Brain-Computer Interfacing
WS 2018/2019 – Vorlesung #09

Benjamin Blankertz

Lehrstuhl für Neurotechnologie, TU Berlin

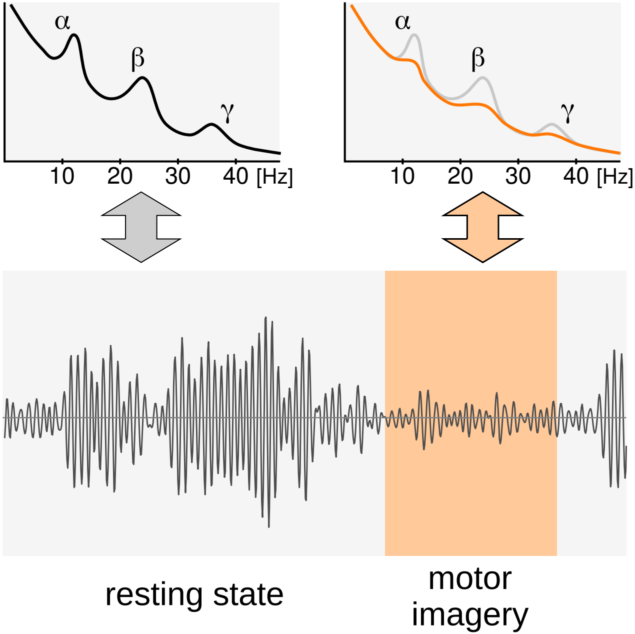
benjamin.blankertz@tu-berlin.de

19 · Dez · 2018

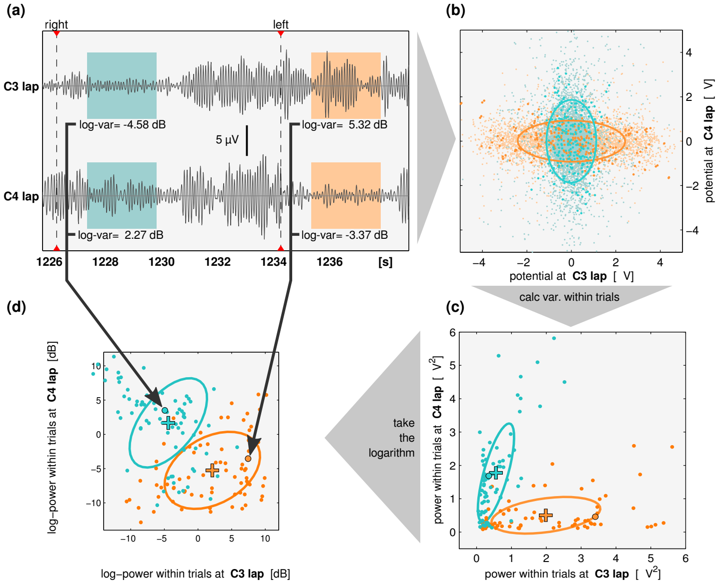
Today's Topics

- Classification of modulations of brain rhythms
- Common Spatial Pattern (CSP) analysis to classify different conditions that are characterized by a modulation of the amplitude of brain rhythms [Blankertz et al, 2008], [Parra et al, 2008; Ramoser et al, 2000].

Modulation of Rhythms – Spectra and Time Series



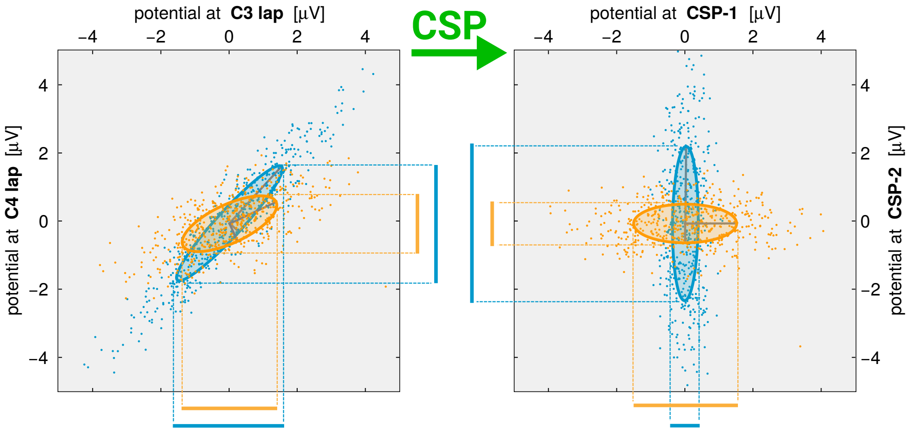
Features for Classification: Log Band-Power



Note that this illustration assumes the ideal case of **unmixed signals**.

The Challenge: Mixing by Volume Conduction

But signals are more mixed across channels than in the previous figure:

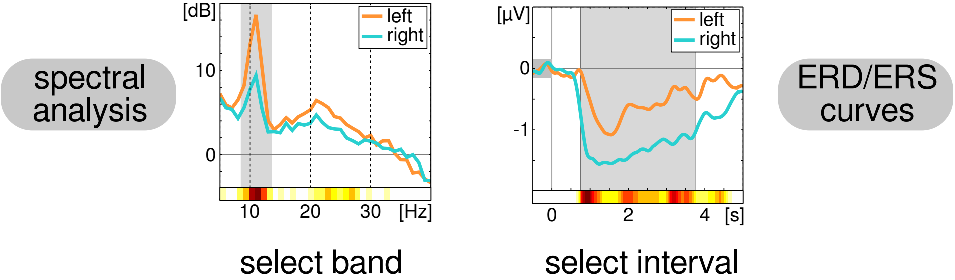


- Calculating band-power features in raw channels (left) would make the mixing of information irreversible for subsequent classification.
- Spatial filtering (e.g. CSP) has to be performed beforehand!

Running Example and Preparations

For illustration of today's methods, we used again the example of left hand and right hand motor imagery, to be performed for about 4 s after presentation of a visual cue.

We assume that an appropriate frequency band was selected, the continuous EEG data has been band-pass filtered, and an appropriate time interval was selected. Appropriate means here providing good discriminability between the two motor imagery conditions.



Typically discriminability measures like $\pm r^2$ or AUC are used for the selection.

Recap: Eigenvalue Decomposition

Given $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ symmetric and pos. definite (satisfied for covariance matrices), there exists an orthonormal matrix $\mathbf{V} \in \mathcal{O}(p)$ and diagonal matrix $\mathbf{D} \in \mathrm{Diag}(p)$, such that

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$$

In view of what comes next, please note the equivalent formulation:

$$\mathbf{V}^\top \mathbf{\Sigma} \mathbf{V} = \mathbf{D} \tag{1}$$

The reason for equivalence is that $\mathbf{V}^\top \mathbf{V} = \mathbf{I} = \mathbf{V}\mathbf{V}^\top$ holds due to the orthonormality of $\mathbf{V}$.

The transformation is also called diagonalization.

A Bit of Math

Today, we assume that all time series have mean zero. This is justified because we work with band-pass filtered data. Anyway, the statements all apply in general, but the derivations would be longer.

Some math: (under zero mean assumption)

- $\bar{\mathbf{x}} \in \mathbb{R}^{1\times T} \quad \Rightarrow \quad \mathrm{var}\,(\bar{\mathbf{x}}) = \frac{1}{T-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^\top = \frac{1}{T-1}\sum_{t=1}^{T}(\bar{\mathbf{x}}(t))^2$
- $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_K \in \mathbb{R}^{1\times T} \quad \Rightarrow \quad \mathrm{var}\,([\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_K]) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{var}(\bar{\mathbf{x}}_k)$

Here, the term $[\bar{\mathrm{x}}_1, \ldots, \bar{\mathrm{x}}_K]$ denotes the concatenation of the signals $\bar{\mathrm{x}}_k$ (row vectors).

The second equation means that the variance calculated across concatenated trials is the same as the average of the single-trial variances under the zero-mean assumption.

Recap: Variance and Covariance of Projected Data

Let samples $\mathbf{x}(1), \ldots, \mathbf{x}(T) \in \mathbb{R}^P$ and a vector $\mathbf{w} \in \mathbb{R}^P$ be given.

- **What is the variance of $\mathbf{X}$ in the direction of $\mathbf{w}$?**

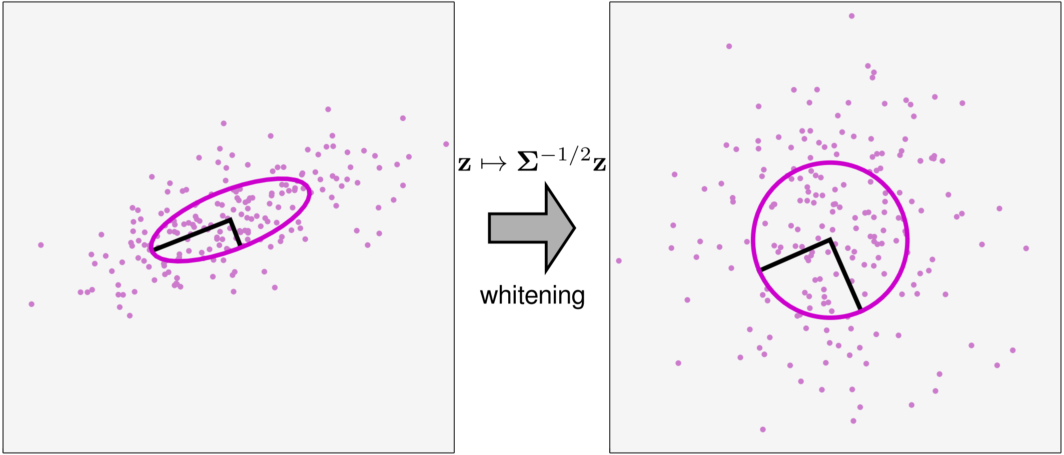
$$\operatorname{var}(\mathbf{w}^\top \mathbf{X}) = \tfrac{1}{T-1}\mathbf{w}^\top \mathbf{X}(\mathbf{w}^\top \mathbf{X})^\top = \tfrac{1}{T-1}\mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} = \mathbf{w}^\top \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{w} \tag{2}$$

!!! If $\mathbf{w}$ is an Eigenvector of $\boldsymbol{\Sigma}_{\mathbf{X}}$, then $\operatorname{var}(\mathbf{w}^\top \mathbf{X})$ is equal to the corresponding Eigenvalue (cf. eqn. (1)).

- **What is the cov. of $\mathbf{X}$ after transformation with matrix $\mathbf{P}$?**

$$\boldsymbol{\Sigma}_{\mathbf{P}^\top \mathbf{X}} = \tfrac{1}{T-1}\mathbf{P}^\top \mathbf{X}(\mathbf{P}^\top \mathbf{X})^\top = \tfrac{1}{T-1}\mathbf{P}^\top \mathbf{X}\mathbf{X}^\top \mathbf{P} = \mathbf{P}^\top \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{P} \tag{3}$$

Recap: Illustration of Whitening



Whitening is an invertible linear transformation for a given set of data points, such that the variance in all directions in the transform space is one. The resulting factors are uncorrelated.

Generalized Eigenvalue Decomposition

Generalized Eigenvalue Decomposition denotes the following theorem, which is also called *Simultaneous Diagonalization*:

Given $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}$ *symmetric and pos. definite (satisfied for covariance matrices), there exists an invertible matrix* $\mathbf{W} \in \mathbb{R}^{p \times p}$ *and a diagonal matrix* $\mathbf{D} \in$ *Diag*(p)*, such that*

$$\mathbf{AW} = \mathbf{BWD} \quad \& \quad \mathbf{W}^\top \mathbf{BW} = \mathbf{I} \tag{4}$$

A proof of that theorem can be found on `wikipedia.org`: Positive-definite matrix → Simultaneous diagonalization.

Multiplying the first equation by $\mathbf{W}^\top$ from the left, and then using the second equation we obtain

$$\mathbf{W}^\top \mathbf{AW} = \mathbf{W}^\top \mathbf{BWD} = \mathbf{D} \tag{5}$$

When choosing $\mathbf{B} = \mathbf{I}$ you obtain plain EVD.

CSP Solved by Generalized Eigenvalue Decomposition

CSP analysis is performed by a generalized Eigenvalue decomposition wrt. the matrices $(\mathbf{A} =)\ \mathbf{\Sigma}_1$ and $(\mathbf{B} =)\ \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$, cf. eqns (4), and (5):

$$\mathbf{W}^\top \mathbf{\Sigma}_1 \mathbf{W} = \mathbf{D}_1 \quad \& \quad \mathbf{W}^\top (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2) \mathbf{W} = \mathbf{I} \tag{6}$$

This gives a CSP filter matrix $\mathbf{W}$ (backward model; typically not orthogonal) which is the simultaneous diagonalizer of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$:

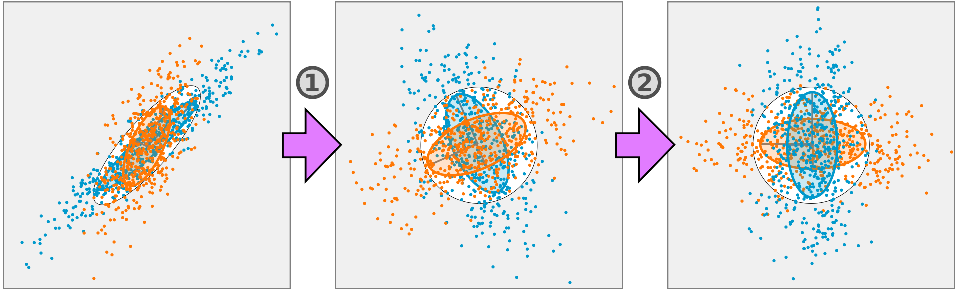
$$\begin{aligned} \mathbf{W}^\top \mathbf{\Sigma}_1 \mathbf{W} &= \mathbf{D}_1 \\ \mathbf{W}^\top \mathbf{\Sigma}_2 \mathbf{W} &= \mathbf{D}_2, \qquad \text{with } \mathbf{D}_2 := \mathbf{I} - \mathbf{D}_1 \end{aligned} \tag{7}$$

In particular, the scaling is such that $\mathbf{D}_1 + \mathbf{D}_2 = \mathbf{I}$.

In Python this can be done by

```
» d, W = scipy.linalg.eigh(a=Sigma1, b=Sigma1+Sigma2).
```

Illustration of CSP in 2D: Simultaneous Diagonalization

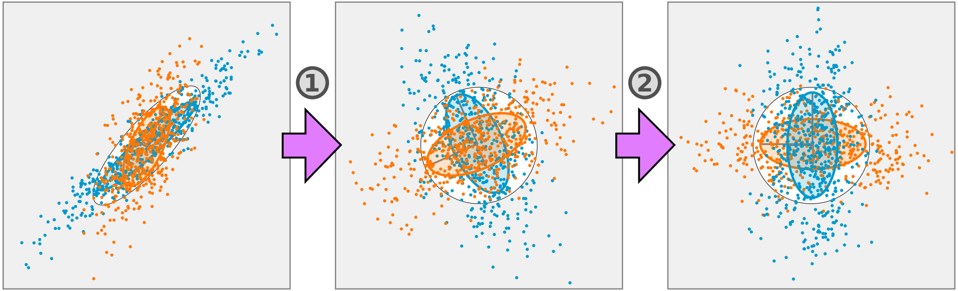


Left: Blue and orange ellipsoids refer to the class conditional covariance matrices, while the covariance sum $\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$ is depicted in white.

Central: Data distribution after whitening with respect to the covariance sum.

Right: After a final rotation, the variance along the horizontal direction is maximal for the orange class, while it is minimal for the blue class and vice versa along the vertical direction.

Illustration of CSP in 2D: Simultaneous Diagonalization



① $z \mapsto \mathbf{P}^\top z$ with $\mathbf{P}$ performing a whitening wrt. to $\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$

② $z \mapsto \mathbf{R}^\top z$ with rotation $\mathbf{R}$ to diagonalize both distributions
(Note, that this rotation preserves the achievement of whitening.)

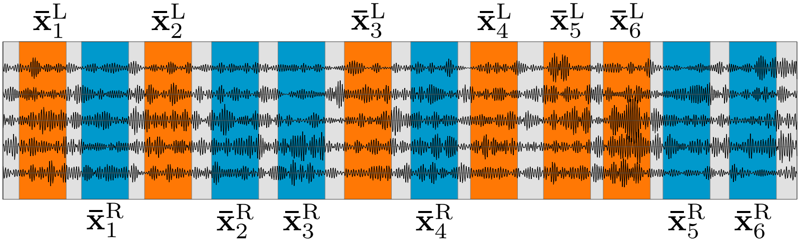
The combined transform $z \mapsto \mathbf{W}^\top z$ with $\mathbf{W} := \mathbf{PR}$ satisfies

$$\mathbf{W}^\top \boldsymbol{\Sigma}_1 \mathbf{W} = \mathbf{D}_1 \quad \& \quad \mathbf{W}^\top (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \mathbf{W} = \mathbf{I}.$$

Note: (1) $\mathbf{W}$ is typically **not** orthogonal and (2) $\mathbf{W}$ can directly be obtained by GEVD as discussed before.

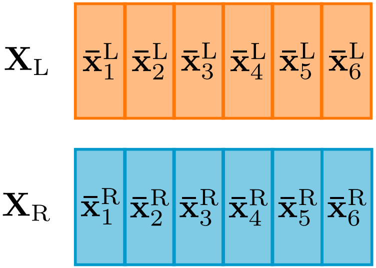
How to Apply CSP to the EEG Data

From band-pass filtered signals, we take the trials $\bar{\mathbf{x}}_k^{\mathrm{L}}$ of left and $\bar{\mathbf{x}}_k^{\mathrm{R}}$ of right hand motor imagery ...



and concatenate trials class-wise ...

$$\mathbf{X}_{\mathrm{L}} = [\bar{\mathbf{x}}_1^{\mathrm{L}}, \ldots, \bar{\mathbf{x}}_{K_L}^{\mathrm{L}}]$$
$$\mathbf{X}_{\mathrm{R}} = [\bar{\mathbf{x}}_1^{\mathrm{R}}, \ldots, \bar{\mathbf{x}}_{K_R}^{\mathrm{R}}]$$



and calculate the covariance matrices:

$$\mathbf{\Sigma}_{\mathrm{L}} = \frac{1}{T_{\mathrm{L}}-1}\mathbf{X}_{\mathrm{L}}\mathbf{X}_{\mathrm{L}}^{\top}, \quad \mathbf{\Sigma}_{\mathrm{R}} = \frac{1}{T_{\mathrm{R}}-1}\mathbf{X}_{\mathrm{R}}\mathbf{X}_{\mathrm{R}}^{\top}$$

According to eqn (7), $\mathbf{W}$ holds Eigenvectors of both, $\mathbf{\Sigma}^{\mathrm{L}}$ and $\mathbf{\Sigma}^{\mathrm{R}}$. Moreover the corresponding Eigenvalues in $\mathbf{D}^{\mathrm{L}}$ and $\mathbf{D}^{\mathrm{R}}$ sum to 1.

We obtain for the signals that are projected with an Eigenvector $\mathbf{w}_i$:

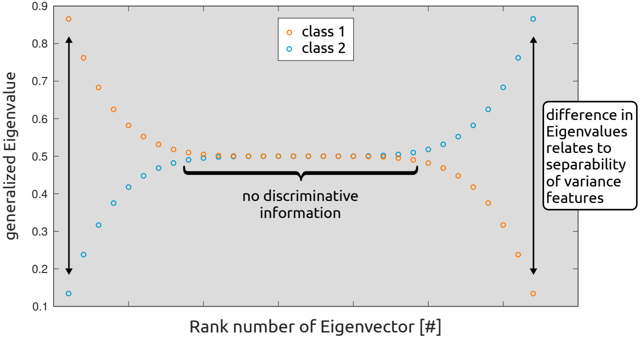
$$d_i^{\mathrm{L}} = \mathbf{w}_i^\top \mathbf{\Sigma}_{\mathbf{X}_{\mathrm{L}}} \mathbf{w}_i = \mathrm{var}\left(\mathbf{w}_i^\top \mathbf{X}_{\mathrm{L}}\right) = \frac{1}{K_{\mathrm{L}}} \sum_{k=1}^{K_{\mathrm{L}}} \mathrm{var}\left(\mathbf{w}_i^\top \bar{\mathbf{x}}_k^{\mathrm{L}}\right)$$

Hence, the variance of projected signals is (at least on average):

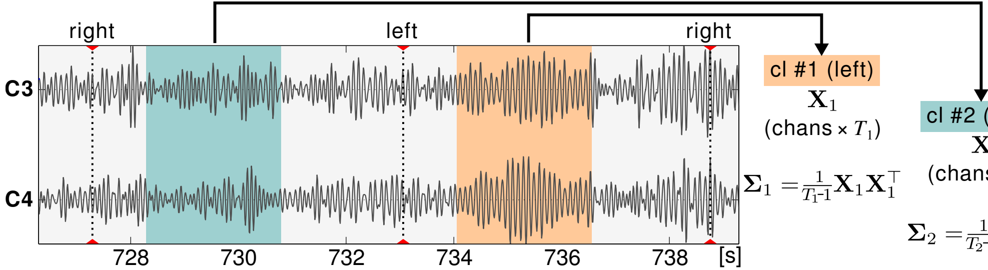
$$\mathrm{var}\left(\mathbf{w}_i^\top \bar{\mathbf{x}}_k^{\mathrm{L}}\right) \approx d_i^{\mathrm{L}}$$

and

$$\begin{aligned}\mathrm{var}\left(\mathbf{w}_i^\top \bar{\mathbf{x}}_k^{\mathrm{R}}\right) &\approx d_i^{\mathrm{R}} \\ &= 1 - d_i^{\mathrm{L}}\end{aligned}$$



EEG-signals during **motor imagery**, band-pass filtered (here 9–13 Hz):



$\mathbf{X}_1$ (chans × T_1)

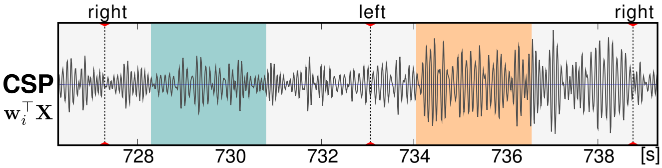
$\mathbf{X}_2$ (chans × T_2)

$\boldsymbol{\Sigma}_1 = \frac{1}{T_1\text{-}1}\mathbf{X}_1\mathbf{X}_1^\top$

$\boldsymbol{\Sigma}_2 = \frac{1}{T_2\text{-}1}\mathbf{X}_2\mathbf{X}_2^\top$

$$\mathbf{W}^\top\boldsymbol{\Sigma}_1\mathbf{W} = \mathbf{D} \quad \& \quad \mathbf{W}^\top(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{W} = \mathbf{I}$$

1) choose eigenvector $\mathbf{w}_i$ from $\mathbf{W}$ having a **large** eigenvalue d_i w.r.t. $\boldsymbol{\Sigma}_1$.

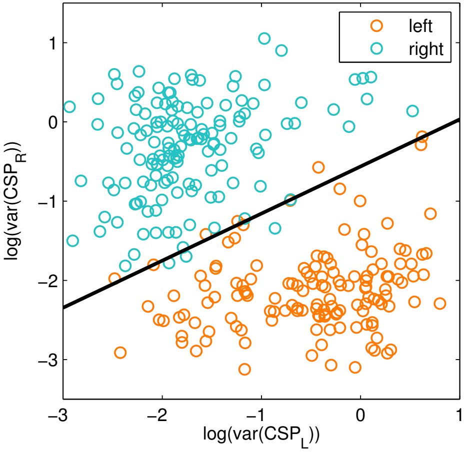


$\mathrm{var}(\mathbf{w}_i^\top\mathbf{X}_1) = d_i$ large

$\mathrm{var}(\mathbf{w}_i^\top\mathbf{X}_2) = 1 - d_i$ small

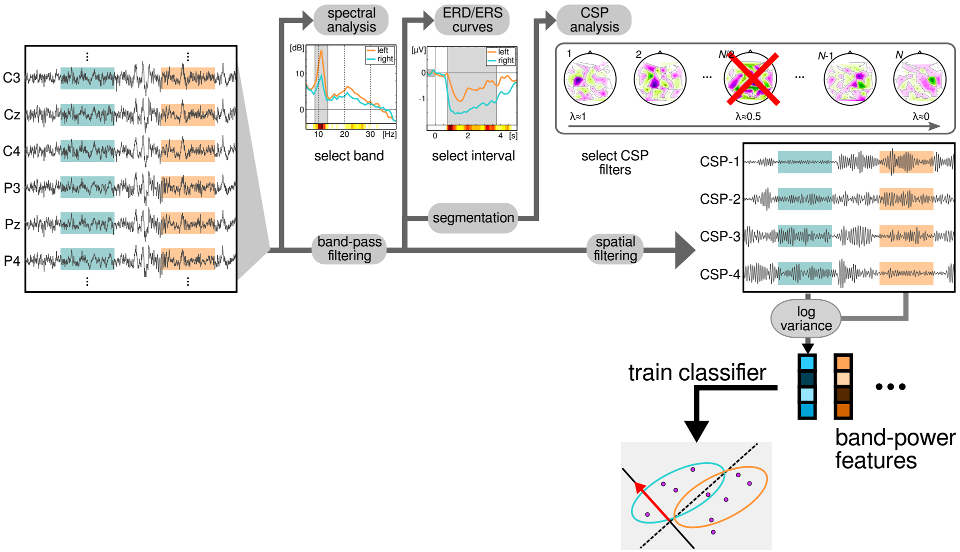
Final Step: CSP-based Features and Classification

From each trial of the CSP filtered signals **log band power features** are calculated. This is a scatter plot of the resulting CSP features for two selected dimensions:



Since the features are low dimensional, shrinkage is typically not necessary.

Calibration of CSP-based classification:



Calibration of CSP-based Classification

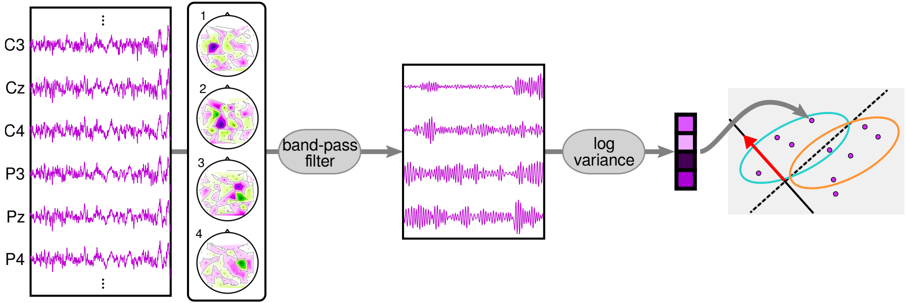
- Determine most discriminative frequency band,
- band-pass filter EEG in that band,
- extract single trials using the time interval in which ERD/ERS is expected (or determine discriminative interval from the data),
- calculate and select (typically 2-6) CSP filters,
- and apply them to EEG single trials,
- calculate the log variance within trials.

This results in a low dimensional feature vector for each trial (dimensionality equals number of selected CSP filters).

- Train a linear classifier like LDA on the features.

Applying CSP-based Classification Online

- band-pass filter the raw EEG
- and apply the spatial CSP filters (or vice versa),
- calculate the variance in short windows (e.g. last 500 ms),
- take the logarithm,
- and apply the classifier (weighting and bias).

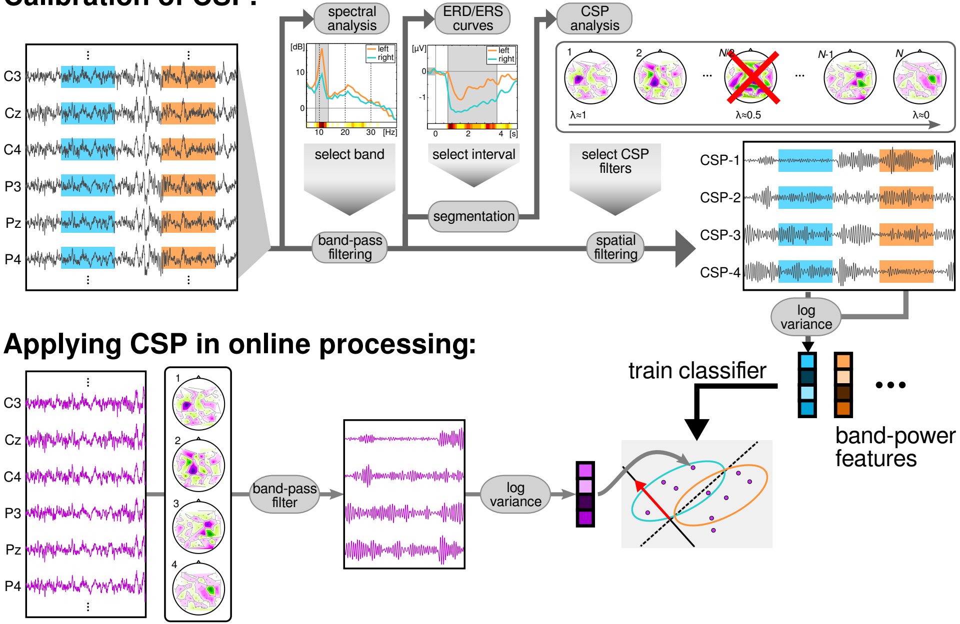


For more details on CSP see [Blankertz et al, IEEE Sig Proc Mag 2008].

This Completes the Picture of the CSP Work Flow

Calibration of CSP:

Applying CSP in online processing:



Validating a CSP-based Classification Method

Crucial measure: **generalization performance**, i.e., the accuracy obtained when the classifier is applied to new data, which have not been seen before, see also [Lemm et al, 2011].

This is particularly important for methods that use class labels like CSP!

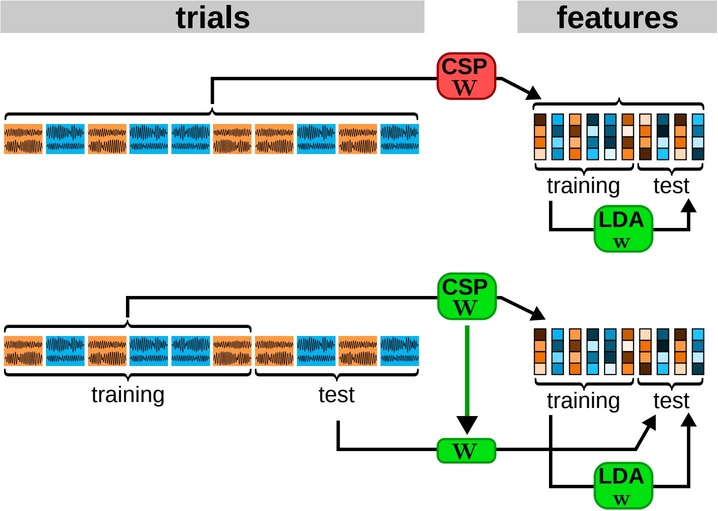
In practice: Estimation of CSP filter must be performed on training data only! Take the spatial filter obtained by CSP on training data and apply it to the test data.

For cross-validation this means, that CSP has to be applied in each fold on the training set and transfered to the test set.

Validating a CSP-based Classification Method

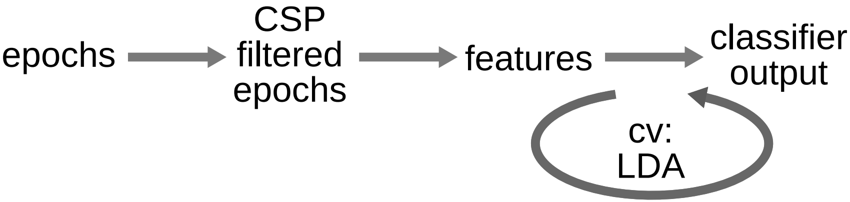
faulty: test samples (and its labels) are used to optimize CSP filters

valid: CSP and LDA both only use training samples in the optimization

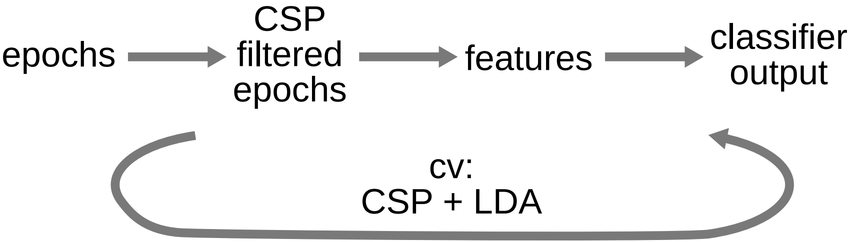


Cross-Validating a CSP-based Classification Method

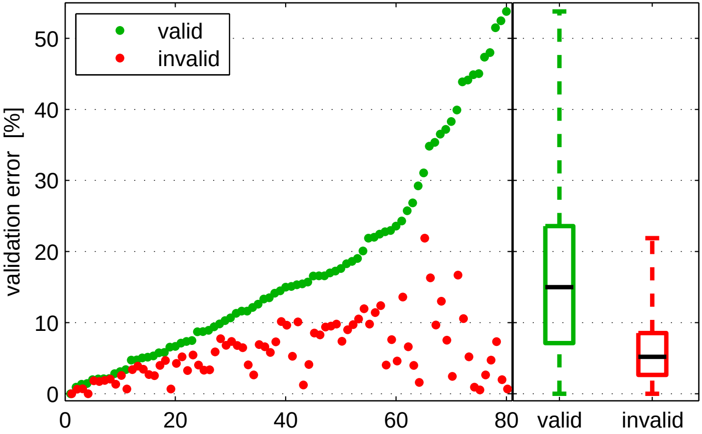
faulty:



valid:



Faulty and Valid Validation in CSP-based Classification



On data sets of 80 volunteers performing motor imagery, CSP-based classification was validated in a proper way (CSP within cross-validation) and in one incorrect way, where CSP filters have been calculated from the whole data set and only classification was cross-validated.

An Alternative Approach

Next, we discuss an alternative approach to derive CSP analysis which provides useful background information.

The previously introduced approach to obtain CSP filters via generalized Eigenvalue decomposition emerged like deus ex machina. The approach presented here will show, how one arrives from the idea of the optimization at its solution.

The approach of formalizing an optimization objective as Rayleigh coefficient is widely used and very useful.

It is a stroke of luck, if a problem can be cast into the maximization (or minimization) of a Rayleigh coefficient, because it provides an easy and effective solution.

Optimization with the Rayleigh Coefficient

We define the Rayleigh coefficient wrt the sym. matrices $\mathbf{A}$ and $\mathbf{B}$ as

$$R_{\mathbf{A},\mathbf{B}}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{A}\mathbf{w}}{\mathbf{w}^\top \mathbf{B}\mathbf{w}}.$$

The minimum and maximum of R can be obtained by the following theorem:

*The **Min-Max Theorem** states: $d_1 \leq R_{\mathbf{A},\mathbf{B}}(\mathbf{w}) \leq d_C$, if $d_1 \leq \cdots \leq d_C$ are the **generalized Eigenvalues** of $\mathbf{A}$ and $\mathbf{B}$.*

Let $\mathbf{w}_i$ be the corresponding Eigenvectors (i.e., $\mathbf{A}\mathbf{w}_i = \mathbf{B}\mathbf{w}_i d_i$). Then

$$R_{\mathbf{A},\mathbf{B}}(\mathbf{w}_i) = \frac{\mathbf{w}_i^\top \mathbf{A}\mathbf{w}_i}{\mathbf{w}_i^\top \mathbf{B}\mathbf{w}_i} = \frac{\mathbf{w}_i^\top \mathbf{B}\mathbf{w}_i d_i}{\mathbf{w}_i^\top \mathbf{B}\mathbf{w}_i} = d_i$$

Accordingly, the min (max) of R is attained for $\mathbf{w}_1$ (for $\mathbf{w}_C$).

CSP Optimization Problem

Let $\mathbf{X}_i \in \mathbb{R}^{C \times T_i}$ be the concatenation of all band-pass filtered trials of class i along the time dimension (T_i is the total number of time points of all trials of class i, and C being the number of channels).

$\mathbf{\Sigma}_i = \frac{1}{T_i - 1} \mathbf{X}_i \mathbf{X}_i^\top \in \mathbb{R}^{C \times C}$ are the corresponding covariance matrices (mean does not need to be subtracted – it is zero anyway, due to band-pass filtering).

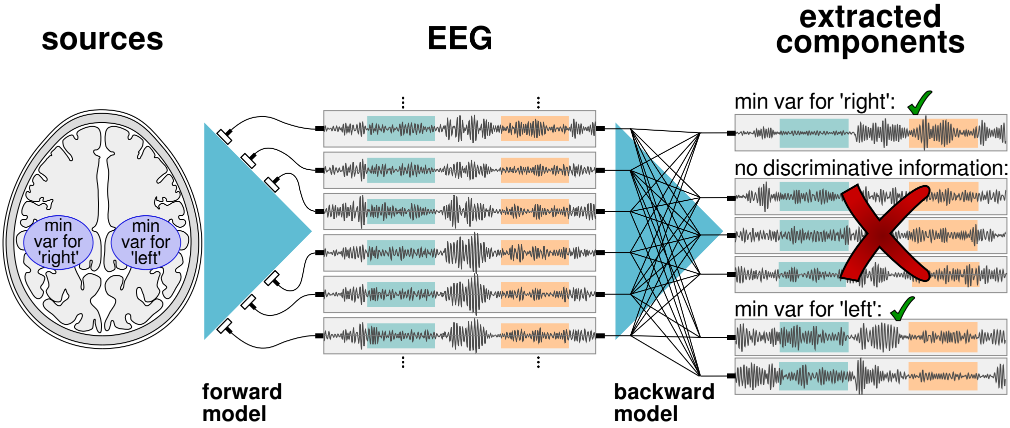
Then the **CSP** filter $\mathbf{w}_1$ that maximizes variance for class 1 is determined by the following optimization:

$$\underset{\mathbf{w} \in \mathbb{R}^C}{\operatorname{argmax}} \frac{\operatorname{var}(\mathbf{w}^\top \mathbf{X}_1)}{\operatorname{var}(\mathbf{w}^\top \mathbf{X}_1) + \operatorname{var}(\mathbf{w}^\top \mathbf{X}_2)} = \underset{\mathbf{w} \in \mathbb{R}^C}{\operatorname{argmax}} \frac{\mathbf{w}^\top \mathbf{\Sigma}_1 \, \mathbf{w}}{\mathbf{w}^\top (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2) \, \mathbf{w}} \tag{8}$$

According to the Min-Max Theorem this optimization can be solved by generalized Eigenvalue decomposition. So we arrive at the same solution as before.

CSP in the Framework of the Linear Model

CSP analysis can be interpreted in terms of our linear model of the EEG:



Note: For CSP analysis it does not matter, whether the extracted components correspond to single sources. The aim is just to extract most discriminative components.

Lessons Learnt

After this lecture you should

- have the idea about the aim of CSP analysis,
- be capable deriving the variance properties of CSP-filtered signals,
- be aquainted with generalized Eigenvalues,
- know the steps for
  - calibrating CSP-based classification, and
  - applying CSP-based classification
- be familiar with the derivation via the Rayleigh coefficient as a general approach to optimization problems,
- be aware of the special attention required for the validation of CSP-based classification.

References I

- Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K.-R. (2008).
  Optimizing spatial filters for robust EEG single-trial analysis.
  IEEE Signal Process Mag, 25(1):41–56.
- Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011).
  Introduction to machine learning for brain imaging.
  NeuroImage, 56:387–399.
- Parra, L., Christoforou, C., Gerson, A., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M., and Sajda, P. (2008).
  Spatiotemporal linear decoding of brain state.
  IEEE Signal Process Mag, 25(1):107–115.
- Ramoser, H., Müller-Gerking, J., and Pfurtscheller, G. (2000).
  Optimal spatial filtering of single trial EEG during imagined hand movement.
  IEEE Trans Rehabil Eng, 8(4):441–446.