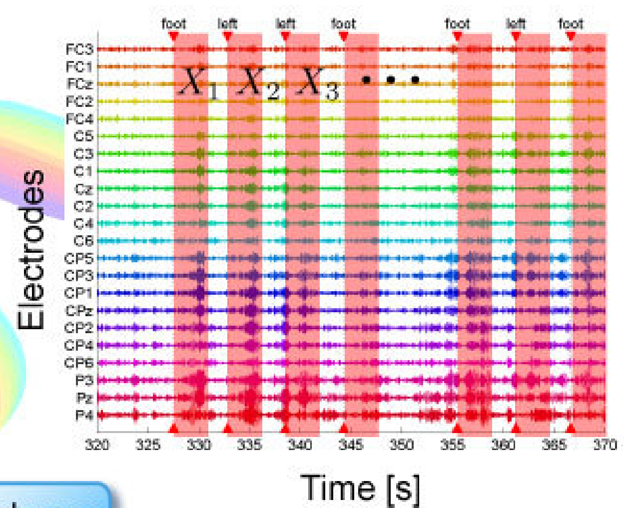
Logistic Regression for Single-trial EEG Classification

Ryota Tomioka[1,3], Kazuyuki Aihara[1], & Klaus-Robert Müller[2,3]

[1]The University of Tokyo, [2]Technical University of Berlin, & [3]Fraunhofer FIRST.IDA



Band-pass 7-30Hz

Epoched EEG: $d \; X_i$ ($d \times T$)

Sensor covariance of the epoched EEG: $d \; X_i X_i^\top$ ($d \times d$) (i=1,...,n)

d: #channels
T: #sampled time-points
n: #examples

Optimality?
Number of components?
Selecting components?

Conventional approach: Classifying with CSP

Classwise covariance matrix

$$\Sigma^{(c)} = \frac{1}{|\mathcal{I}^c|}\sum_{i\in\mathcal{I}^c} X_i X_i^\top \qquad (c \in \{+,-\}),$$

Simultaneous diagonalization, [Fukunaga 1972, Koles 1991]
i.e., a Generalized Eigenvalue Problem (GEP).

$$\Sigma^{(+)}\boldsymbol{w} = \lambda \Sigma^{(-)}\boldsymbol{w}$$

LDA on the log-power features

$$f_{\rm CSP}\left(X; \{\boldsymbol{w}_j\}_{j=1}^J, \{c_j\}_{j=0}^J\right) = \sum_{j=1}^J c_j \log \boldsymbol{w}_j^\top X X^\top \boldsymbol{w}_j + c_0$$

How is it related to a generative model?

Zero-mean Gaussian with no temporal correlation

$$P(X|c) \propto \prod_{t=1}^T \exp\left\{-\frac{1}{2}\boldsymbol{x}_t^\top {\Sigma^{(c)}}^{-1}\boldsymbol{x}_t\right\} = \exp\left\{-\frac{1}{2}{\rm tr}\left[{\Sigma^{(c)}}^{-1} X X^\top\right]\right\} \qquad (c\in\{+,-\})$$

leads to the discriminative model we use!

$$\log\frac{P(y=+1|X)}{P(y=-1|X)} = \frac{1}{2}{\rm tr}\left[\left(-{\Sigma^{(+)}}^{-1} + {\Sigma^{(-)}}^{-1}\right) X X^\top\right] + {\rm const.}$$

Proposed method: Logistic Regression (LR) classifier

We model the symmetric logit transform of the conditional probability

$$\log\frac{P(y=+1|X)}{P(y=-1|X)} = f(X;\boldsymbol{\theta})$$

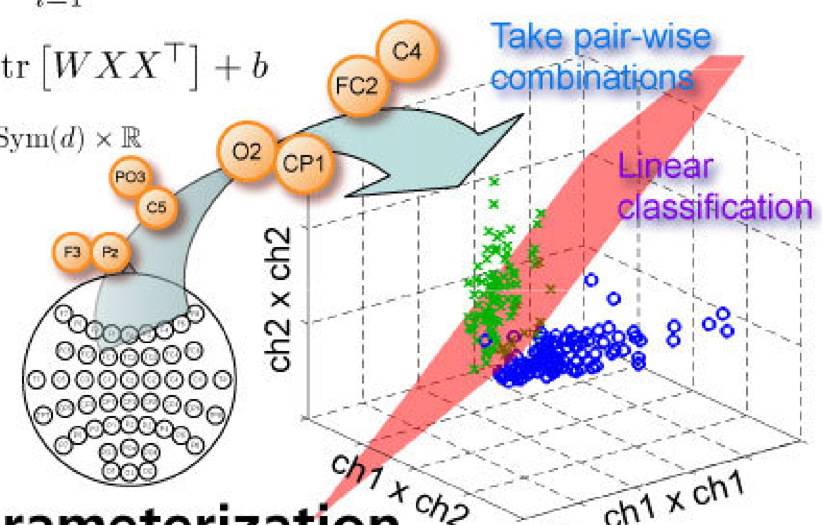
Full-rank symmetric matrix parameterization

Linearly classify w.r.t. the sensor-covariance coefficients

$$\min_{W\in{\rm Sym}(d), b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \log\left(1+e^{-y_i f(X_i;\boldsymbol{\theta})}\right) + \frac{C}{2}{\rm tr}\Sigma W \Sigma W$$

$$f(X;\boldsymbol{\theta}) = {\rm tr}\left[W X X^\top\right] + b$$

$$\boldsymbol{\theta} := (W, b) \in {\rm Sym}(d) \times \mathbb{R}$$



Rank=2 parameterization

The classifier is approximated with the difference of two rank=1 matrices

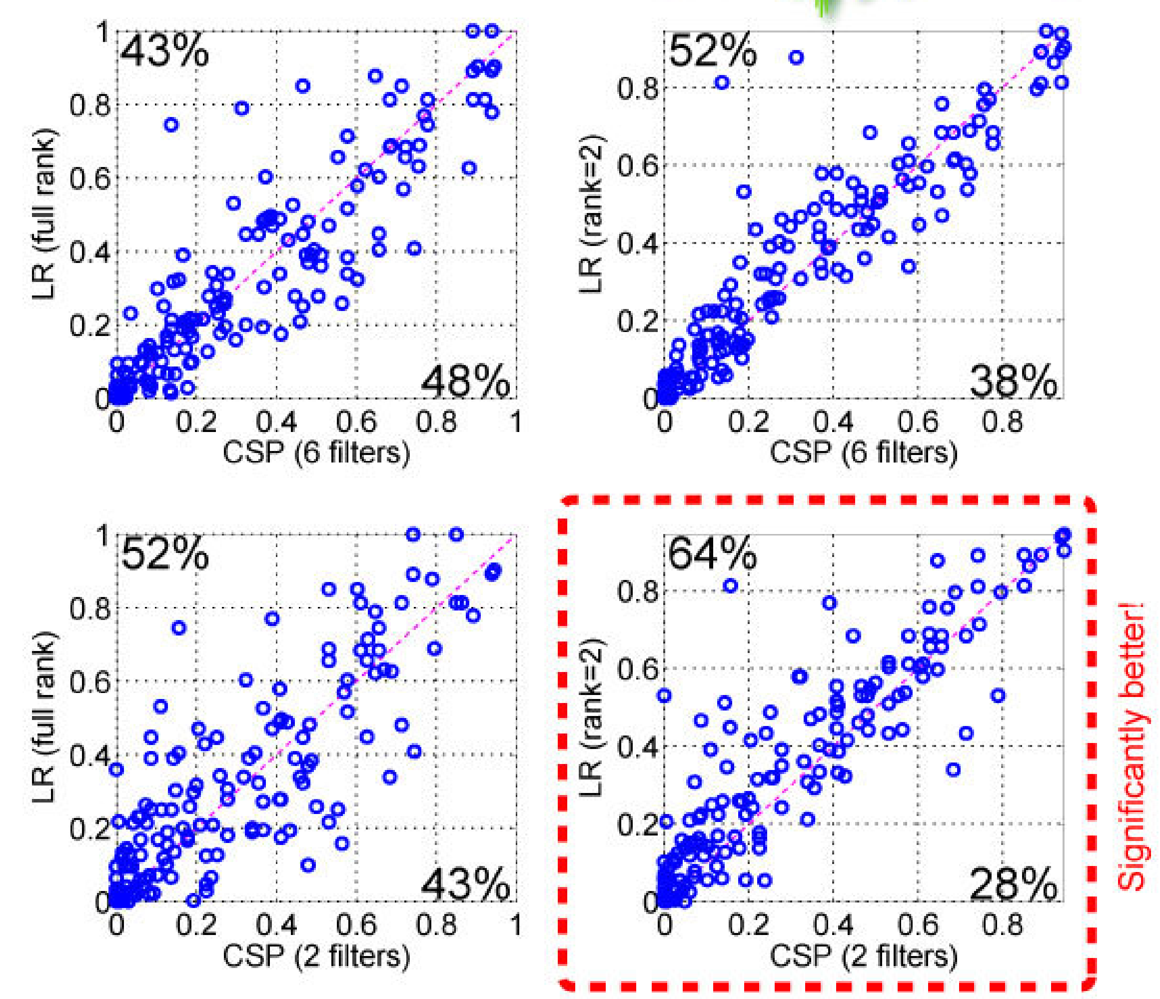
$$\min_{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \log\left(1+e^{-y_i \bar{f}(X_i;\bar{\boldsymbol{\theta}})}\right) + \frac{C}{2}\left(\boldsymbol{w}_1^\top\Sigma\boldsymbol{w}_1 + \boldsymbol{w}_2^\top\Sigma\boldsymbol{w}_2\right) \qquad (1)$$

The pooled covariance matrix is introduced to ensure invariance to linear transformations. $\Sigma = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$

$$\bar{f}(X;\bar{\boldsymbol{\theta}}) := \frac{1}{2}{\rm tr}\left[\left(-\boldsymbol{w}_1\boldsymbol{w}_1^\top + \boldsymbol{w}_2\boldsymbol{w}_2^\top\right) X X^\top\right] + b,$$

$$\bar{\boldsymbol{\theta}} := (\boldsymbol{w}_1, \boldsymbol{w}_2, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$$

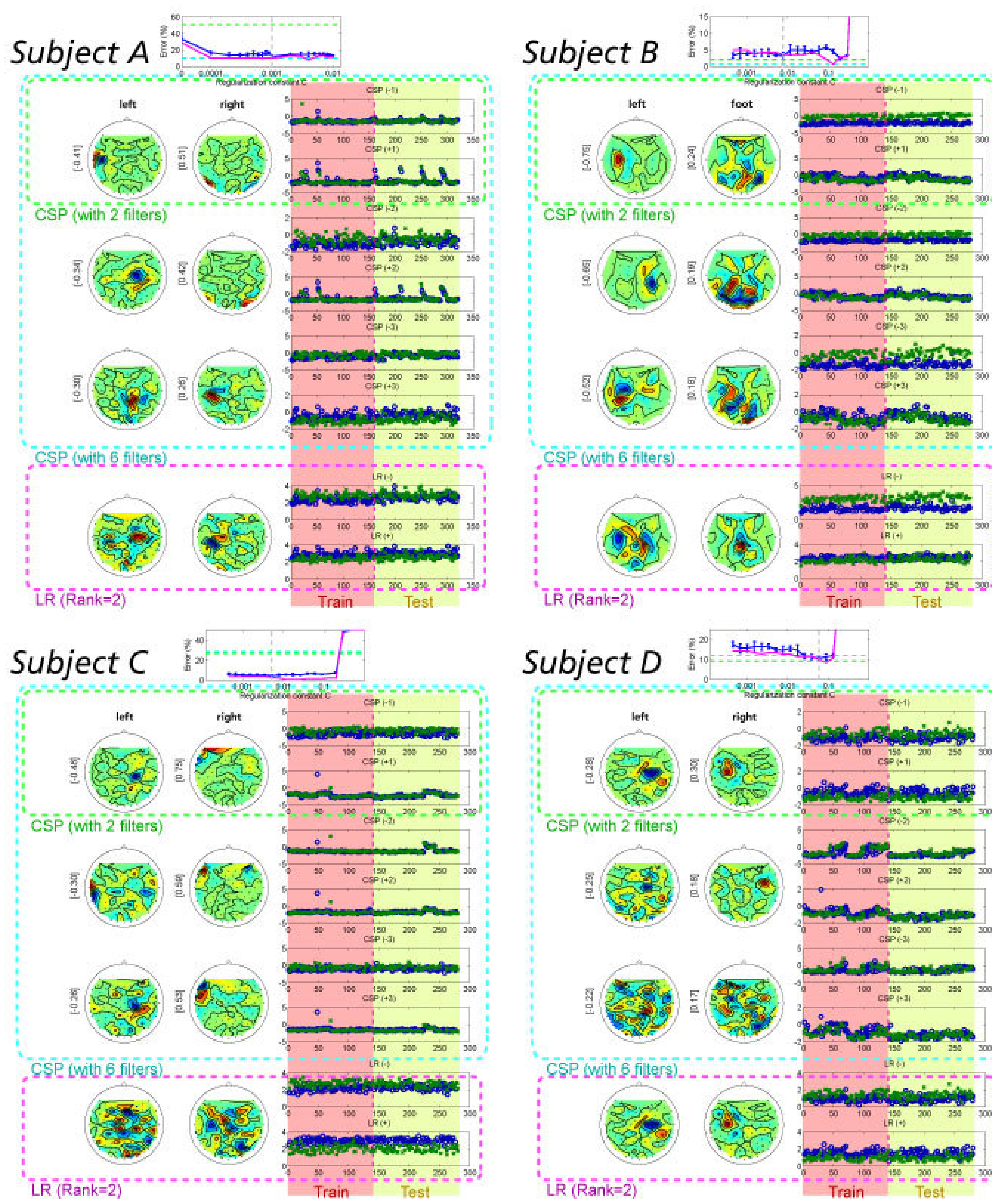
Classification performance



Experimental setups

Paradigm: motor imagination: "left hand", "right hand" or "foot".
Instruction: visual cue "L", "R", or "F" on the screen.
Feedback: no feedback (calibration measurements).
Sampling: recorded at 1000Hz. down-sampled to 100Hz.
Electrodes: 32, 64, or 128 channels. #subjects=29. #datasets=162.
Preprocessing: band-pass filtered at 7-30Hz.
Training and testing: chronological split, i.e., all methods are trained on the first half and tested on the second half.
Performance measure: bits per decision:

$$1-\left(p_{\rm err}\log_2\frac{1}{p_{\rm err}} + (1-p_{\rm err})\log_2\frac{1}{1-p_{\rm err}}\right)$$



Connection between the rank=2 parametrization and CSP

Differentiating (1) w.r.t. $\boldsymbol{w}_j$ at the optimum $\bar{\boldsymbol{\theta}}^*$

$$\pm\frac{1}{n}\sum_{i=1}^n \frac{e^{-z_i}}{1+e^{-z_i}} y_i X_i X_i^\top \boldsymbol{w}_j^* + C\Sigma\boldsymbol{w}_j^* = 0 \qquad (j=1,2)$$

where $z_i := y_i \bar{f}(X_i;\bar{\boldsymbol{\theta}}^*)$.

Recalling that $\Sigma = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$, we obtain a pair of GEPs (find eigenvectors with unit eigenvalues):

$$\Sigma^{(-)}(\bar{\boldsymbol{\theta}}^*, 0)\boldsymbol{w}_1^* = \Sigma^{(+)}(\bar{\boldsymbol{\theta}}^*, C)\boldsymbol{w}_1^*$$

$$\Sigma^{(+)}(\bar{\boldsymbol{\theta}}^*, 0)\boldsymbol{w}_2^* = \Sigma^{(-)}(\bar{\boldsymbol{\theta}}^*, C)\boldsymbol{w}_2^*$$

where we define the uncertainty weighted covariance matrix as:

$$\Sigma^{(\pm)}(\bar{\boldsymbol{\theta}}^*, C) = \sum_{i\in\mathcal{I}_\pm} \frac{e^{-z_i}}{1+e^{-z_i}} X_i X_i^\top + C\sum_{i=1}^n X_i X_i^\top$$

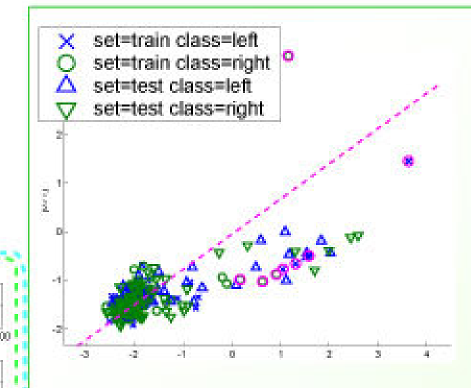
$= 1 - P(y=y_i|X=X_i)$ | Regularization

Small weight — Classifier is certain about the decision.

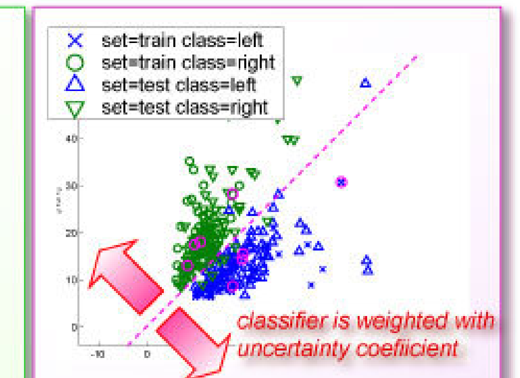
Large weight — Classifier is uncertain about the decision.

because,

$$z_i = y_i\bar{f}(X_i;\bar{\boldsymbol{\theta}}^*) = \log\frac{P(y=y_i|X=X_i)}{P(y\neq y_i|X=X_i)}$$



Low dimensional feature space using CSP for Subject A. "outliers" are marked magenda.



Low dimensional feature space using logistic regression (rank=2) for Subject A. "outliers" are marked magenda.

Conclusion

Logistic regression for the classification of single trial motor-imagery EEG
- Classifies linearly in the space of variance and covariances
- a discriminative model, i.e., the marginal density P(X) is left unestimated.
- the classifier output has a probabilistic interpretation.
- the classification problem is formulated under a single optimization problem.

(a) *Full-rank symmetric matrix parametrization*
- convex optimization.
- connection to a generative model.

(b) *Rank=2 approximation*
- improved physiological interpretability.
- connection to CSP [Koles, 1991].
- comparable or favorable performance to CSP with 6 filters or 2 filters, respectively.