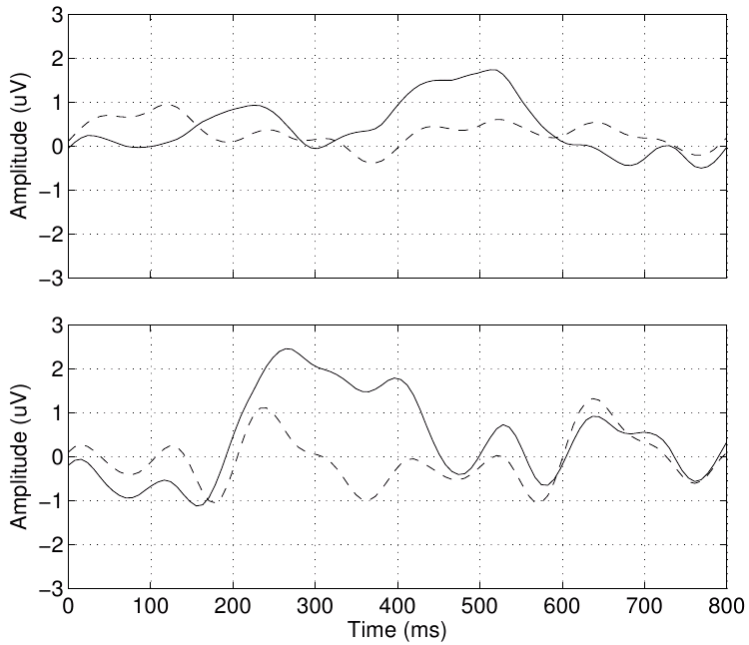
LRP : variation in electrical activity at the surface of the brain, that reflect the preparation of motor activity on a certain side of the body



Other examples of well-known neurophysiology results

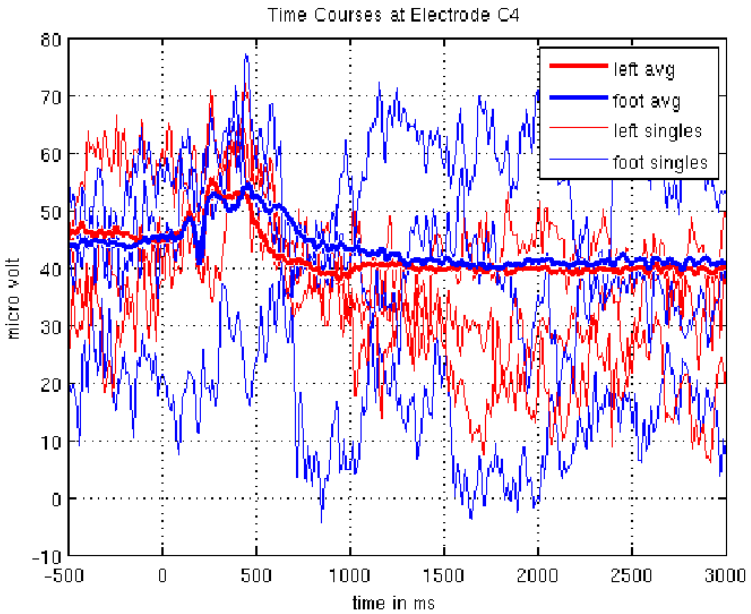
Averaged P300 Oddball

averaged waveforms on Pz for disabled (top) and enabled (bottom) people.



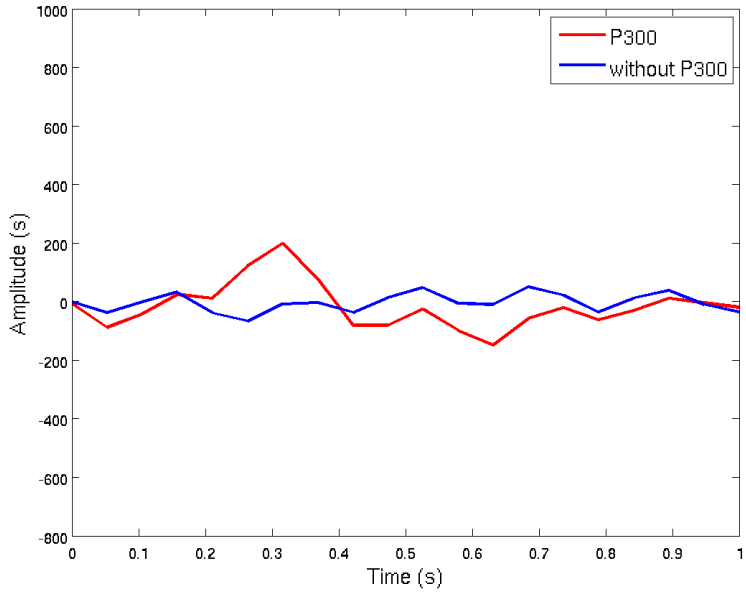
Why do we need machine learning ?

- Neurophysiology gives information only about the average brain
- EEG Signals have a low signal-to-noise ratio
- BCI paradigm considers only single specific brain but
  - intra-subject variability
  - inter-subject variability



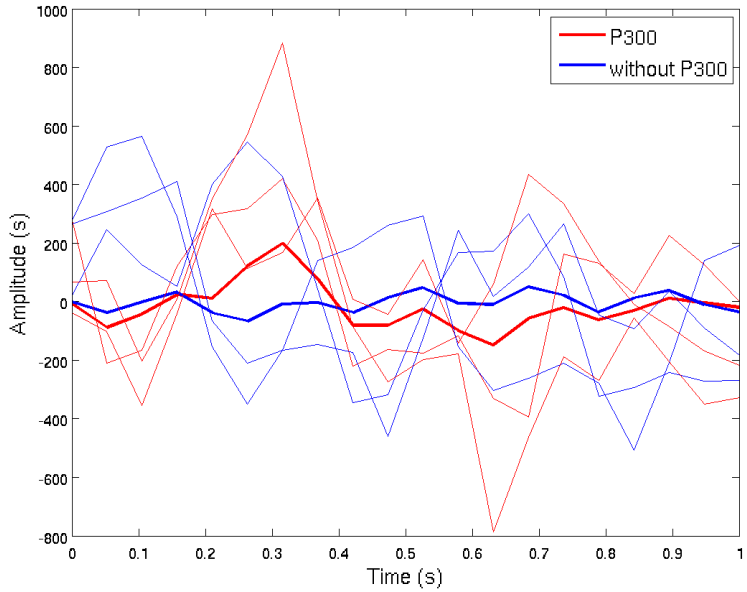
Why do we need machine learning ?

- Neurophysiology gives information only about the average brain
- EEG Signals have a low signal-to-noise ratio
- BCI paradigm considers only single specific brain but
  - intra-subject variability
  - inter-subject variability



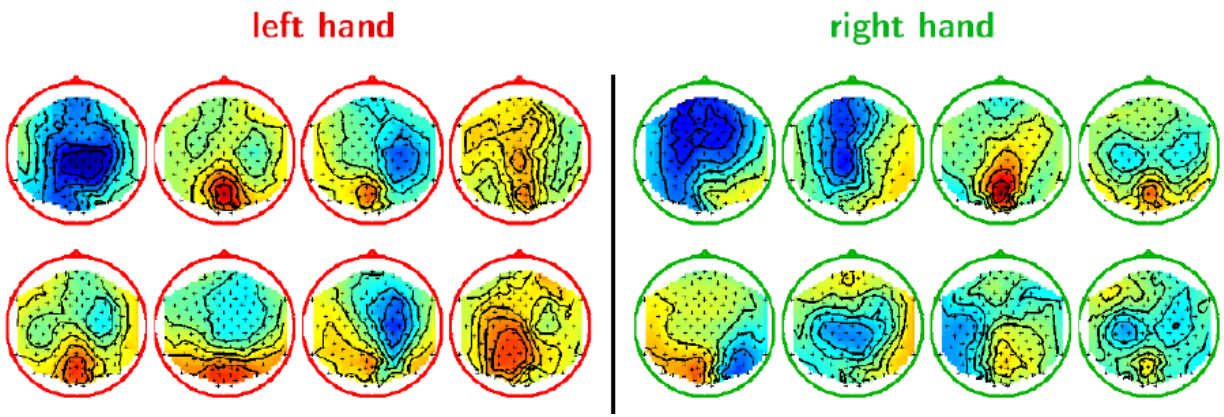
Why do we need machine learning ?

- Neurophysiology gives information only about the average brain
- EEG Signals have a low signal-to-noise ratio
- BCI paradigm considers only single specific brain but
  - intra-subject variability
  - inter-subject variability



BCI Motor imagery : variability in ERD

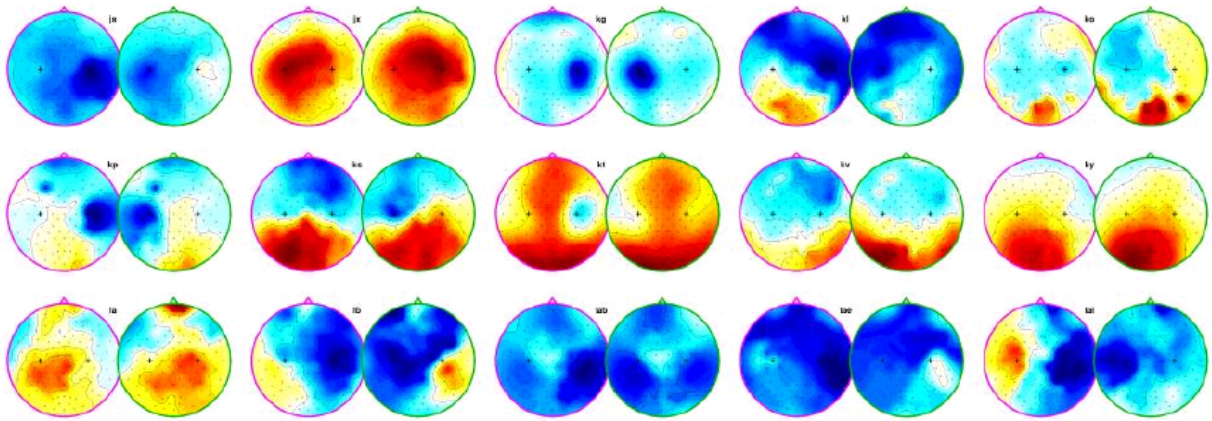
- intra-subject variability : topography power in the α band (around 10 Hz) during trial of 3.5s for a single subject for the same day
- inter-subject variability : left hand vs right hand for different subject



Muller et al.

BCI Motor imagery : variability in ERD

- intra-subject variability : topography power in the α band (around 10 Hz) during trial of 3.5s for a single subject for the same day
- inter-subject variability : left hand vs right hand for different subject



Muller et al.

BCI Motor imagery : variability in ERD

- intra-subject variability : topography power in the α band (around 10 Hz) during trial of 3.5s for a single subject for the same day
- inter-subject variability : left hand vs right hand for different subject

Machine learning and signal processing at the rescue for handling

- low signal-to-noise ratio
- variability

Machine Learning in a nutshell

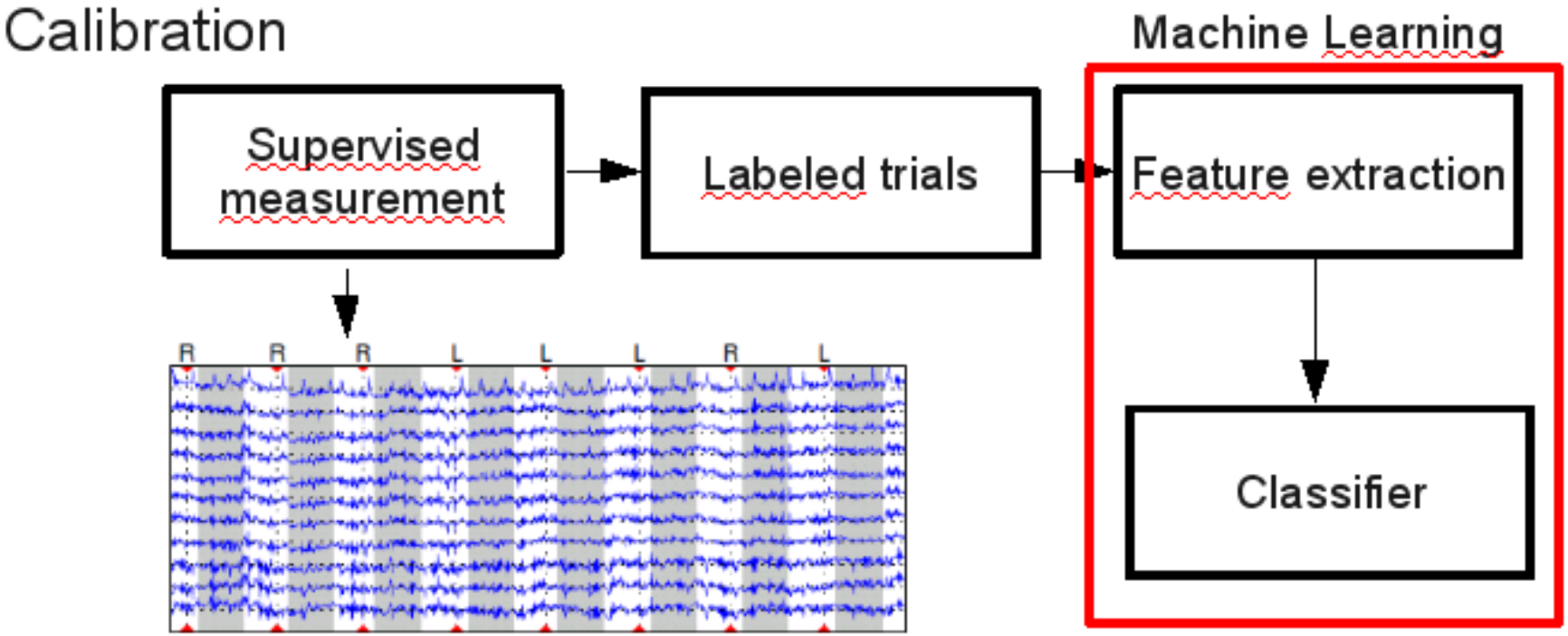
Objective

learn to recognize complex patterns and make decisions based on data

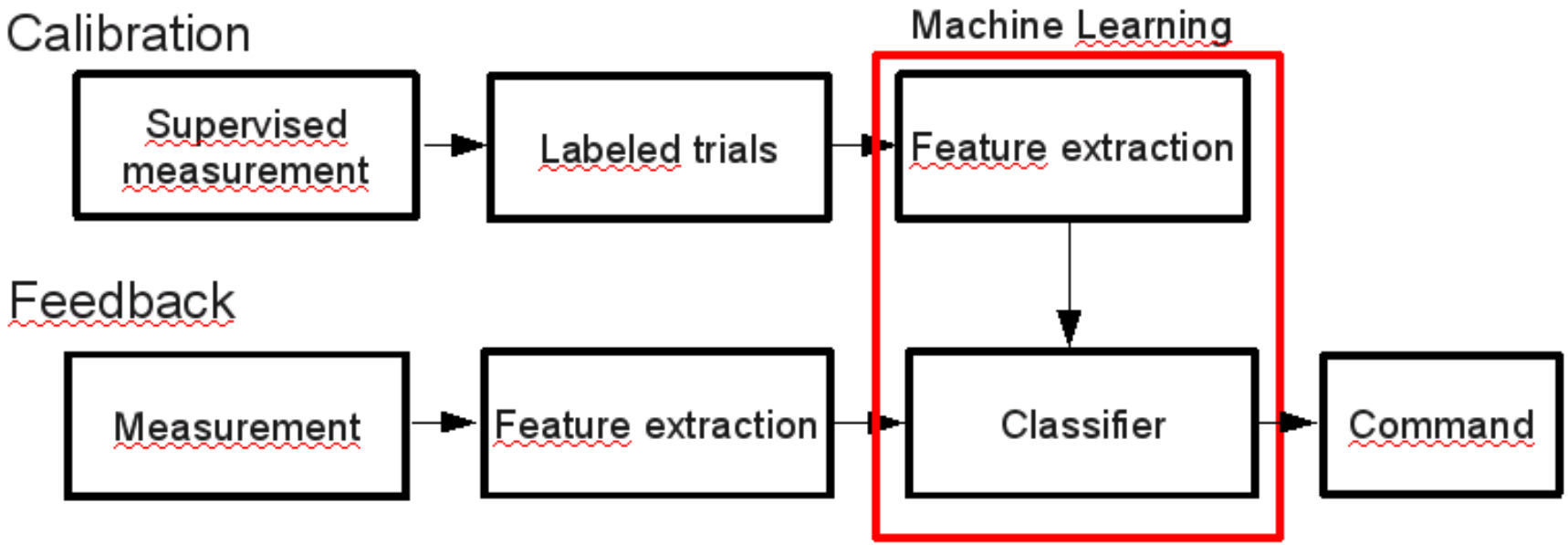
Usual ML tasks of interest for BCI

- feature extraction (CSP, ICA, ...)
- feature selection (channel selection)
- supervised learning of a classifier (LDA, SVM ...)
- online learning

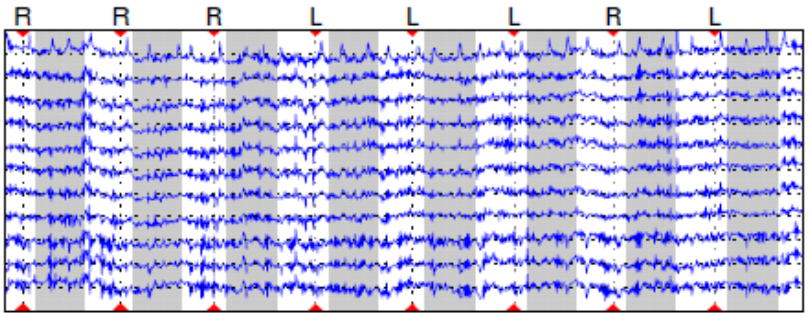
Machine learning in play in BCI



Machine learning for BCI : operating mode - Feedback

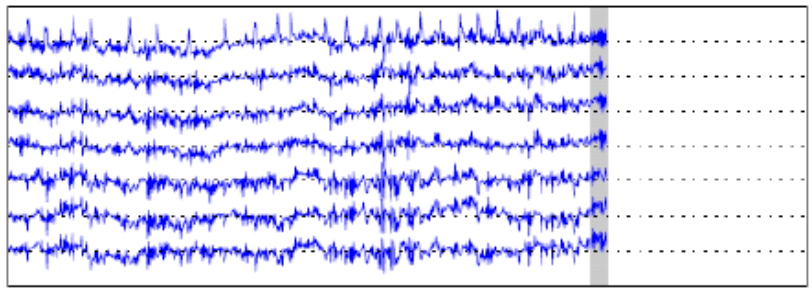


offline: calibration (10–20 minutes)



collect training samples

online: feedback (up to 6 hours)



classification of sliding windows ($\leq$ 1s)

Specific machine learning tasks in BCI

Feature extraction and selection

- learn filters (spatial or temporal)
- channel selection
- subject-specific features

Classifier

- mental state classification
- subject-independent mental state classification

BCI P300 Speller

Details

- Sequence = 12 rows/columns are flashed randomly
- Sequences are repeated a certain number of times of a letter spelling

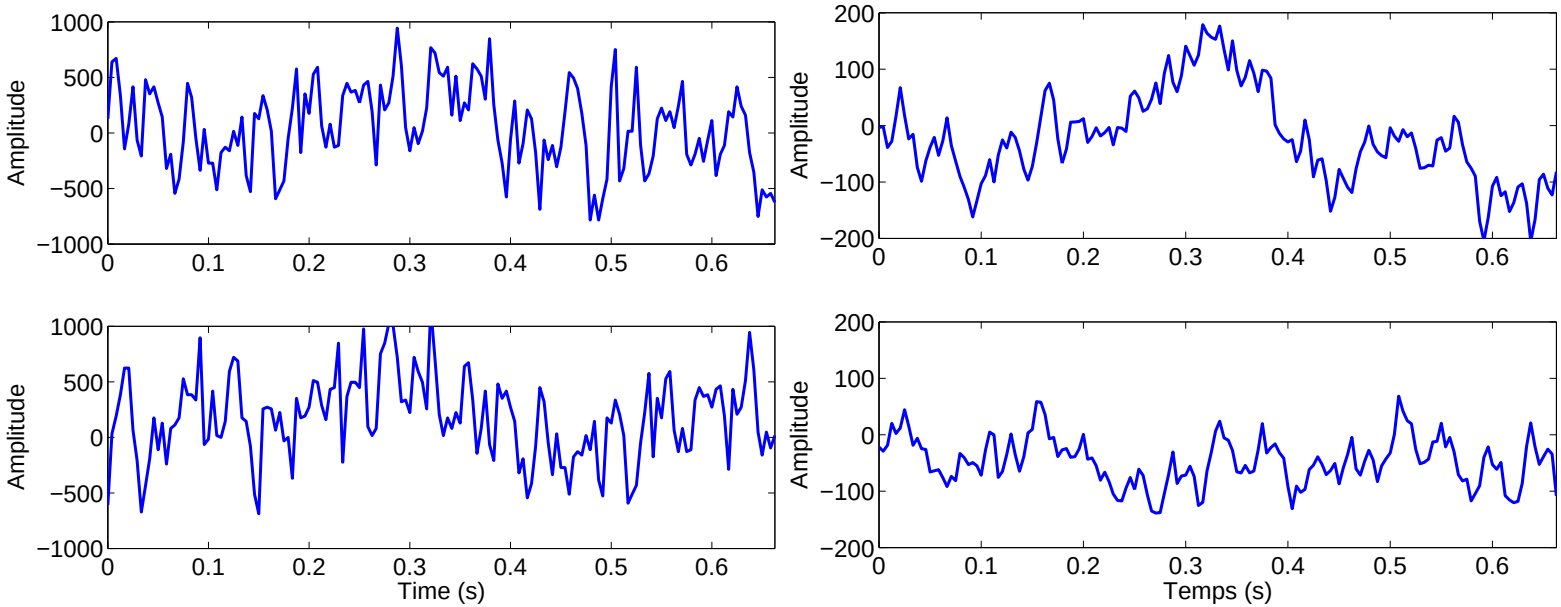
Machine Learning Objective

Classify a post-stimulus EEG signal as containing or not a P300 evoked potential.

The EEG Signal

Characteristics

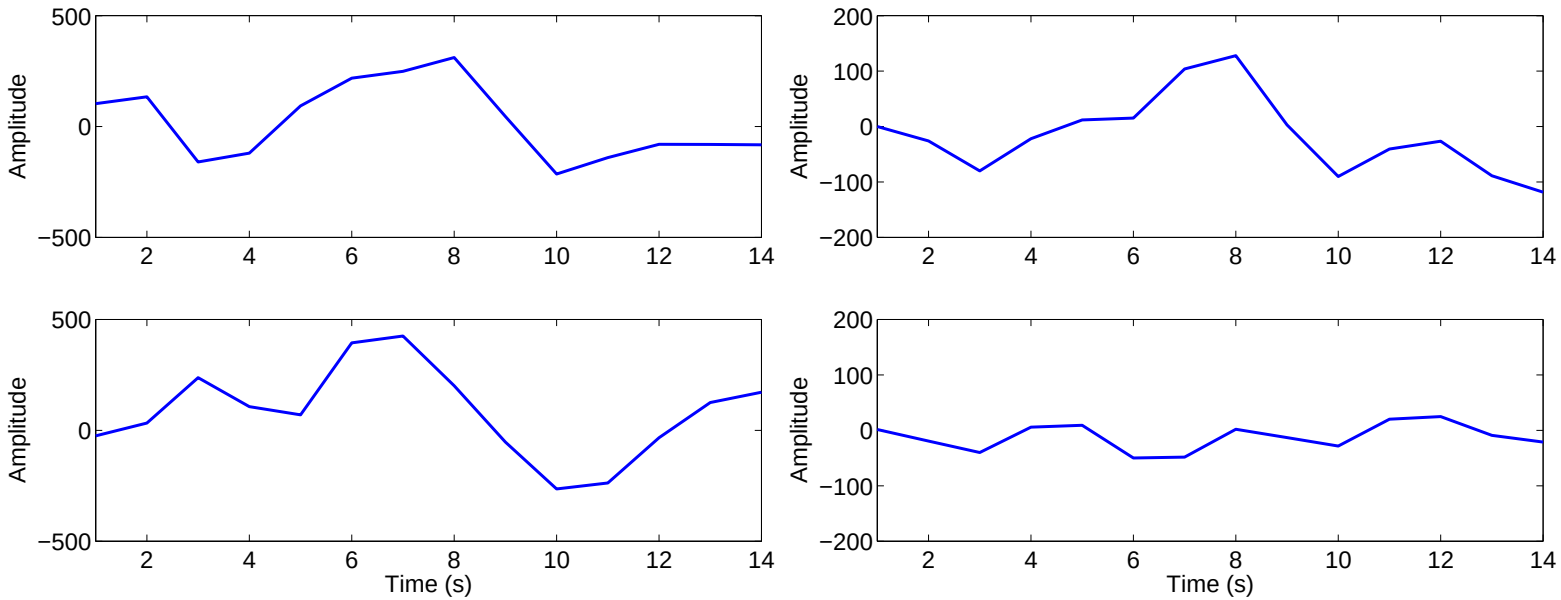
- Time windows : 666 ms after visual stimuli
- Problem Dimensionality : $64 \times 160 = 10240$
- Low signal to noise ratio
- Non-stationary
- Visualization of P300 in averaged signals.



Signal and Preprocessing

On each channel

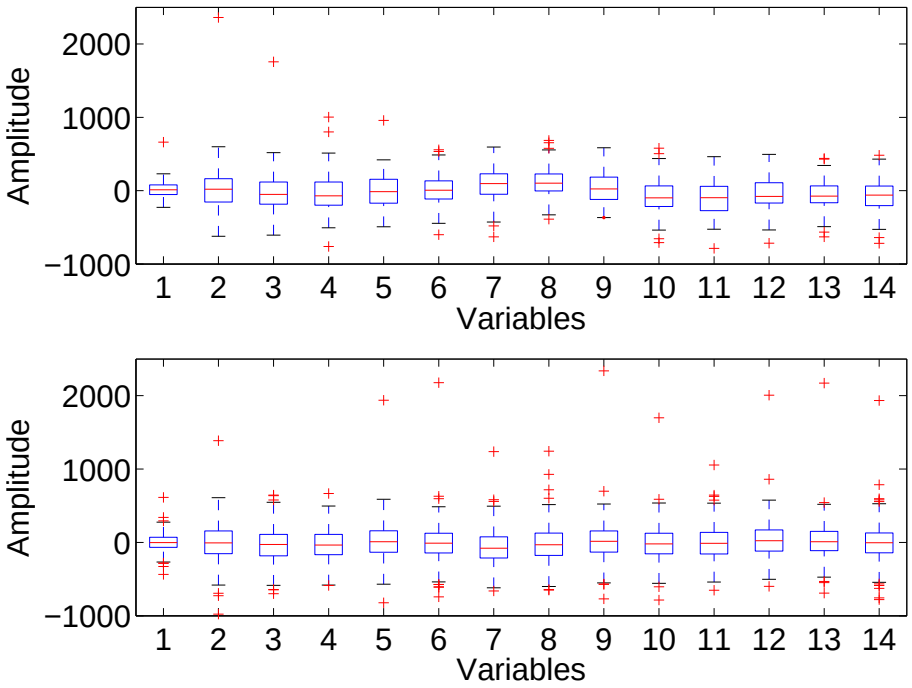
- BandPass filtering [0.1, 10] Hz
- Decimation : 14 time samples
- Problem Dimensionality : $64 \times 14 = 896$

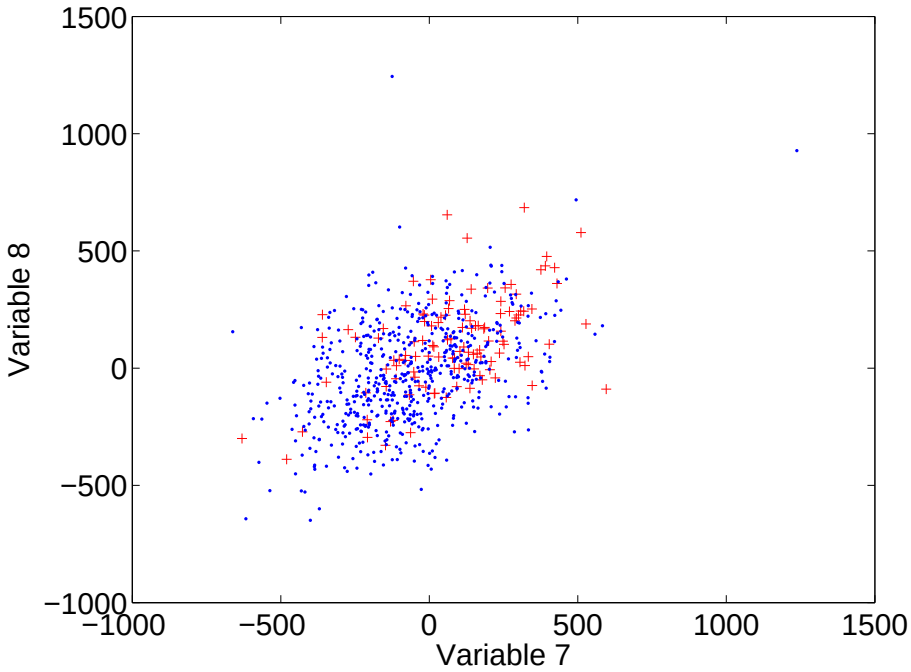


The obstacles for an easy classification

Difficulties

- Low Signal to noise ratio : classes are highly mixed
- Signal variability

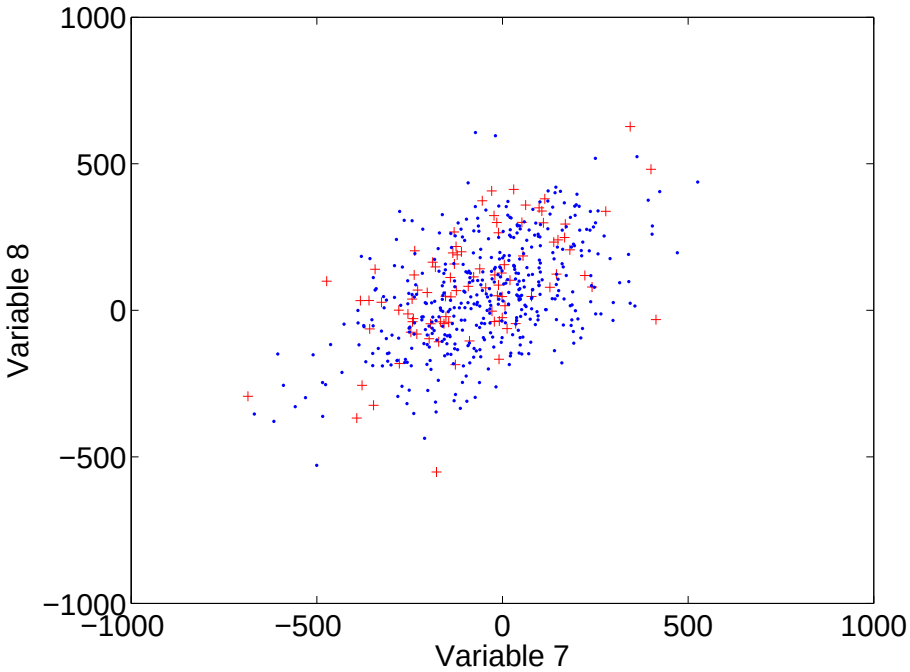


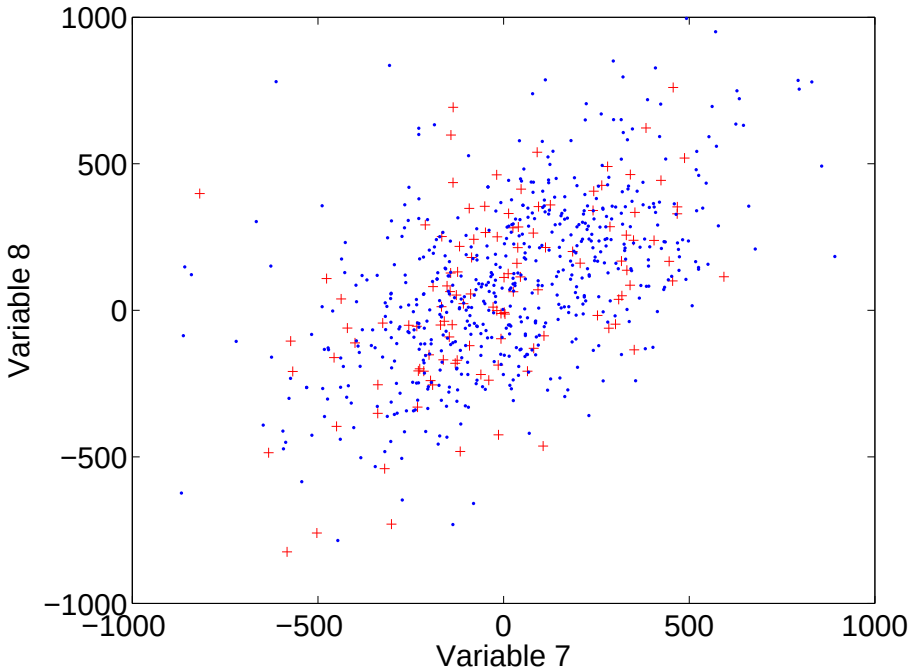


The obstacles for an easy classification

Difficulties

- Low Signal to noise ratio : classes are highly mixed
- Signal variability





Classification strategy

Difficulties

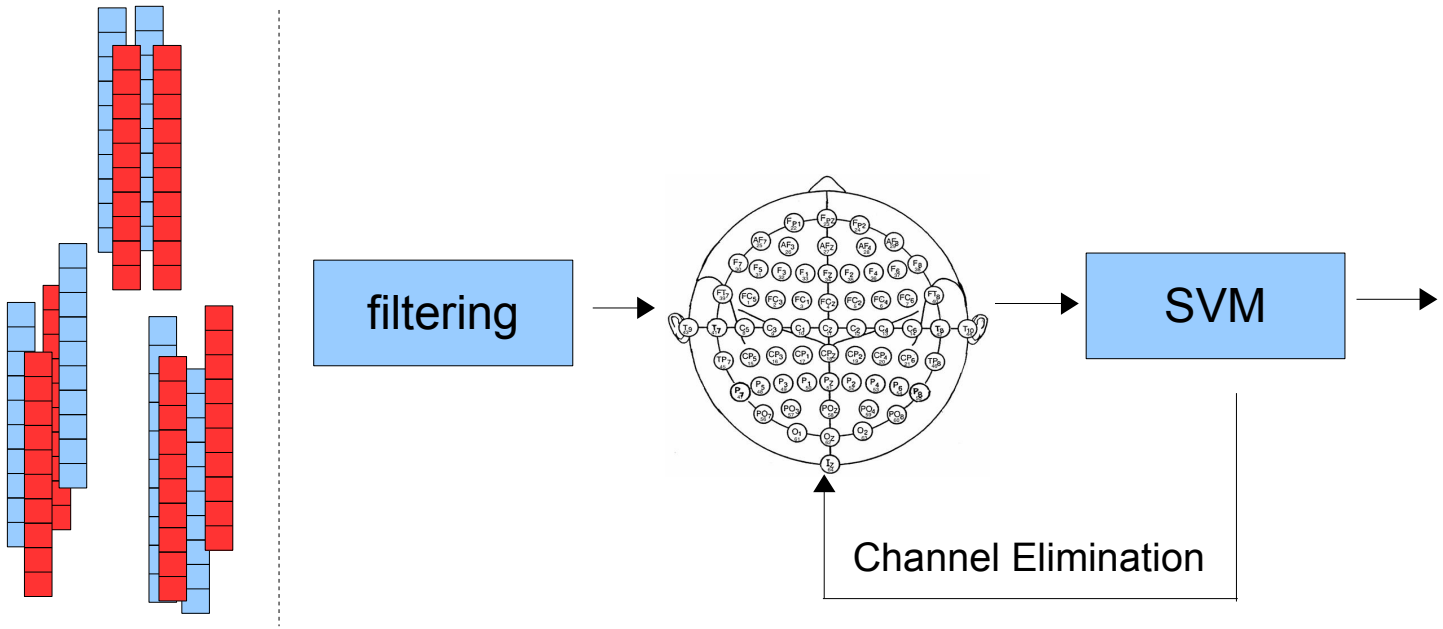
- Brain activities and noise non-related to the BCI task
- Slow non-stationary phenomena

Solutions

- reduce variabilities by separating training sets in "homogenous" sets.
- Learn several classifiers for each set and perform channel selection.
- average SVMs outputs.

Learning Strategy

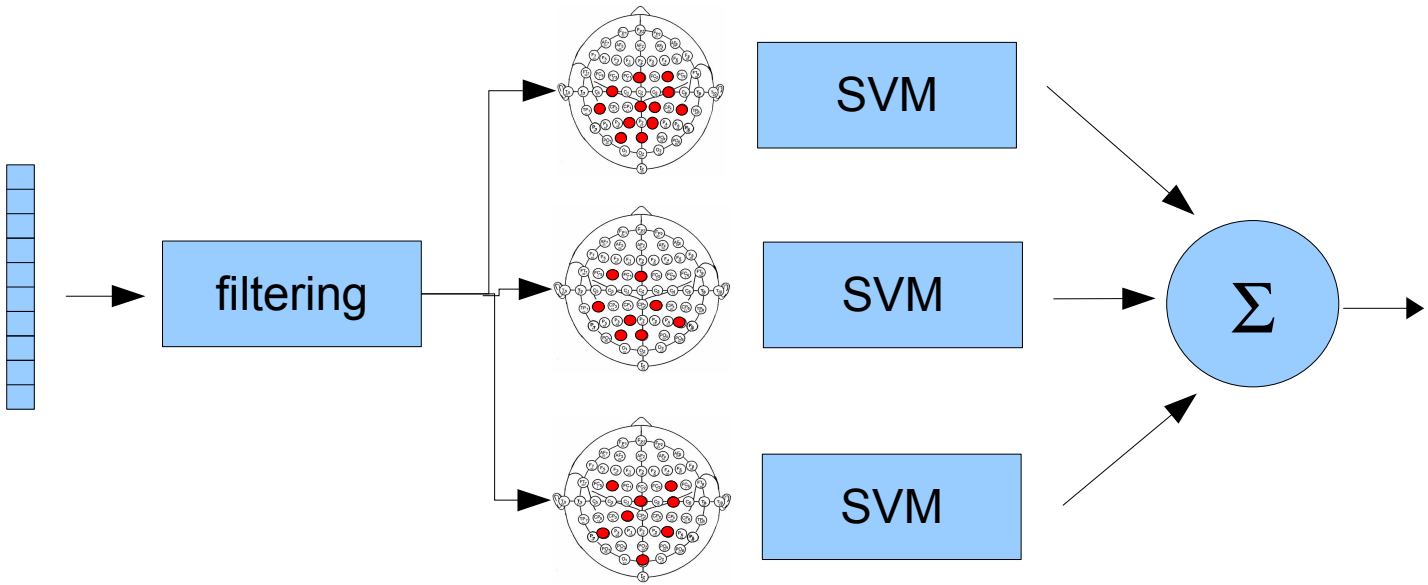
Training



Test Strategy

Procedure

- The object to classify is a post stimulus signal corresponding to a row or a column
- Object = signals from the 64 channels
- Filtering, decimation and selection of the channels
- SVMs outputs



Selecting the spelled letter

Procedure

- a post stimulus signal (row or column)
- Processing + SVMs outputs
- Increments score associated to the row/column

$$S_{lig} = S_{lig} + \sum_{i=1}^{N} f_i(x) \qquad S_{col} = S_{col} + \sum_{i=1}^{N} f_i(x)$$

Final selection

- After k sequences : $12 \times k$ stimuli
- symbol coordinates in matrix

$$\arg\max_{lig} S_{lig} \qquad \arg\max_{col} S_{col}$$

Backward Channel Elimination

Outline

- Rank channels by a decreasing order of importance
- Selection criterion : performance

$$\frac{TP}{TP + FP + FN}$$

  evaluated on validation data
- Backward Elimination Algorithm

Application to BCI Competition III

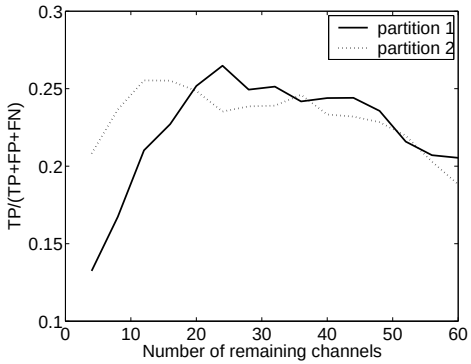
Experimental set-up

- 2 Subjects using P300 speller
- EEG have been collected on a 64-channel scalp
- 85 spelled symbols for training, 100 for evaluating performances
- for each single spelled symbol, we have 180 trials, 30 of which should contain a P300.
- cluster of 5 letters (900 signals)
- linear classifier on each partition

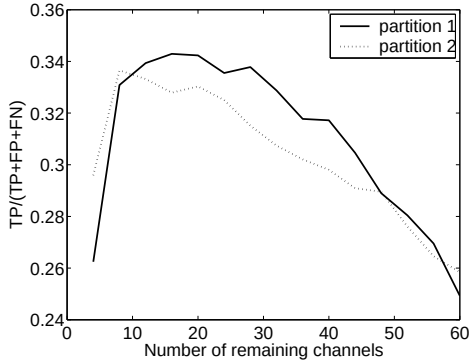
Application to BCI Competition III

- Optimal number of channels to select

Subject A

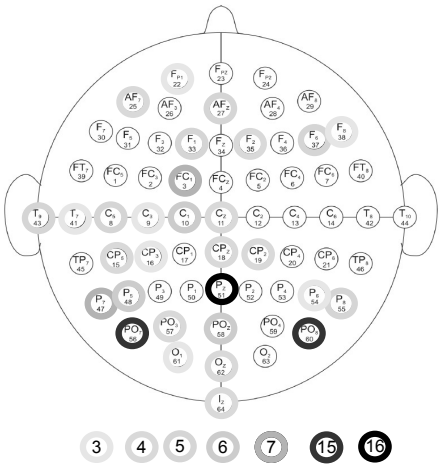


Subject B

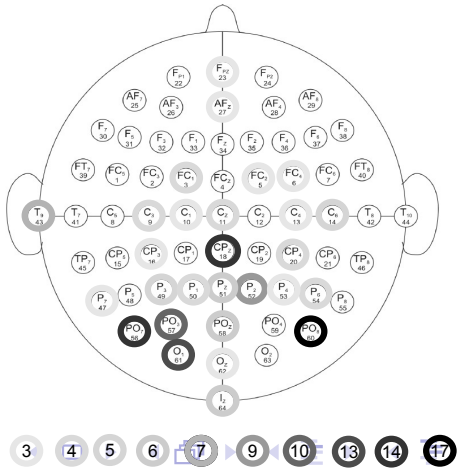


- Channels selected

Subject A



Subject B



Application to BCI Competition III

- Effect of channel selection

	Performances after 5 and 15 sequences					
Dataset	Optimal channels		64 channels		8 channels	
A_1	26	66	22	55	24	60
A_2	41	69	22	61	15	54
A_3	28	64	19	59	5	27
A_4	36	81	24	56	17	53
A_5	39	75	27	69	23	52
B_1	62	93	52	80	41	76
B_2	61	90	49	73	31	54
B_3	56	81	45	65	36	65
B_4	57	89	49	81	47	65
B_5	53	89	59	88	33	70

- Compared performances on Competition

	Nb of sequences	
Algorithms	5	15
this algorithm	**73.5**	**96.5**
2nd ranked algorithm	55.0	90.5
3rd ranked algorithm	59.5	90

Dealing with the BCI P300 Speller

Lessons learned

- for P300, simple temporal features do the job
- do channel selection (for better adaptation)
- take care of variabilities
- perform classifier output's averaging for better robustness to variabilities (see also Krusienski, 2009)
- find more efficient strategy for channel selection
- use better preprocessing : xDAWN, spatial filtering

Performance

- Winner of the BCI Competition III
- Third place of the BCI P300 MLSP 2010 competition with (81.9% recognition rate vs 82.1%)

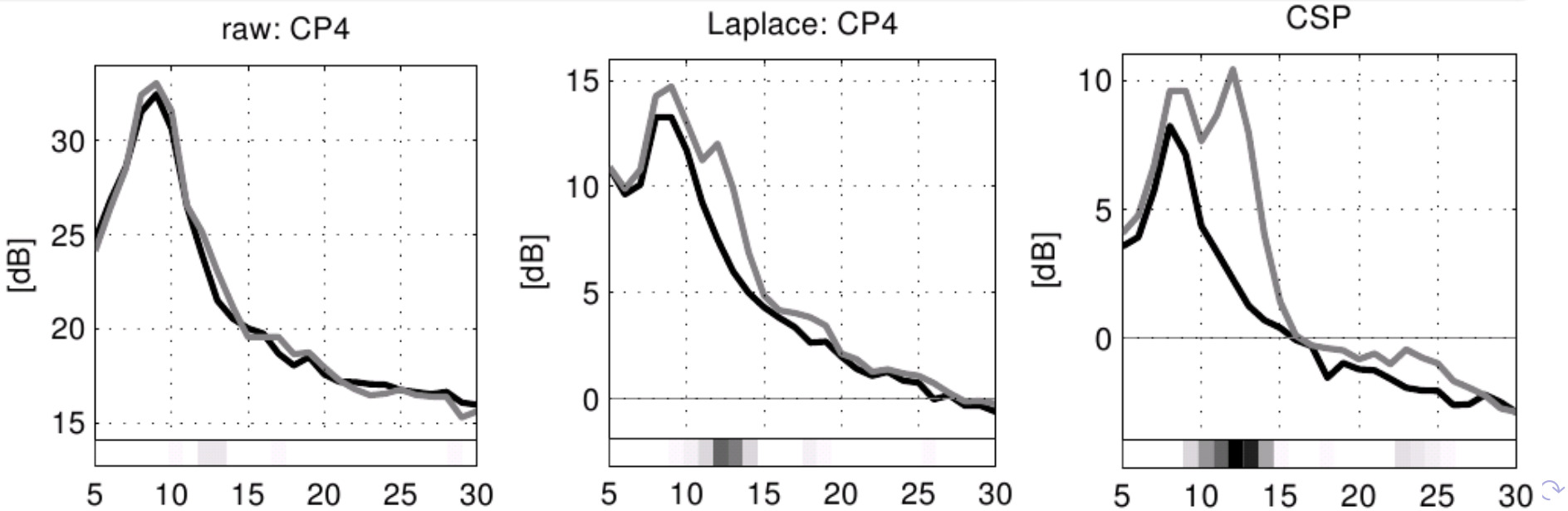
Learn spatial filters for BCI : Common Spatial Patterns

Why is it important ?

- EEG potentials have poor spatial resolution
- nearby sources contribute to each scalp electrode
- strong signals can interfere with the weak signal of interest

Illustration of effectiveness (Blankertz,2008)

spectra from left (dark) and right (light) hand motor imagery.



Learning spatial filters for BCI

The picture

- Labeled multi-channel EEG data : $S_i \in \mathbb{R}^{K \times T}$, $y_i = \{+1, -1\}$, $i = 1, \cdots, n$, (T number of samples , K number of channels). S_i is supposed to be band-pass filtered.
- Look for a spatial filter $w \in \mathbb{R}^K$ so that $S_i^t w$ has high power spectra for one class and low power spectra for the other class
- Approximation of power spectra with signal variance

$$\|S_i^t w\|^2 = w^t S_i S_i^t w$$

- Overall variance for positive and negative class signals

$$\sum_{i:y_i=1} w^t S_i S_i^t w \quad \text{and} \quad \sum_{i:y_i=-1} w^t S_i S_i^t w$$

Learning spatial filters for BCI

- Now, the problem can be formulated as

$$\max_w \frac{\sum_{i:y_i=1} w^t S_i S_i^t w}{\sum_{i:y_i=-1} w^t S_i S_i^t w}$$

- Owing to scale invariance in w, it can be translated into

$$\begin{aligned} max_w \quad & w^T \Sigma_{+1} w \\ & w^T \Sigma_{-1} w = 1 \end{aligned}$$

  where Σ_{+1} and Σ_{-1} are the covariance matrix for positive and negative classes.

- Lagrangian and optimality give

$$(\Sigma_{-1})^{-1} \Sigma_{+1} w = \lambda w$$

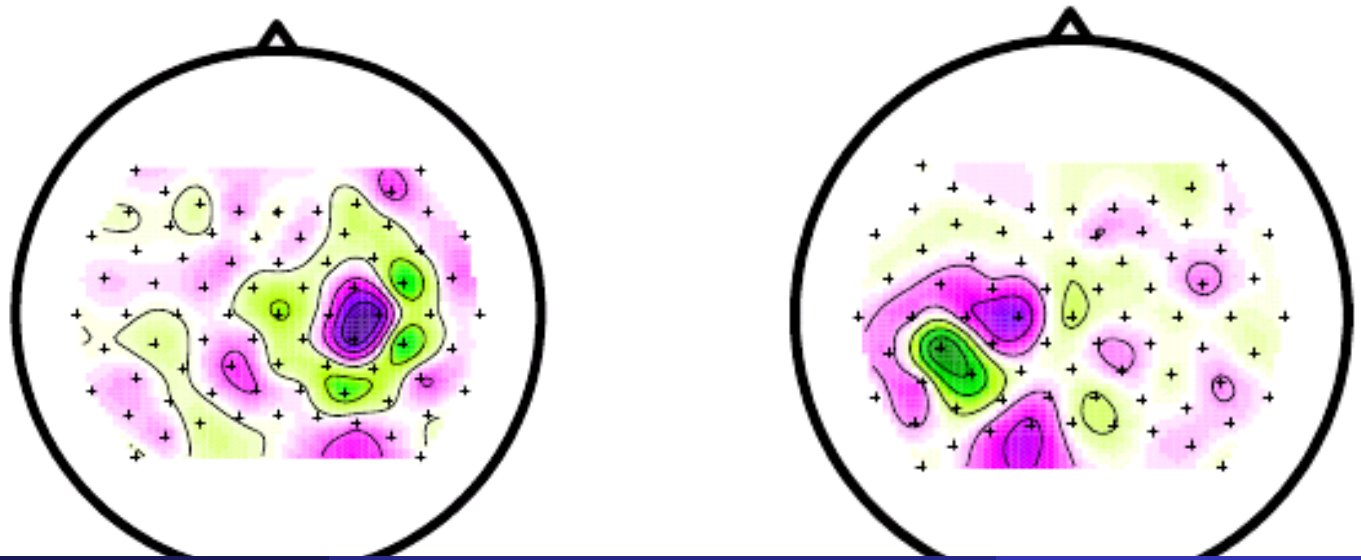
  with λ Lagrangian multiplier. We have a generalized eigenvalue problem

Learning spatial filters for BCI

The resulting filter

- the spatial filters are the eigenvector of $\Sigma_{-1}^{-1}\Sigma_{+1}$
- symmetry of the problem : use extreme eigenvalues/eigenvectors
- number of spatial filters to use : cross-validation if necessary or pick few.

Visualizing the filter (Blankertz,2008)



Applying spatial filter for BCI Motor imagery

Experimental set-up (Blankertz,2006)

- 6 subjects, MI : left hand, right hand, foot
- 32 channels [7, 30] Hz band-pass filtered
- training set 420 trials (140 for each class)
- feedback : 1D cursor control (hit the correct target)
- 6 spatial filters + LDA with log-variance as features.

		calibration	**feedback**	
		accuracy	**accuracy**	**duration**
sbj	**classes**	**[%]**	**[%]**	**[s]**
al	LF	98.0	98.0 ± 4.3	2.0 ± 0.9
ay	LR	97.6	95.0 ± 3.3	1.8 ± 0.8
av	LF	78.1	90.5 ± 10.2	3.5 ± 2.9
aa	LR	78.2	88.5 ± 8.1	1.5 ± 0.4
aw	RF	95.4	80.5 ± 5.8	2.6 ± 1.5
au	—	–	–	–
mean		89.5	90.5 ± 7.6	2.3 ± 0.8

Learn spatio-temporal filter : CSSSP (Dornhege, 2006)

Why spatio-temporal filters ?

- improve subject-specific BCI by focusing on appropriate spectral band frequency.
- improve spatial filters by integrating temporal filters.

Set-up and approach

- classify EEG motor imagery 2 classes trials
- Define FIR filter
- Define filtered signal version $S_{i,b} = S_i + \sum_{k=2}^{T} b_k S_i^k$
- find w and $\{b\}$ that maximizes spatial filter criterion
- build power spectra features and use linear classifier

Learn spatio-temporal filter : CSSSP

The optimization problem

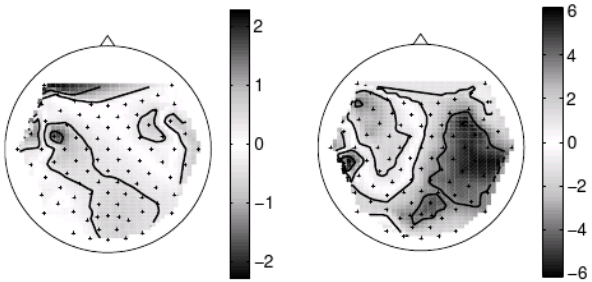
$$\max_b \max_w w^T \quad \left(\sum_{k=0}^{T-1}(\sum_{j=1}^{T-k} b_j b_{k+j})\Sigma_1^k\right) w + \frac{C}{T}\sum_k |b_k|$$
$$w^T \left(\sum_{k=0}^{T-1}(\sum_{j=1}^{T-k} b_j b_{k+j})(\Sigma_1^k + \Sigma_2^k)\right) w = 1$$

Outline

- sparse filter induced by the ℓ_1 penalization of b.
- same optimization problem as in CSP for fixed b.
- subgradient descent optimization for optimizing b

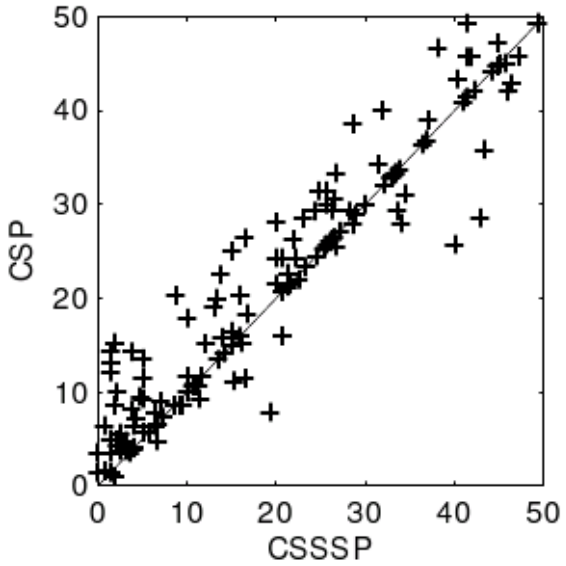
Effectiveness of CSSSP

- example of patterns CSP vs CSSSP



The discriminative pattern extracted by CSSSP has a much clearer and plausible topography (centered around C3 and C4)

- error comparison CSP vs CSSSP on test data



- 60 datasets from 22 subjects (120 to 200 trials)
- 50/50 train/test split
- 32 channels [7, 30] Hz band-pass filtered
- subject-specific preprocessing leads to better mental states recognition

Learn smooth spatial filters

why smooth filter ?

- CSP ignores the spatial location of EEG electrodes
- neighboring neurons tend to have similar functions.
- neighboring electrodes measure similar signals

Spatially regularized CSP (Lotte,2010)

- Smooth spatial filter
- Formulation of regularized spatial filter :

$$J(w) = max_w \frac{w^T \Sigma_{+1} w}{w^T \Sigma_{-1} w + \alpha P(w)}$$

where $P(w)$ is a function that measures the spatial smoothness of w (the smaller the more regular)

Spatially regularized CSP

Spatial regularizer : Laplacian penalty

- Laplacian penalty : let $G(i,j) = Sim(Elec_i, Elec_j)$ be a similarity measure on electrodes position (e.g $G(i,j) = 1$ if electrodes i and j are considered as neighbors), then we define

$$P(w) = \sum_{i,j} G_{i,j}(w_i - w_j)^2 = w^t(D - G)w$$

  with $D, G \in \mathbb{R}^{K \times K}$ and D is a diagonal matrix such that $D_{ii} = \sum_j G_{i,j}$

- Interpretation :
  - $P(w)$ becomes larger as neighbor electrodes get some different weights.
  - small value of $P(w)$ encourages smooth spatial filter

Spatially regularized CSP : solution

Eigenvalue problem

- the filter is given by the eigenvectors of

$$(\Sigma_{-1} + \alpha(D - G))^{-1}\Sigma_{+1}$$

corresponding to largest eigenvalues.

- the problem is not symmetric. For the second class, the filter is related to

$$(\Sigma_{+1} + \alpha(D - G))^{-1}\Sigma_{-1}$$

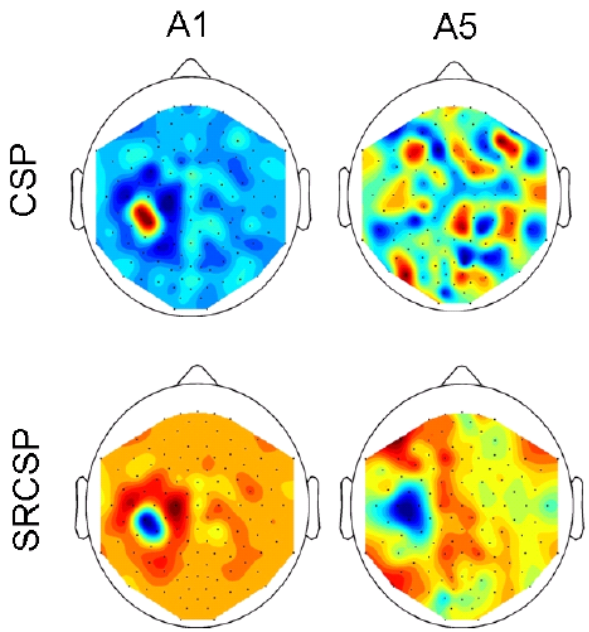
- Hyperparameters : α and parameters of the similarity measure.

Effectiveness of SRCSP (Lotte, 2010)

Experimental setup

- Dataset from BCI Competition III
- 5 subjects, mental imagery : left and right hand
- 280 trials per subject, 118 channels, [8-30] Hz filtering
- compare spatial filter and SRCSP recognition rate

	BCI competition III, data set IVa				
Subject	A1	A2	A3	A4	A5
CSP	66.07	96.43	47.45	71.88	49.6
SRCSP	**72.32**	96.43	**60.2**	**77.68**	**86.51**



Summary on spatial filters

the good

- made untrained subjects able to control a BCI
- considerable improvements of performance wrt to state of the art
- provide interpretable filters

caveats and questions

- need to estimate a covariance matrix
- sensitive to outliers and noise
- is it the most appropriate criterion to optimize ? Large Margin approach (see Flamary et al.)
- how to adapt the filter to signal non-stationarities ? ML issue : online resolution of an eigenvalue problem

Novel challenges

what new since Donchin 88 ?

- proof of concept of BCI to different applications (P300 Speller, prothesis control, wheelchair command)
- more robust and efficient BCI
- study of ALS in-home use (Sellers, 2008)
- dry electrodes for easier use

what are the next big issues for a Machine Learning ?

- reduce calibration or zero training (covariate shift)
- adaptation (how to cope with user's variabilities (fatigue, mental states ...)
- find invariant feature and subject-independent classifier

Reduced calibration and zero training approaches

Multi-task learning (Almagir, 2010)

Use data from available subjects to learn a generic classifier that can be applied to a novel subject without training

Sequential Subject Selection (Lotte, 2010)

Among data from available subjects use those that are most helpful for classifying novel subject data.

Ensemble of BCI Classifier (Fazli,2010)

Among a large set of classifiers trained with available data, select a few of them that minimize empirical loss.

Sequential Subject Selection (Lotte, 2010)

Framework

- motor imagery
- Typically used BCI algorithms consider CSP and LDA
- Both algorithms need estimation of covariance matrices of the data.

Heuristic

- Mix data from available subjects and novel subject for estimating covariance matrix

$$\hat{C}_t = (1-\lambda)C_t + \frac{\lambda}{|S|}\sum_{i \in S} C_i$$

Sequential Subject Selection (Lotte, 2010)

algorithm for subject selection

- Forward-Backward selection

  Start from empty subject set
  Look for the subject which added dataset maximizes accuracy
  Among already selected subject, remove subject's dataset that maximizes accuracy

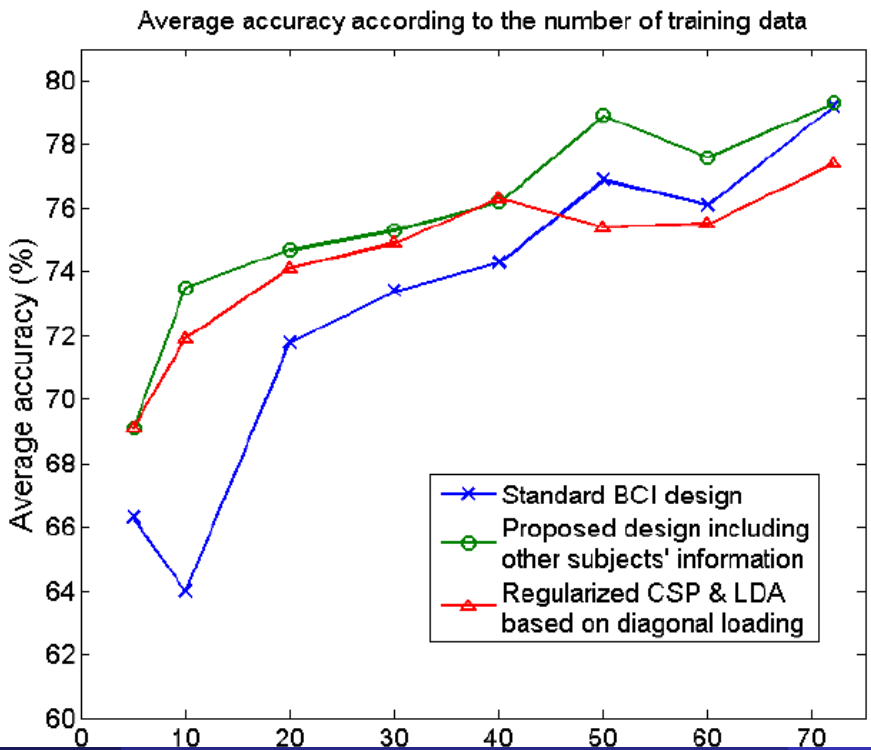
Comments

- can be time-consuming
- still need some few data

Sequential Subject Selection (Lotte, 2010)

Experimental set-up

- Motor imagery dataset (left and right hand), 9 subjects
- signals are band-pass filtered (8-30 Hz)
- 72 trials per class on time-segment of 2 seconds for training and testing



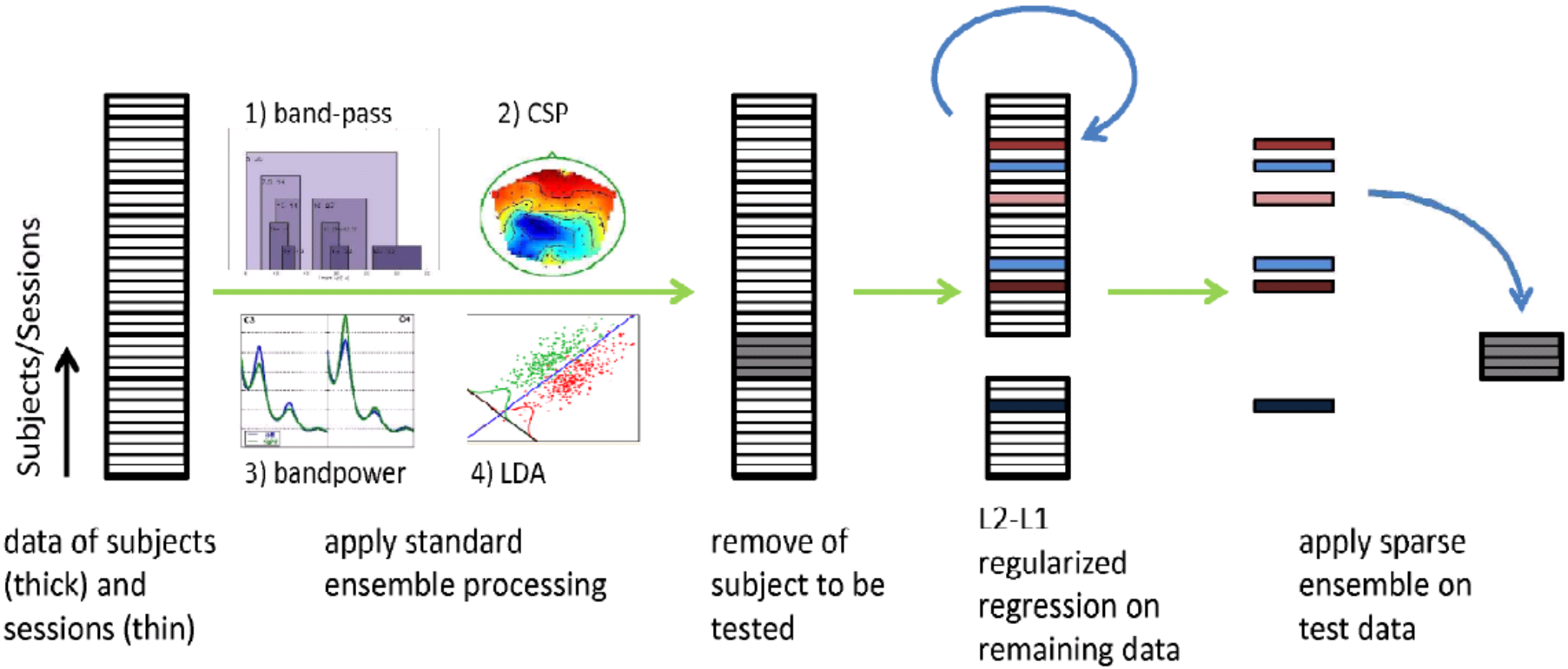
Ensemble of Classifier (Fazli,2010)

Idea

- zero training approach for a novel subject
- construct an ensemble method from a large set of weak classifiers (mimic idea from boosting)
- weak classifier :
  - single subject/session dataset
  - particular frequency band
  - CSP + LDA

Ensemble of Classifier (Fazli,2010)

Flowchart



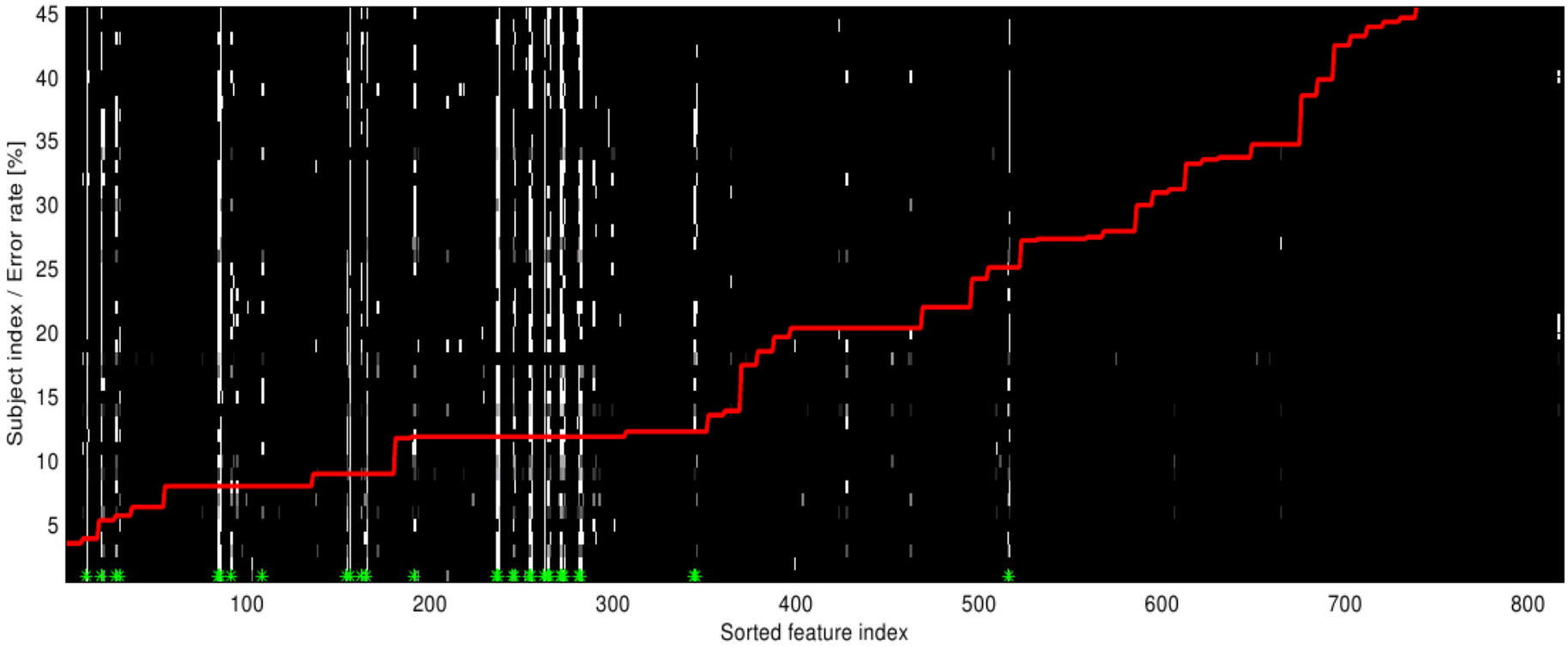
The final gating function used for testing on subject

$$\begin{array}{ll} \min_{w_{i,j}} & \sum_{x \in X} (h(x) - y(x))^2 + \lambda \sum_{i,j} |w_{i,j}| \\ \text{with} & h(x) = \sum_{i,j \in S} w_{i,j} c_{i,j}(x) + b \end{array}$$

Ensemble of Classifier (Fazli,2010)

Feature selection results during cross-validation

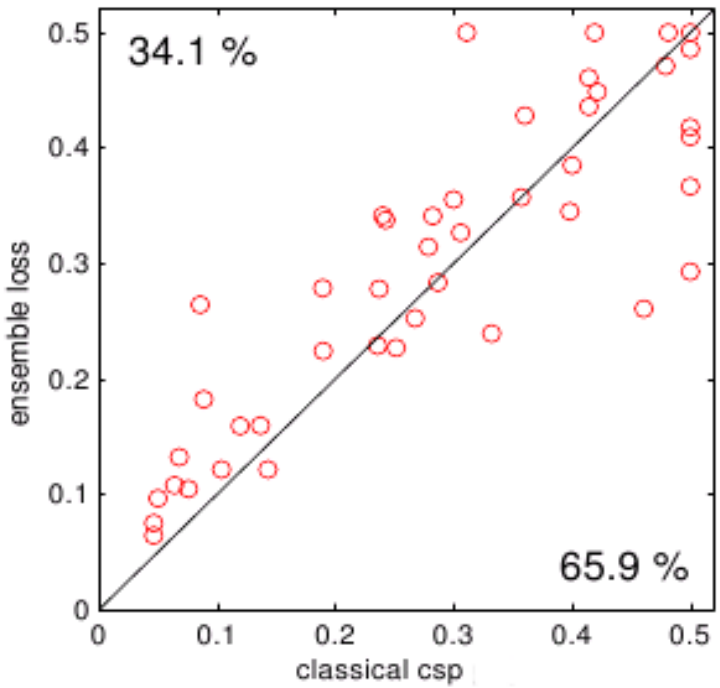
- 45 subjects, 90 sessions, 9 band-pass filters
- leave-one-subject out

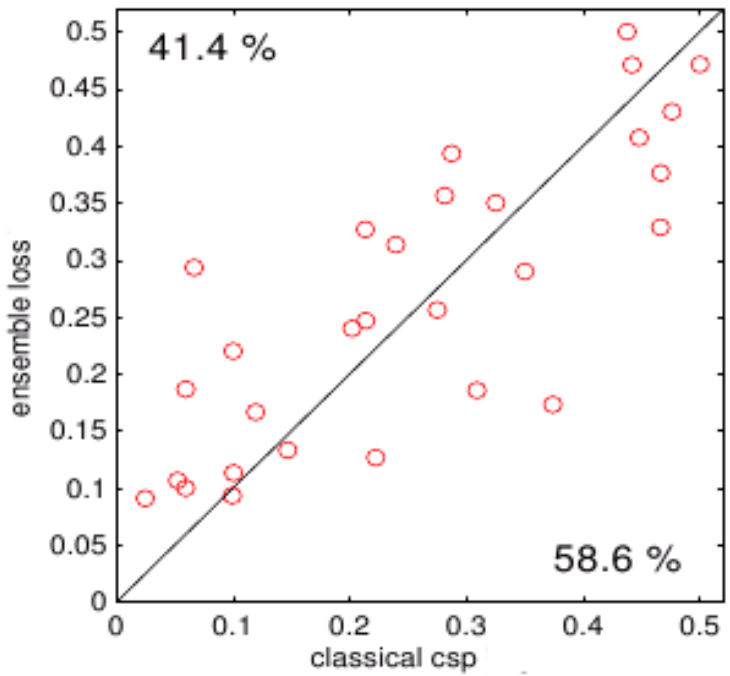


Ensemble of Classifier (Fazli,2010)

Results

- 45 subjects, 90 sessions, 9 band-pass filters
- leave-one-subject out (left) and hold-out (right)
- compare with subject-dependent auto-band CSP





Multi-Task Learning

Hypothesis

- learning related tasks jointly can improve performances on each single task
- some knowlegde can be shared across tasks :
  - models are similar
  - all tasks share the same features

Example of applications

- Character recognition (Obozinski, 2009)
- Medical image recognition (Xiong,2007)

Multi-Task Learning : general framework

Framework

- T tasks. For each task, we have samples$\{x_{i,t}, y_{i,t}\}_{i=1}^{m_t}$ coming from the same space and drawn accordinfg to P_t
- Distributions P_t are different for each task but related
- Learn decision function $f_t(x_{i,t})$ that predicts accurately $y_{i,t}$

$$\min_{f_1,\cdots,f_T} \sum_{t=1}^{T} \sum_{i=1}^{m_t} L_t(y_{i,t}, f(x_{i,t})) + \lambda \Omega(f_1, \cdots, f_T)$$

where $L_t(\cdot,\cdot)$ are some loss functions, Ω a regularizer that should reflect tasks-relatedness and λ a hyperparameter

Typical prior for the regularizer

- models are similar
- shared feature selection

Multi-Task Learning with Joint linear Least-Squares

Framework

- linear model for each task t : $w_t^T x$
- sparse shared features : define $W = [w_1 w_2 \cdots w_T]$,

$$min_{w_t} \frac{1}{2} \sum_t \|X_t w_t - y_t\|^2 + \lambda \sum_k \|W_{k,.}\|_2$$

- weights are similar :

$$min_{v_t, w_0} \frac{1}{2} \sum_t \|X_t w_t - y_t\|^2 + \|w_0\|^2 + \sum_t \|v_t^2\|$$

with $w_t = w_0 + v_t$

- weights are distributed according to a Gaussian distribution

$$min_{w_t, \mu, \Sigma} \frac{1}{2} \sum_t \|X_t w_t - y_t\|^2 + \sum_t (w_t - \mu)^t \Sigma^{-1} (w_t - \mu) + \frac{T}{2} \log \det(\Sigma)$$

Multi-Task Learning with Joint linear Least-Squares

Solving MTL with Gaussian prior on w_t (Alamgir,2010)

- The problem

$$\min_{w_t,\mu,\Sigma} \frac{1}{2\lambda}\sum_t \|X_t w_t - y_t\|^2 + \frac{1}{2}\sum_t (w_t-\mu)^t \Sigma^{-1}(w_t-\mu) + \frac{T}{2}\log\det(\Sigma)$$

- alternate optimization wrt to W and (μ, Σ)

Update

- weights w_t

$$w_t = (\frac{1}{\lambda}X_t^t X_t + \Sigma^{-1})^{-1}(\frac{1}{\lambda}X_t^t y_t + \Sigma^{-1}\mu)$$

- Gaussian distribution

$$\mu = \frac{1}{T}\sum_t w_t \qquad \Sigma = \frac{1}{K}\sum_t (w_t-\mu)(w_t-\mu)^t$$

Applying MTL with Gaussian prior to BCI

Objective

- zero training BCI
- BCI adaptation to subject as data stream in

How-to ?

- Consider all subjects with already available data as tasks
- learn the shared prior μ and Σ
- start with the weights for the novel user

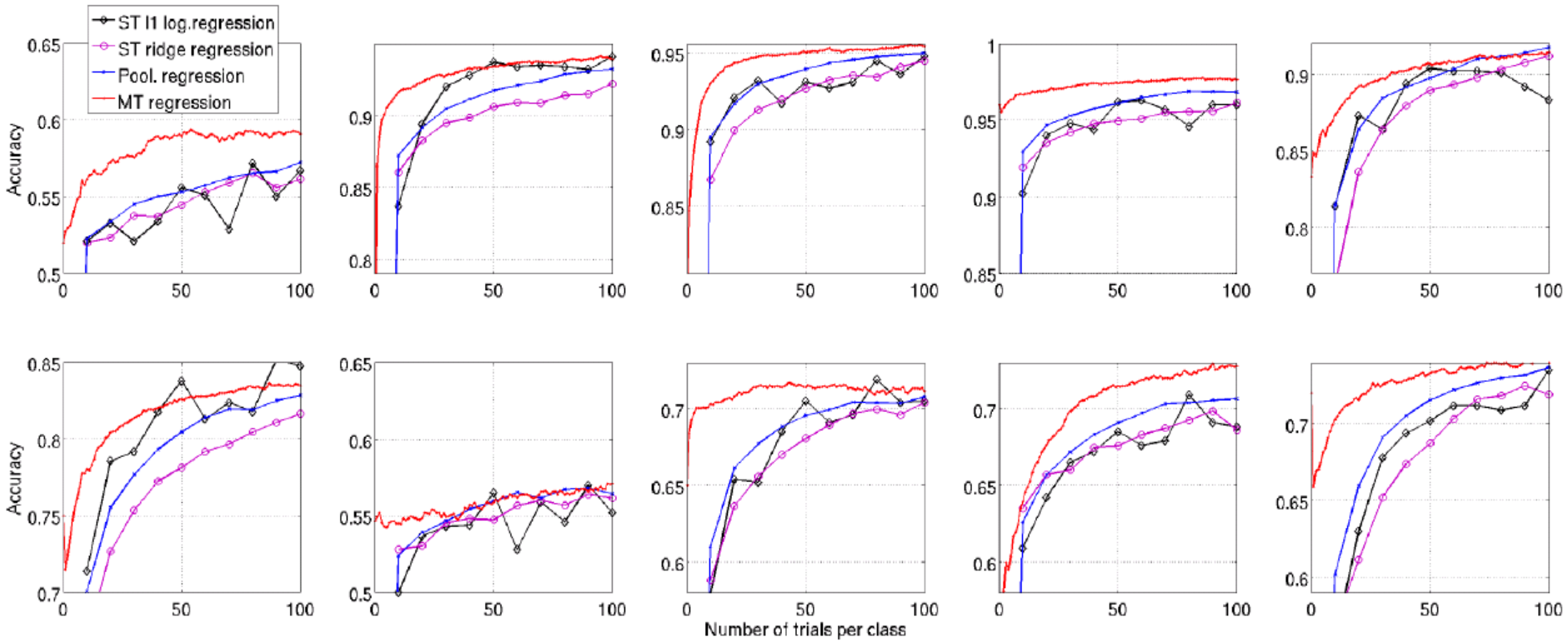
$$w = (\frac{1}{\lambda} X^t X + \Sigma^{-1})^{-1} (\frac{1}{\lambda} X^t y + \Sigma^{-1} \mu)$$

- update matrix X and y as data from the novel user are acquired

Effectiveness of MTL for BCI

Experimental set-up

- 10 subjects, motor imagery : left hand, right hand
- 150 trials. For the novel subject : 100 used for online, 50 for test



Conclusions and challenges

State of the art

- Successful applications of BCI
- pro-eminent role of machine learning
- some issues have been addressed e.g channel selection, feature extraction with spatial filters, tackling inter variability by providing subject-specific automated methods

future Machine challenges for BCI

- improve zero training BCI
- adaptation to evolution of the subject mental states (online learning)
- learn invariant features and subject-independent classifier that generalizes well
- proper addressing of concept drift problem