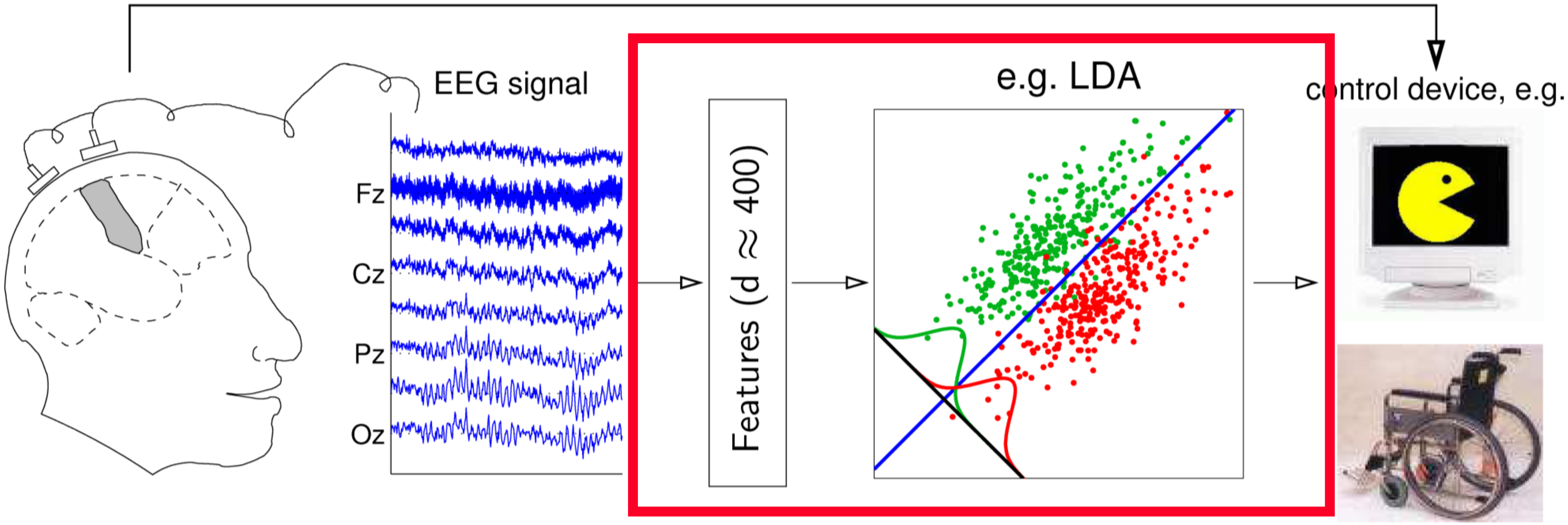
BCI and Nonstationarity

Klaus-Robert Müller, Siamac Fazli, Paul von Bünau, Frank Meinecke, Wojciech Samek, Gabriel Curio, Benjamin Blankertz et al.

Noninvasive Brain-Computer Interface



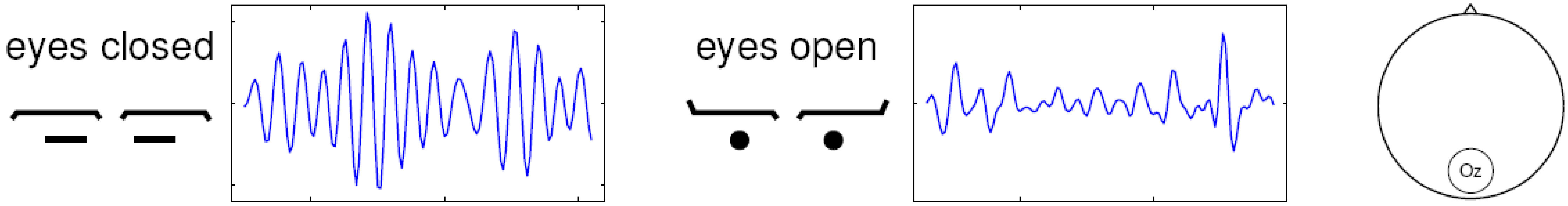
DECODING

BCI: Translation of human intentions into a technical control signal
without using activity of muscles or peripheral nerves

Towards imaginations: Modulation of Brain Rhythms

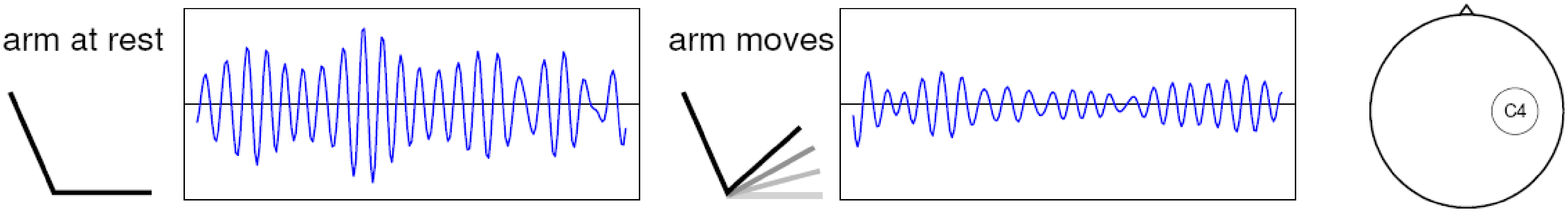
Most rhythms are idle rhythms, i.e., they are **attenuated** during activation.

- α-rhythm (around 10 Hz) in visual cortex:



Single channel

- μ-rhythm (around 10 Hz) in motor and sensory cortex:



IMAGINATION of left arm

BBCI paradigms

Leitmotiv: *›let the machines learn‹*

- healthy subjects *untrained* for BCI

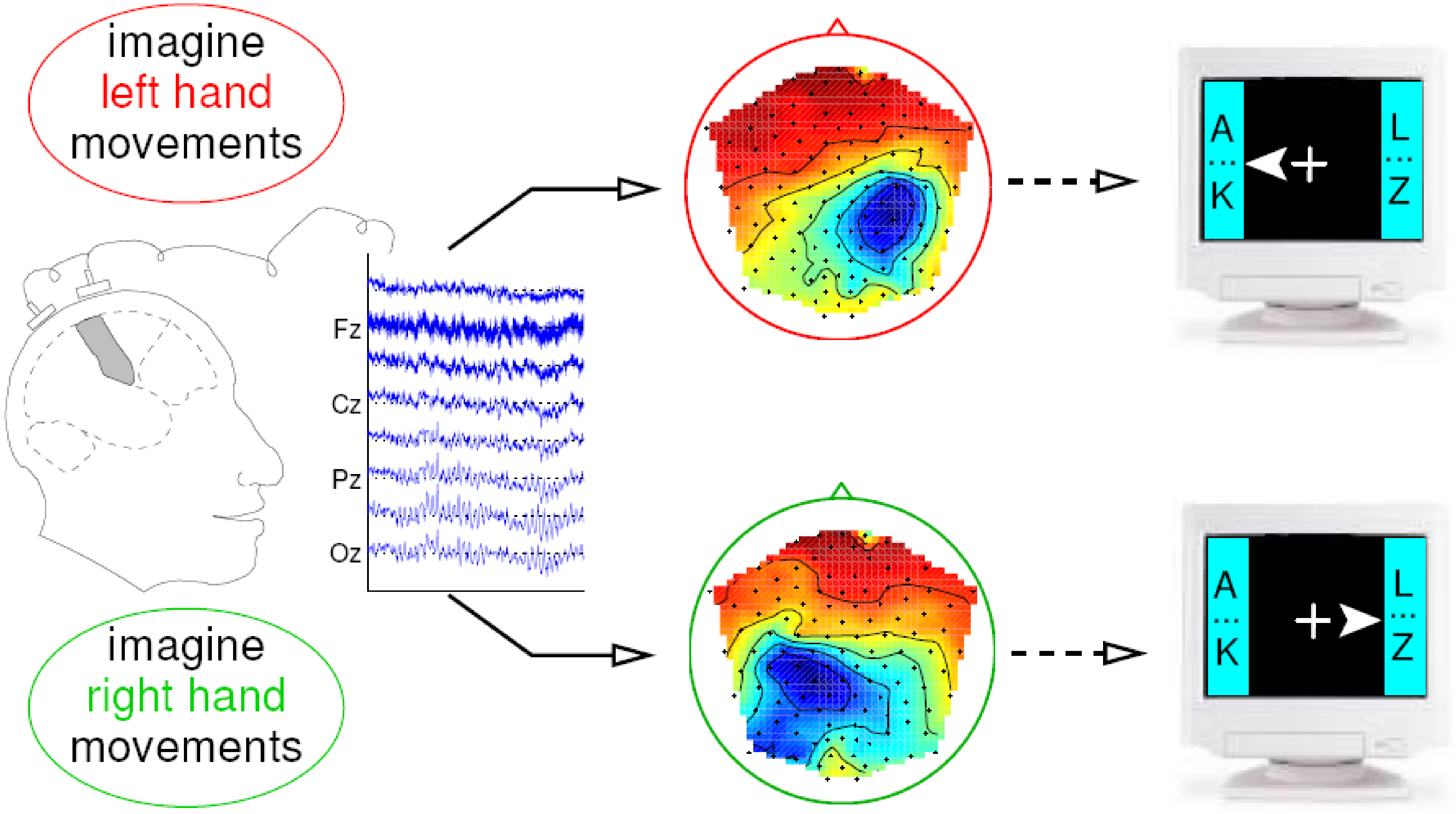
A: training <10min: right/left hand **imagined** movements

→ infer the respective brain acivities (ML & SP)

B: online feedback session

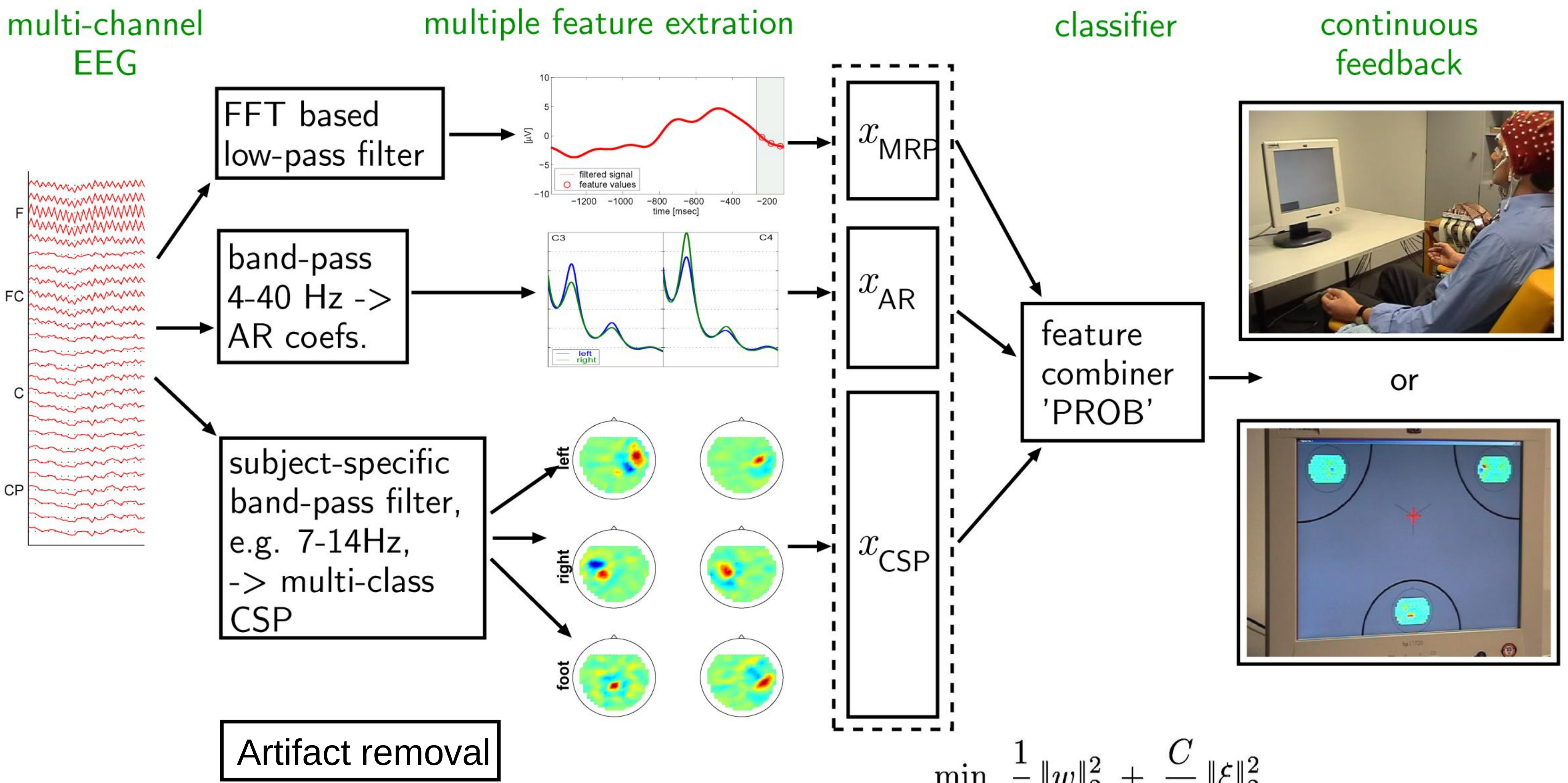
Playing with BCI: training session (20 min)

Machine learning approach to BCI: infer prototypical pattern



Inference by CSP Algorithm

BBCI Set-up

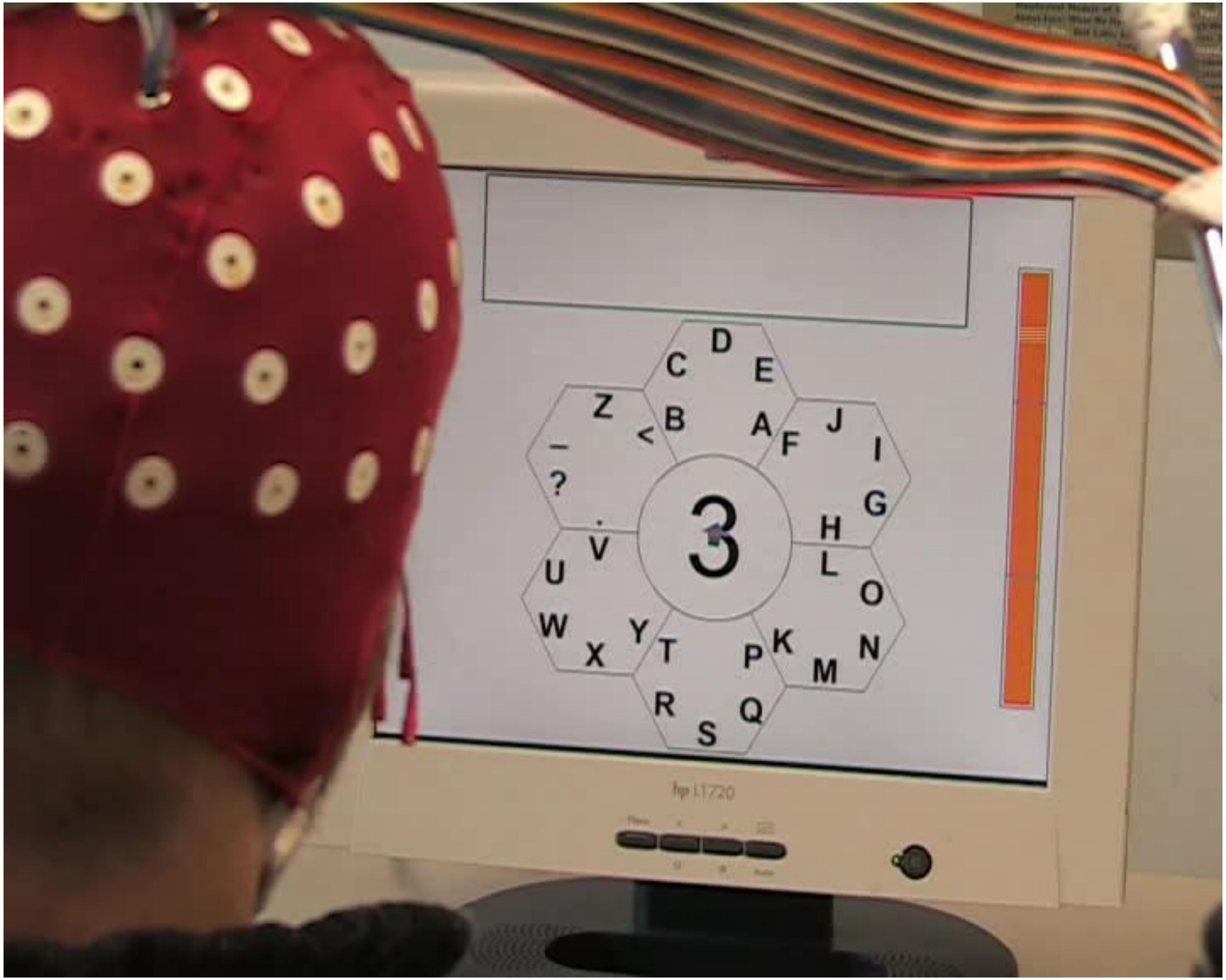


$$\min_{w,b,\xi} \ \frac{1}{2}\,\|w\|_2^2 \ + \ \frac{C}{K}\,\|\xi\|_2^2$$

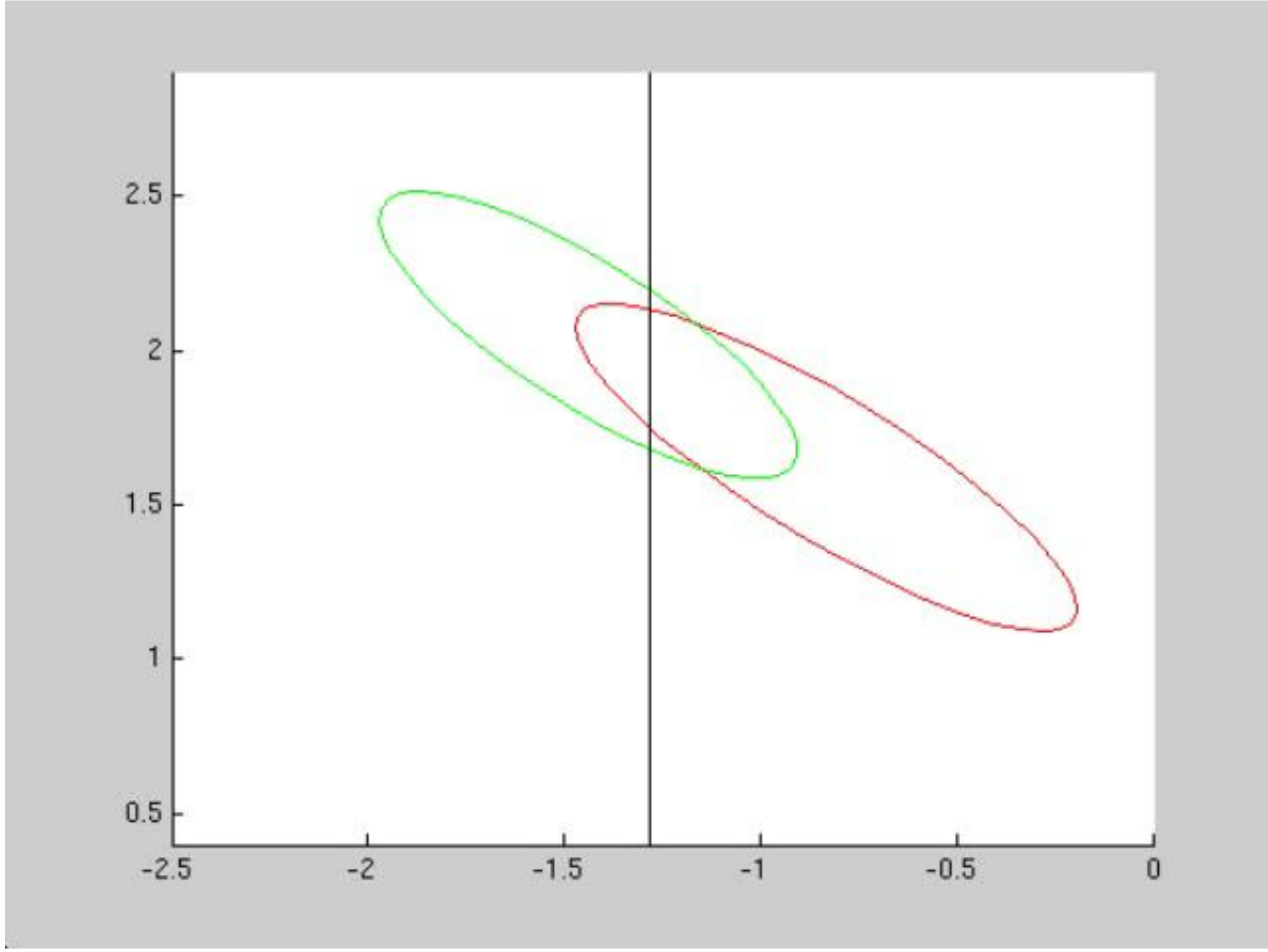
$$\text{subject to} \qquad y_k(w^\top x_k + b) = 1 - \xi_k \quad \text{for } k = 1, \ldots, K$$

[cf. Müller et al. 2001, 2007, 2008, Dornhege et al. 2003, 2007, Blankertz et al. 2004, 2005, 2006, 2007, 2008]

Spelling with BBCI: a communication for the disabled



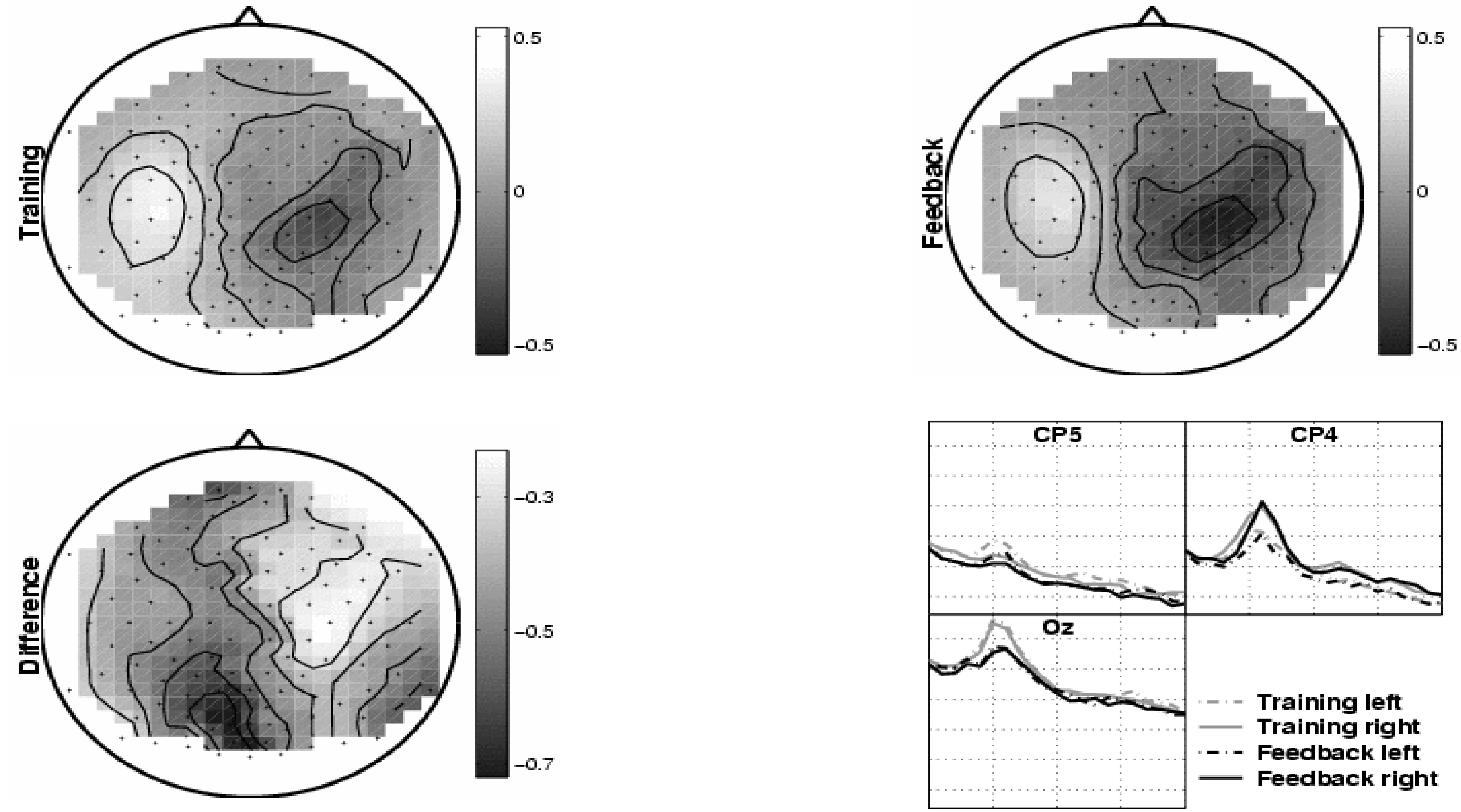
Future Issues: Shifting distributions within experiment



Mathematical flavors of non-stationarity

- Bias adaptation between training and test $f(x) = w\,x + \mathbf{b}$
- Invariant features
- Covariate shift
- SSA: projecting to stationary subspaces
- Nonstationarity due to subject dependence: Mixed effects model
- Transfering nonstationarity
- Co-adaptation …

Neurophysiological analysis



[cf. Krauledat et al. 07]

Weighted Linear Regression for covariate shift compensation

Given training samples

$$\{(\boldsymbol{x}_i, y_i) \mid y_i = f(\boldsymbol{x}_i) + \epsilon_i\}_{i=1}^{n}$$

for some function f and linearly independent basis functions $\Phi = \{\varphi_i(\boldsymbol{x})\}_{i=1}^{p}$, find

$\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_p^*)^\top$ which minimizes

$$\min_{\{\alpha_i\}_{i=1}^{p}} \left[\sum_{i=1}^{n} w(\boldsymbol{x}_i) \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 + \langle \boldsymbol{R}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle \right].$$

$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{p} \alpha_i \varphi_i(\boldsymbol{x})$, choosing $w(\boldsymbol{x}_i) = \frac{p_{fb}(\boldsymbol{x}_i)}{p_{tr}(\boldsymbol{x}_i)}$ yields **unbiased** estimator even under covariate shift

[cf. Sugiyama & Müller 2005, Sugiyama et al. JMLR 2007]

Projections $\longleftrightarrow$ Nonstationary

Source separation paradigms

Principal Component Analysis (PCA)

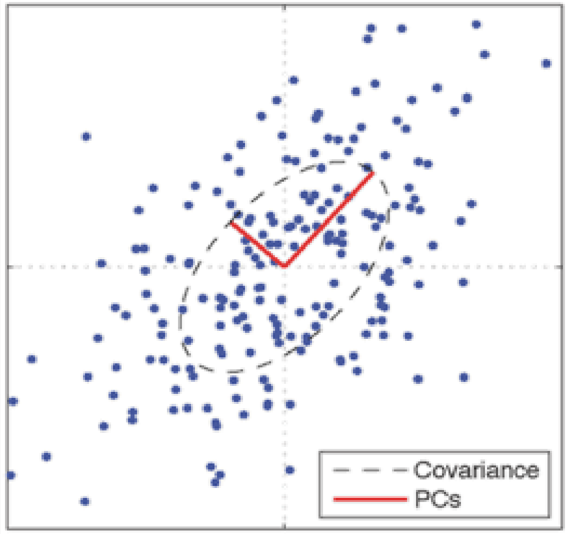
uncorrelated sources

orthogonal mixing

max. variance

$$X = A \begin{bmatrix} S^{(1)} \\ \vdots \\ S^{(d)} \end{bmatrix}$$

min. variance



Source separation paradigms

Principal Component Analysis (PCA)

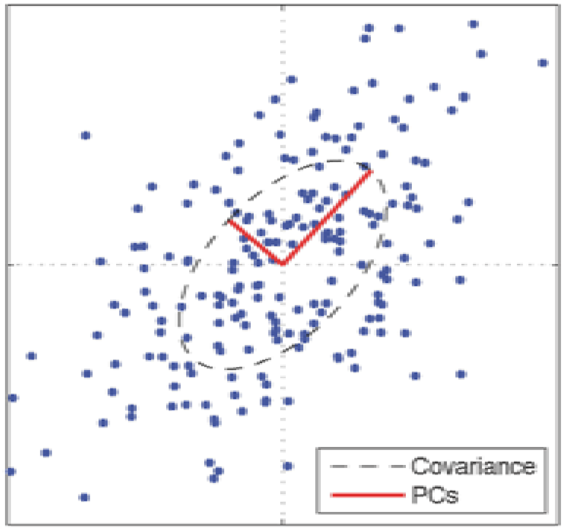
uncorrelated sources

orthogonal mixing

max. variance

$$X = A \begin{bmatrix} S^{(1)} \\ \vdots \\ S^{(d)} \end{bmatrix}$$

min. variance

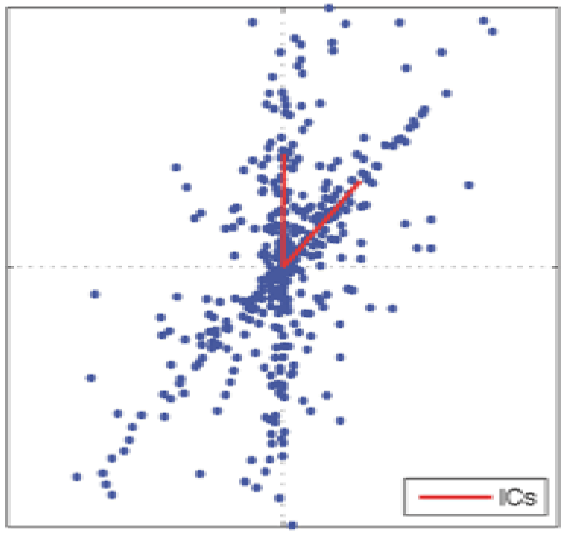


Independent Component Analysis (ICA)

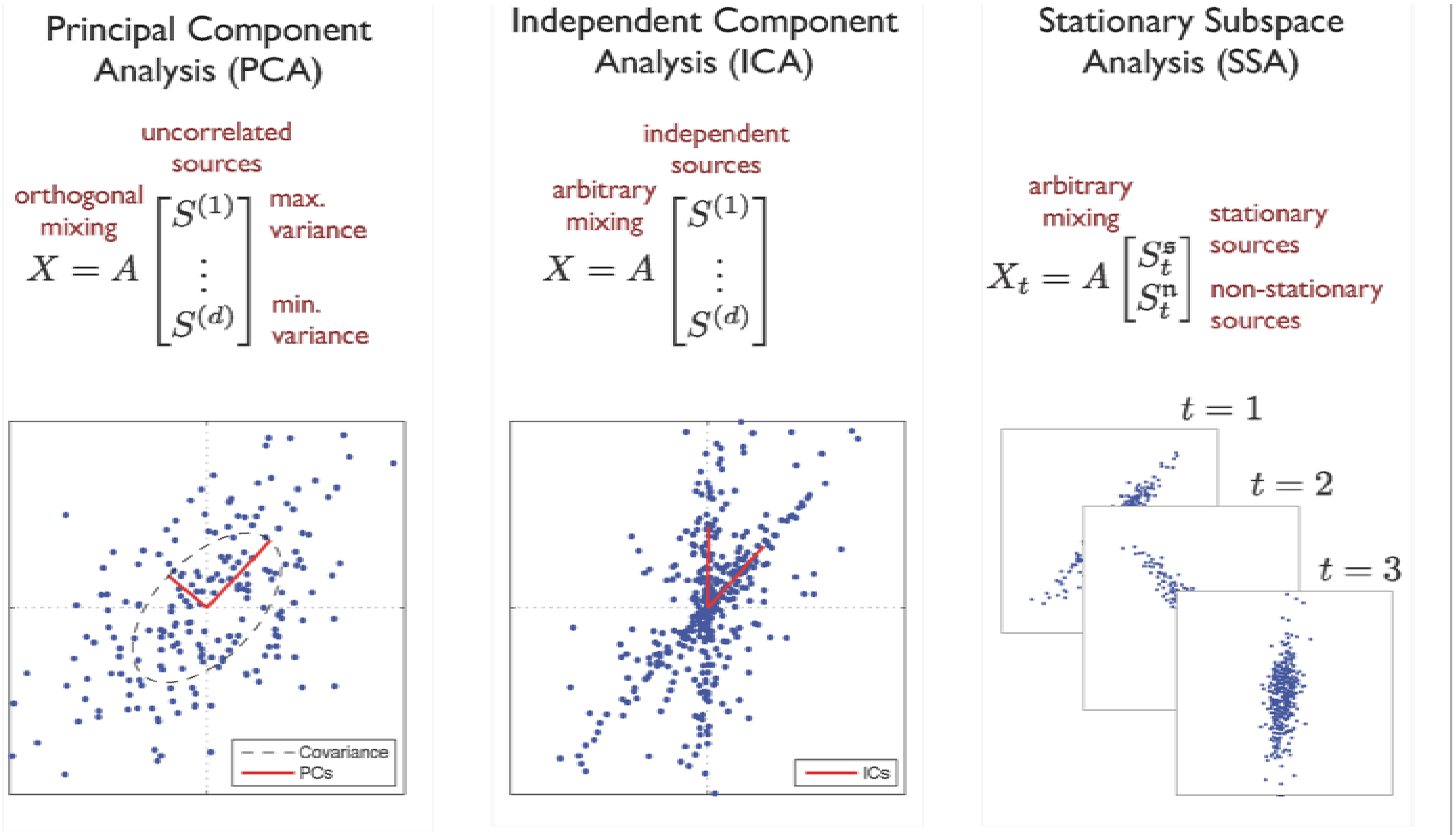
independent sources

arbitrary mixing

$$X = A \begin{bmatrix} S^{(1)} \\ \vdots \\ S^{(d)} \end{bmatrix}$$



Source separation paradigms



The Stationary Subspace Analysis model

Linear mixing of stationary and non-stationary sources

$$
\underbrace{X_t}_{\text{observed D-variate data}} = A \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} \underbrace{A^{\mathfrak{s}}}_{\text{stationary subspace}} & \underbrace{A^{\mathfrak{n}}}_{\text{non-stationary subspace}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix} \begin{matrix} \text{d stationary sources} \\ \text{D-d non-stationary sources} \end{matrix}
$$

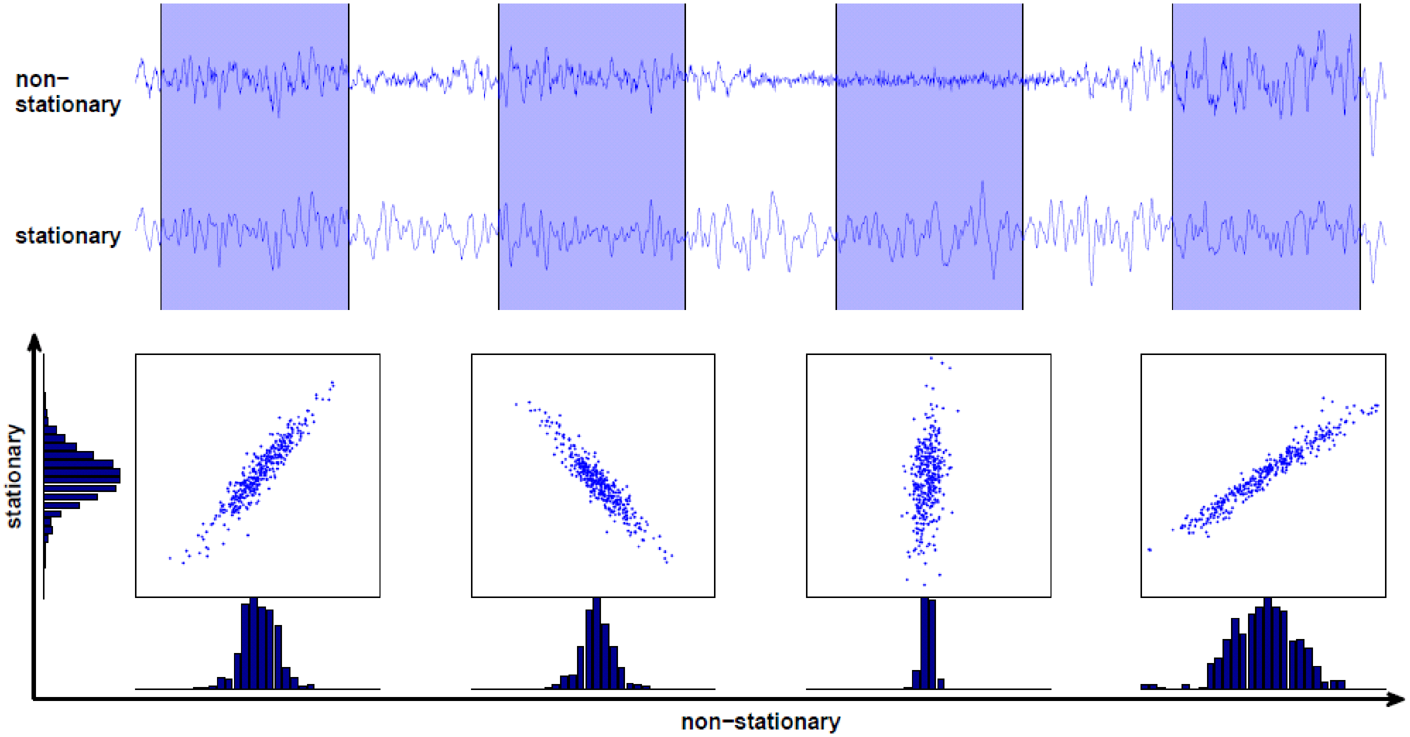
A source is *stationary* if its mean and covariance is constant over time, i.e.

$$\mathbb{E}[S_{t_1}] = \mathbb{E}[S_{t_1}] \quad \text{and} \quad \mathbb{E}[S_{t_1} S_{t_1}^\top] = \mathbb{E}[S_{t_2} S_{t_2}^\top]$$

for all time points t_1, t_2

[von Bünau P, Meinecke F C, Kiraly F J and Müller K-R. *Phys. Rev. Letter*, 2009]

Splitting into stationary and nonstationary subspace: SSA



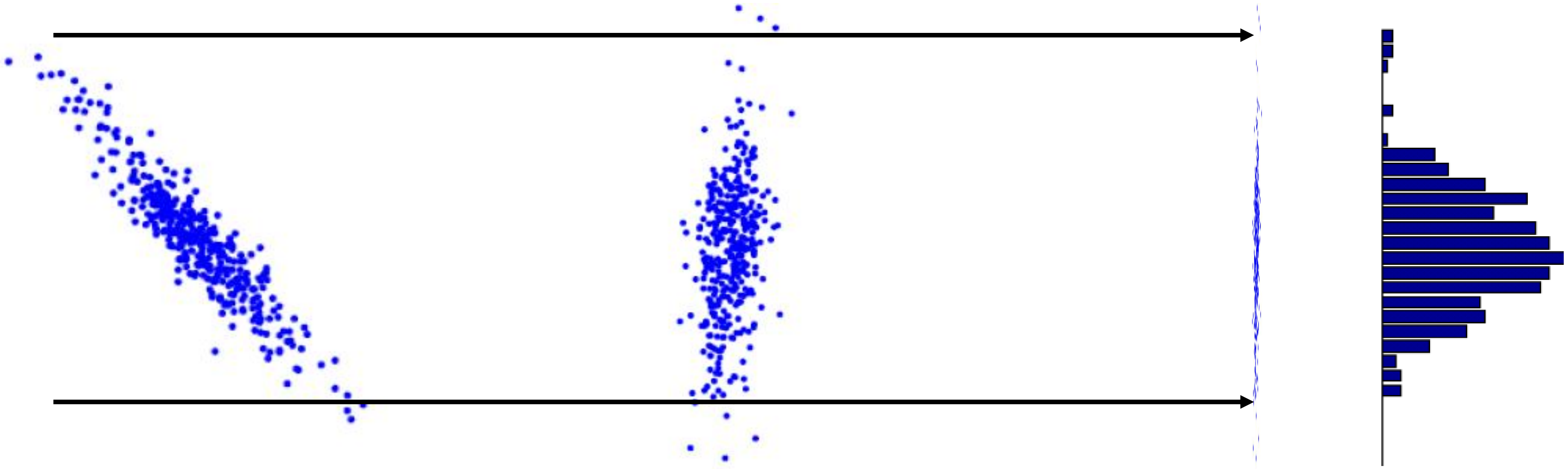
- d stationary source signals $s^{\mathfrak{s}}(t) \in \mathbb{R}^d$
- $D - d$ non-stationary source signals $s^{\mathfrak{n}}(t) \in \mathbb{R}^{(D-d)}$
- Observed signals: instantaneous linear superpositions of sources

$$x(t) = As(t) = \begin{bmatrix} A^{\mathfrak{s}} & A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix}$$

invert

[cf. Bünau, Meinecke, Kiraly, Müller PRL 09]

SSA



given: Epochs X_i of Data points in $\mathbb{C}^n$

wanted: Linear subspace S of $\mathbb{C}^n$ such that marginalized data sets $X_i\,|_S$ look the same

„stationary projection”

Inverting the SSA model

Aim of SSA: find a demixing matrix $\hat{A}^{-1} = \begin{bmatrix} B^{\mathfrak{s}} \\ B^{\mathfrak{n}} \end{bmatrix}$

Projection to the stationary sources

Projection to the non-stationary sources

... that separates the two groups of sources in the observed data.

Is this inverse unique?

$$\begin{bmatrix} \hat{S}_t^{\mathfrak{s}} \\ \hat{S}_t^{\mathfrak{n}} \end{bmatrix} = \hat{A}^{-1} \begin{bmatrix} A^{\mathfrak{s}} & A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} B^{\mathfrak{s}} A^{\mathfrak{s}} & B^{\mathfrak{s}} A^{\mathfrak{n}} \\ B^{\mathfrak{n}} A^{\mathfrak{s}} & B^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix}$$

Estimated sources

De-mixing

Observed data

Latent sources

Inverting the SSA model

Aim of SSA: find a demixing matrix $\hat{A}^{-1} = \begin{bmatrix} B^{\mathfrak{s}} \\ B^{\mathfrak{n}} \end{bmatrix}$

Projection to the stationary sources

Projection to the non-stationary sources

... that separates the two groups of sources in the observed data.

Is this inverse unique?

$$\begin{bmatrix} \hat{S}_t^{\mathfrak{s}} \\ \hat{S}_t^{\mathfrak{n}} \end{bmatrix} = \hat{A}^{-1} \begin{bmatrix} A^{\mathfrak{s}} & A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} B^{\mathfrak{s}} A^{\mathfrak{s}} & \overbrace{B^{\mathfrak{s}} A^{\mathfrak{n}}}^{=0} \\ B^{\mathfrak{n}} A^{\mathfrak{s}} & B^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix}$$

Estimated sources

De-mixing

Observed data

Latent sources

Inverting the SSA model

Aim of SSA: find a demixing matrix

$$\hat{A}^{-1} = \begin{bmatrix} B^{\mathfrak{s}} \\ B^{\mathfrak{n}} \end{bmatrix}$$

Projection to the stationary sources

Projection to the non-stationary sources

... that separates the two groups of sources in the observed data.

Is this inverse unique?

= 0

$$\begin{bmatrix} \hat{S}_t^{\mathfrak{s}} \\ \hat{S}_t^{\mathfrak{n}} \end{bmatrix} = \hat{A}^{-1} \begin{bmatrix} A^{\mathfrak{s}} & A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} B^{\mathfrak{s}} A^{\mathfrak{s}} & B^{\mathfrak{s}} A^{\mathfrak{n}} \\ B^{\mathfrak{n}} A^{\mathfrak{s}} & B^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix}$$

Estimated sources

De-mixing

Observed data

arbitrary!

Latent sources

Arbitrary because:

- “nonstationary + stationary = nonstationary”

Inverting the SSA model

Aim of SSA: find a demixing matrix $\hat{A}^{-1} = \begin{bmatrix} B^{\mathfrak{s}} \\ B^{\mathfrak{n}} \end{bmatrix}$

Projection to the stationary sources

Projection to the non-stationary sources

... that separates the two groups of sources in the observed data.

Is this inverse unique?

$$\underbrace{\begin{bmatrix} \hat{S}^{\mathfrak{s}}_t \\ \hat{S}^{\mathfrak{n}}_t \end{bmatrix}}_{\text{Estimated sources}} = \underbrace{\hat{A}^{-1}}_{\text{De-mixing}} \underbrace{\begin{bmatrix} A^{\mathfrak{s}} & A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S^{\mathfrak{s}}_t \\ S^{\mathfrak{n}}_t \end{bmatrix}}_{\text{Observed data}} = \begin{bmatrix} B^{\mathfrak{s}} A^{\mathfrak{s}} & \overbrace{B^{\mathfrak{s}} A^{\mathfrak{n}}}^{= 0} \\ B^{\mathfrak{n}} A^{\mathfrak{s}} & B^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} \underbrace{\begin{bmatrix} S^{\mathfrak{s}}_t \\ S^{\mathfrak{n}}_t \end{bmatrix}}_{\text{Latent sources}}$$

arbitrary!

Arbitrary because:

- “nonstationary + stationary = nonstationary”
- Linear transformations do not alter stationarity/nonstationarity

Inverting the SSA model

Aim of SSA: find a demixing matrix $\hat{A}^{-1} = \begin{bmatrix} B^{\mathfrak{s}} \\ B^{\mathfrak{n}} \end{bmatrix}$

Projection to the stationary sources

Projection to the non-stationary sources

... that separates the two groups of sources in the observed data.

Is this inverse unique?

$$\begin{bmatrix} \hat{S}_t^{\mathfrak{s}} \\ \hat{S}_t^{\mathfrak{n}} \end{bmatrix} = \hat{A}^{-1} \begin{bmatrix} A^{\mathfrak{s}} & A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} B^{\mathfrak{s}} A^{\mathfrak{s}} & \overset{=0}{B^{\mathfrak{s}} A^{\mathfrak{n}}} \\ B^{\mathfrak{n}} A^{\mathfrak{s}} & B^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix}$$

Estimated sources | De-mixing | Observed data | arbitrary! | Latent sources

Arbitrary because:

- "nonstationary + stationary = nonstationary"
- Linear transformations do not alter stationarity/nonstationarity

Identifiability

$$\begin{bmatrix} \hat{S}_t^{\mathfrak{s}} \\ \hat{S}_t^{\mathfrak{n}} \end{bmatrix} = \hat{A}^{-1} A \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} B^{\mathfrak{s}} A^{\mathfrak{s}} & \overset{=0}{B^{\mathfrak{s}} A^{\mathfrak{n}}} \\ B^{\mathfrak{n}} A^{\mathfrak{s}} & B^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}} \\ S_t^{\mathfrak{n}} \end{bmatrix}$$

Estimated stationary and non-stationary sources

De-mixing

Observed data

arbitrary!

Latent sources

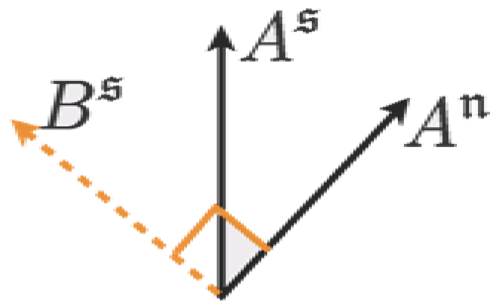
Identifiability

$$\begin{bmatrix}\hat{S}_t^{\mathfrak{s}}\\ \hat{S}_t^{\mathfrak{n}}\end{bmatrix} = \hat{A}^{-1} A \begin{bmatrix} S_t^{\mathfrak{s}}\\ S_t^{\mathfrak{n}}\end{bmatrix} = \begin{bmatrix} B^{\mathfrak{s}} A^{\mathfrak{s}} & \overset{=0}{B^{\mathfrak{s}} A^{\mathfrak{n}}}\\ B^{\mathfrak{n}} A^{\mathfrak{s}} & B^{\mathfrak{n}} A^{\mathfrak{n}}\end{bmatrix} \begin{bmatrix} S_t^{\mathfrak{s}}\\ S_t^{\mathfrak{n}}\end{bmatrix}$$

Estimated stationary and non-stationary sources | De-mixing | Observed data | arbitrary! | Latent sources

We can identify:

- the true non-stationary space
- the true stationary sources (up to linear transformations)



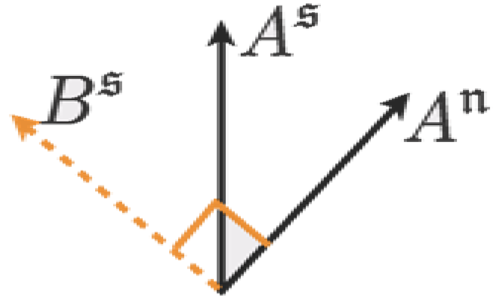
Identifiability

$$\begin{bmatrix}\hat{S}^{\mathfrak{s}}_t \\ \hat{S}^{\mathfrak{n}}_t\end{bmatrix} = \hat{A}^{-1} A \begin{bmatrix}S^{\mathfrak{s}}_t \\ S^{\mathfrak{n}}_t\end{bmatrix} = \begin{bmatrix} B^{\mathfrak{s}} A^{\mathfrak{s}} & \overset{= 0}{B^{\mathfrak{s}} A^{\mathfrak{n}}} \\ B^{\mathfrak{n}} A^{\mathfrak{s}} & B^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix}S^{\mathfrak{s}}_t \\ S^{\mathfrak{n}}_t\end{bmatrix}$$

Estimated stationary and non-stationary sources | De-mixing | Observed data | arbitrary! | Latent sources

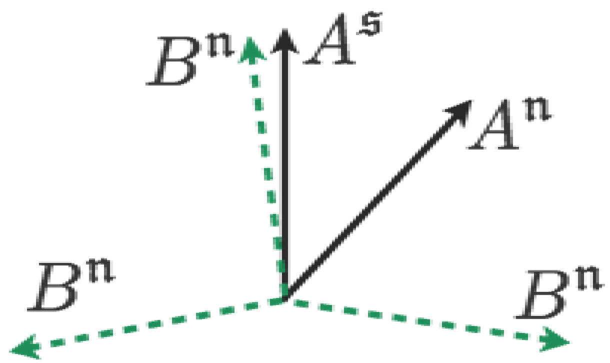
We can identify:

- the true non-stationary space
- the true stationary sources (up to linear transformations)



We cannot identify:

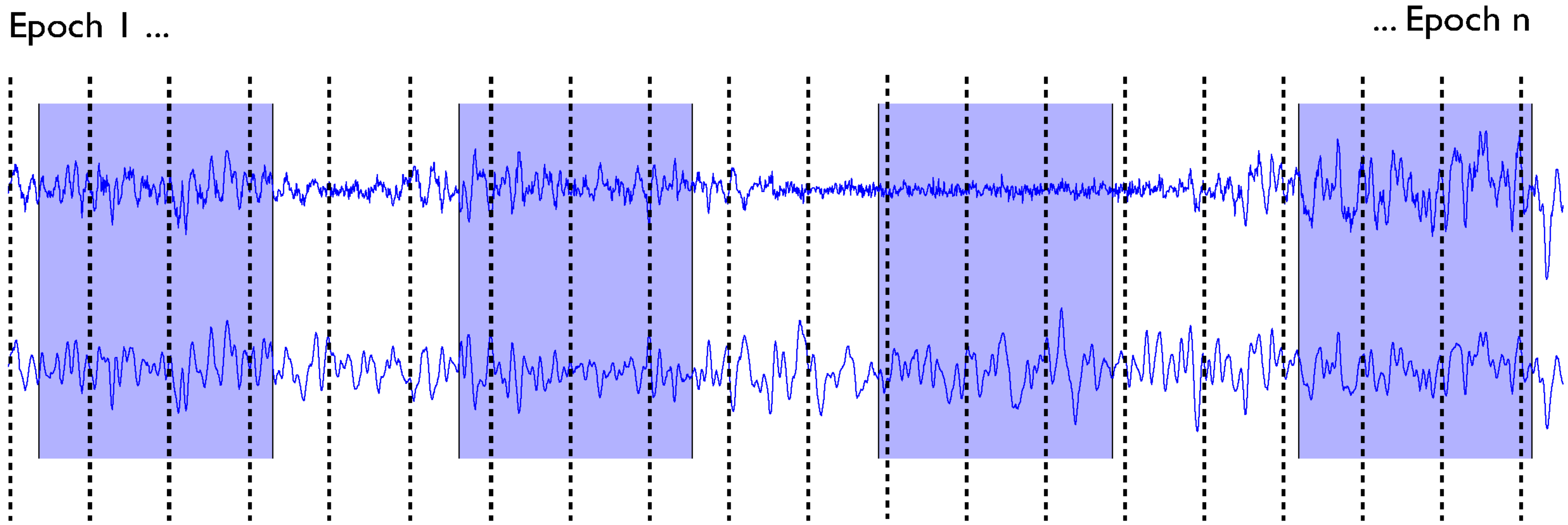
- the true stationary space
- the true non-stationary sources



In practice: find the *most* nonstationary sources!

The SSA algorithm

Divide the data into epochs (consecutive or sliding window)



Estimate the epoch mean and covariance matrix.

μ_1, Σ_1 $\cdots$ μ_n, Σ_n

The algorithm: optimizing stationarity

Find the two projections by minimizing/maximizing a measure of stationarity

$$\hat{A}^{-1} = \begin{bmatrix} B^{\mathfrak{s}} \\ B^{\mathfrak{n}} \end{bmatrix}$$

Projection to the stationary sources

Projection to the non-stationary sources

Measure of non-stationarity: KL-divergence between each epoch and the average epoch using a Gaussian approximation.

$$B^{\mathfrak{s}} = \operatorname*{argmin}_{B} \sum_{i=1}^{n} D_{\mathrm{KL}} \left[\underbrace{\mathcal{N}(B\mu_i, B\Sigma_i B^{\top})}_{\text{Epoch i}}, \underbrace{\mathcal{N}(B\bar{\mu}_i, B\bar{\Sigma}_i B^{\top})}_{\text{Average epoch}} \right]$$

Find $B^{\mathfrak{n}}$ by *maximizing* this loss function.

Simplifying the objective (symmetries!)

Without loss of generality we can:

(a) set the average mean to zero;

(b) whiten the average covariance matrix; and

(c) constraint ourselves to projections with orthogonal rows.

$$B^{\mathfrak{s}} = \underset{B}{\operatorname{argmin}} \sum_{i=1}^{n} D_{\mathrm{KL}} \left[\mathcal{N}(B\mu_i, B\Sigma_i B^\top), \mathcal{N}(B\bar{\mu}_i, B\bar{\Sigma}_i B^\top) \right]$$

(a) (b)

$$= \underset{BB^\top = I}{\operatorname{argmin}} \sum_{i=1}^{n} D_{\mathrm{KL}} \left[\mathcal{N}(B\mu_i, B\Sigma_i B^\top), \mathcal{N}(0, I) \right]$$

(c)

$$= \underset{BB^\top = I}{\operatorname{argmin}} \sum_{i=1}^{n} -\log\det(B\Sigma_i B^\top) + \|B\mu_i\|^2$$

This means: $\hat{A}^{-1} = BW$ where the whitening is $W = \bar{\Sigma}^{-\frac{1}{2}}$

rotation whitening

Optimizing in the special orthogonal group

Multiplicative update of the rotation part

$$B^{\text{new}} \leftarrow R B^{\text{old}}$$

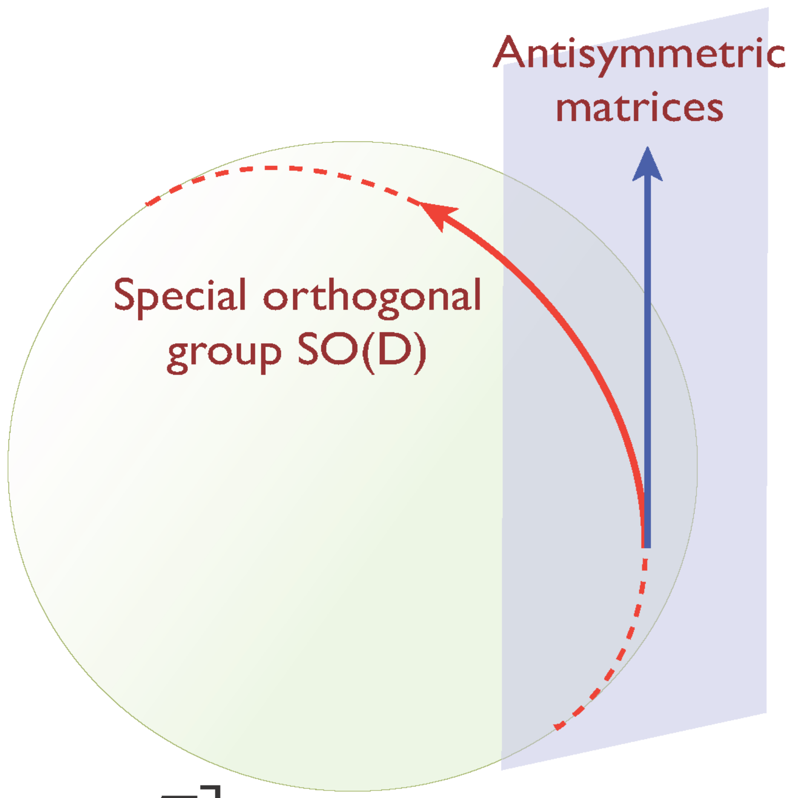
update rotation

Parametrize the update R as the matrix exponential of an antisymmetric matrix M

$$R = \exp(M) \text{ with } M^\top = -M$$

Interpretation: M_{ij} rotation angle of axis i towards axis j

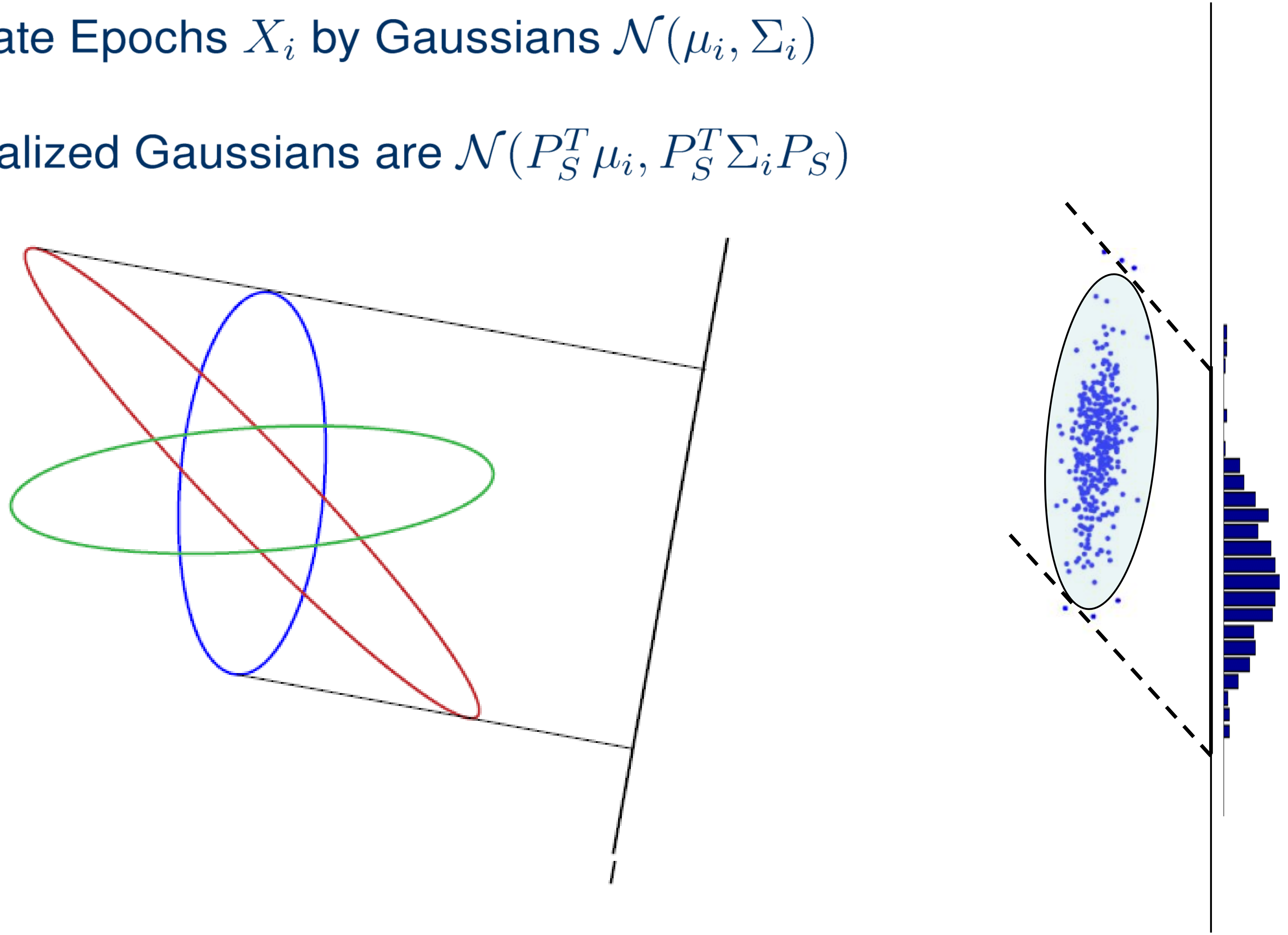
This leads to a gradient of the form: $\left.\dfrac{\partial L_{B^{\text{old}}}}{\partial M}\right|_{M=0} = \begin{bmatrix} 0 & Z \\ -Z^\top & 0 \end{bmatrix}$



SSA: how many epochs?

Estimate Epochs X_i by Gaussians $\mathcal{N}(\mu_i, \Sigma_i)$

Marginalized Gaussians are $\mathcal{N}(P_S^T \mu_i, P_S^T \Sigma_i P_S)$



Identifiability: theoretical results

Theorem

If the non-stationarity affects *both the mean and the covariance matrix*, then we need

number of non-stat. directions

$$n > \frac{D-d}{2} + 1 \quad \text{epochs}$$

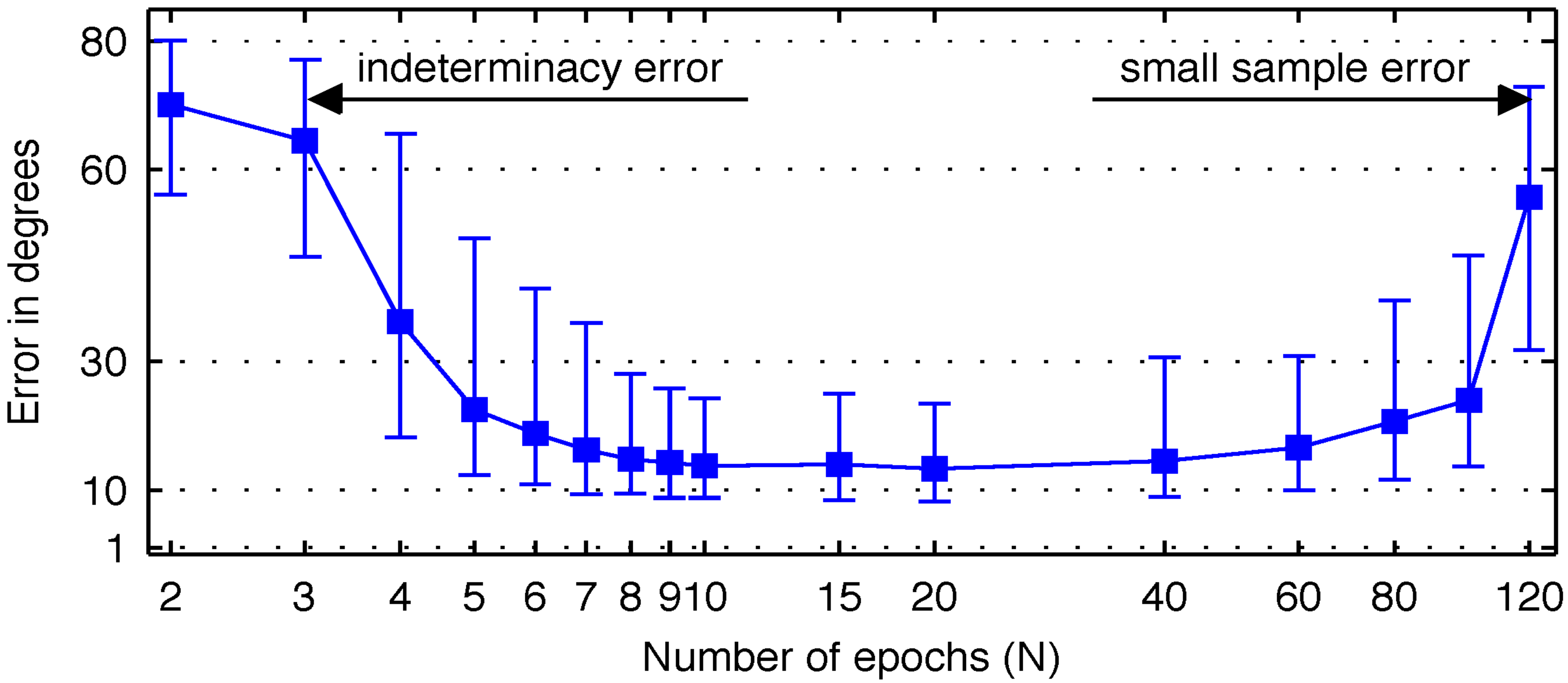
in order to guarantee that there are no spurious stationary directions.

If the *mean is constant* we need

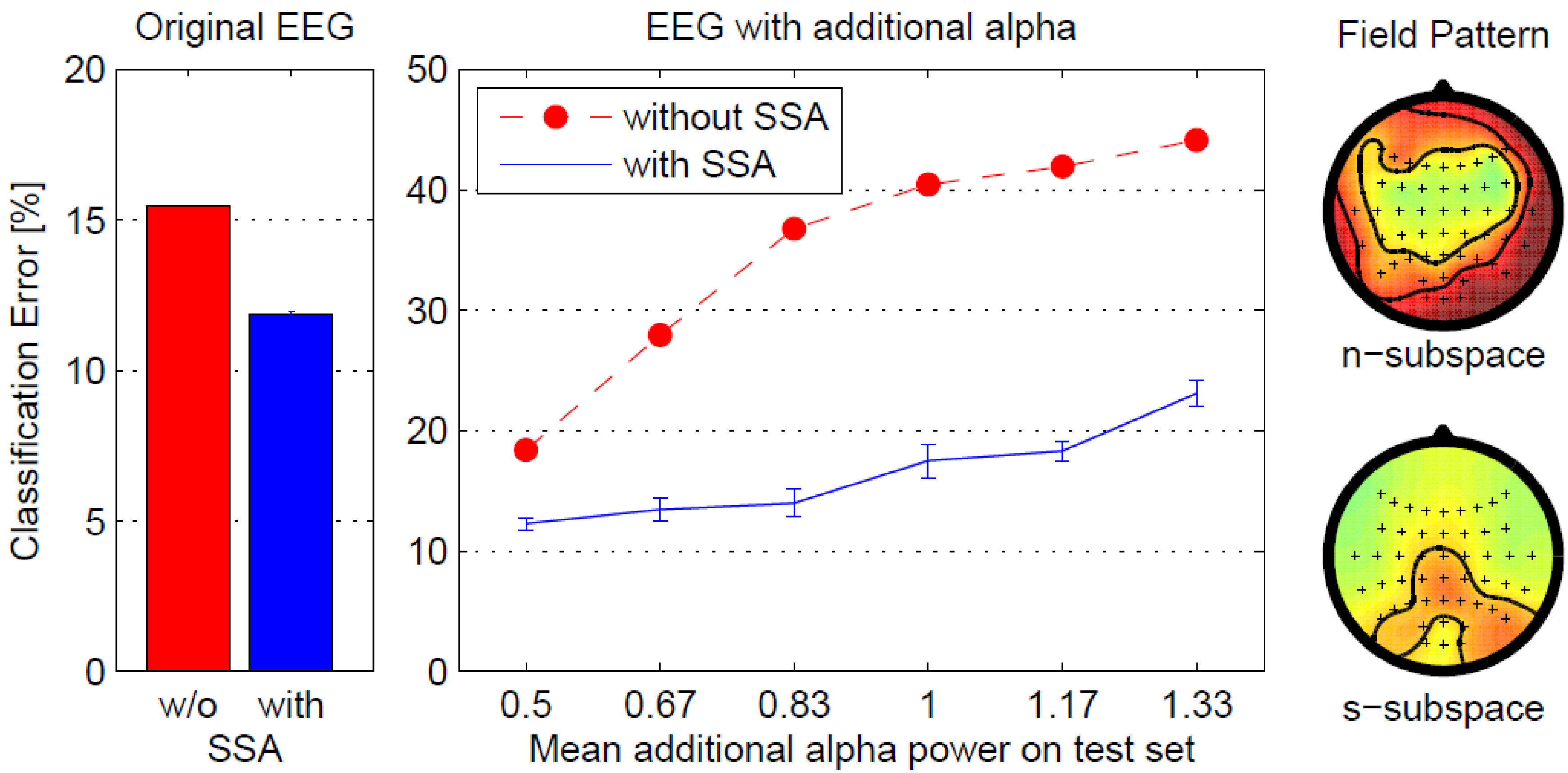
$$n > D - d + 1 \quad \text{epochs.}$$

Simulations on synthetic data

- Number of dimensions D=8 with four stationary sources d=4
- Total number of samples: 1000
- Error measure: subspace angle between the true and the found non-stationary subspace

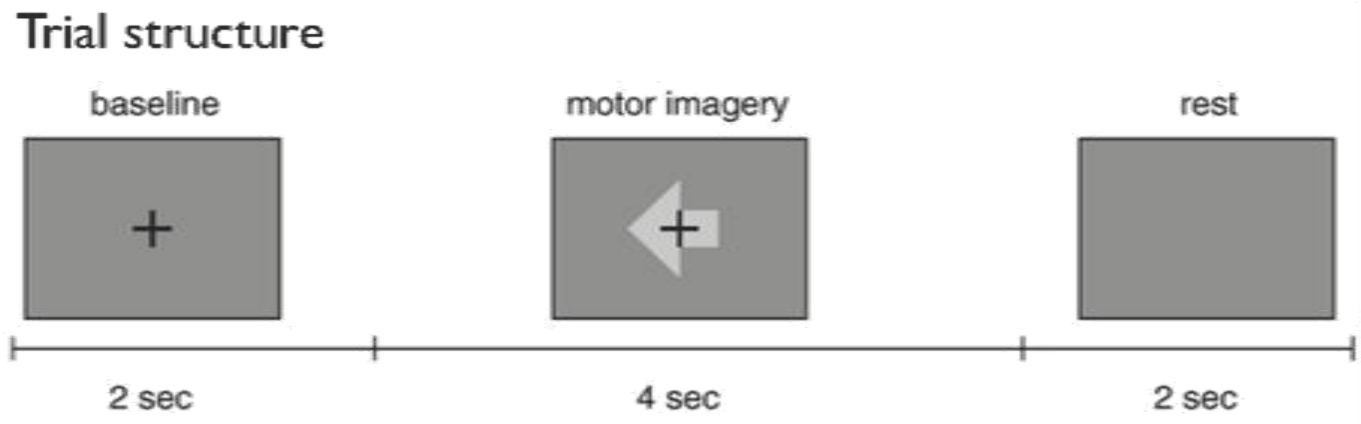


Application to Brain-Computer-Interfacing



Application to EEG analysis

Brain-Computer-Interfacing experiment: *imagined* movements leading to event-related-desynchronization (ERD)



Dataset

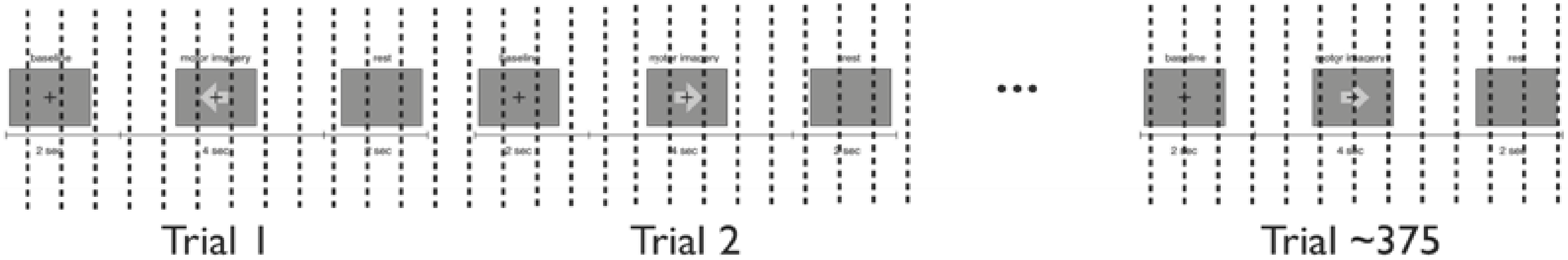
- 40 subjects
- Classes: left/right/foot
- ~125 trials per class
- 88 EEG channels

[Blankertz, B,. Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.-R. *IEEE Signal Processing*, 2008]

What are the strongest changes in the data?

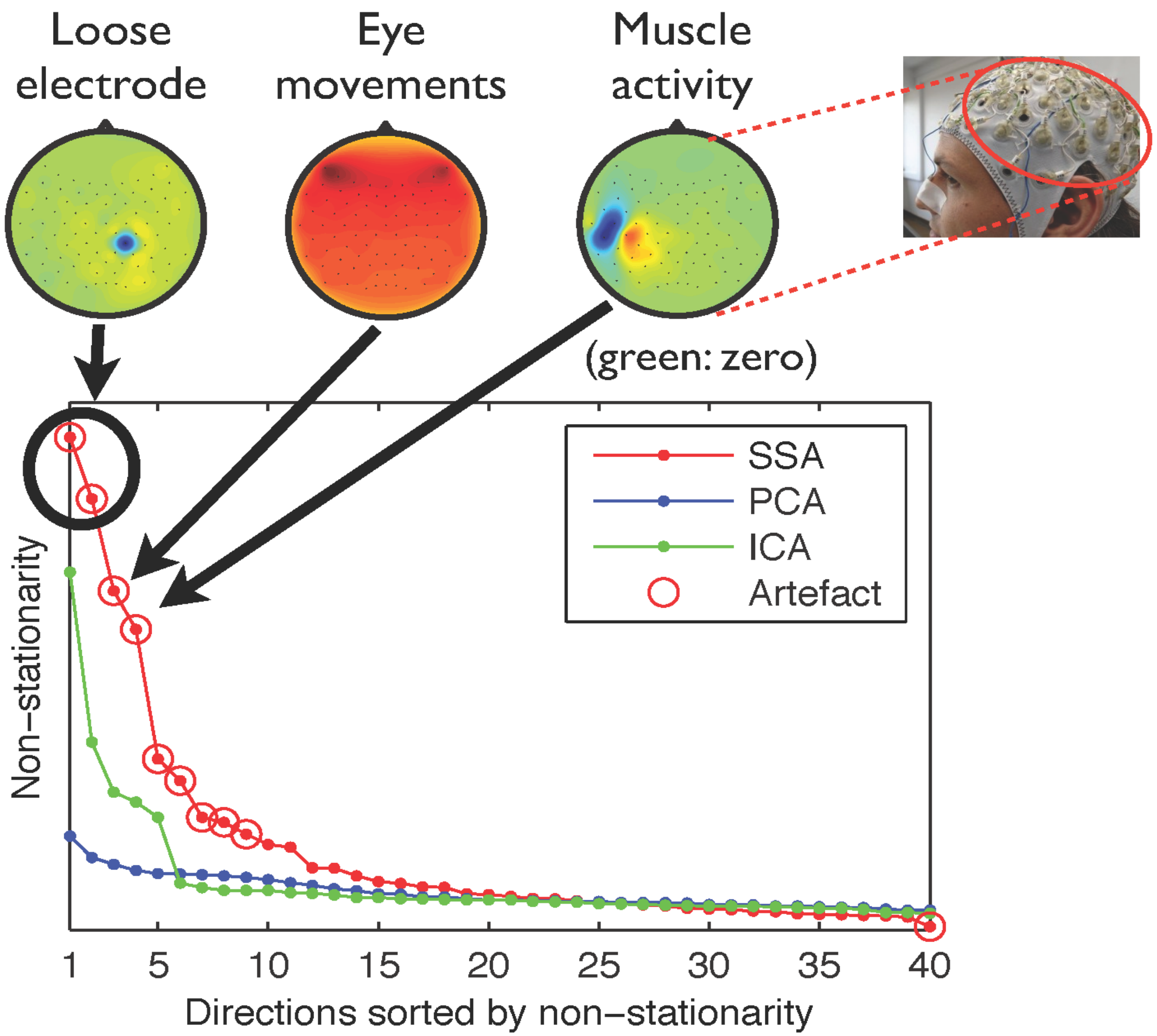
- What are the strongest changes?
- And could we have found them using ICA or PCA?

Setup: concatenate all trials of one subject; divide the data into 0.5s epochs.



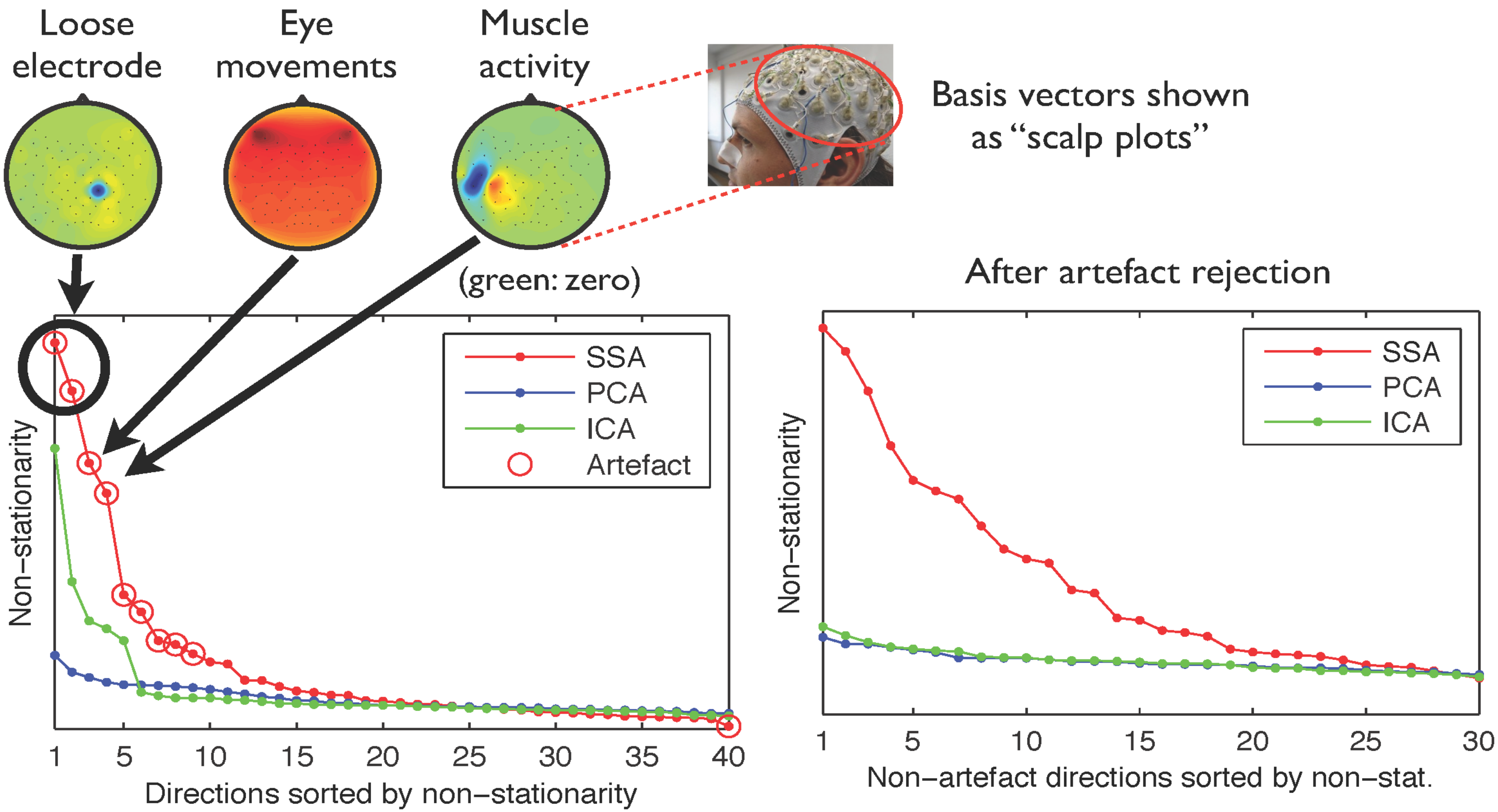
Apply SSA to find the most non-stationary sources

Results on one subject



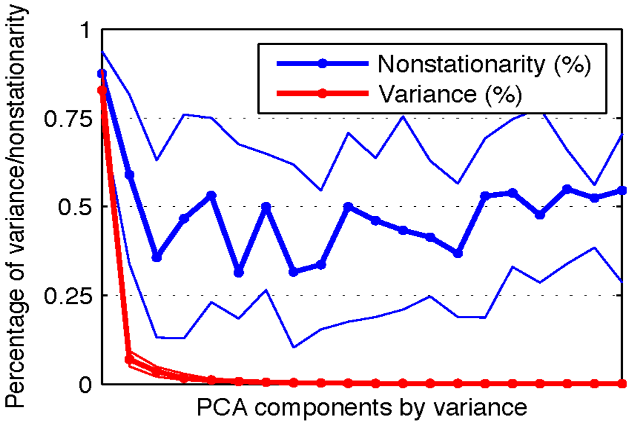
Basis vectors shown as “scalp plots”

Results on one subject

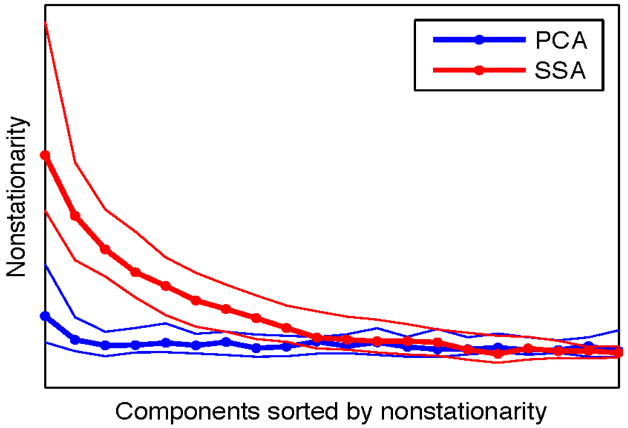


PCA and ICA do not find nonstationarities

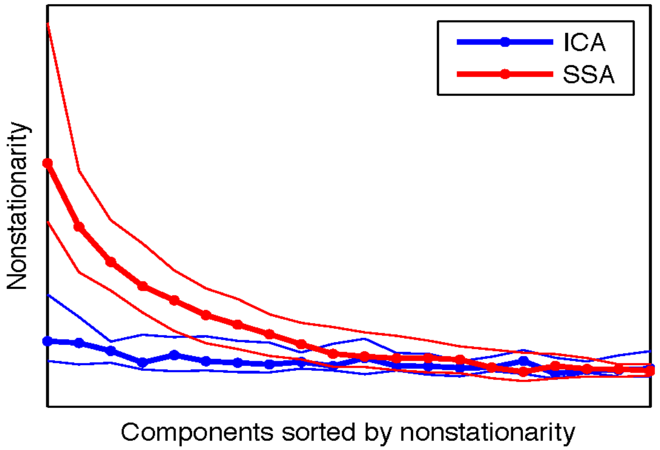
Results over all 40 subjects after artefact rejection



Variance (signal power) is not associated with the strength of nonstationarities



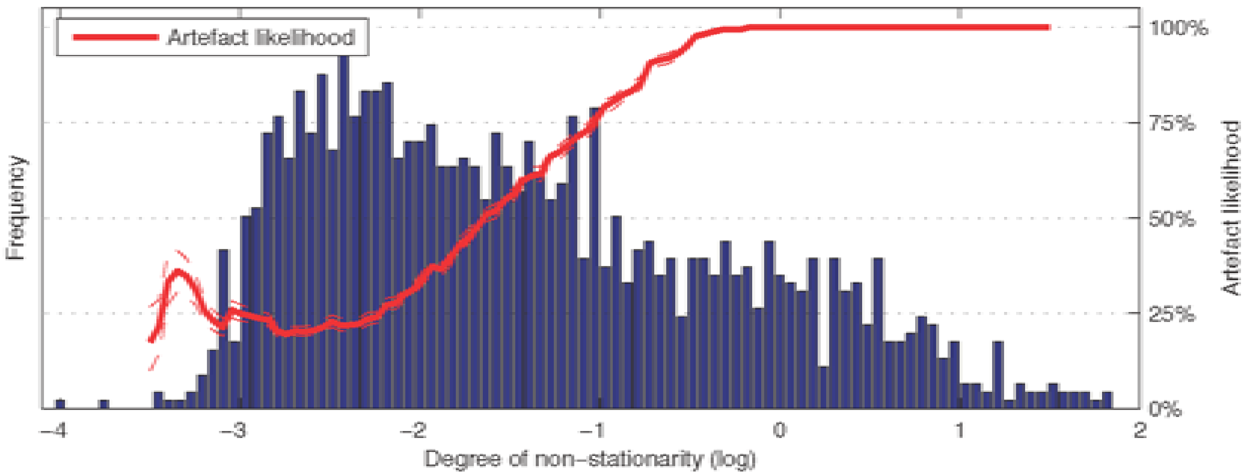
PCA basis is not optimal w.r.t nonstationarity



ICA basis is not optimal w.r.t nonstationarity

Classification of SSA directions

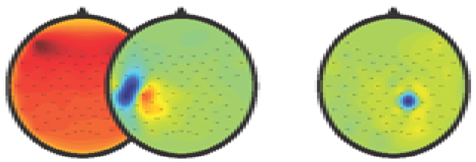
Distribution of non-stat. score over all 40 subjects (= 1600 SSA components)



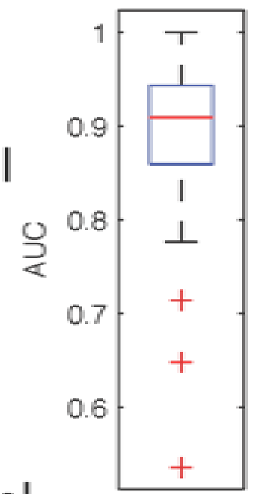
Nonstationarity is correlated with artefact likelihood

Muscle artefacts/eye movements

Loose electrodes



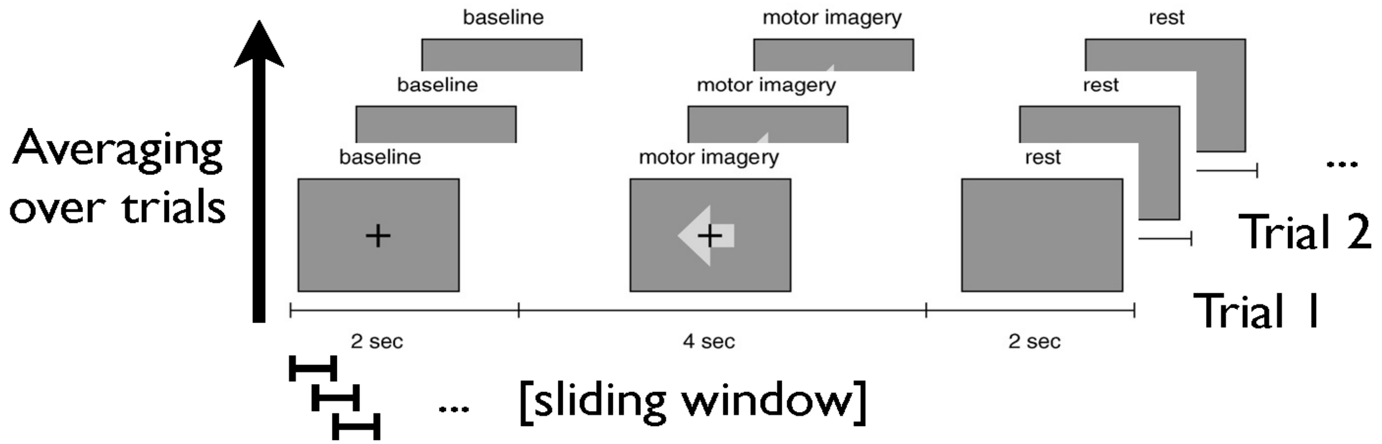
AUC of nonstat. score for artefact classification over all subjects



Ground truth: manual classification by an expert

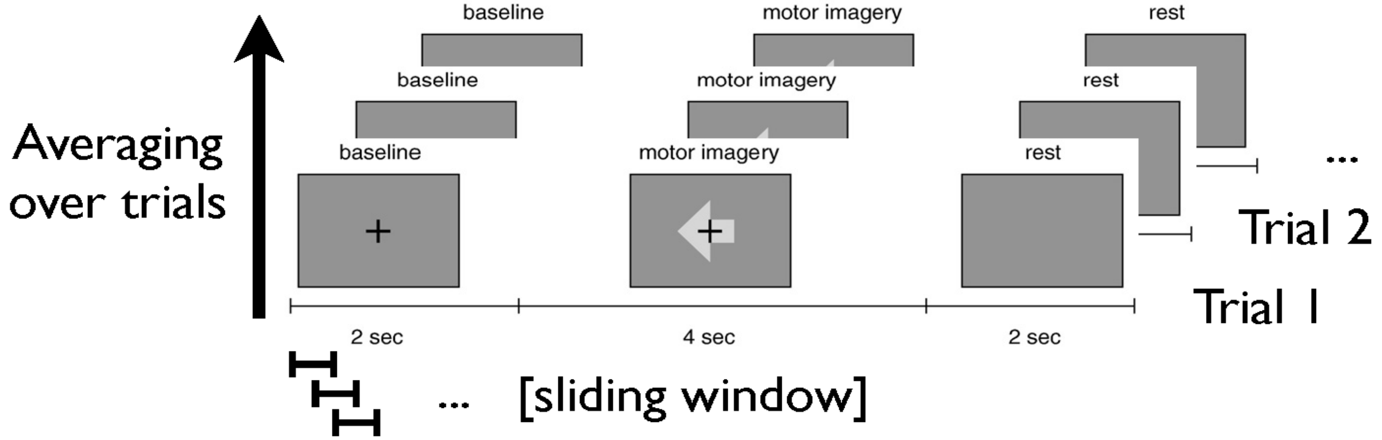
What happens during a trial? (on average)

Setup: sliding window (0.5s) averaged over all trials of of a class for one subject

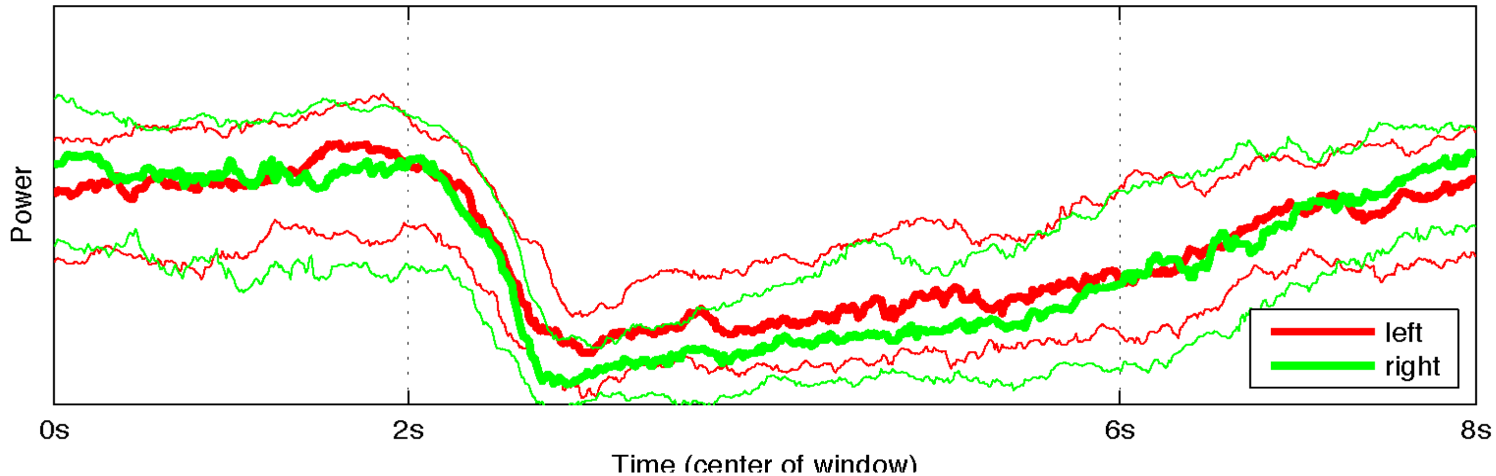


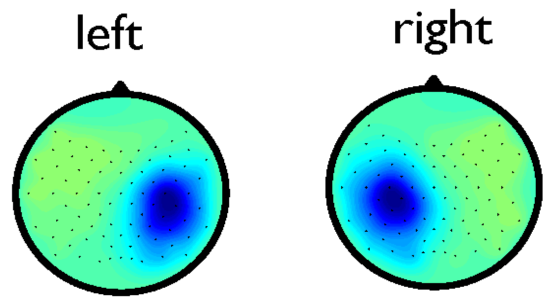
What happens during a trial? (on average)

Setup: sliding window (0.5s) averaged over all trials of of a class for one subject



Most non-stationary source for left and right class





Basis vectors shown as "scalp plots"

Summary: stationary subspace analysis

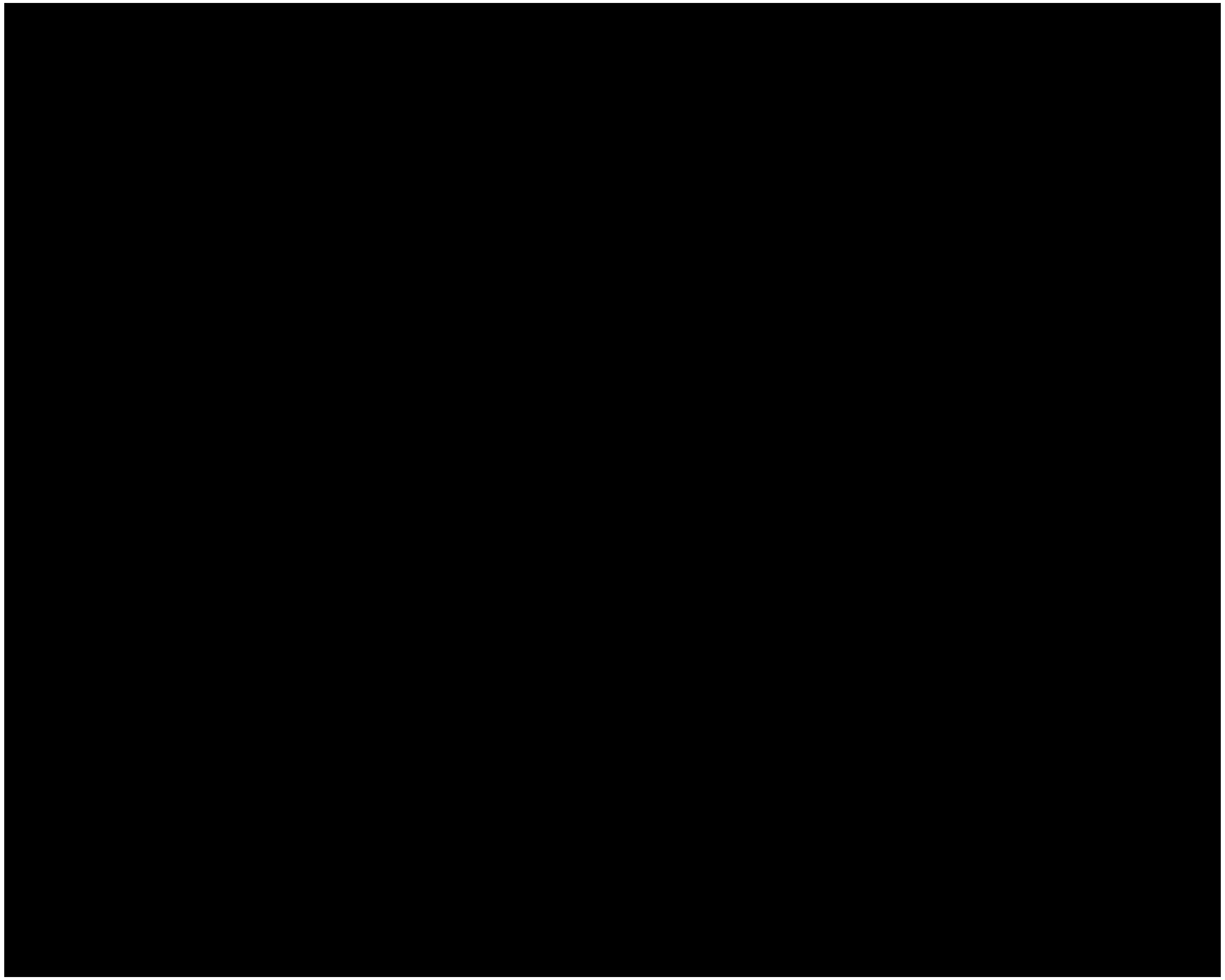
- SSA finds subspaces in which the sources are stationary/nonstationary.

- Important open questions:

  - How to deal with distribution changes in higher-order moments or temporal structure?

  - Model selection: how to choose the number of stationary/non-stationary sources?

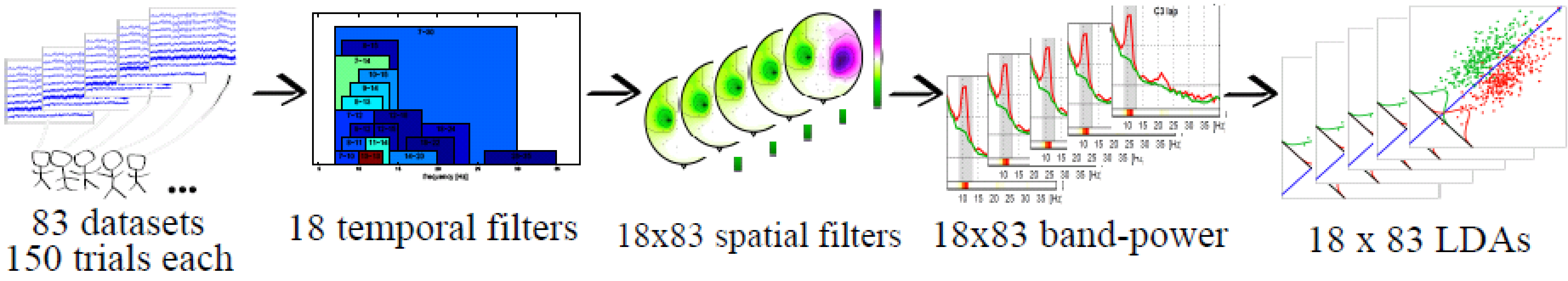
Real Man Machine Interaction



Multimodal $\longleftrightarrow$ Nonstationary

Towards a subject independent BCI decoder

- we end up with **1494 features** and $83 \cdot 150 =$ **12450 trials**
- to find a **subject-independent BCI**, we can perform ℓ_1-regularized regression (or others like LMM) using **leave-one-subject-out cross-validation**
- note that our trials have a **grouping** structure



Model formulation

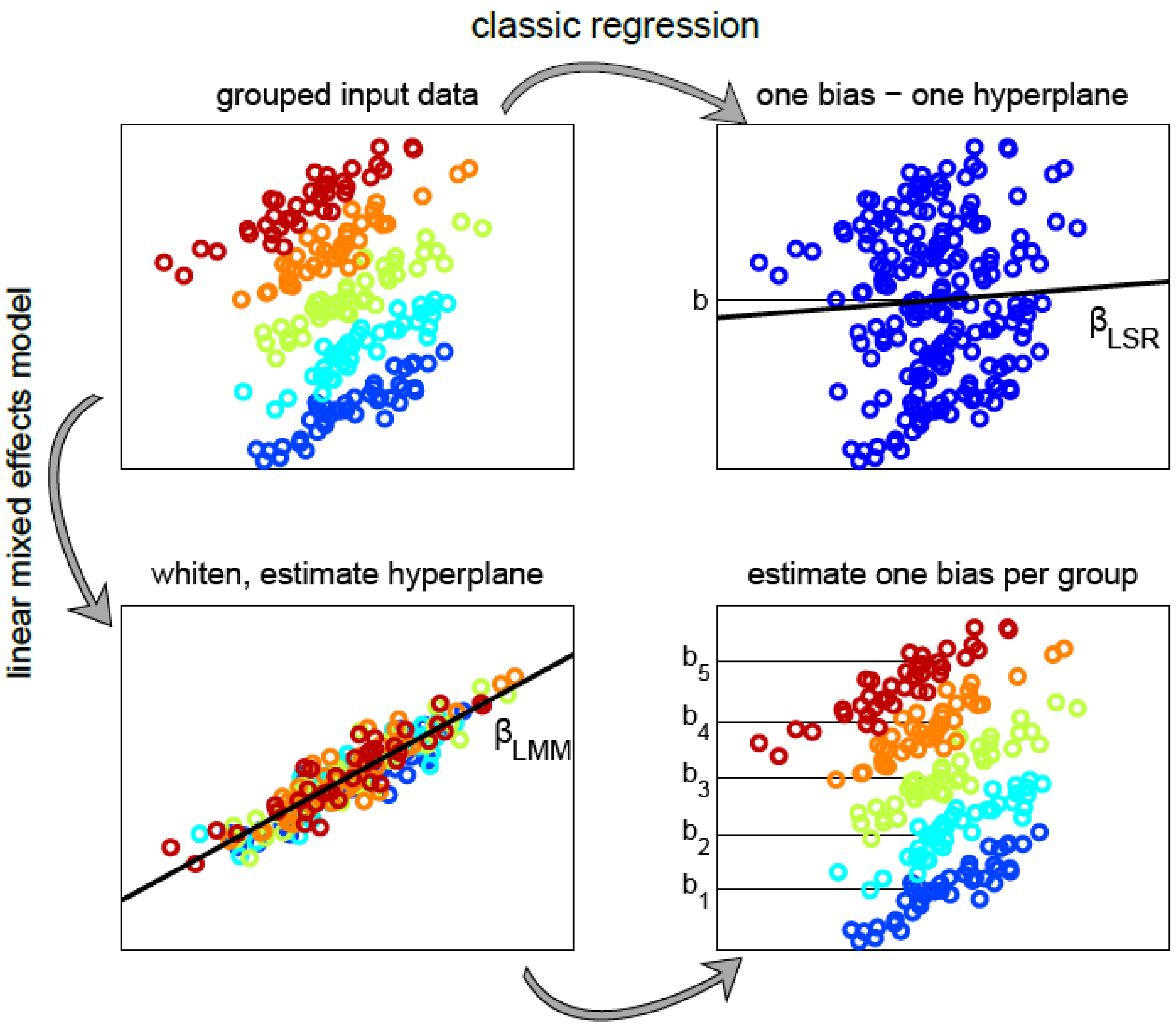
- Reminder – Linear regression:
  - $\mathbf{y} = \mathbf{X}\beta + \varepsilon$

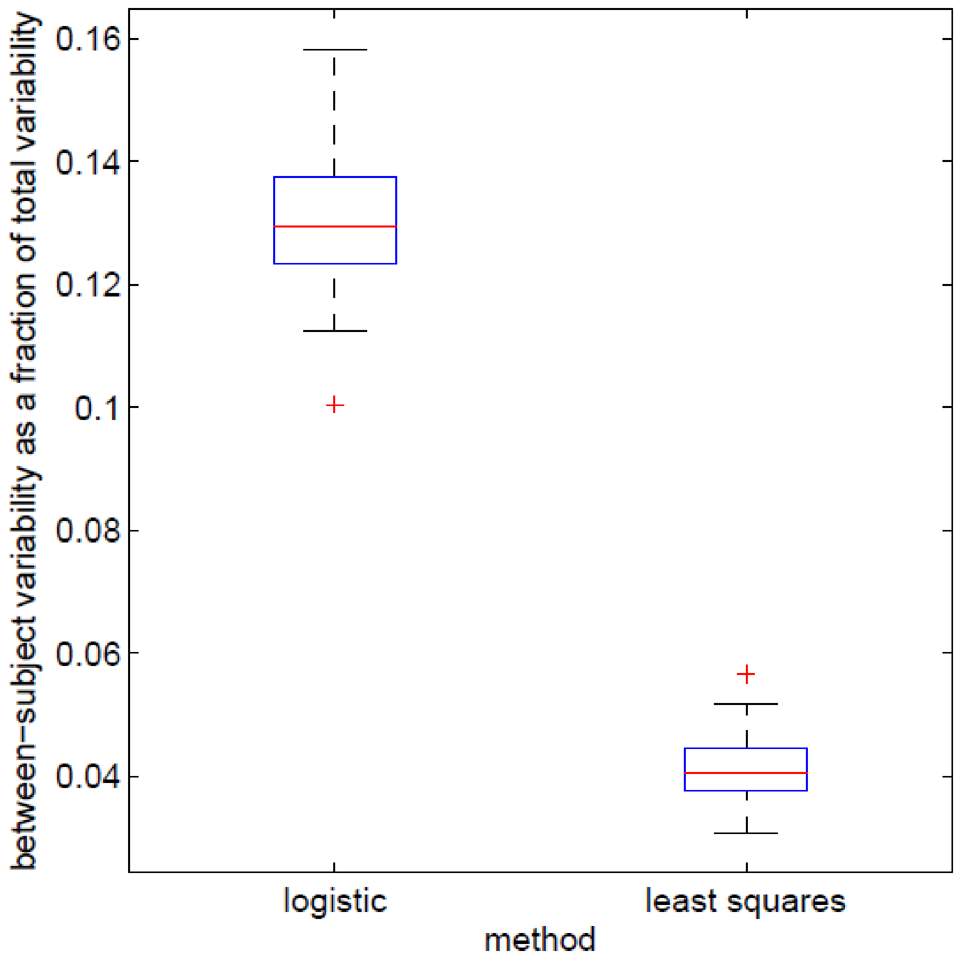
$$b_i \sim \mathcal{N}_q(0, \tau^2 I_q)$$

$$\varepsilon_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$$

- Mixed effects model with n groups:
  - $\mathbf{y}_i = \mathbf{X}\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i \quad \forall i \in \{1 \ldots n\}$
  - Consists of n simultaneous equations, one for each group
  - The equations are coupled by the common term $\mathbf{X}\beta$
  - Each equation has a group-dependent term $\mathbf{Z}_i\mathbf{b}_i$
  - In our case, each $\mathbf{Z}_i$ is simply a vector of ones, i.e. the corresponding $\mathbf{b}_i$ is scalar and represents the bias of group i
  - So-called **random intercepts model**
- Since we expect our features to be redundant and are aiming for better interpretability, we enforce sparsity by adding an ℓ_1 penalty

Linear Mixed Effects Model: intuition

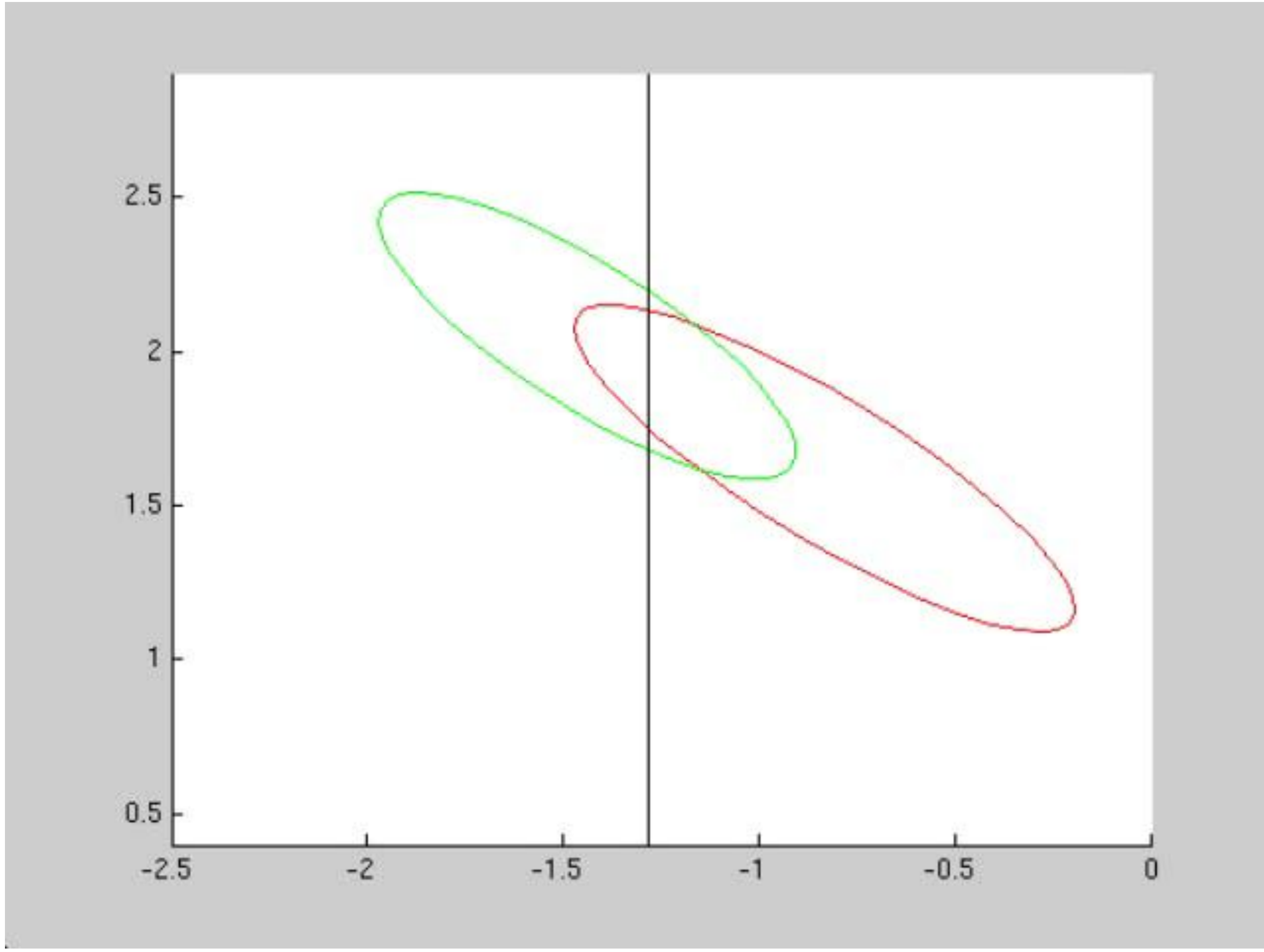




[Fazli, Müller et al. 2011]

Multimodal $\longleftrightarrow$ Nonstationary

Motivation: Shifting distributions within experiment



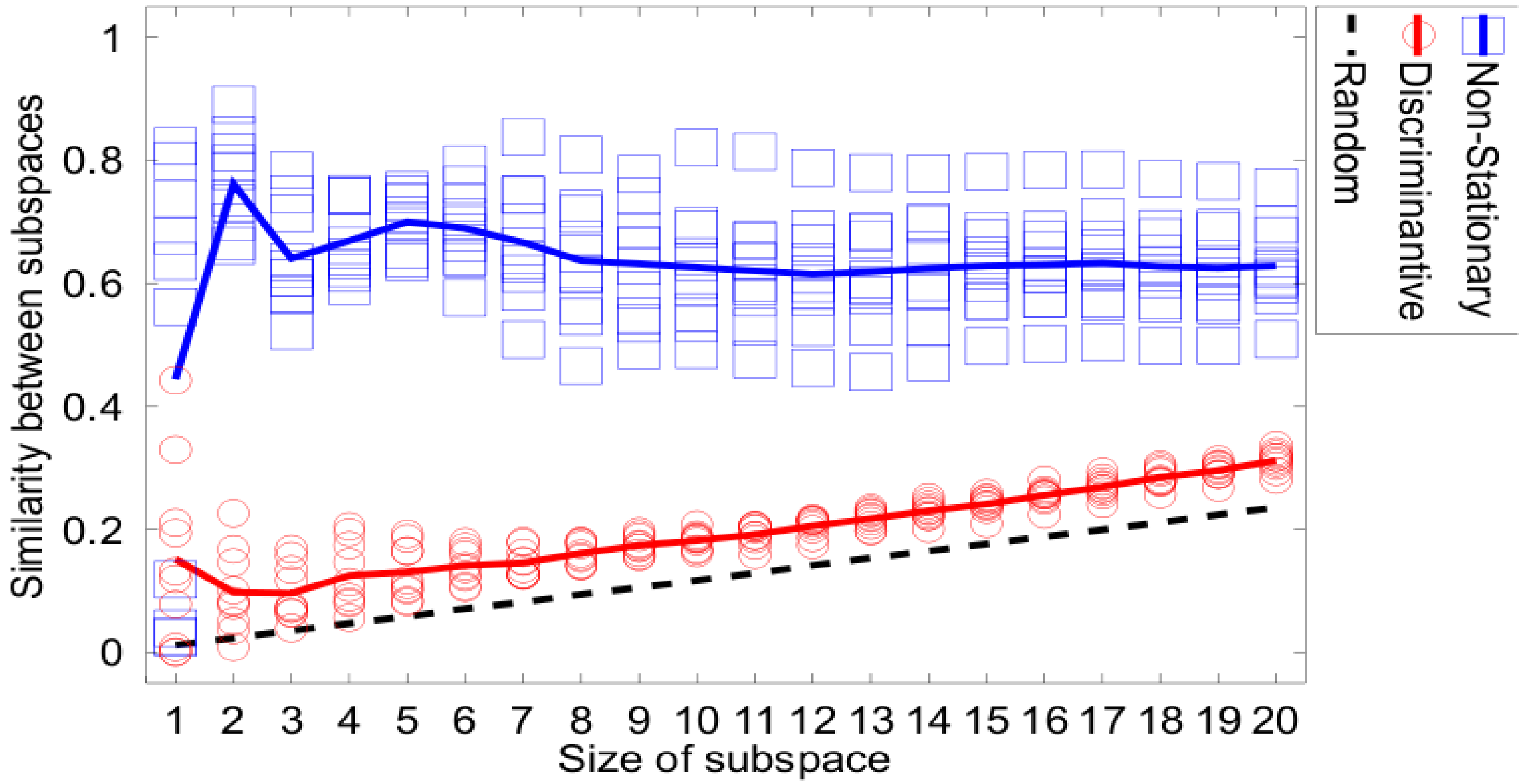
But: Is the nonstationarity different between subjects, i.e. could we learn it from other subjects?

Changes are similar !

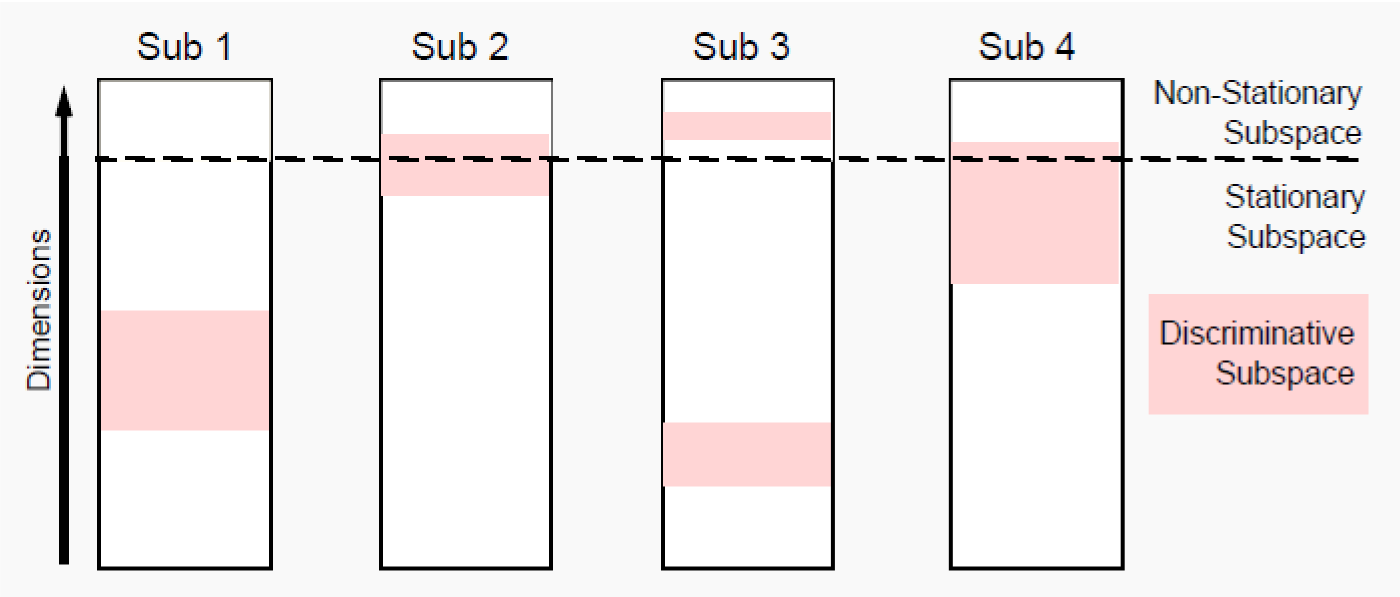
Modalities = Other Subjects

Changes between training and test data are similar between users.

Other multi-subject methods, e.g. cov matrix shrinkage, may improve estimation quality but do not reduce non-stationarities.



Cartoon: learn from adverse nonstationary subspace across subjects



Usually discriminative information is transfered between subjects.

Algorithm

(1) For each subject $i = 1 \ldots n$, $i \neq i^*$ compute the eigenvectors $\mathbf{v}_i^{(1)} \ldots \mathbf{v}_i^{(d)}$ of $\mathbf{\Sigma}_i^{train} - \mathbf{\Sigma}_i^{test}$.

(2) For each subject i select the l eigenvectors with largest absolute eigenvalues.

(3) Aggregate the vectors into a matrix P.

(4) Apply PCA to reduce the dimensionality of the non-stationary subspace $\mathcal{S}_P = \mathrm{span}(P)$ to ν.

(5) Compute the projection matrix $P^{\perp}$ to the orthogonal complement of $\mathcal{S}_P$.

(6) Make i^*s data invariant to the changes by projecting out non-stationarities $\tilde{\mathbf{X}} = (P^{\perp})^T P^{\perp} \mathbf{X}$.

(7) Compute spatial filters from $\tilde{\mathbf{X}}$ using CSP.

Results

Two data sets with different stimulus cues in training and test

1. visual cue in training & auditory cue in test
2. letters in training & moving objects in test

The size of the non-stationary subspace is determined by CV in a leave-one-subject-out manner on the other users.

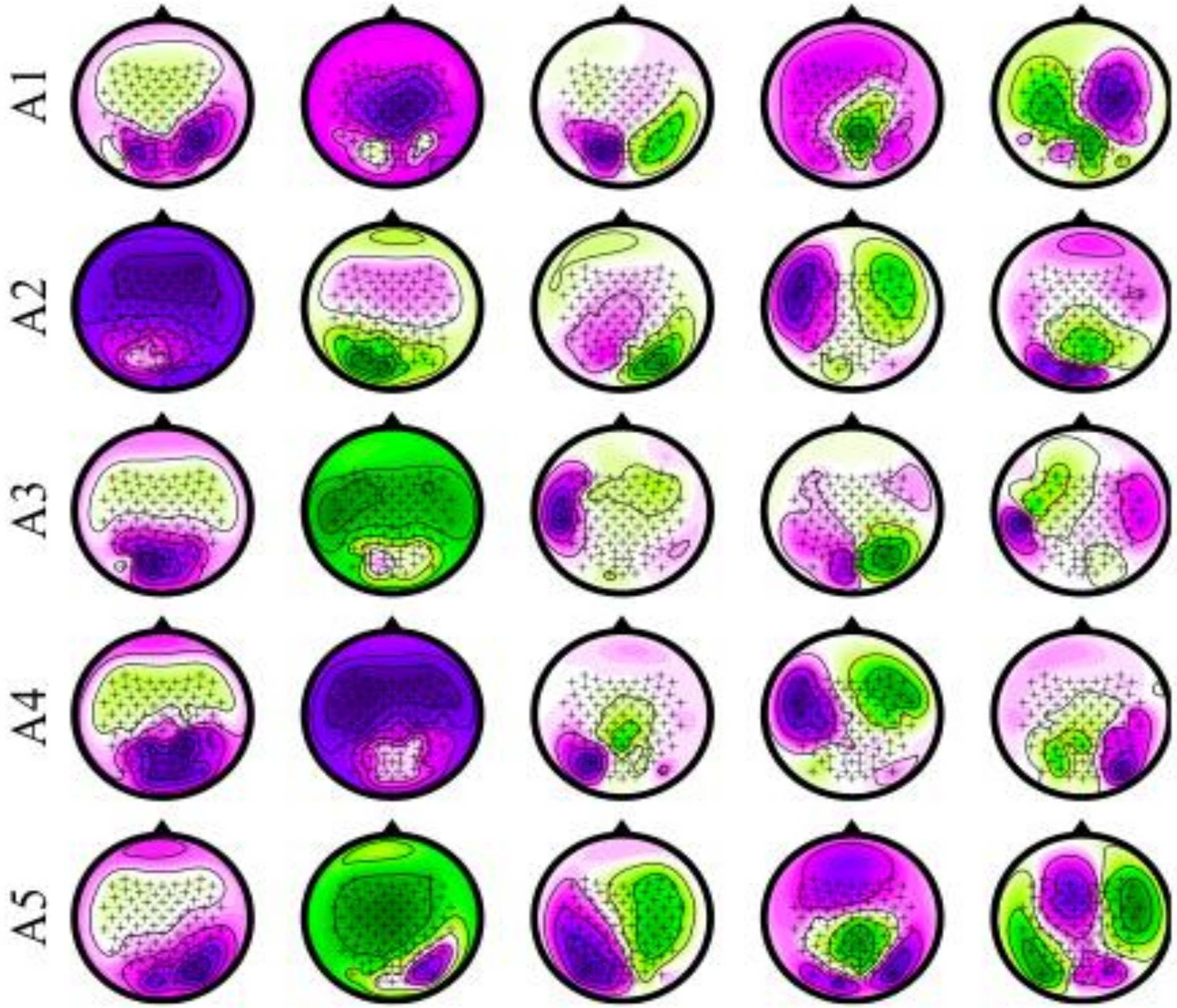
	Audio-Visual Data Set					BCI Competition III					Overall		
Subject	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	Mean	Median	Std
CSP	79.5	80.0	65.8	59.2	94.2	66.1	96.4	58.2	88.8	81.0	76.9	79.8	14.0
ssCSP	87.1	80.8	67.5	65.8	93.3	67.0	94.6	58.2	89.3	85.7	78.9	83.3	13.1

ssCSP: stationary subspace CSP

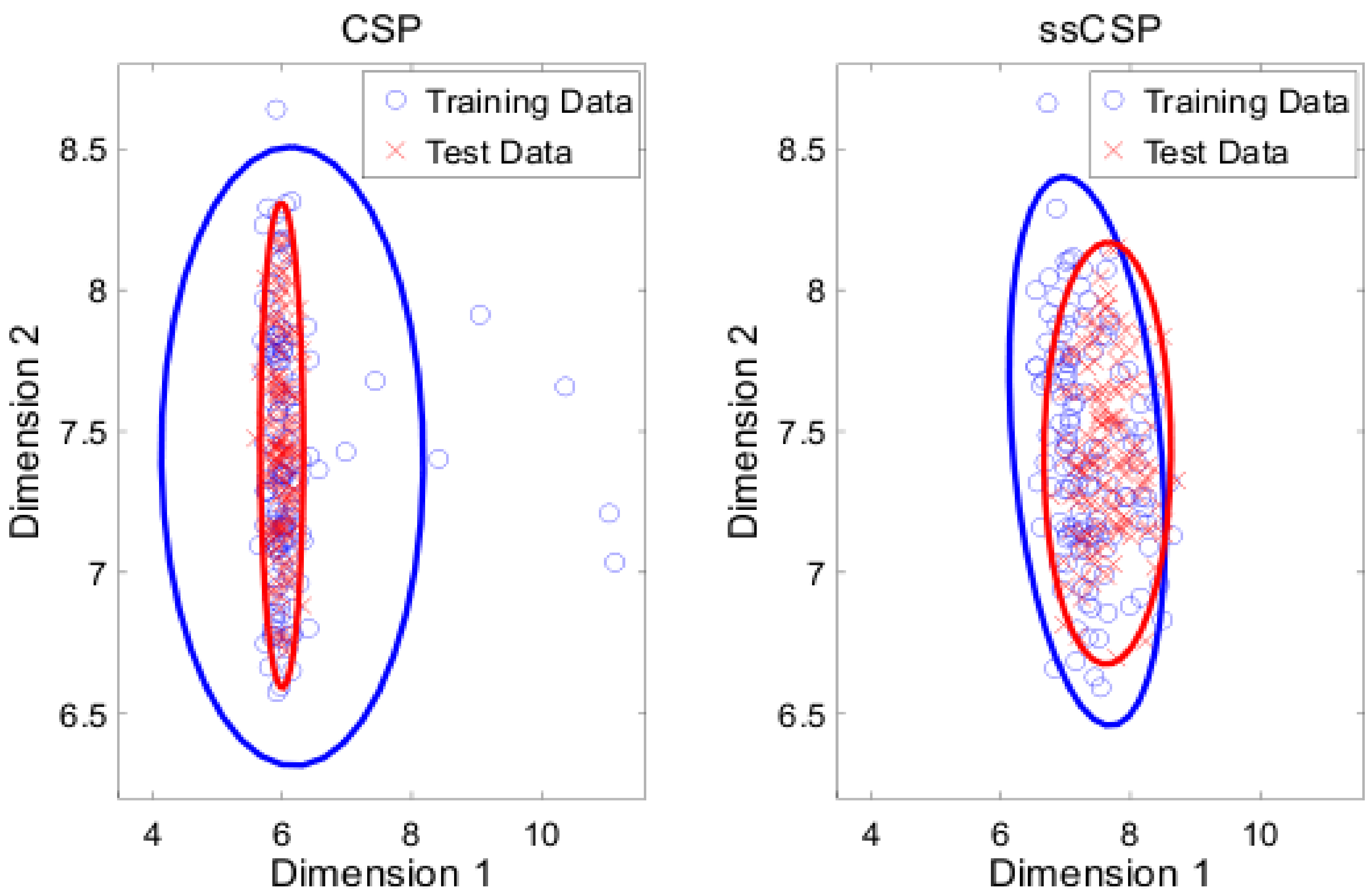
Interpretation

The most non-stationary directions are very similar between users.

Activity in occipital and temporal areas is penalized as these regions are mainly responsible for visual and auditory processing.



Feature distribution becomes stationary



Summary II

- Novel “multi-modal“ approach to reduce non-stationarities in data

- In contrast to other multi-subject methods it does NOT transfer discriminative information, thus is more robust if subject similarity is low.

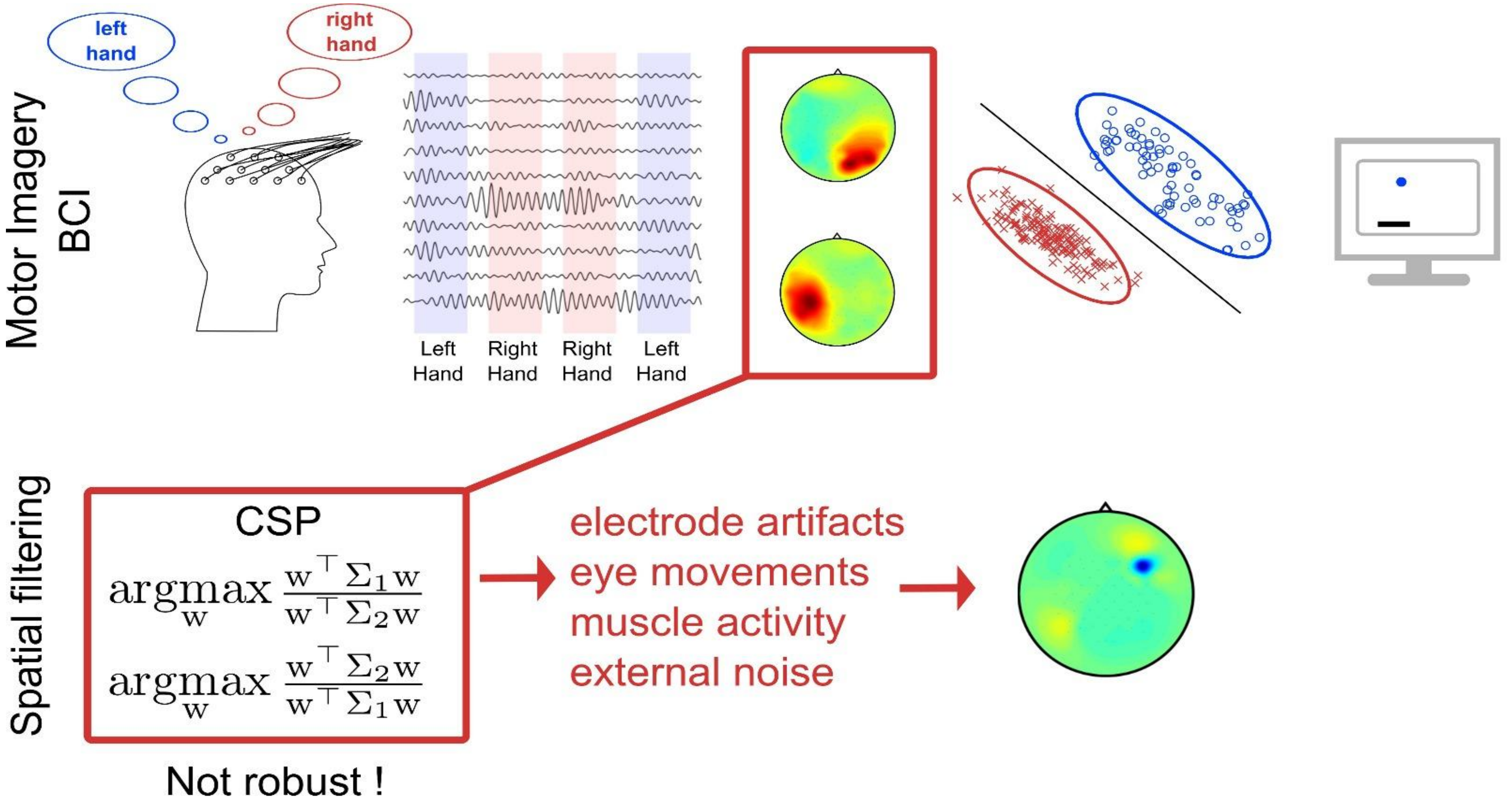
- Non-stationary information appears physiologically interpretable and meaningful.

- The idea of transfering stationary subspaces between subjects can be applied to many other problems.

Multimodal ⟷ Nonstationary

[Samek, Kawanabe, Müller IEEE Rev BME 2014, Nips 2013]

BCI Pipeline



Divergence CSP Framework

Theorem: *Let* $\boldsymbol{W} \in R^{D \times d}$ *be CSP filter and* $\boldsymbol{V} \in R^{D \times d}$ *be a matrix that can be decomposed into a whitening projection and an orthogonal projection. Then*

$$\begin{aligned} \operatorname{span}(\mathbf{W}) &= \operatorname{span}(\mathbf{V}^*) \\ \text{with } \mathbf{V}^* &= \underset{\mathbf{V}}{\operatorname{argmax}}\, \tilde{D}_{kl}\left(\mathcal{N}(\mathbf{0}, \mathbf{V}^\top \mathbf{\Sigma}_1 \mathbf{V}) \,||\, \mathcal{N}(\mathbf{0}, \mathbf{V}^\top \mathbf{\Sigma}_2 \mathbf{V})\right). \end{aligned}$$

Proof: Samek et al. IEEE Rev Bio Med Eng, 2014, in press

Symmetric KL-divergence $\int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx + \int q(x) \log\left(\frac{q(x)}{p(x)}\right) dx$

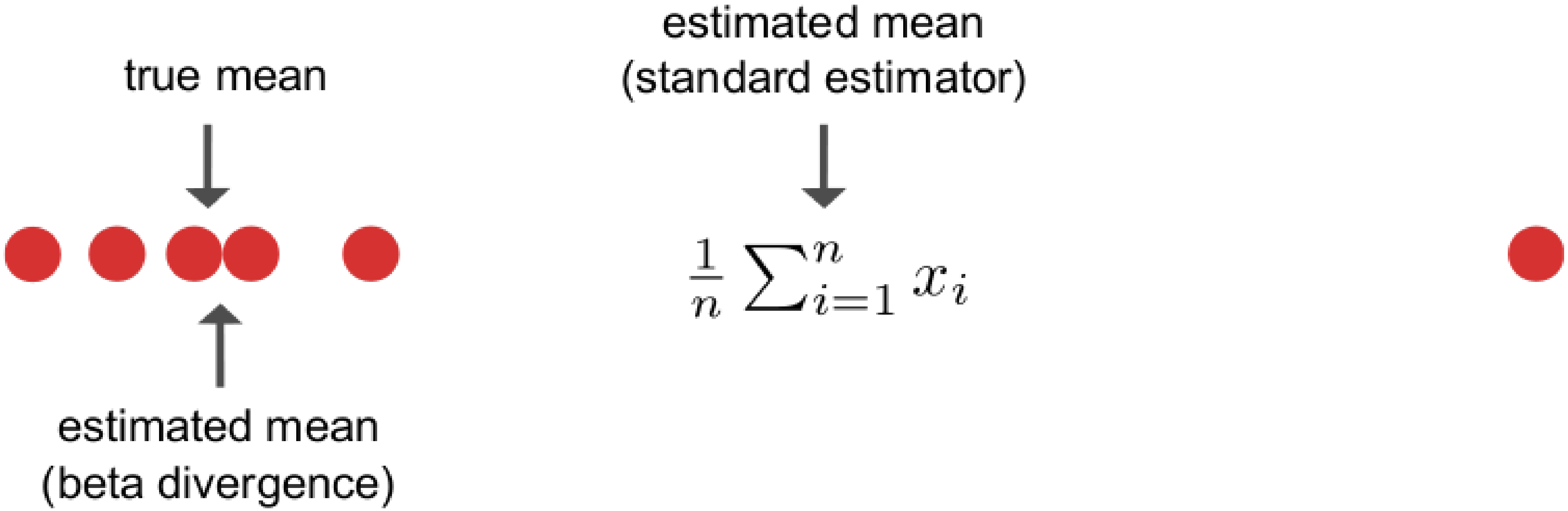
Robustness through Beta Divergence

Use the same mathematical formulation, but a different divergence → "similar to kernel trick"

Beta divergence is generalization of KL-divergence and is robust ($\beta = 0 \rightarrow D_\beta = D_{kl}$)

$$D_\beta(p(x), q(x)) = \frac{1}{\beta}\int (p(x)^\beta - q(x)^\beta)\, p(x)\, dx - \frac{1}{\beta+1}\int (p(x)^{\beta+1} - q(x)^{\beta+1})\, dx$$

Robustness Property



$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{1}{n}\sum_{i=1}^{n} w_i x_i$$

Beta divergence CSP

CSP $=$ kl-divCSP $\approx$ sum kl-divCSP $\xrightarrow{\text{robust}}$ β-divCSP

CSP:

$$\frac{\mathbf{w}^T \mathbf{\Sigma}_1 \mathbf{w}}{\mathbf{w}^T \mathbf{\Sigma}_2 \mathbf{w}}$$

kl-divCSP:

$$\tilde{D}_{KL}\left(\mathbf{w}^T \mathbf{\Sigma}_1 \mathbf{w} \;||\; \mathbf{w}^T \mathbf{\Sigma}_2 \mathbf{w}\right)$$

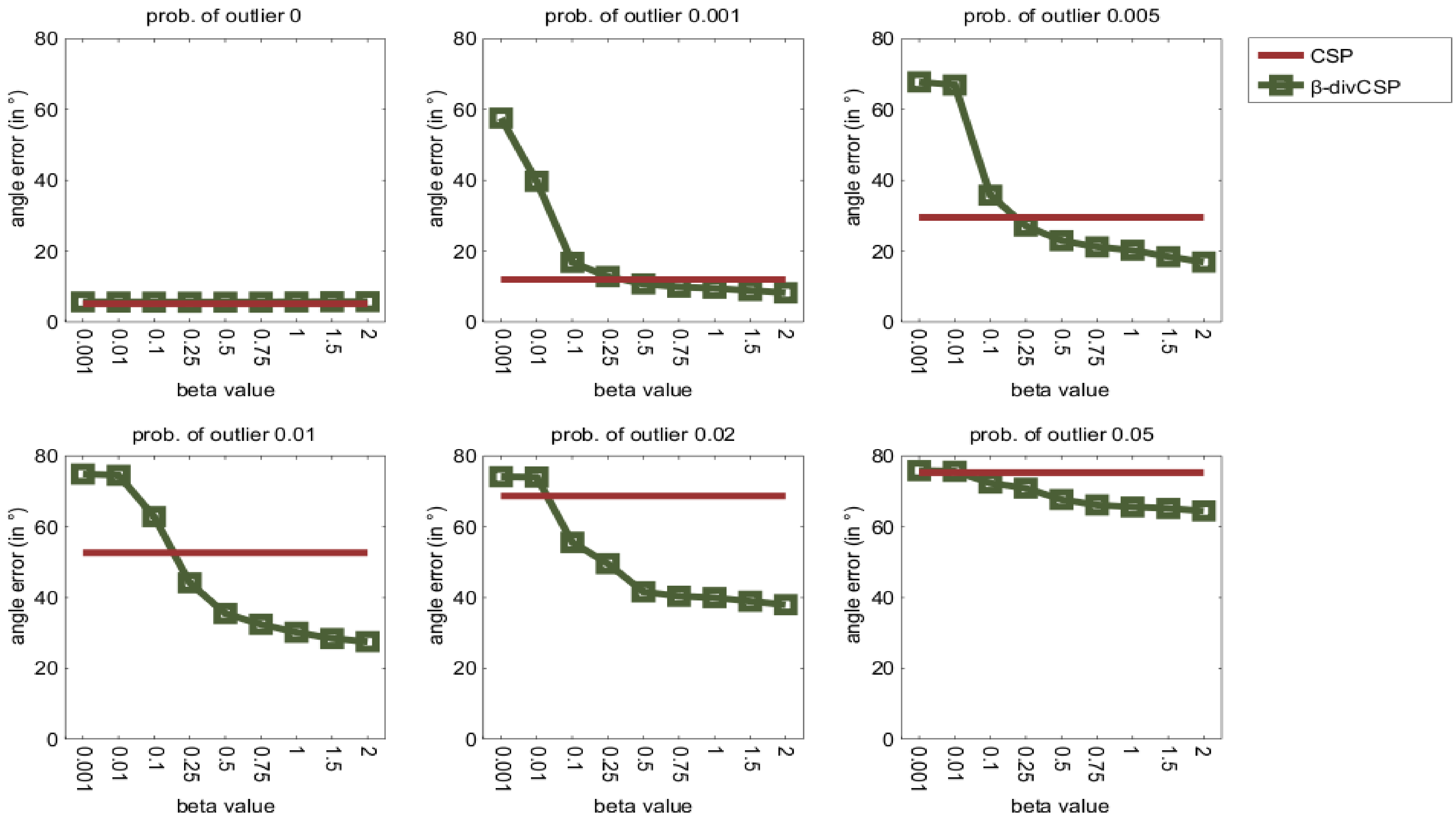
sum kl-divCSP:

$$\sum_i \tilde{D}_{KL}\left(\mathbf{w}^T \mathbf{\Sigma}_1^i \mathbf{w} \;||\; \mathbf{w}^T \mathbf{\Sigma}_2^i \mathbf{w}\right)$$

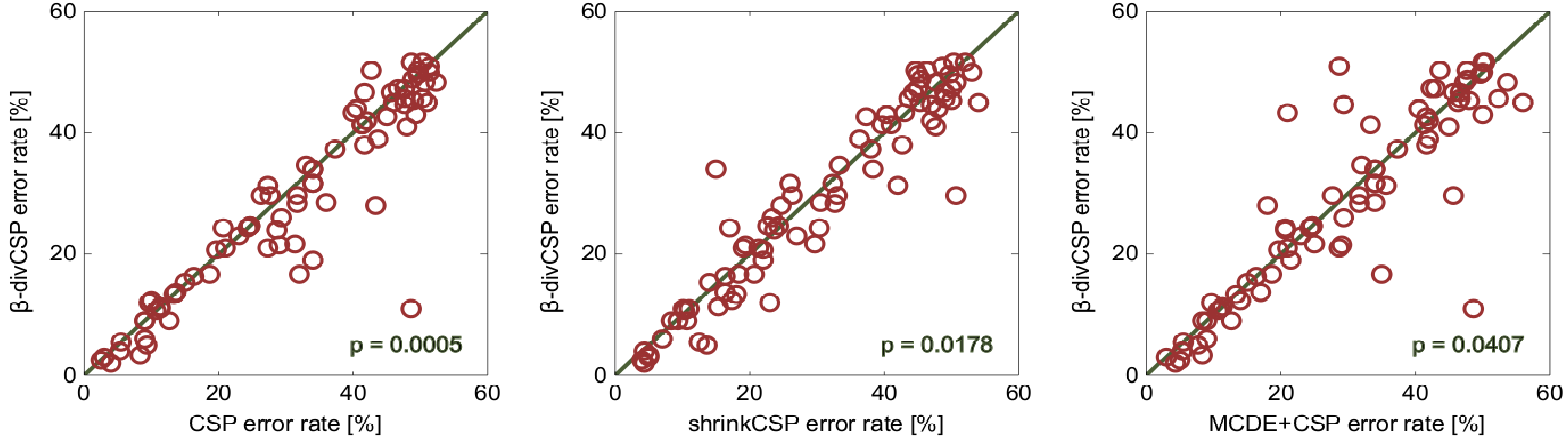
β-divCSP:

$$\sum_i \tilde{D}_{\beta}\left(\mathbf{w}^T \mathbf{\Sigma}_1^i \mathbf{w} \;||\; \mathbf{w}^T \mathbf{\Sigma}_2^i \mathbf{w}\right)$$

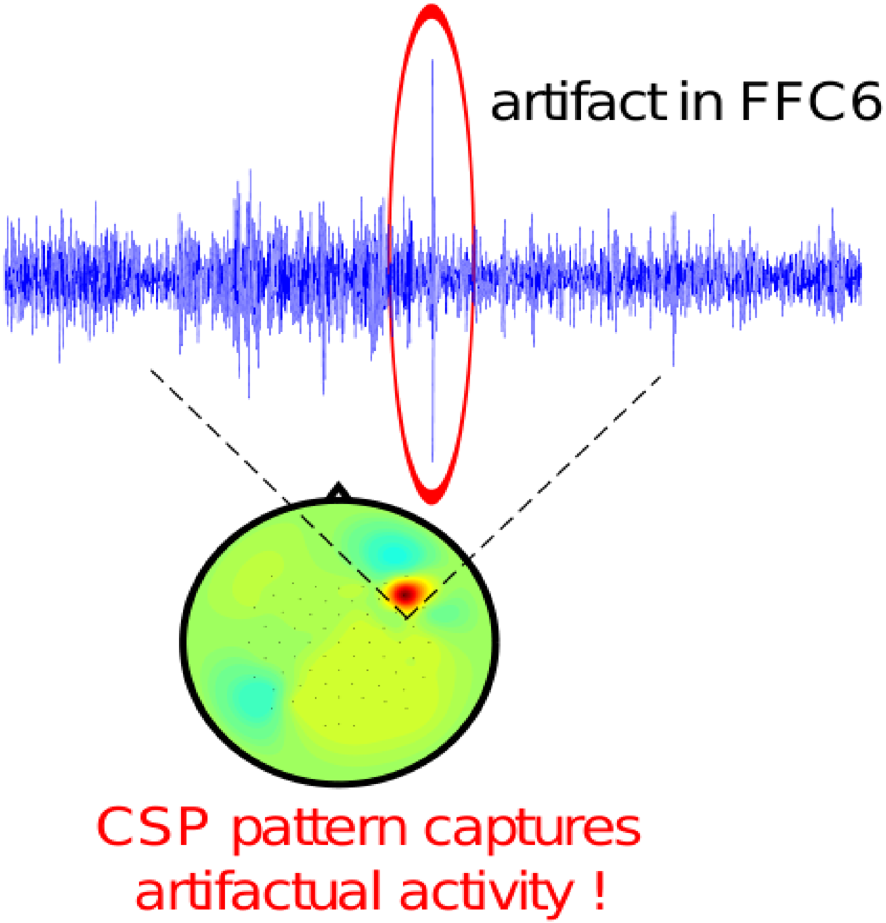
Simulations

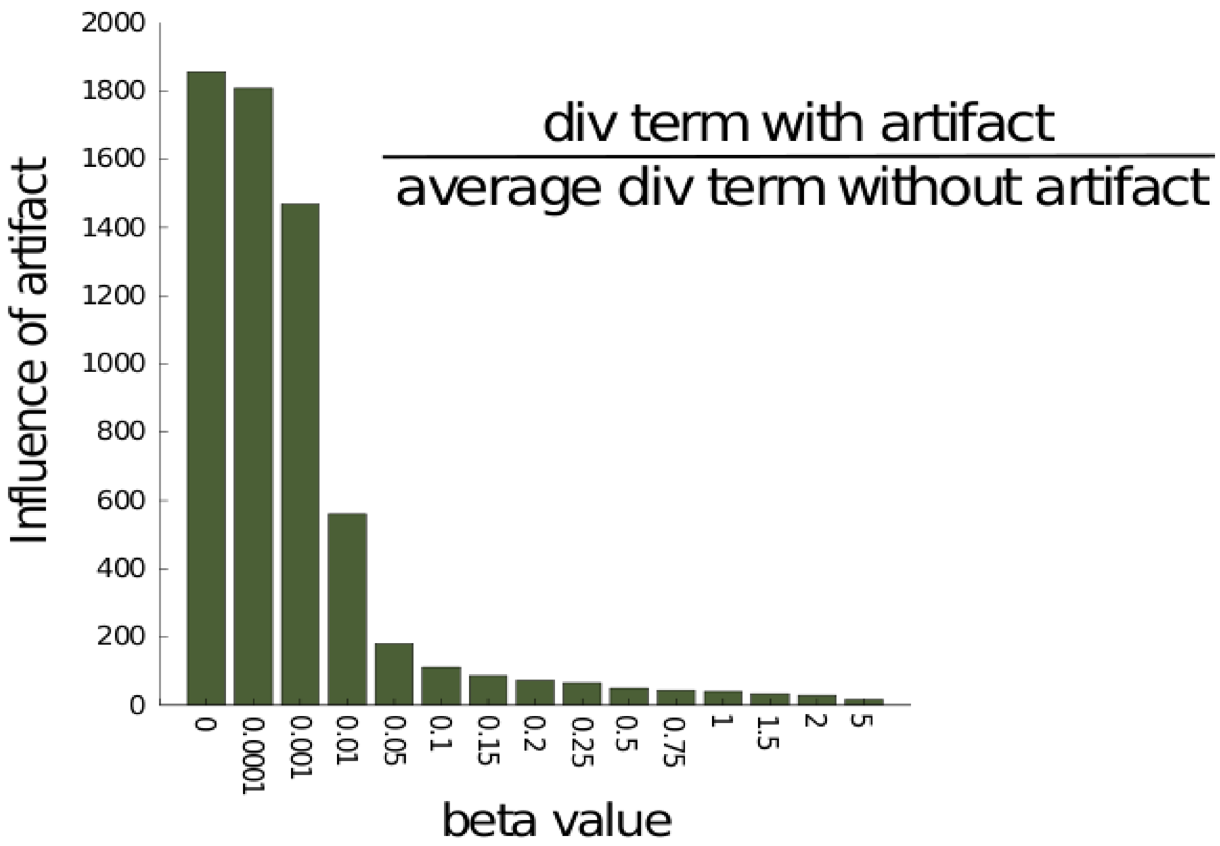


Results


β-divCSP error rate [%]
CSP error rate [%]
p = 0.0005
shrinkCSP error rate [%]
p = 0.0178
MCDE+CSP error rate [%]
p = 0.0407
0
20
40
60

Results


artifact in FFC6
CSP pattern captures
artifactual activity !


2000
1800
1600
1400
1200
1000
800
600
400
200
0
Influence of artifact
div term with artifact
average div term without artifact
0
0.0001
0.001
0.01
0.05
0.1
0.15
0.2
0.25
0.5
0.75
1
1.5
2
5
beta value

Invariance Through Regularization

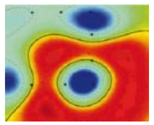
Maximizing variance-ratio not the only objective
→ add regularization term

$$\mathcal{L}(\mathbf{V}) \;=\; \underbrace{(1-\lambda)\tilde{D}_{kl}\left(\mathbf{V}^\top\boldsymbol{\Sigma}_1\mathbf{V} \;\|\; \mathbf{V}^\top\boldsymbol{\Sigma}_2\mathbf{V}\right)}_{\text{CSP Term}} \;-\; \underbrace{\lambda\Delta}_{\text{Regularization Term}}$$

Deflation (one-by-one) and Subspace (all-at-once) optimization algorithm.

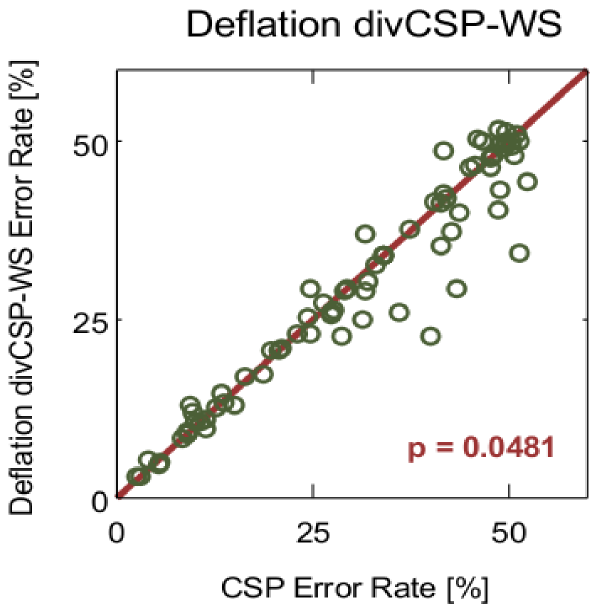
Different Kinds of Regularization

Regularization term $\boldsymbol{\Delta}$
Within-Session (WS) $\boldsymbol{\Delta} = \frac{1}{2N} \sum_{c=1}^{2} \sum_{i=1}^{N} D_{kl} \left(\mathbf{V}^\top \boldsymbol{\Sigma}_c^i \mathbf{V} \;\|\; \mathbf{V}^\top \boldsymbol{\Sigma}_c \mathbf{V} \right)$
Between-Session (BS) $\boldsymbol{\Delta} = \frac{1}{2K} \sum_{c=1}^{2} \sum_{k=1}^{K} \tilde{D}_{kl} \left(\mathbf{V}^\top \boldsymbol{\Sigma}_{tr,c}^k \mathbf{V} \;\|\; \mathbf{V}^\top \boldsymbol{\Sigma}_{te,c}^k \mathbf{V} \right)$
Across-Subject (AS) $\boldsymbol{\Delta} = \frac{1}{2K} \sum_{c=1}^{2} \sum_{k=1}^{K} \tilde{D}_{kl} \left(\mathbf{V}^\top \boldsymbol{\Sigma}_{tr,c}^\ell \mathbf{V} \;\|\; \mathbf{V}^\top \boldsymbol{\Sigma}_{tr,c}^k \mathbf{V} \right)$
Multi-Subject (MS) $\boldsymbol{\Delta} = -\frac{1}{K} \sum_{k=1}^{K} \tilde{D}_{kl} \left(\mathbf{V}^\top \boldsymbol{\Sigma}_1^k \mathbf{V} \;\|\; \mathbf{V}^\top \boldsymbol{\Sigma}_2^k \mathbf{V} \right)$

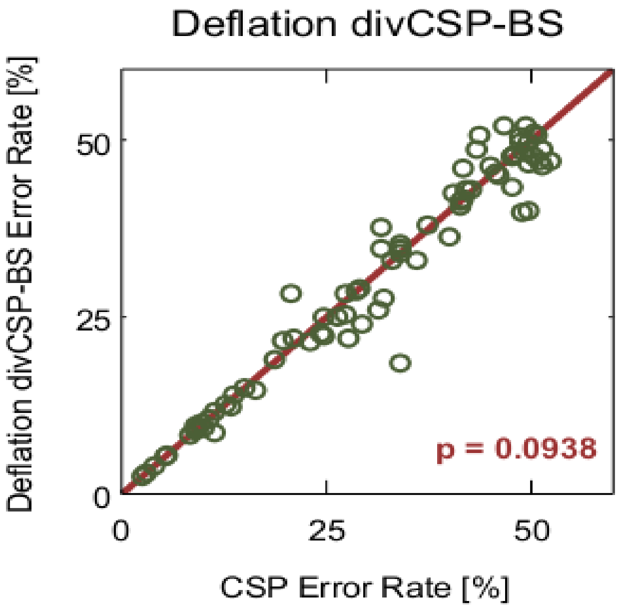


Results

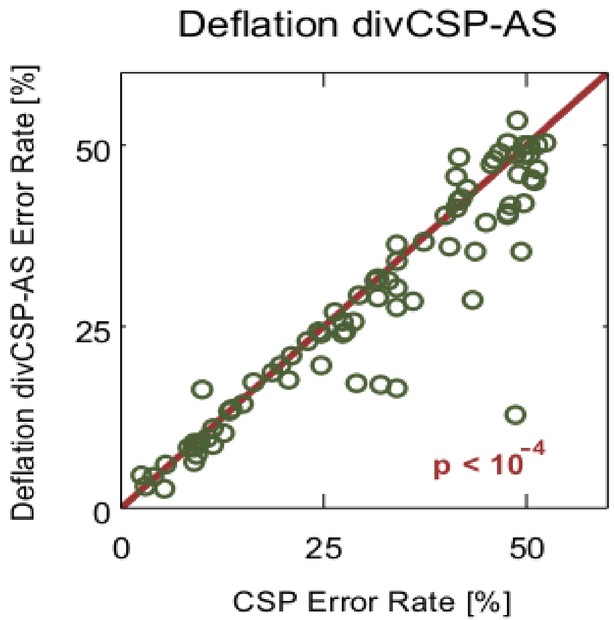
Within-Session Stationarity



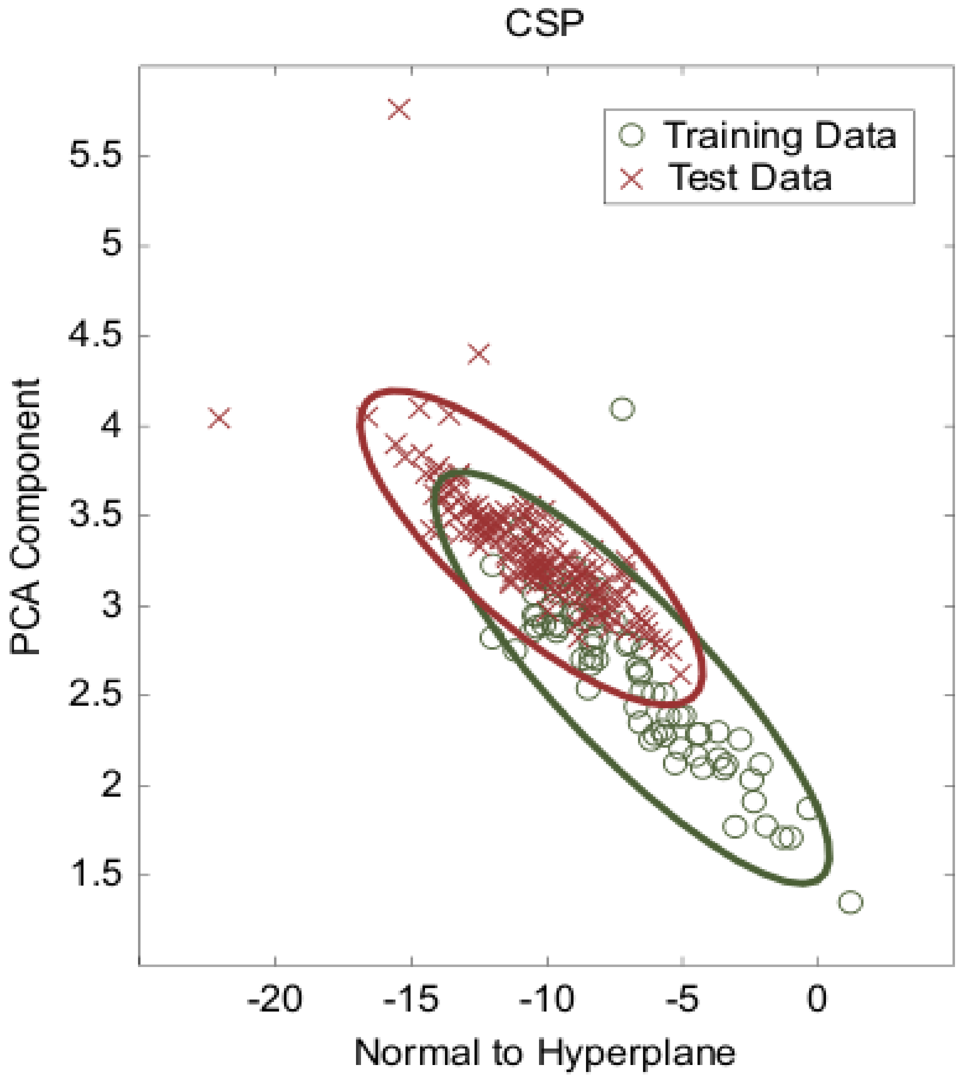
Between-Session Stationarity

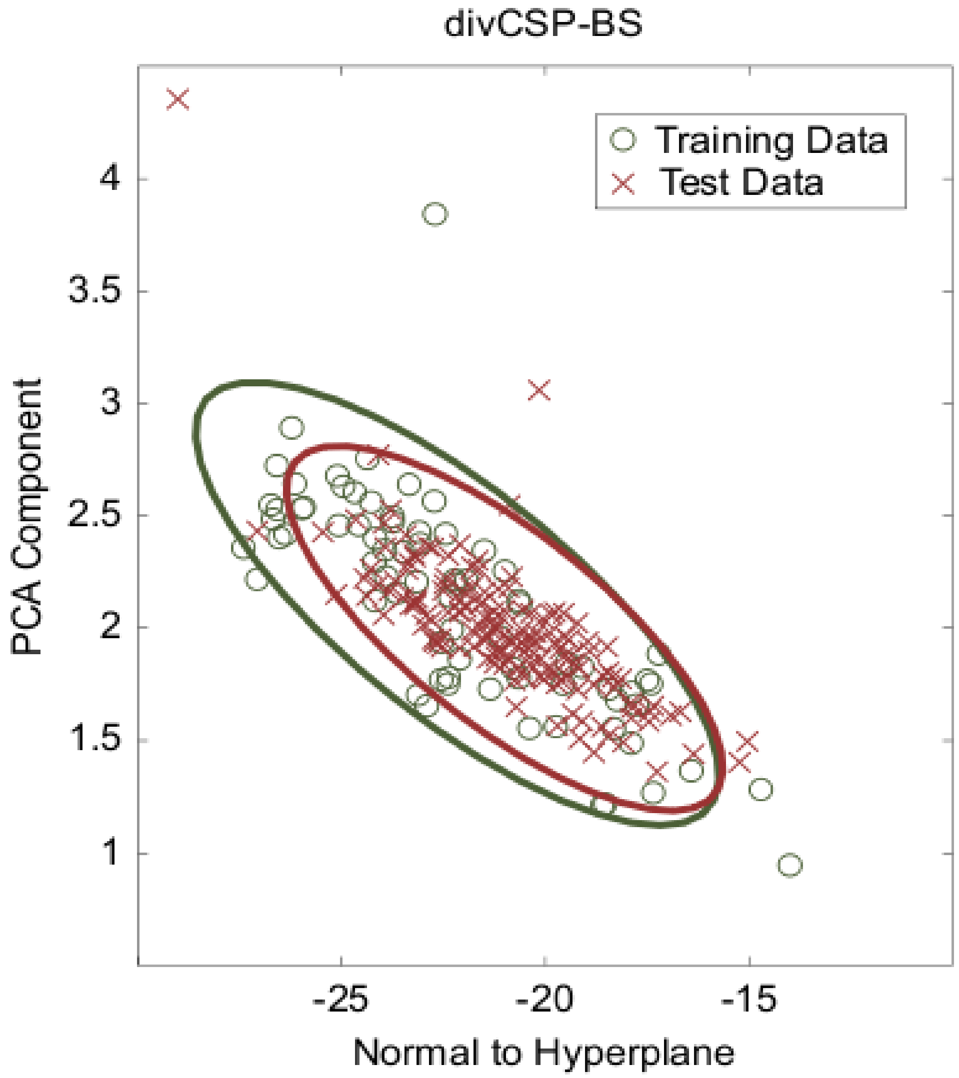


Across-Subject Stationarity



Reducing Shift between Training and Test



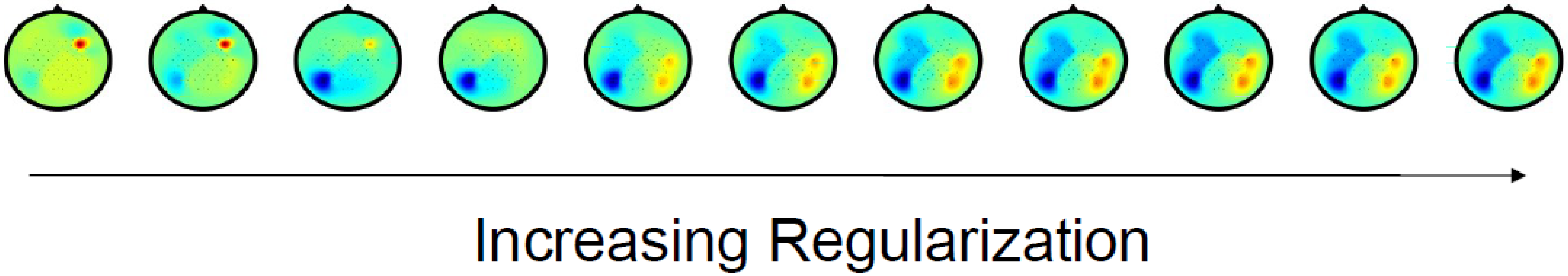


Regularization Towards other Subjects

CSP is affected by artifact in FFC6

This artifact is not present in other subjects data

→ Regularization towards other subjects penalizes spatial Filters that focus on this electrode



Summary III

Divergence CSP Framework

- Integrates many CSP variants in a principled manner

- Common optimization method, comparability, interpretability

- Easily allows to develop novel CSP variants and to integrate information from multiple sources

- “Divergence Trick”

All code is available at:
www.divergence-methods.org

Illiterates ⟷ Nonstationarity

[Vidaurre, Sannelli, Müller & Blankertz Neural Computation 2011]

Approach to „Cure“ BCI Illiteracy

Runs 1 - 3 | fixed Laplace | supervised

- Direct feedback -> Unspecific LDA classifier.
- Each trial, perform adaptation of the cls.
- Features: log band power (alpha and beta).
- Laplacian channels C3, C4 and Cz.

Runs 4 - 6 | CSP + sel.Lap. | supervised

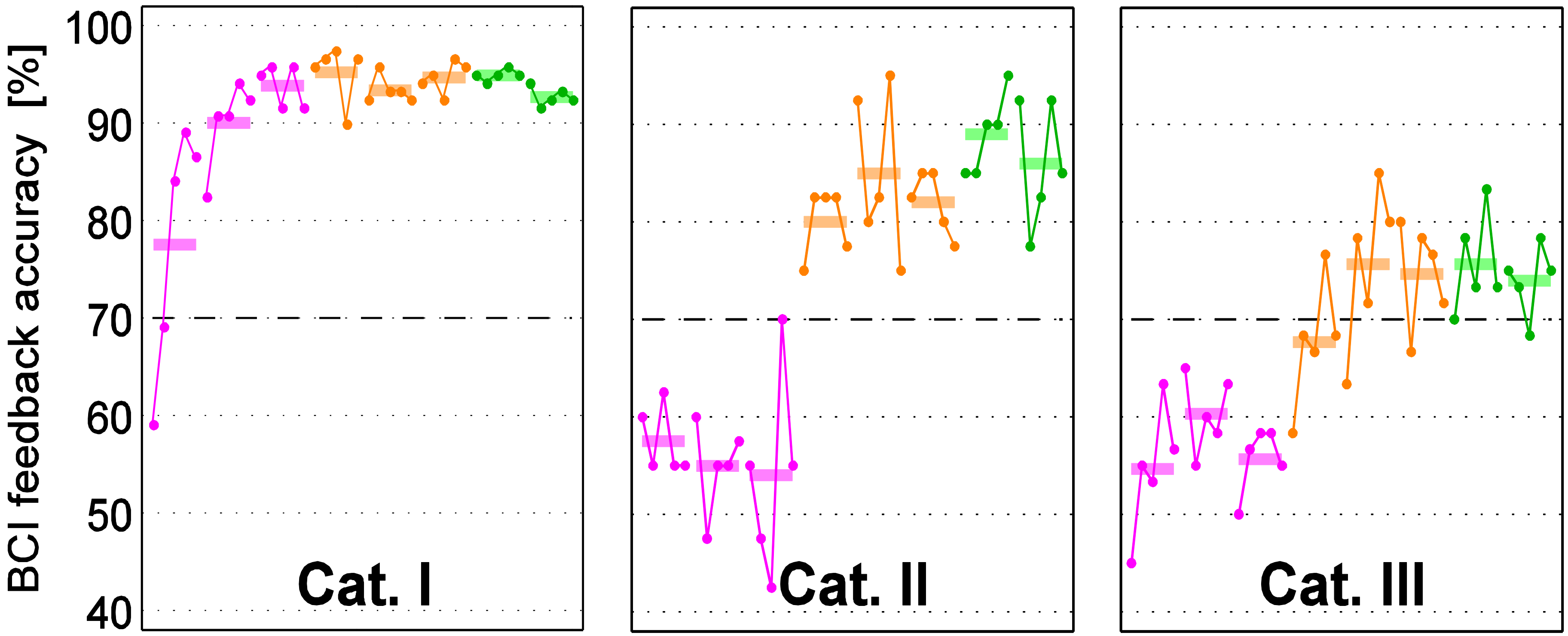
- Compute CSP and sel. Laps. from runs 1-3.
- Fixed CSP filters, automated laps. selection.
- Each trial retrain the classifier.

Runs 7 - 8 | CSP | unsupervised

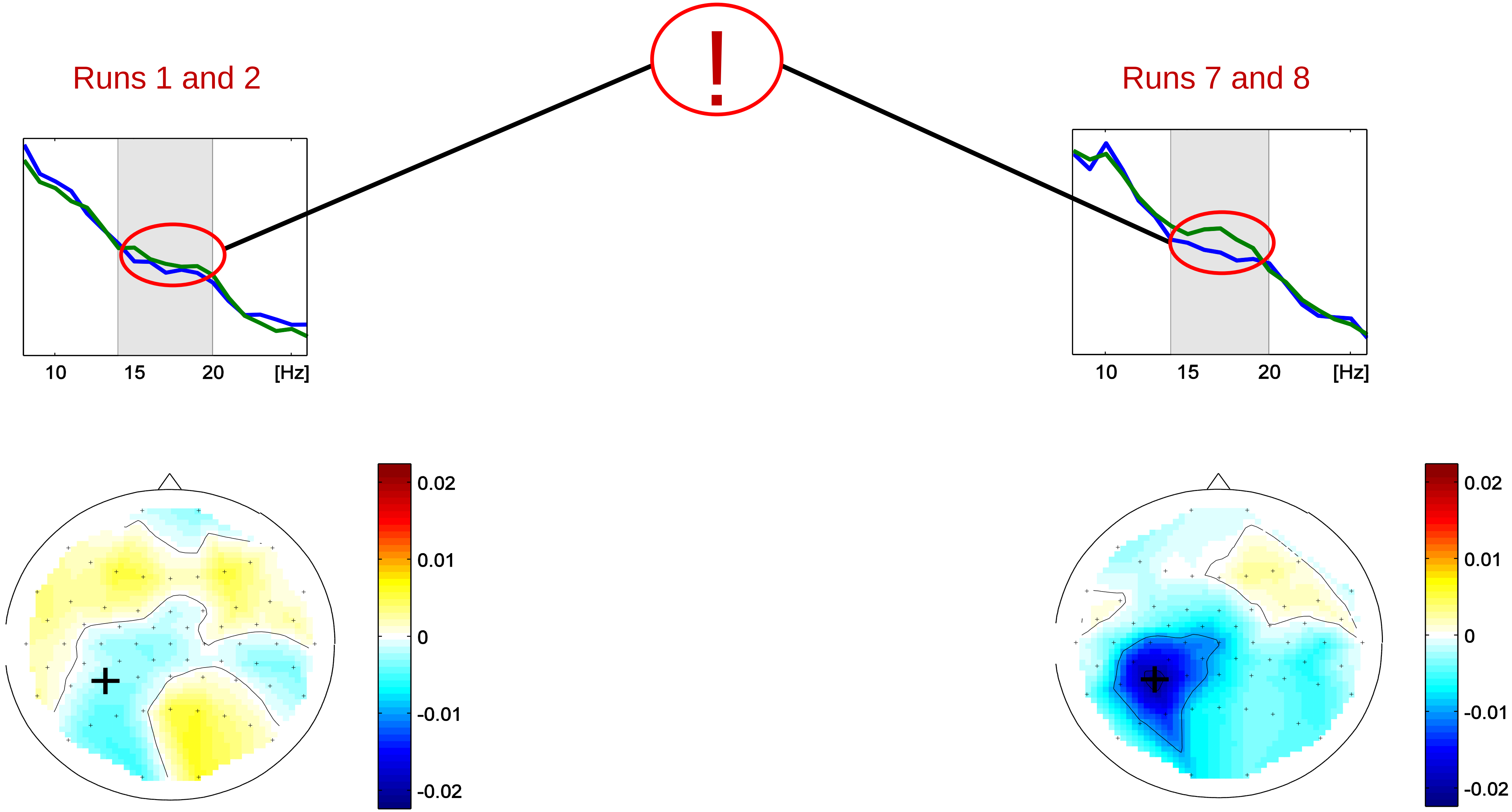
- Compute CSP from runs 4-6.
- Perform unsupervised adaptation of pooled mean.
- Update the bias of the classifier.

[cf. **Vidaurre**, Blankertz, Müller et al. Neural Comp. to appear]

Results (Grand Averages)



Example: one subject of Cat. III



[cf. Vidaurre, Blankertz, Müller et al. 2009]

Conclusion

- BBCI: Untrained, Calibration < 10min, data analysis <<5min, BCI experiment
- 5-8 letters/min mental typewriter CeBit 06,10. Brain2Robot@Medica 07, lNdW 09
- Machine Learning and modern data analysis is of central importance for BCI **et al**
- Important issue of this talk: How to learn under **nonstationarity**?
- Solutions:
- SSA, i.e. project on stationary subspace and learn there, linear, sound & fast
- Modeling: covariate shift based CV: special
- mixed effects model
- co-adaptation, Multimodal
- tracking, invariant features etc

FOR INFORMATION SEE:
www.bbci.de

Thanks to BBCI core team:

Gabriel Curio
Florian Losch
Volker Kunzmann
Frederike Holefeld
Jan Mehnert
Vadim Nikulin@Charite

berlin
brain computer
interface

Florin Popescu
Motoaki Kawanabe
Guido Nolte
Michael Tangermann
Thorsten Dickhaus@academia

Benjamin Blankertz
Claudia Sannelli
Carmen Vidaurre
Siamac Fazli
Martijn Schreuter
Andreas Ziehe
Laura Acqualagna
Wojciech Samek
Sven Dähne
Paul von Bünau

Matthias Krauledat
Guido Dornhege
Frank Meinecke
Steven Lemm
Stefan Haufe
Yakob Badower
Felix Biessmann
Marton Danozci
Roman Krepki@industry

Collaboration with: U Tübingen, Bremen, Albany, KU, TU Graz, EPFL, Daimler, Siemens, MES, MPIs, U Tokyo, TIT, RIKEN, Bernstein Center for Computational Neuroscience Berlin, Columbia, CUNY
Funding by: EU, BMBF and DFG