Introduction to Single-Trial EEG analysis & Brain-Computer Interfacing

Benjamin Blankertz

Neurotechnology Group, Technische Universität Berlin

benjamin.blankertz@tu-berlin.de

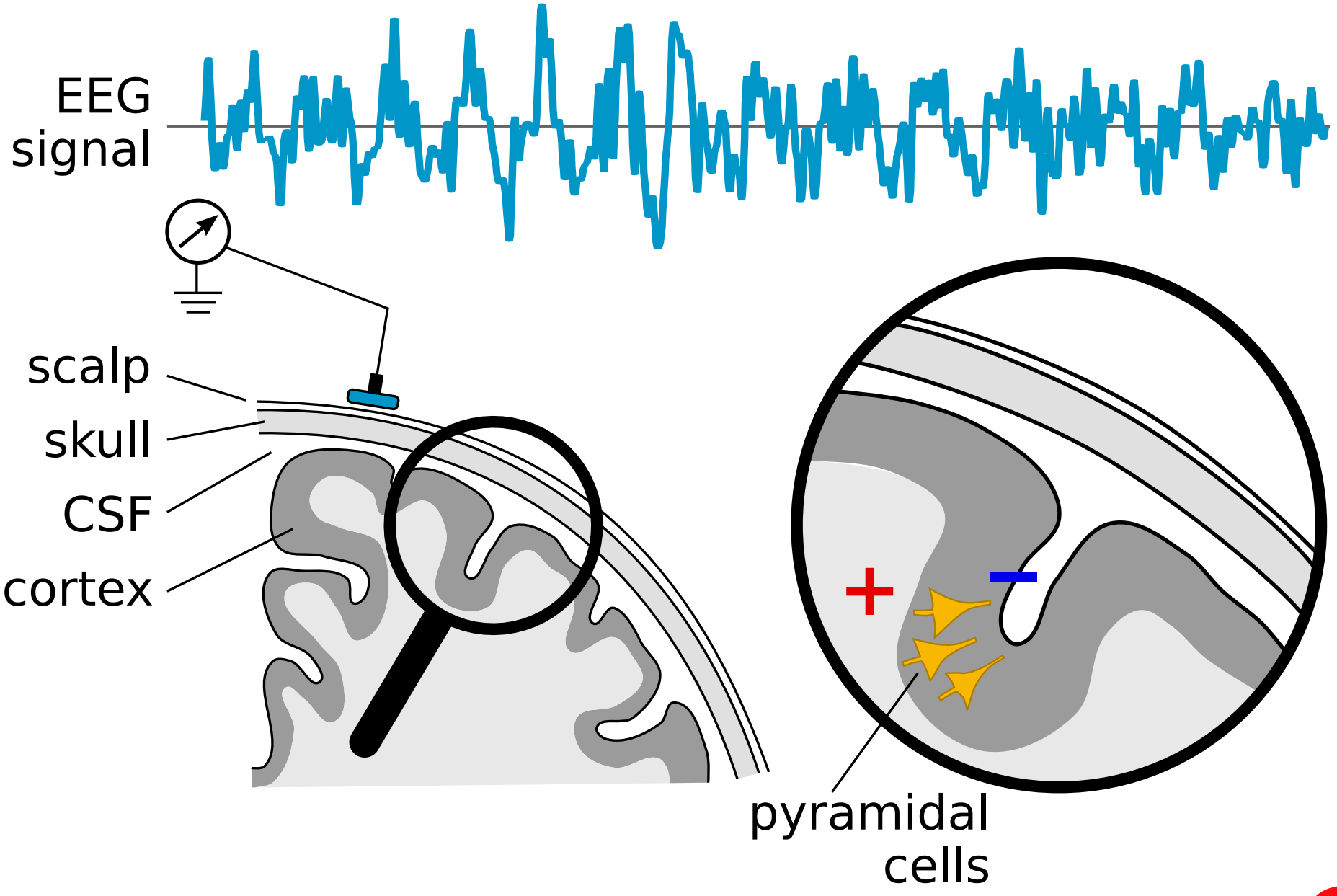
http://www.user.tu-berlin.de/blanker

25|Aug|2013

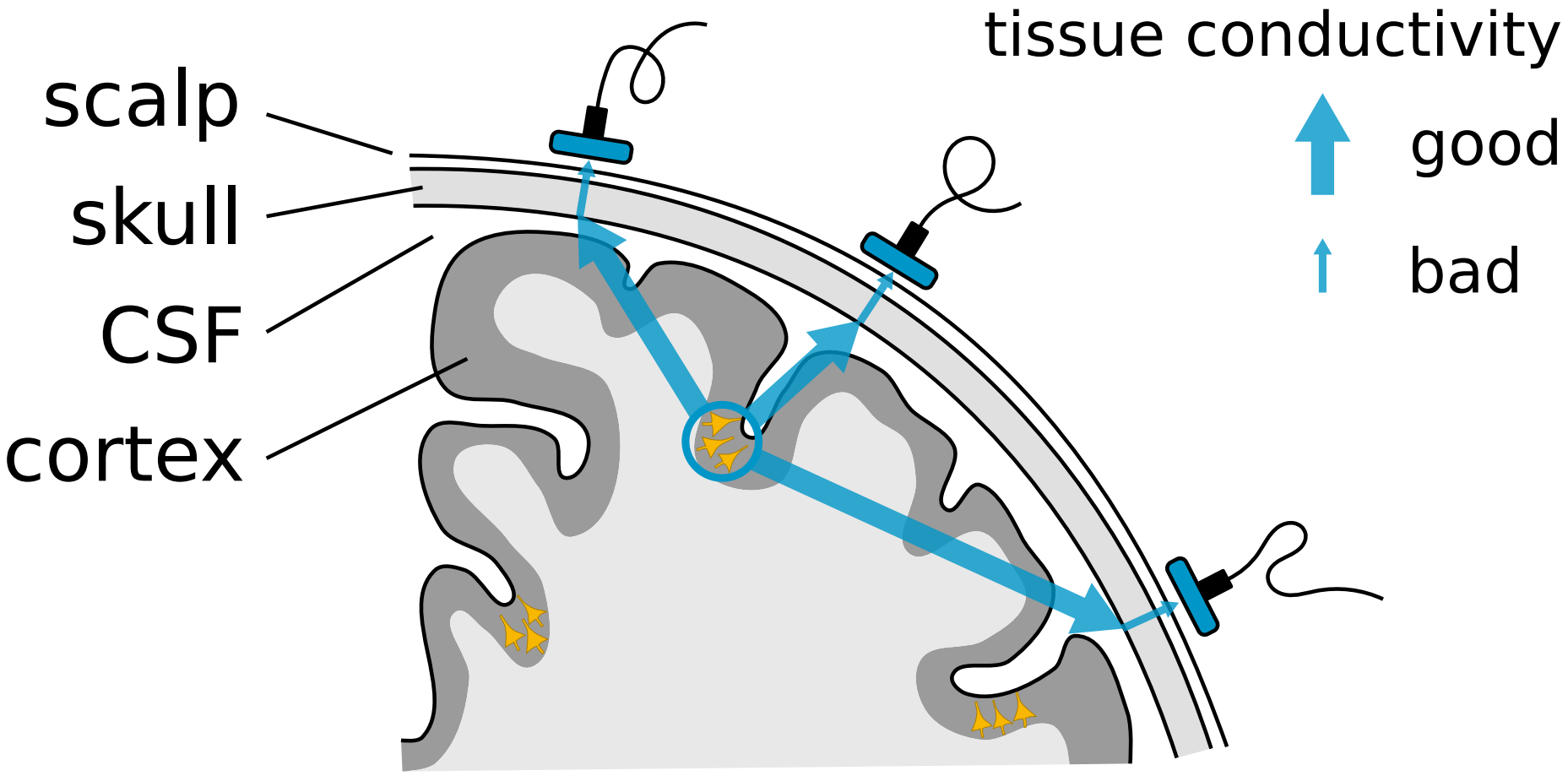
Part I: Introduction

Part I: EEG, ERPs, and from Uni- to Multivaritate Features

Generation of EEG Signals



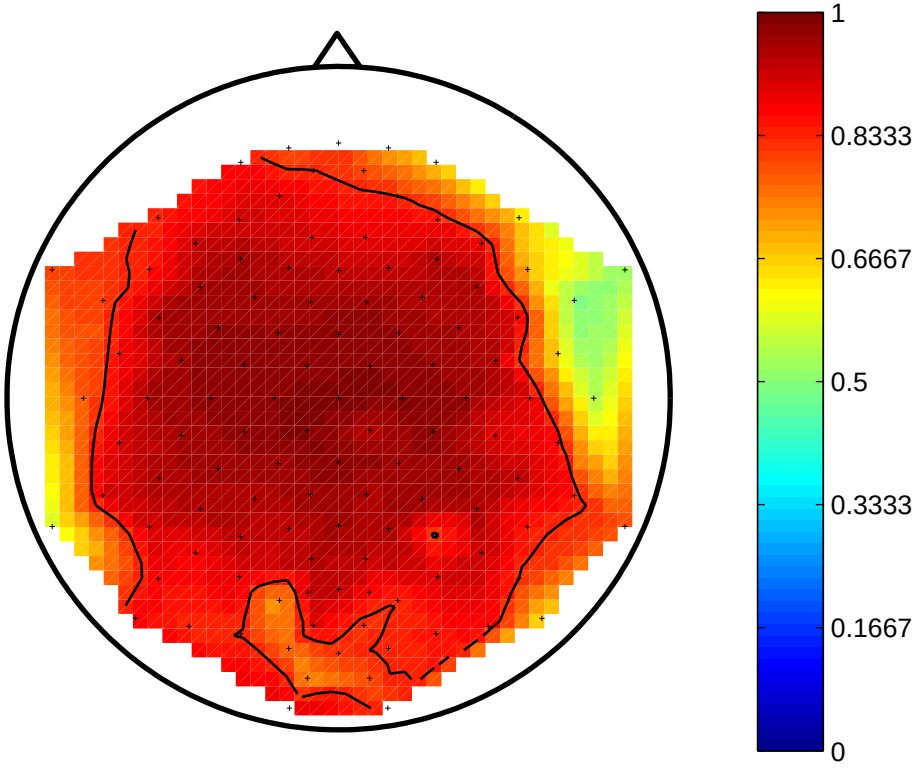
Volume Conduction in EEG



The signal arrives with almost equal intensity at different scalp locations due to the different tissue conductivities.

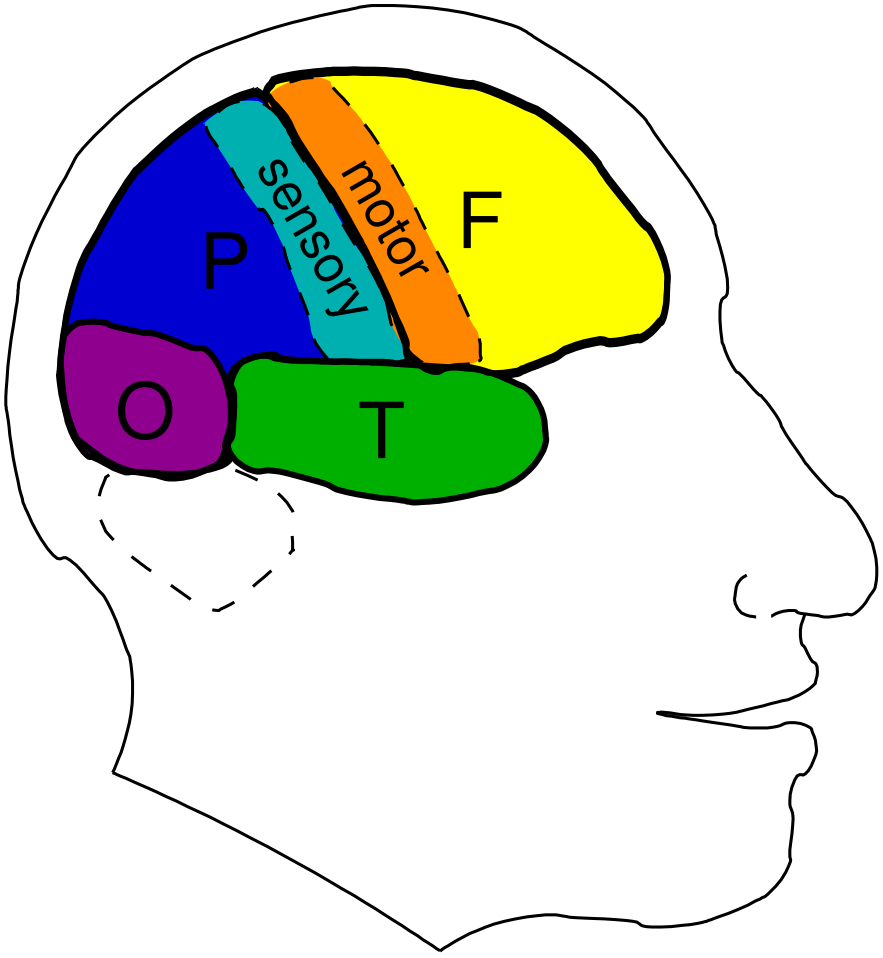
Mind Spatial Smearing!

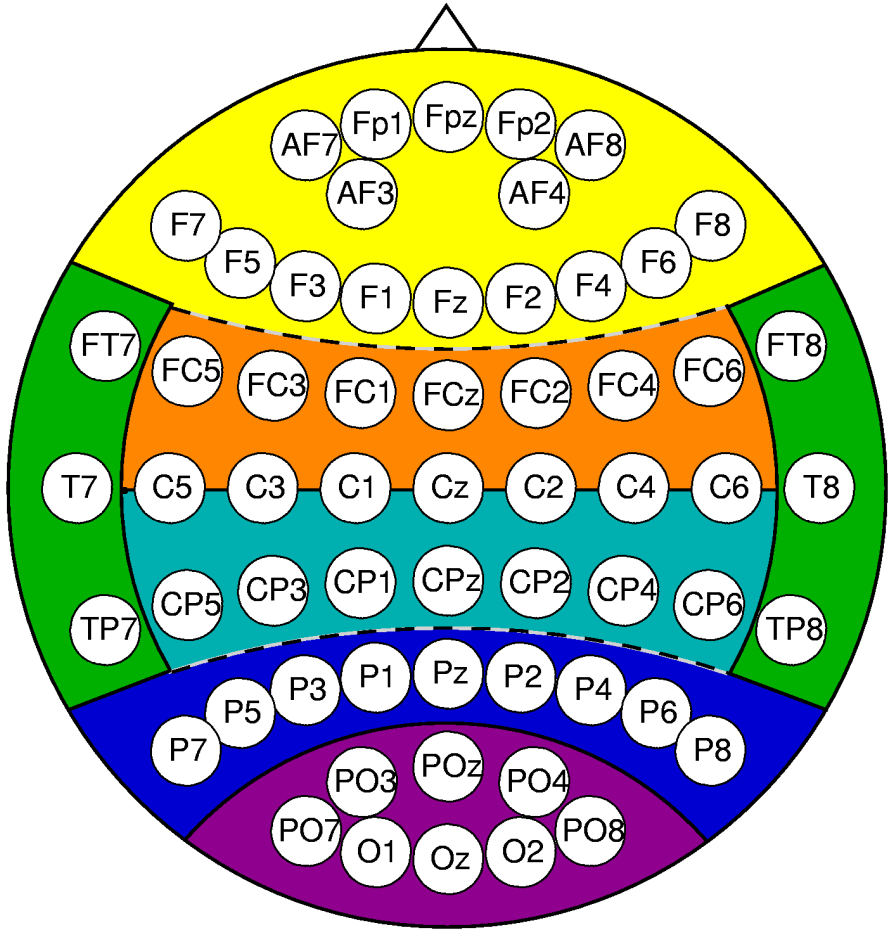
- Raw EEG scalp potentials are known to be associated with a large spatial scale owing to volumne conduction.
- In this typical example data set, most of the channels are highly correlated:



The map shows the correlation coefficient of each channel with channel Cz in the center.

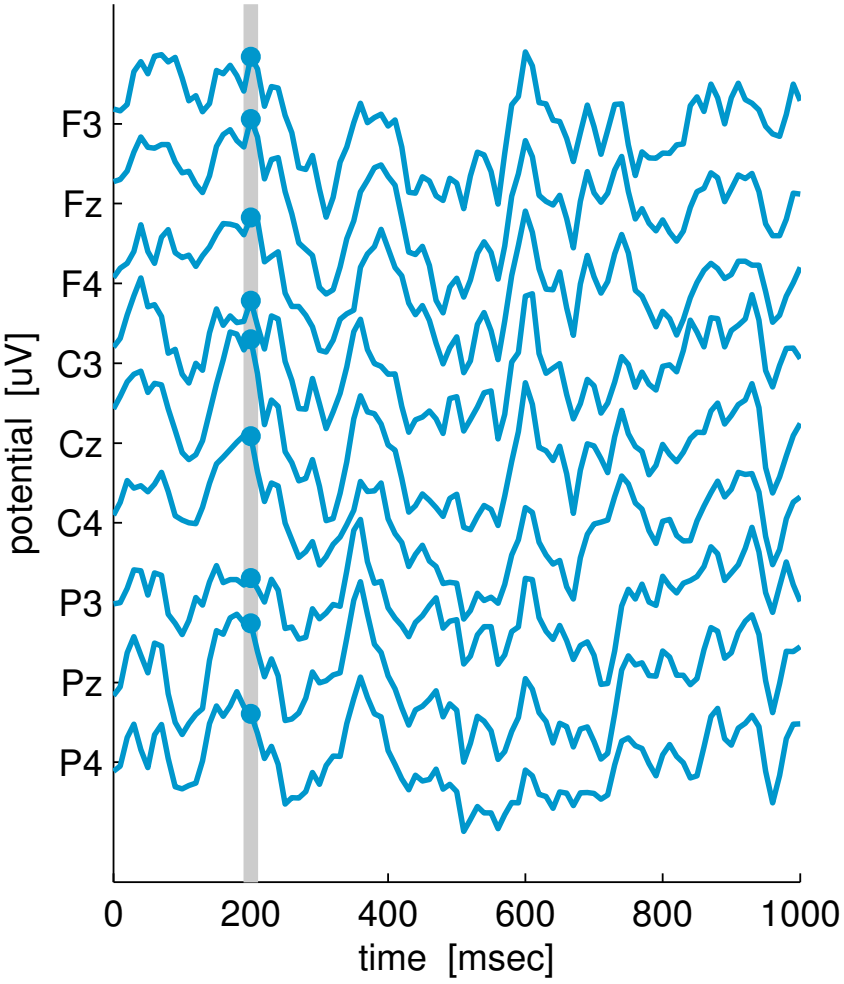
Areas of the Brain

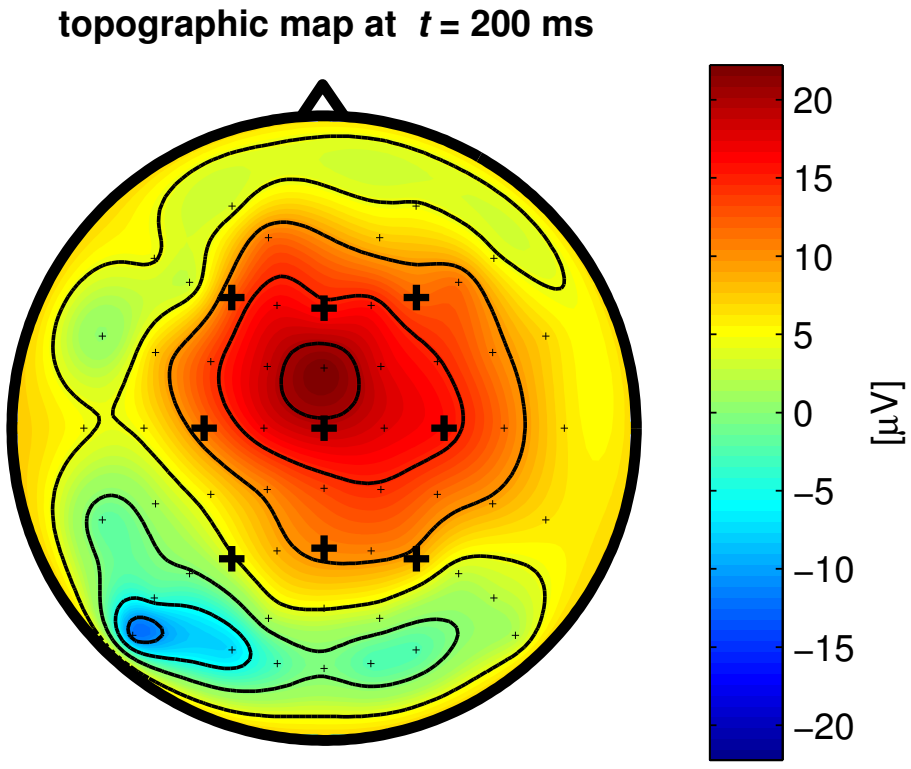




Brain lobes: **F**rontal, **P**arietal, **T**emporal, **O**ccipital.

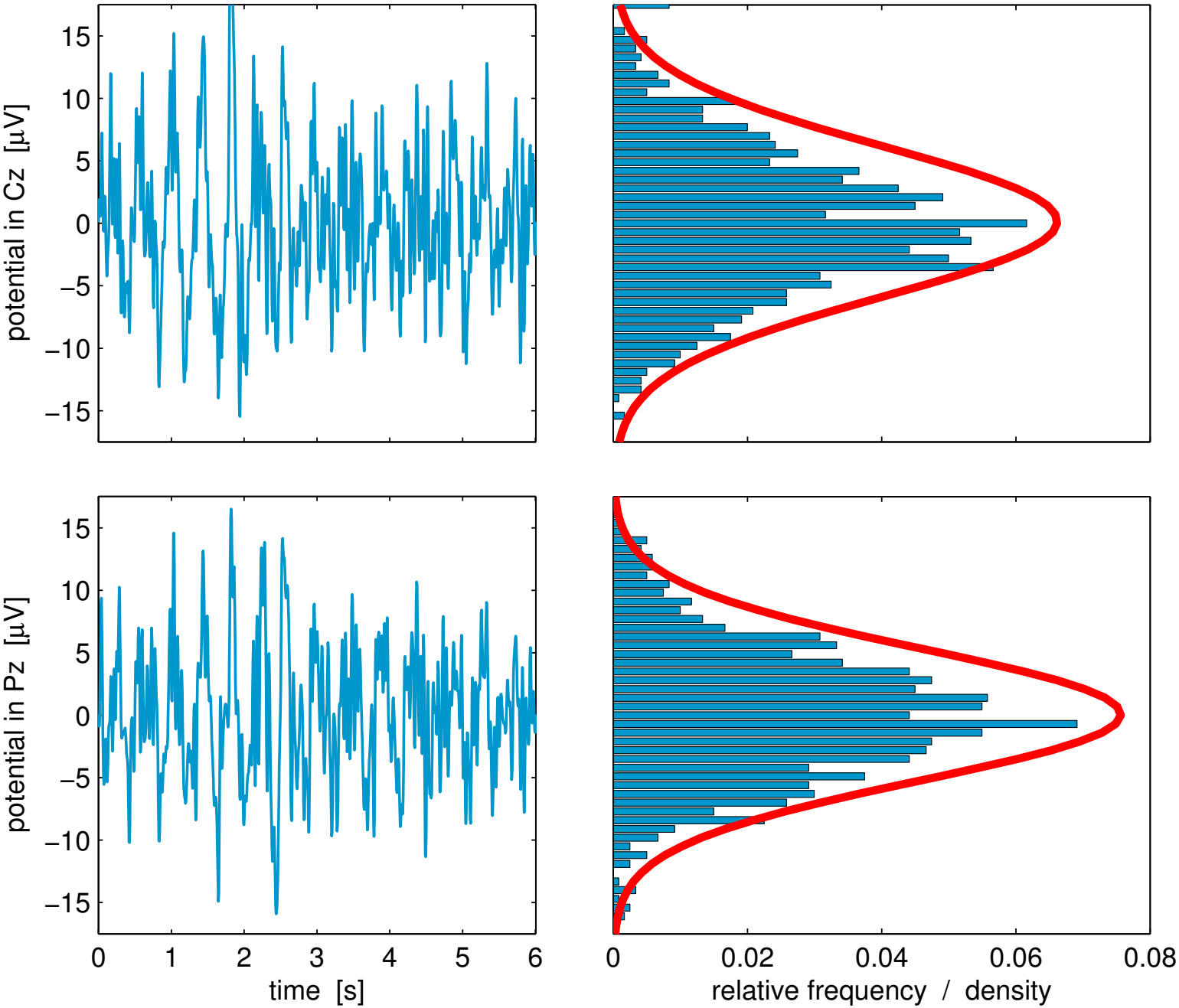
Visualizing Potentials of Multichannel EEG as Maps





Similarly, distributions of band-power across channels can be displayed as topographic maps.

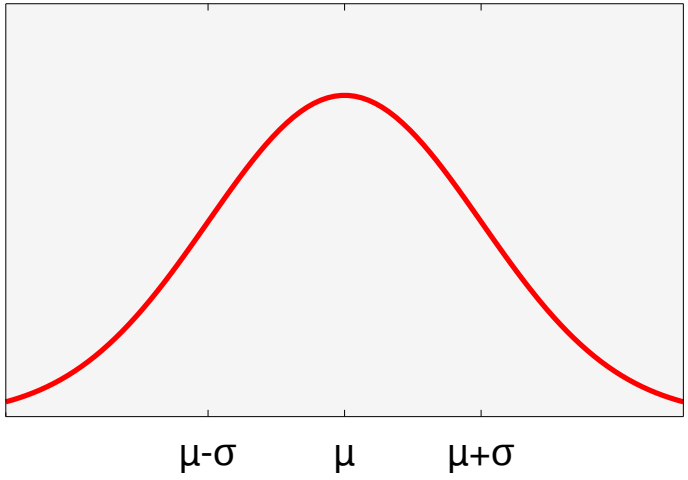
Univariate Distributions of Single-Channel EEG



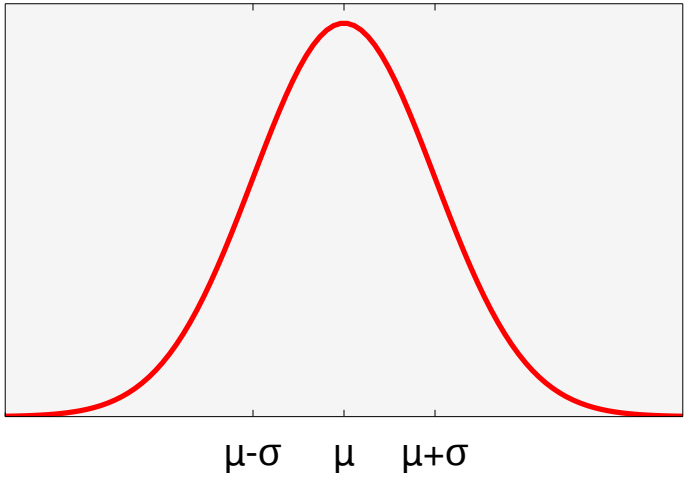
Two Univariate Gaussian Distributions

In the absence of artifacts, the distributions in each single channel is often close to a Gaussian.

Component #1

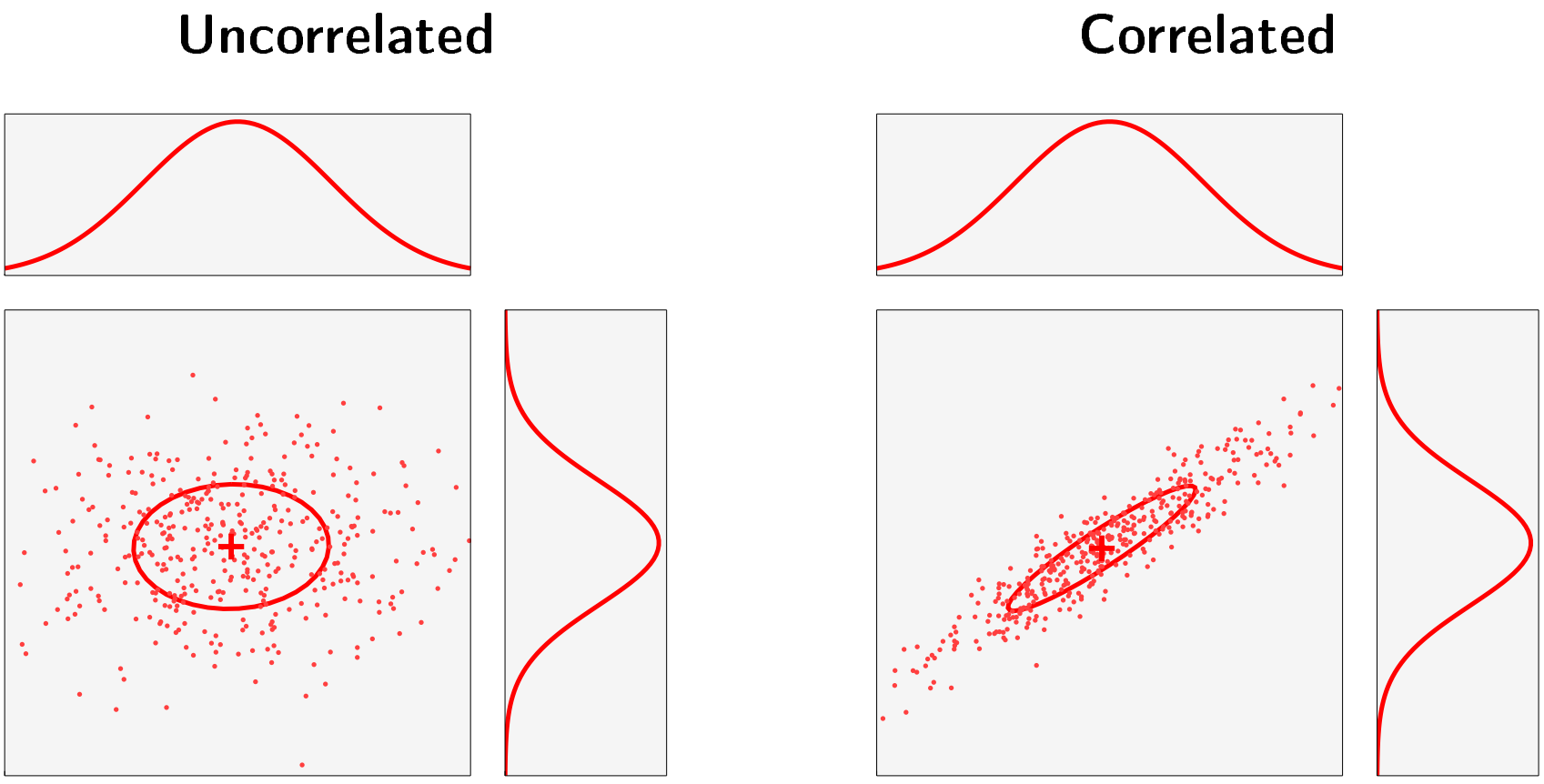


Component #2



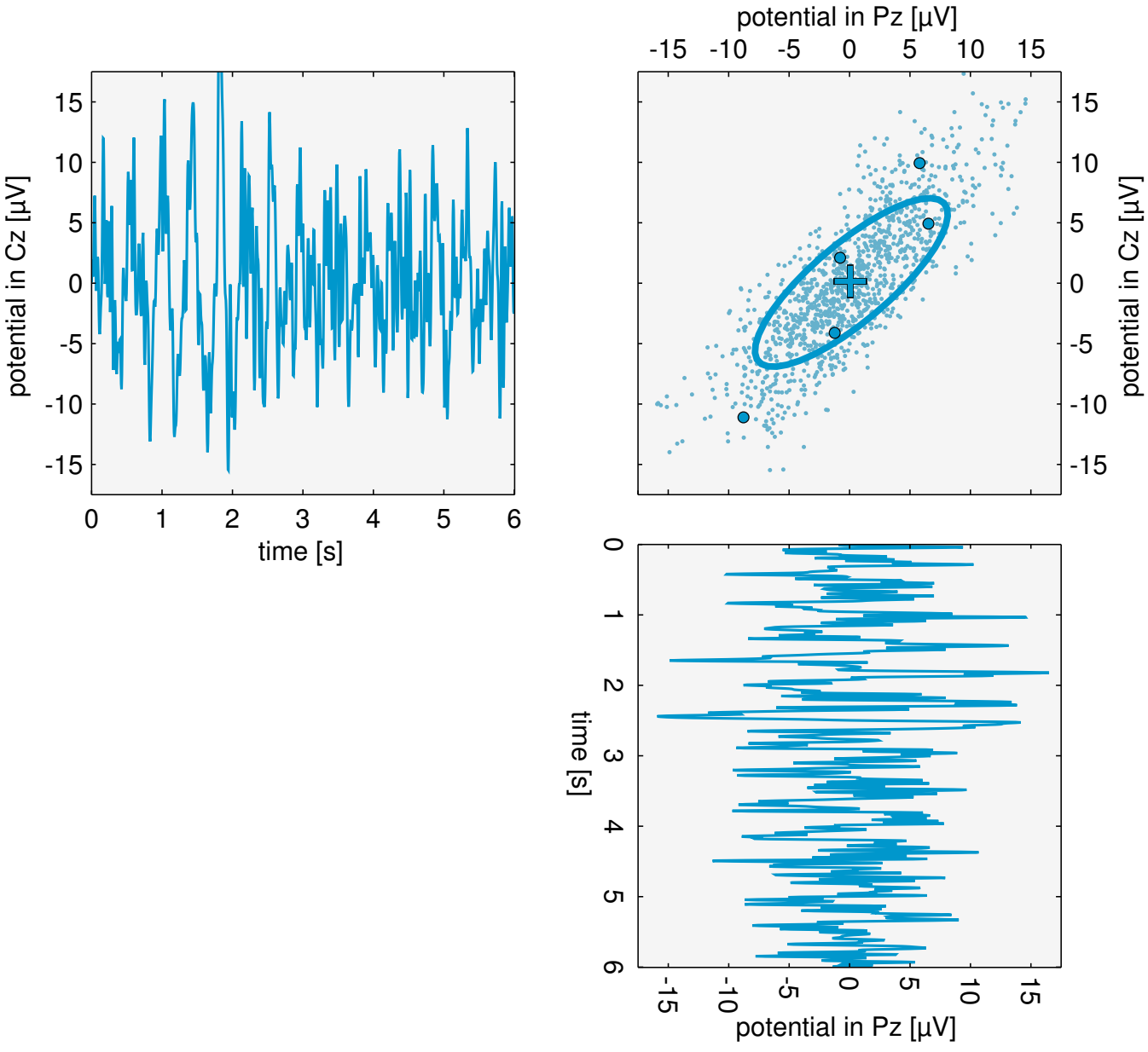
But what might their joint multivariate distribution look like?

Two-Dimensional Gaussians - Correlated or Uncorrelated



- Two-dimensional Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ may have uncorrelated ($\boldsymbol{\Sigma}$ diagonal) or correlated components.
- This cannot be decided from the marginal distributions (univariate components).
- In EEG: remember spatial smearing! $\Rightarrow$ Strong Correlation.

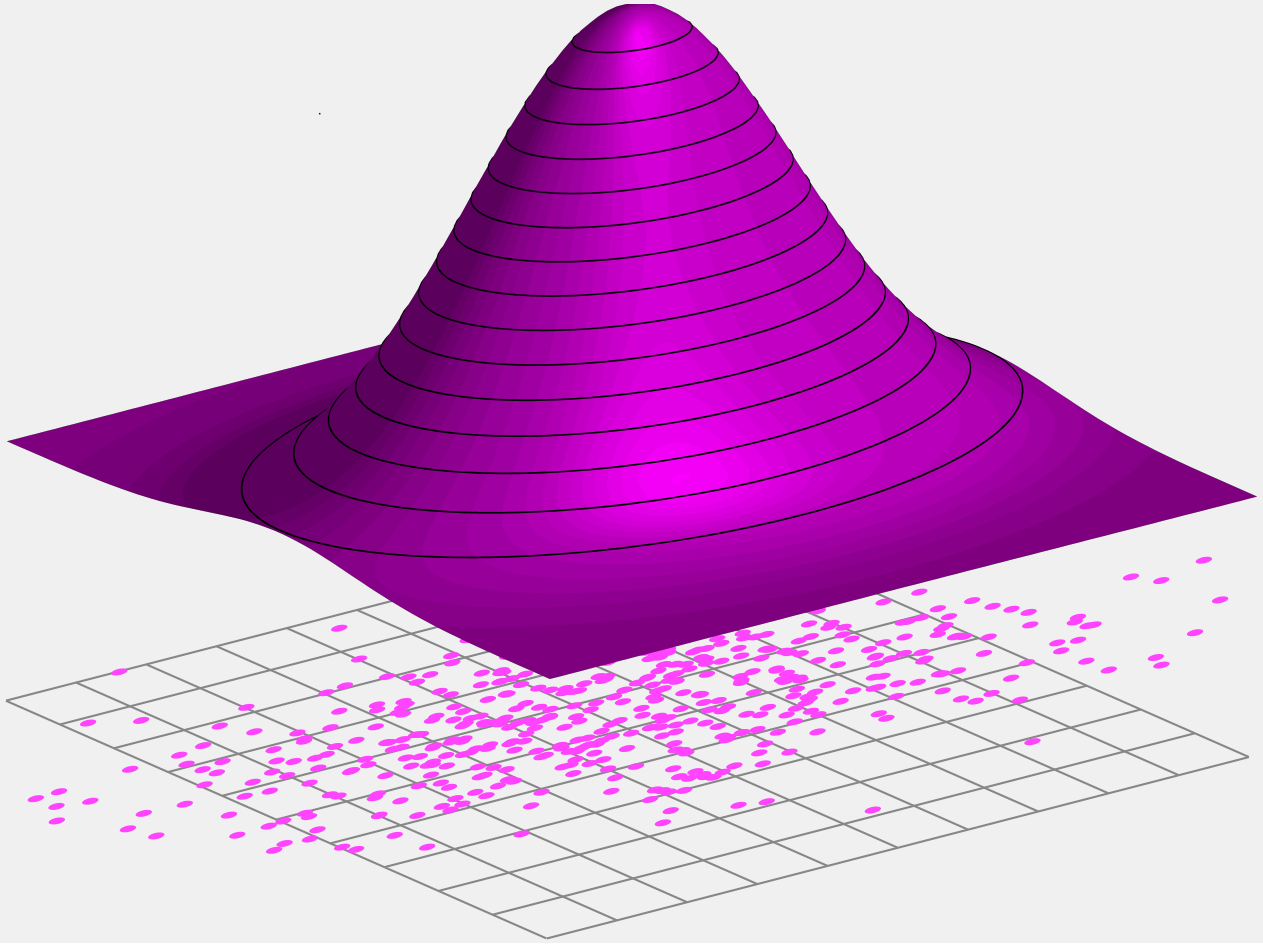
Visualizing Two Channel EEG as Scatter Plot



Multivariate Gaussian Distributions

(a)

$$g(\mathbf{x}) = \frac{1}{2\pi\sqrt{det(\mathbf{\Sigma})}} exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^{\top}\right)$$

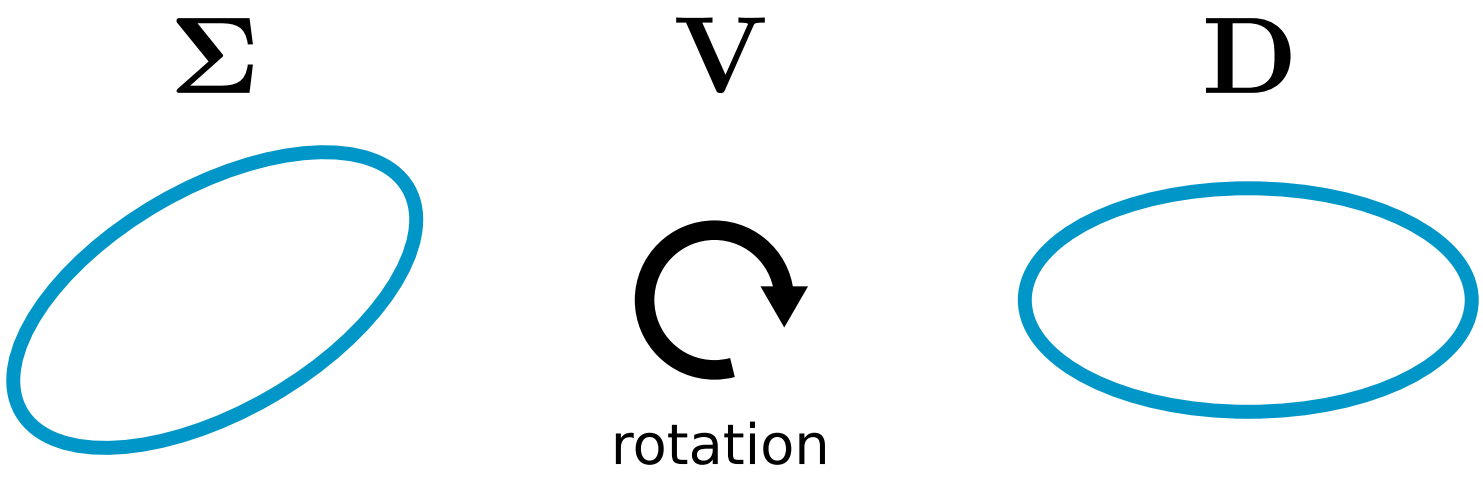


Eigenvalue Decomposition (EVD)

Given $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ symmetric and pos. definite, there exists an orthonormal matrix $\mathbf{V} \in \mathcal{O}(p)$ of **Eigenvectors** and a diagonal matrix $\mathbf{D} \in \mathrm{Diag}(p)$ of **Eigenvalues**, such that

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

In our case, $\mathbf{\Sigma}$ is the covariance matrix of EEG signals $\mathbf{X} \in \mathbb{R}^{p \times T}$.



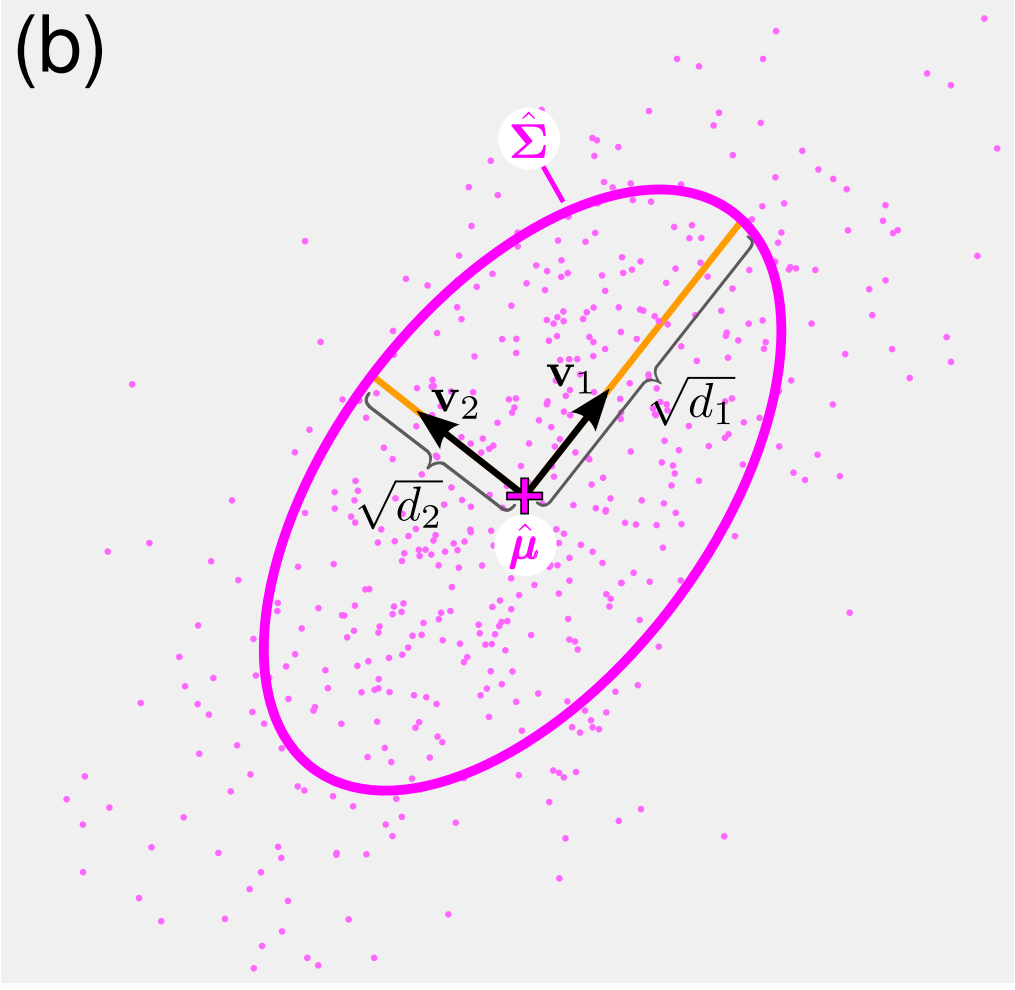
- The eigenvectors corresponding to the m largest eigenvalues allow a representation of data $\mathbf{X}$ in an m-dimensional subspace with minimum projection error.

Characterization of Gaussian Distributions

Assume samples $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathbb{R}^p$ are modeled as $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$.

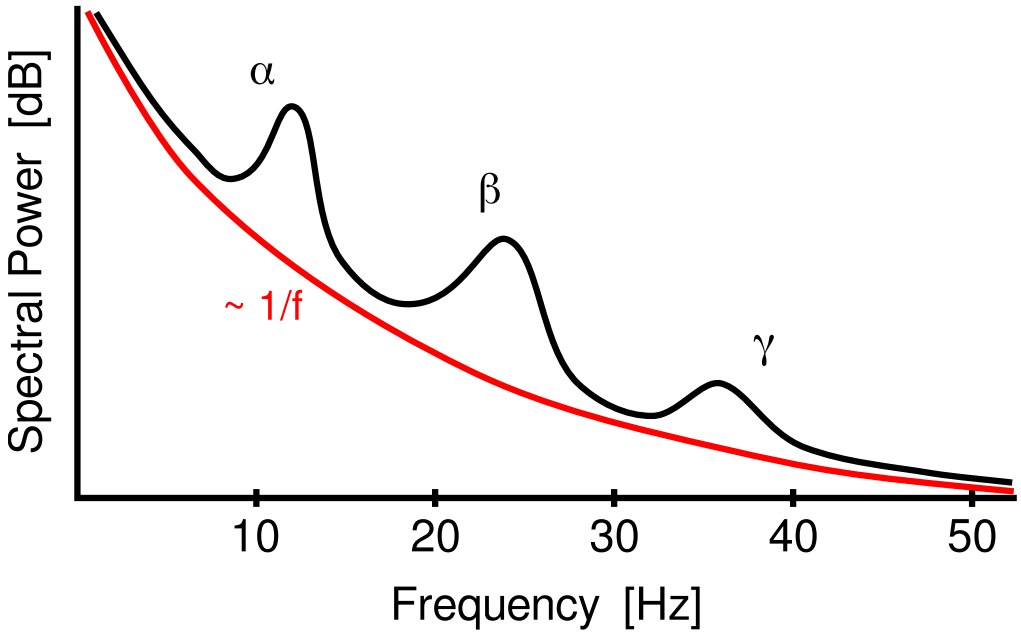
Eigenvalue decomposition of the covariance $\hat{\boldsymbol{\Sigma}}$ of $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{V}\mathbf{D}\mathbf{V}^\top, \quad \text{with orthonormal } \mathbf{V} \text{ and diagonal } \mathbf{D}.$$



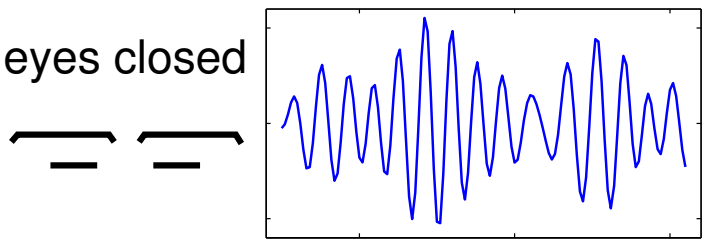
- Eigenvectors are columns of $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$.
- Eigenvalues are diagonal elements d_i of $\mathbf{D}$.
- $\sqrt{d_i} = \mathrm{std}(\mathbf{v}_i^\top \mathbf{X})$
- In $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ typically $\boldsymbol{\mu}$ is considered to be the ideal true value (signal) and $\boldsymbol{\Sigma}$ noise.
- The vector of Eigenvalues is called *Eigenvalue spectrum*

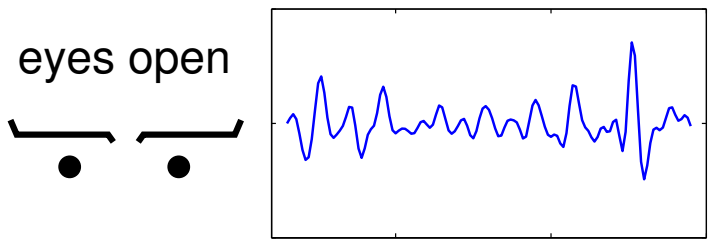
Ongoing Brain Activity – Brain Rhythms

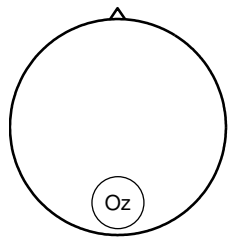


The figure shows an idealized spectrum of EEG signals.

Most brain rhythms are idle rhythms, i.e., they are **attenuated** during activation, e.g., the α-rhythm (around 10 Hz) in visual cortex:







Evoked and Event-Related Potentials

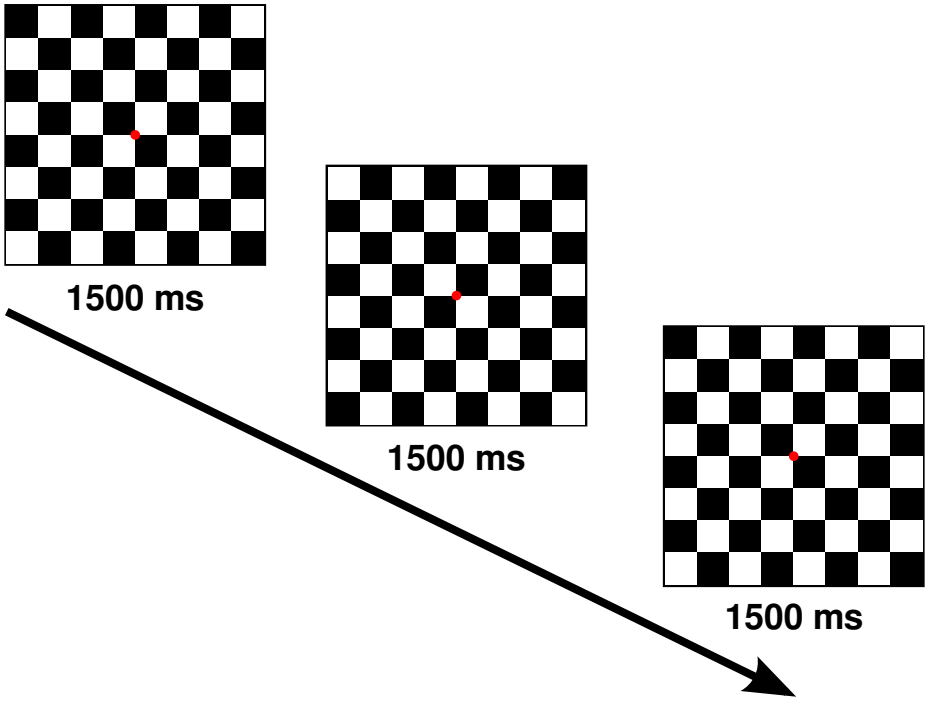
Event-Related Potentials (ERPs) are brain responses that are time-locked to some *event*. The event may be an external sensory stimulus or internal, associated with the execution of a motor, cognitive, or psychophysiologic task.

A subclass are the **evoked potentials (EPs)** which reflect the processing of the physical stimulus, rather than 'higher' processes, that might involve memory, expectation, or attention.

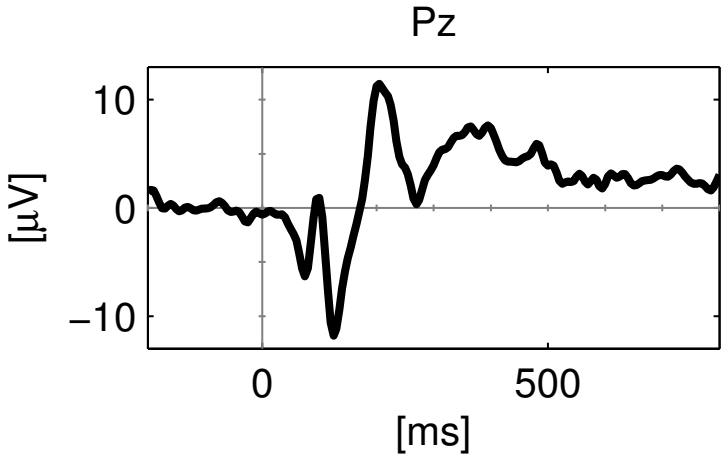
Visual Evoked Potential (VEP)

Visual evoked potentials (VEPs) are ERPs that are caused by stimulation of a subject's visual field. Commonly used are, e.g., checkerboard stimlui that flip at an inter-stimulus interval of 1 to 3s. More specifically, these are called *Flash VEPs* in contrast to many different variants of VEPs [Odom et al, 2004].

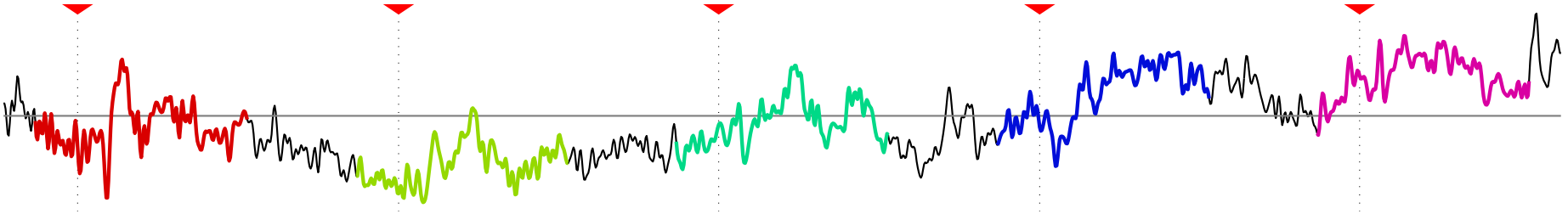
VEPs can be observed regardless of attention, but the amplitude depends strongly on the focality of the stimulus.



Continuous Signal and Event-Related Segments



Continous Signal (with markers):



Segments (epochs) around stimulus markers:

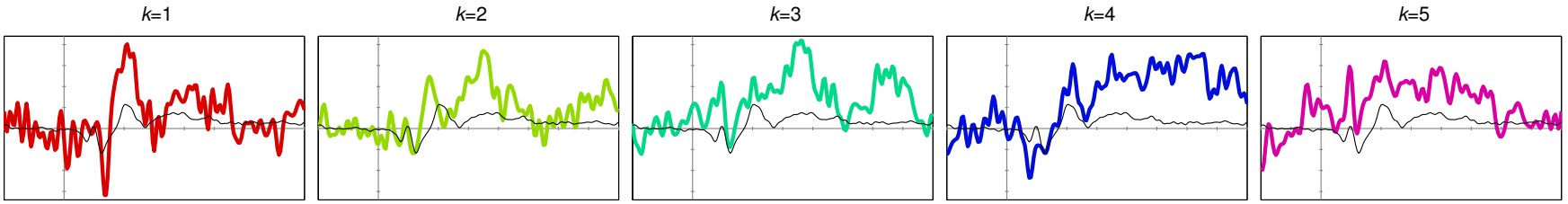
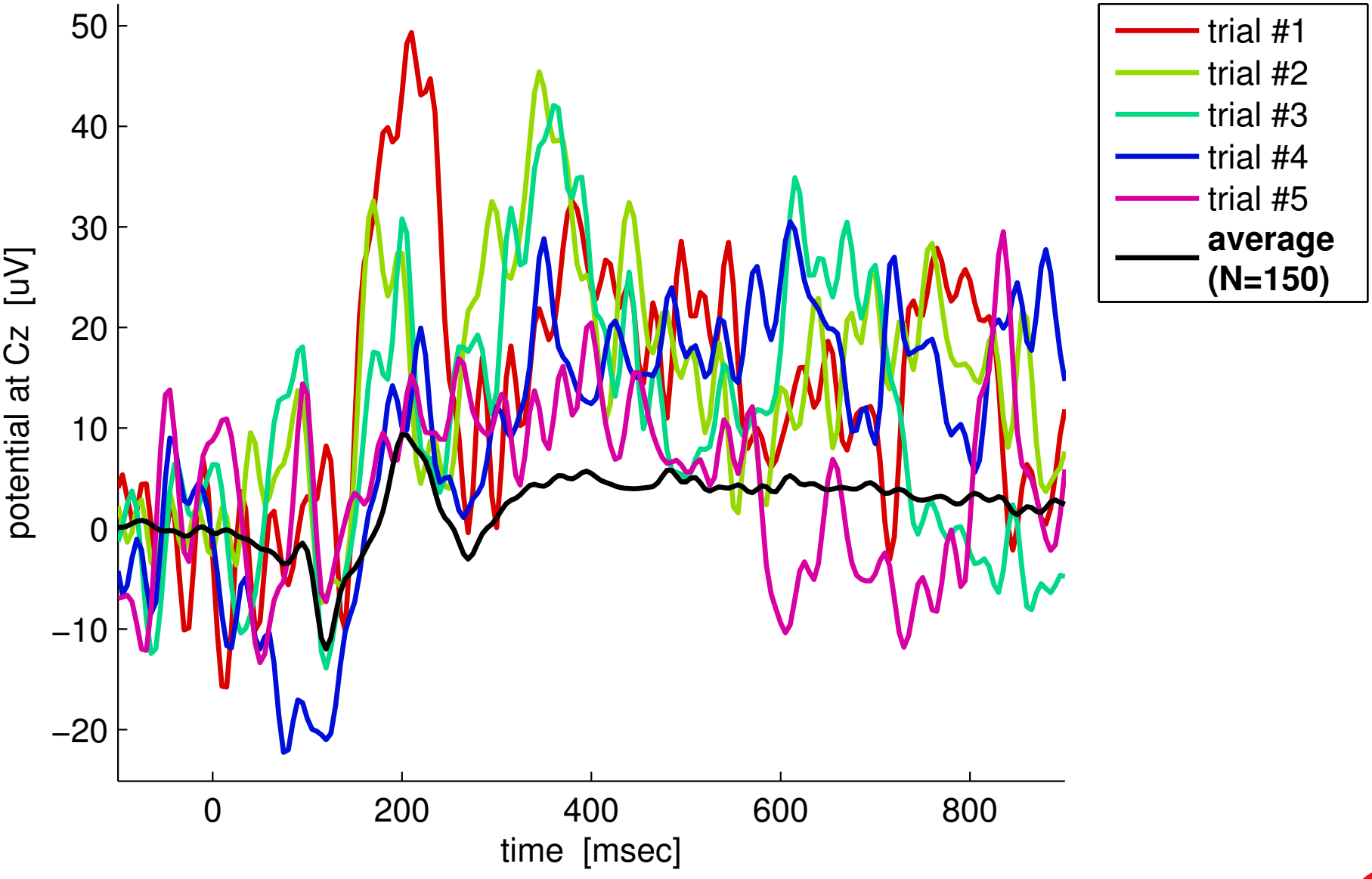
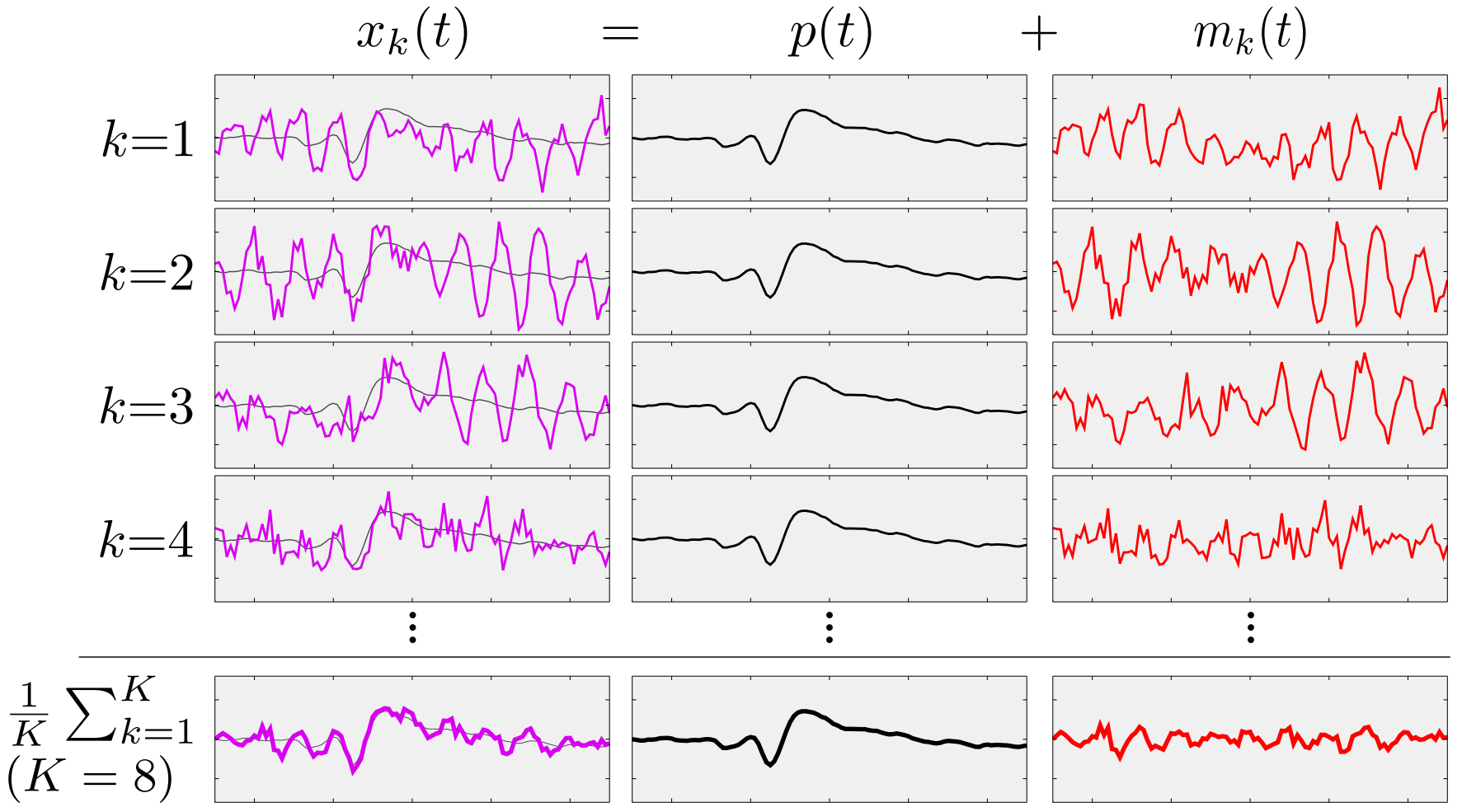


Illustration: Single-Trials and ERPs



Averaging Across Trials

Let us assume the ERP $p(t)$ is constant in each trial k ($k = 1, \ldots, K$), while the 'noise' $m_k(t)$ is iid $\mathcal{N}(0, \sigma_m^2)$ distributed (for a fixed t):



ERP-based Brain-Computer Interfaces (BCIs)

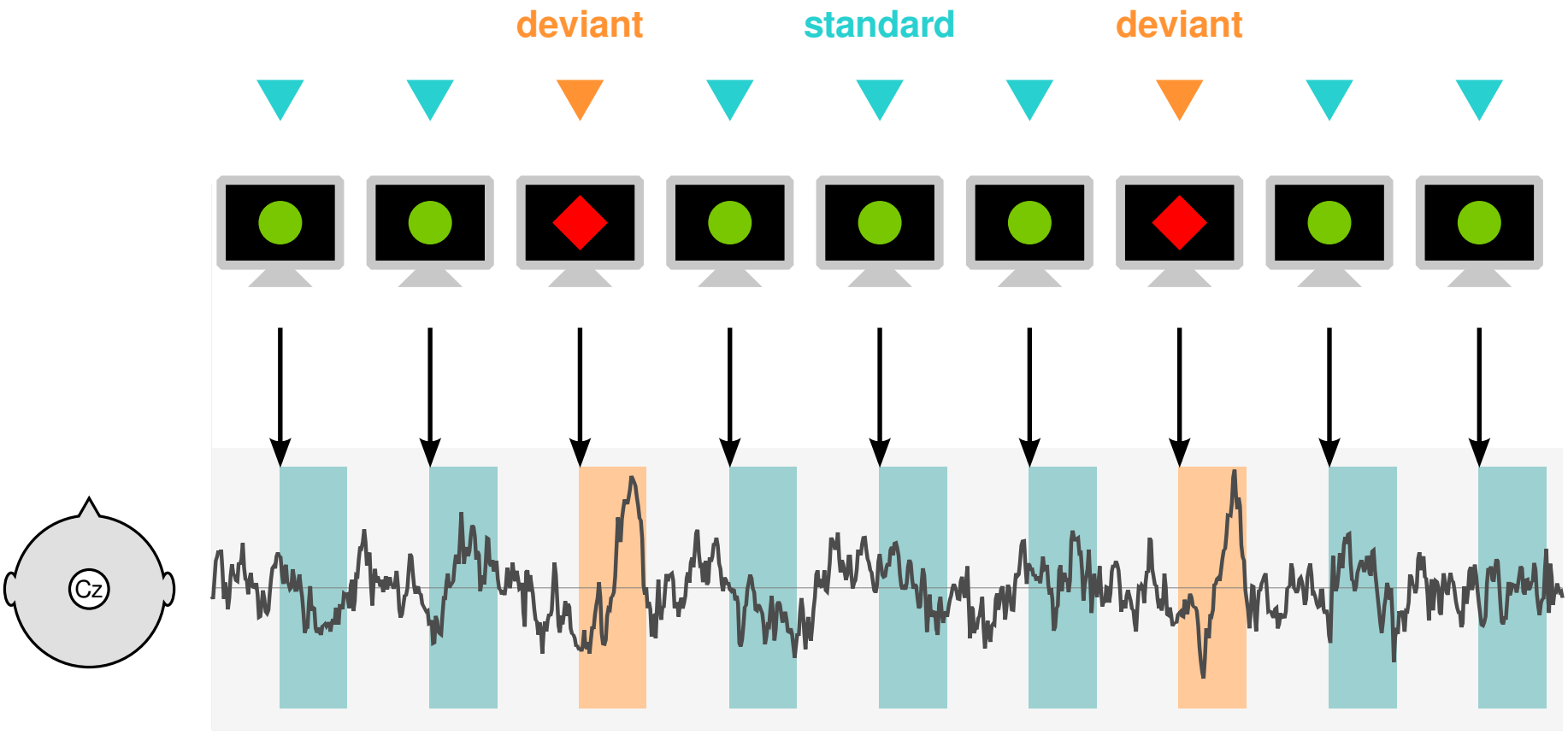
ERPs can be used in the control of a BCI. But

- there must be at least **two conditions** that the user can voluntarily attain, and
- they need to be discriminable in **single-trials**, or at least with just very little averaging.

Oddball Paradigm

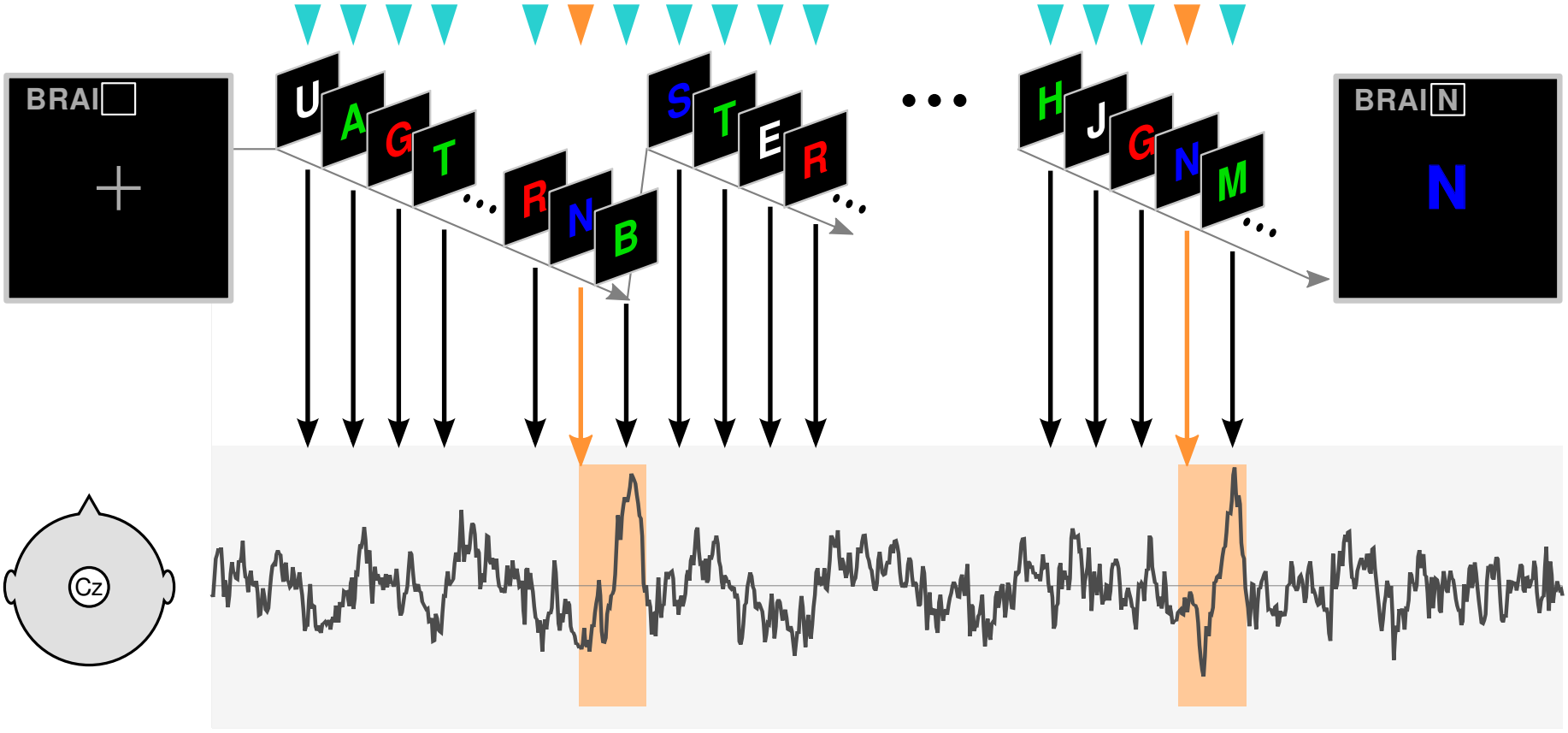
- In the classical *oddball paradigm* (experiment) there are two kinds of stimuli (e.g., low and high tones; or green circles and red squares).
- Stimuli are presented at regular intervals in a random sequence.
- One kind of stimuli is more frequent than the the other one, e.g., with a ratio of 80:20.
- The frequent stimuli are called *standards* and the infrequent stimuli *deviants*.
- The test person has the task to 'detect' the deviant stimuli and, e.g., to count their occurences silently.

Basics: Oddbal Paradigm, P300, BCI Speller



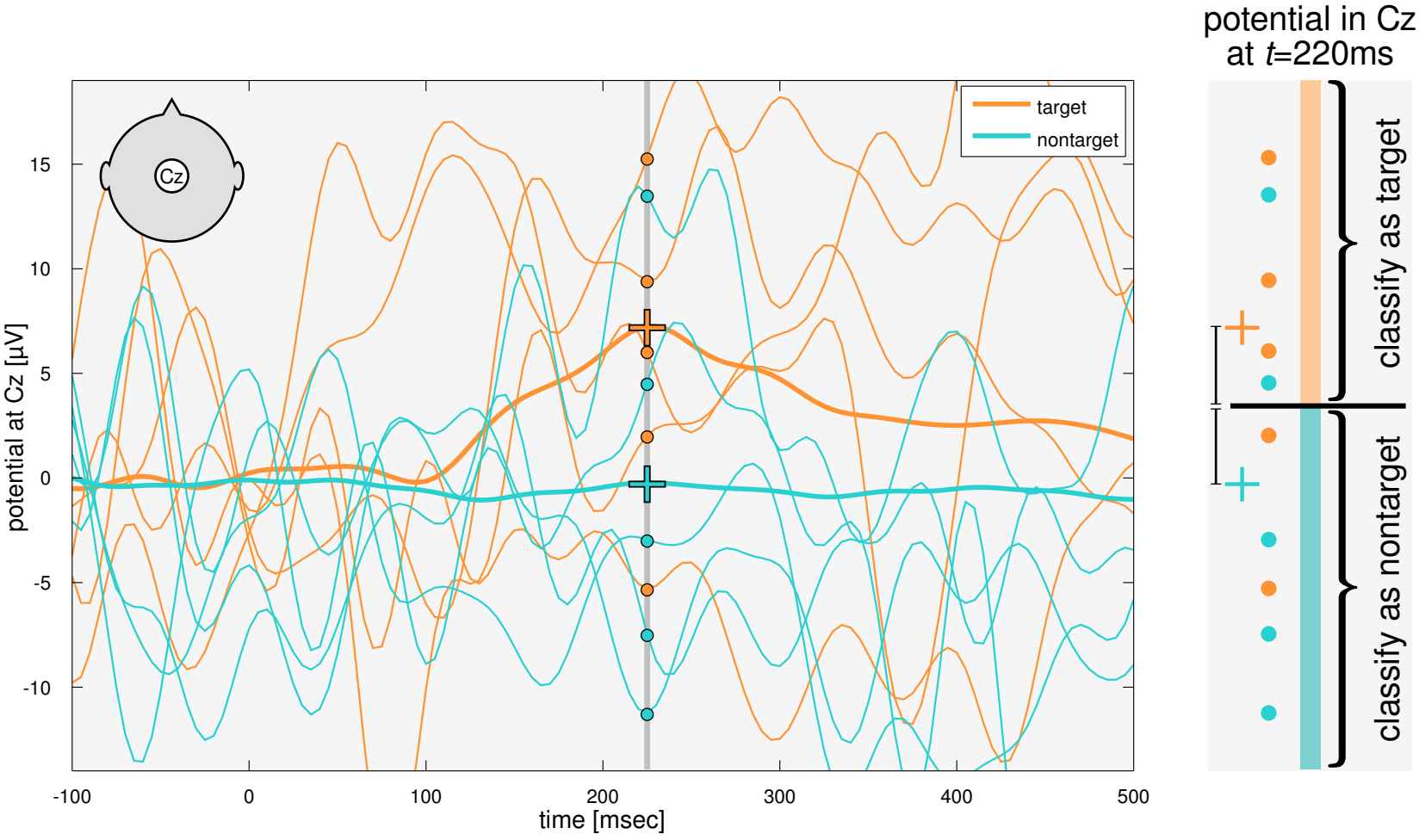
- Segments of the signals are called *epochs* or *single-trials*.

Basics: Oddbal Paradigm, P300, BCI Speller



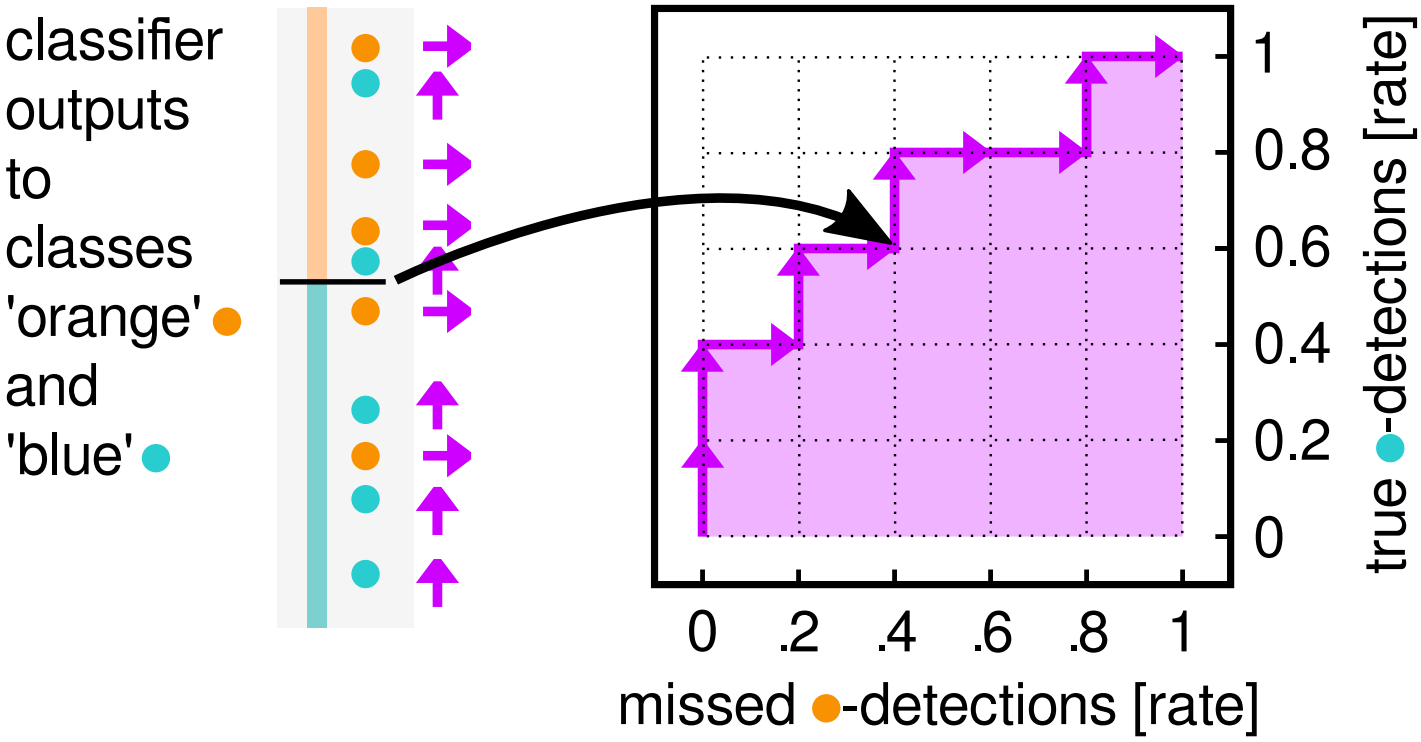
- Segments of the signals are called *epochs* or *single-trials*.
- In BCI epochs are typically strongly overlapping. (Non-target epochs are not shaded in this figure.)

Univariate Features: Averages and Single-Trials



- ERPs can be voluntarily modulated according to an experimental condition, here selective attention to certain *target* stimuli.
- The potential measured 220ms post-stimulus at **Cz** is a one-dimensional observation variable: a *univariate* feature.

Area under the Curve (AUC) as Measure of Seperation



area under the ROC curve:

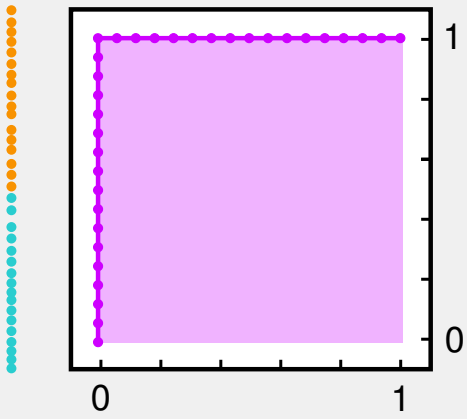
$$\text{AUC} = \frac{18}{25}$$

$= p$ that a random is ranked higher than a random

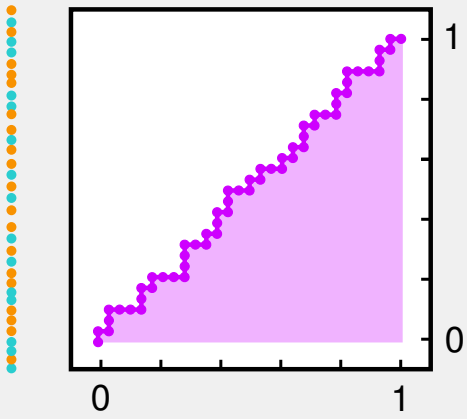
- Area Under the ROC Curve (AUC): **Measure of separation** of two univariate distributions
- Applied to output of a binary classifer: AUC is a bias-independent performance measure.

Examples for ROC Curves and AUC Values

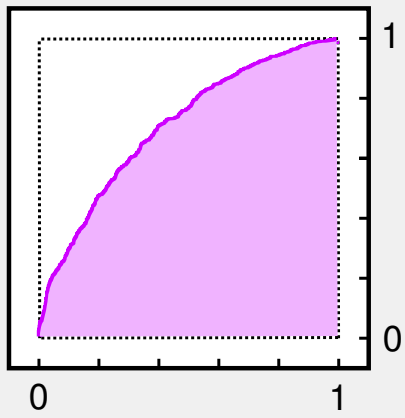
perfectly separated distributions:
AUC = 1 (or = 0)



random distributions:
AUC ≈ 0.5



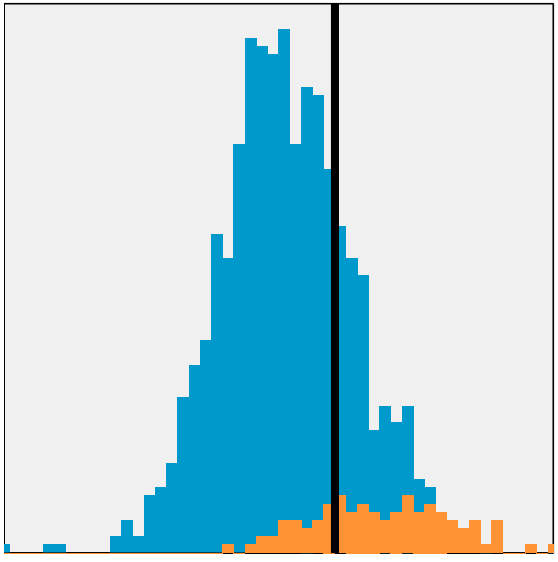
classifier outputs from our example data
AUC ≈ 0.7



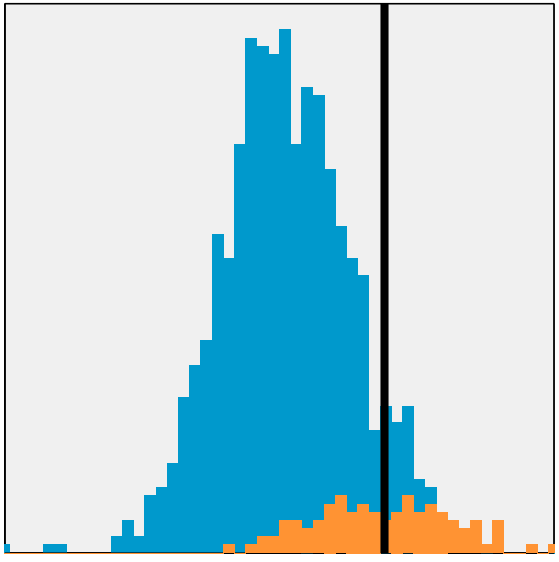
Side Note: Loss Functions for Unbalanced Classes

Blue class: $N_1 = 900$ samples, orange class: $N_2 = 100$ samples.
Weighted error: $\mathrm{err}_{\mathrm{weighted}} = \frac{1}{2}\left(\mathrm{err}|_{\mathrm{class\ 1}} + \mathrm{err}|_{\mathrm{class\ 2}}\right)$

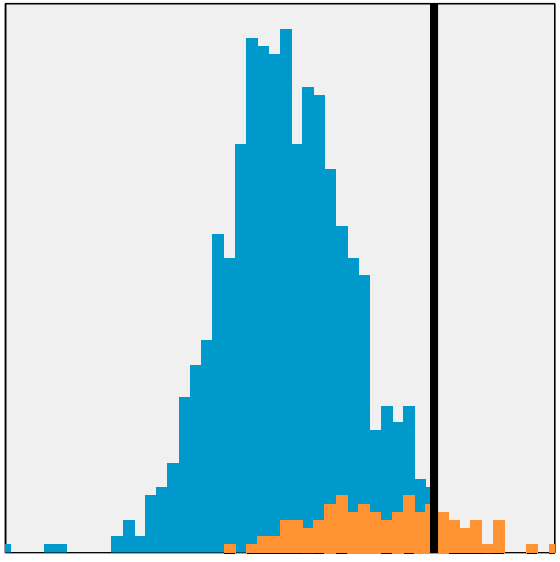
Examples of weighted and unweighted error – bias of classifier is varied:



Error rate
Unweighted: 23.6%
Weighted: 25.1%
AUC-based: 16.6%



Error rate
Unweighted: 12.8%
Weighted: 30.0%
AUC-based: 16.6%



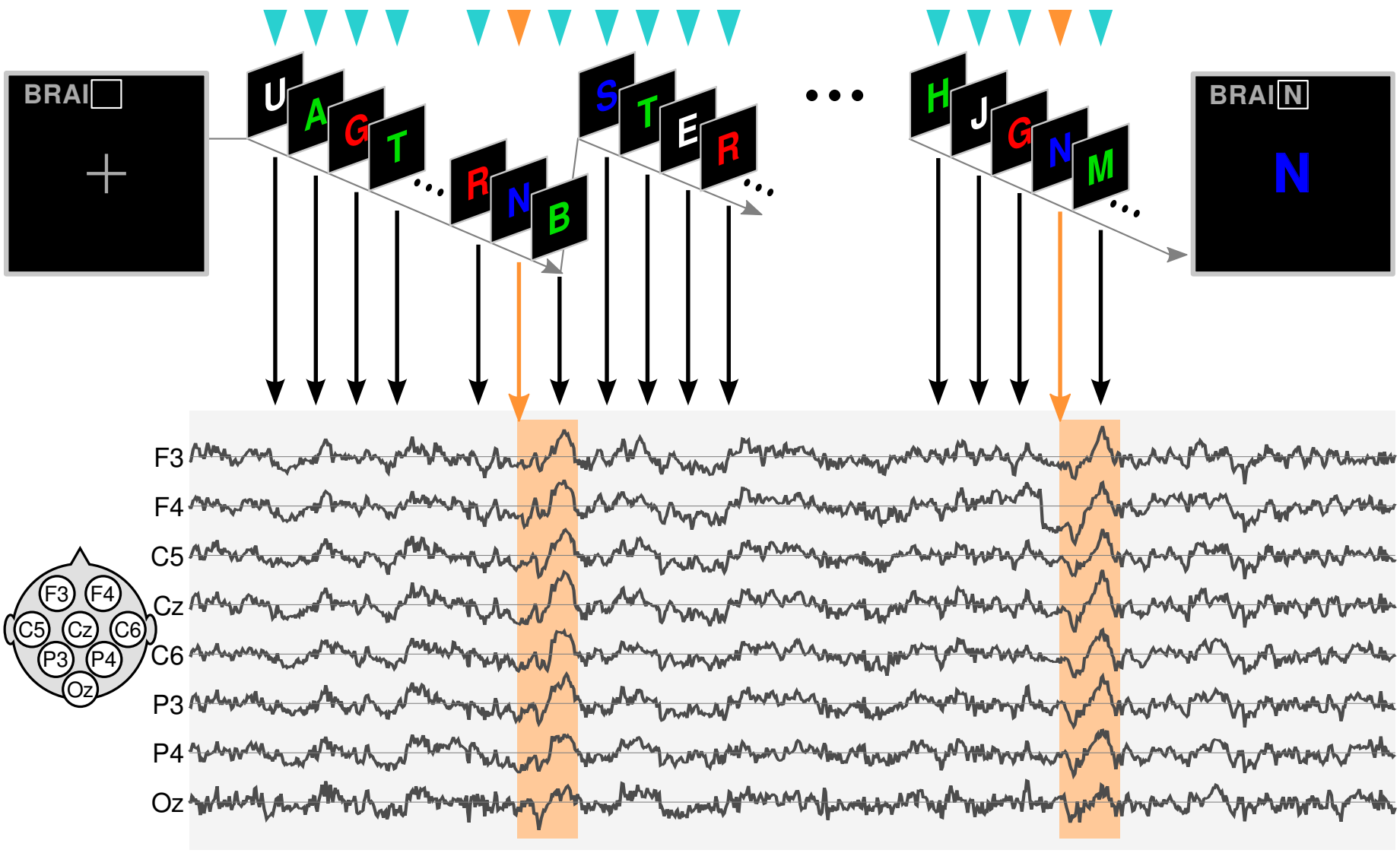
Error rate
Unweighted: 9.5%
Weighted: 39.5%
AUC-based: 16.6%

From Uni- to Multivariate Features

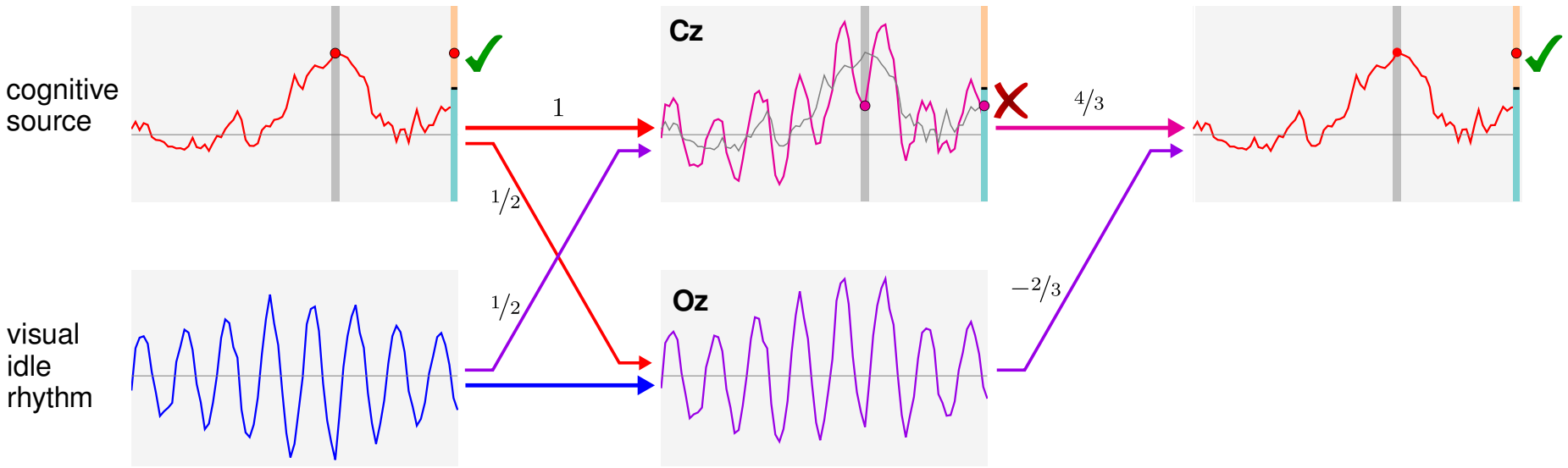
For improved classification of EEG single-trials, we need to accumulate more information in the features.

- sample ERP signals at *multiple* time points/intervals
  $\rightarrow$ *temporal feature*
- join signals from *multiple* channels
  $\rightarrow$ *spatial feature*
- do both things
  $\rightarrow$ *spatio-temporal feature*

Multi-Channel Epochs

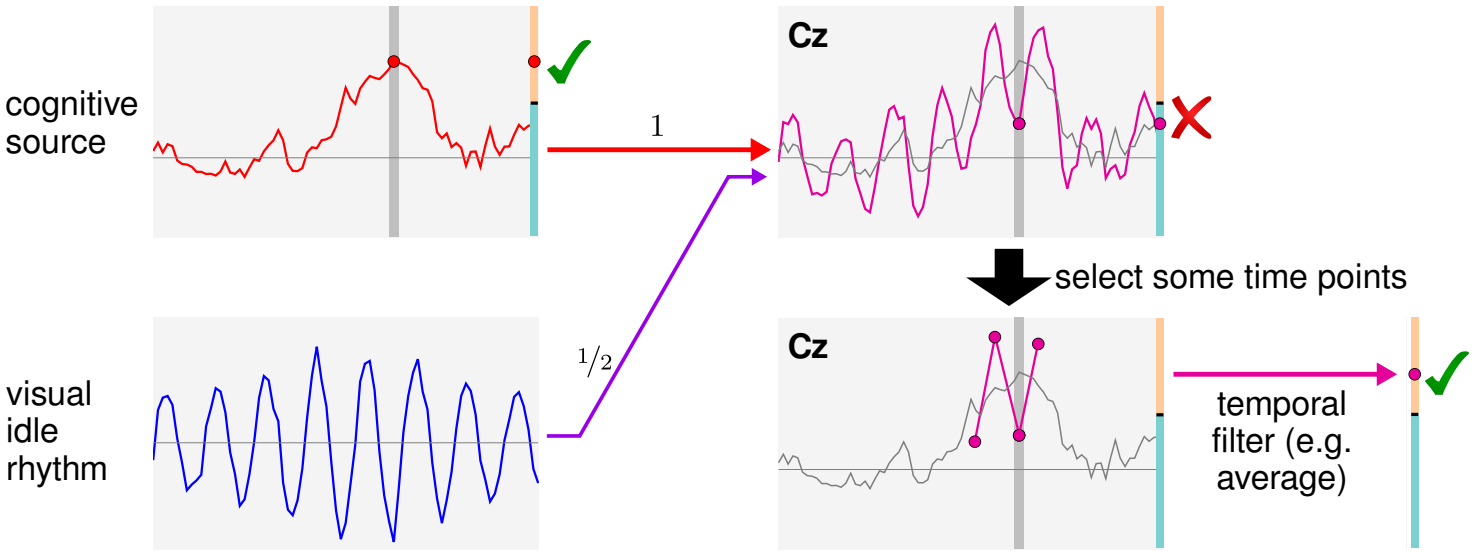


The Virtue of Multivariate Spatial Features

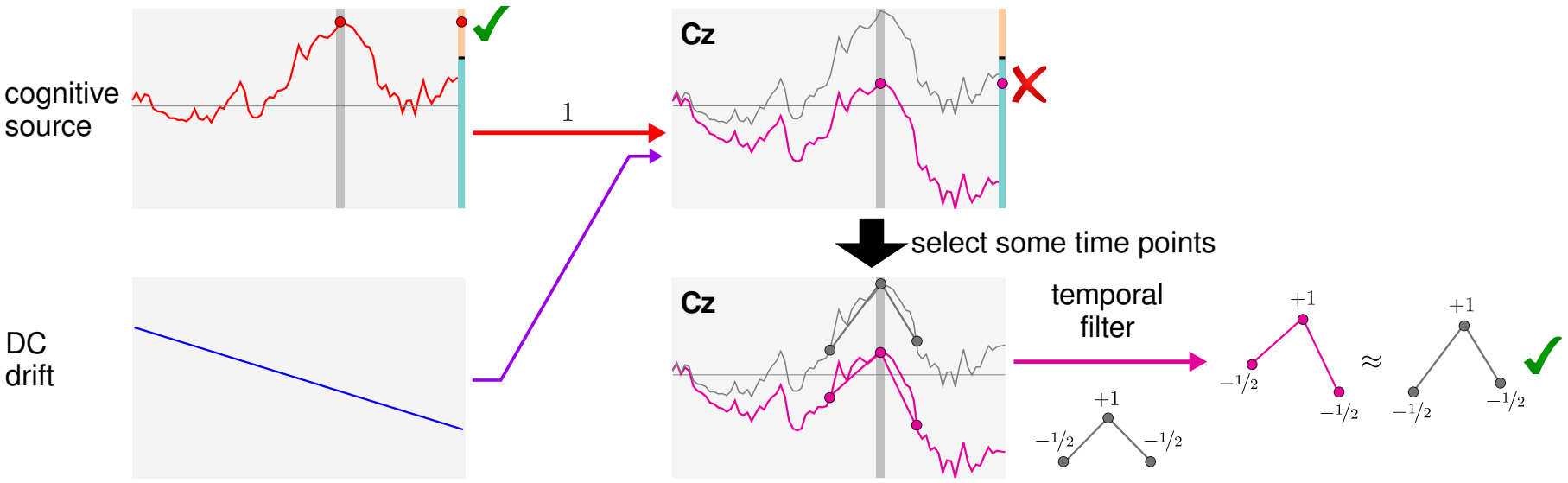


- Here, $\mathbf{w} = [4/3 \quad -2/3]^\top$ is a simple spatial filter.

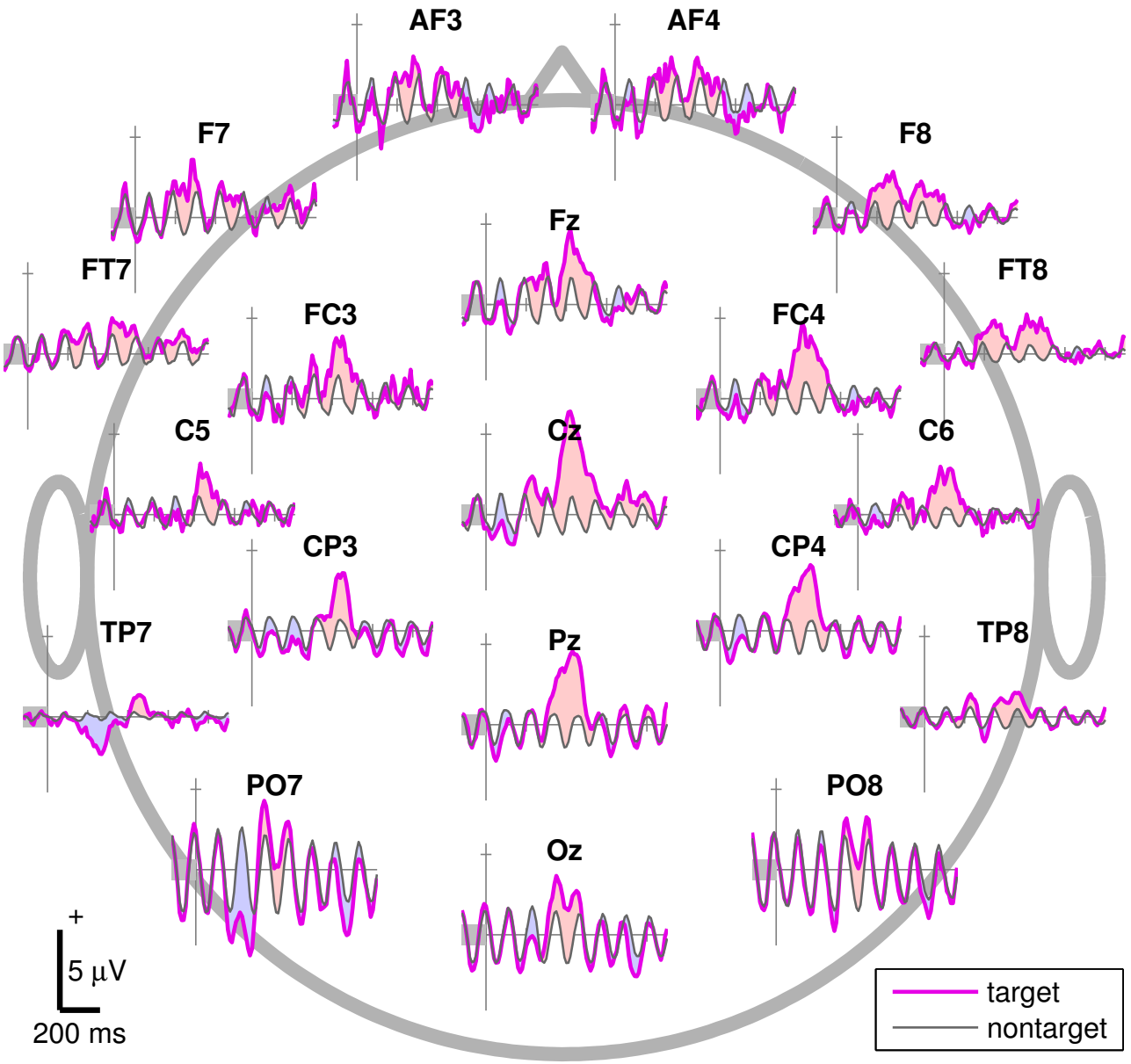
The Virtue of Multivariate Temporal Features



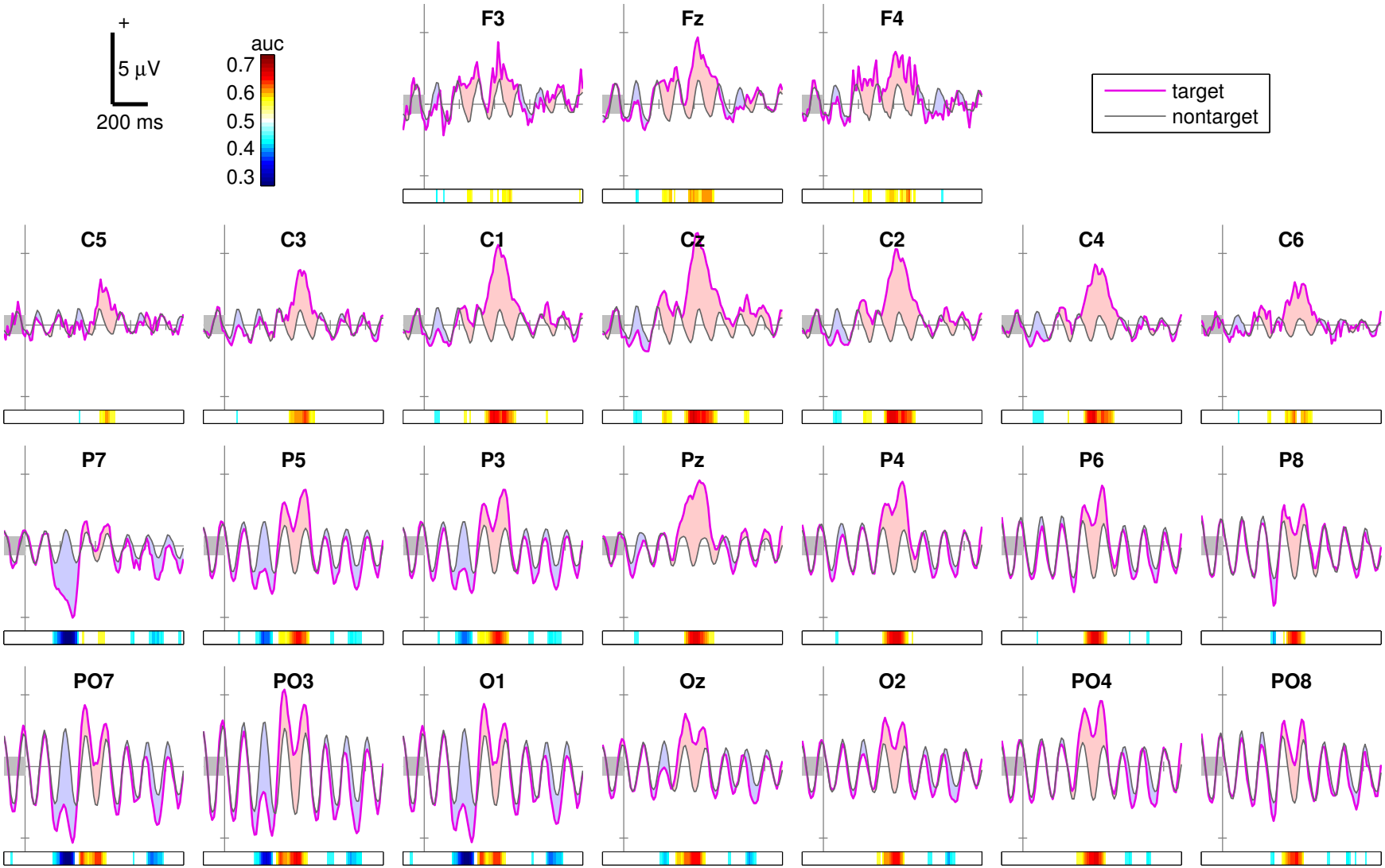
The Virtue of Multivariate Temporal Features



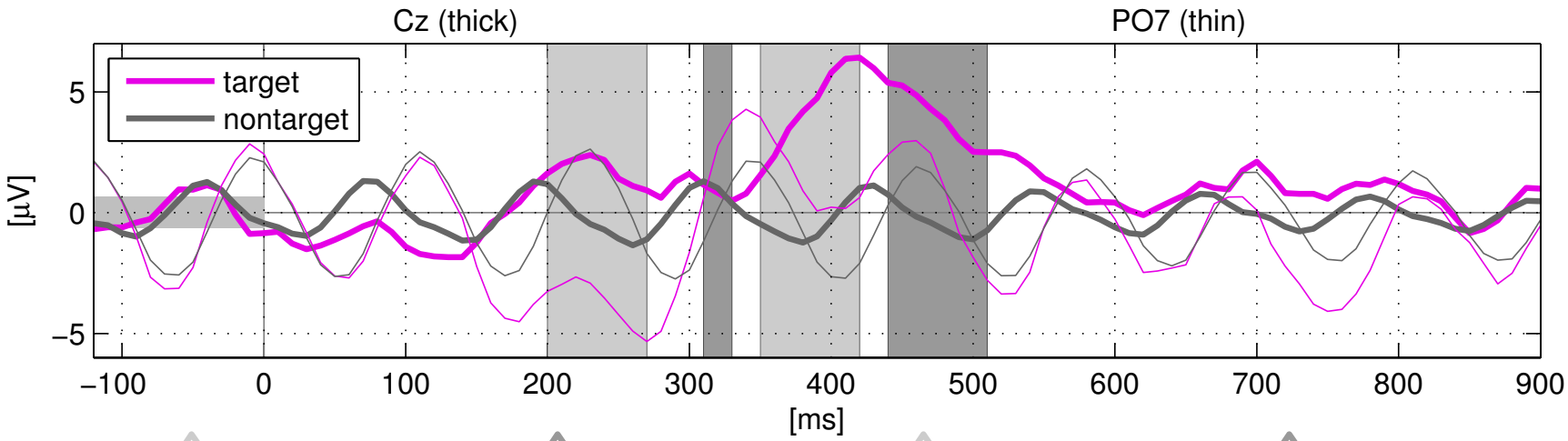
ERPs in a Head Plot

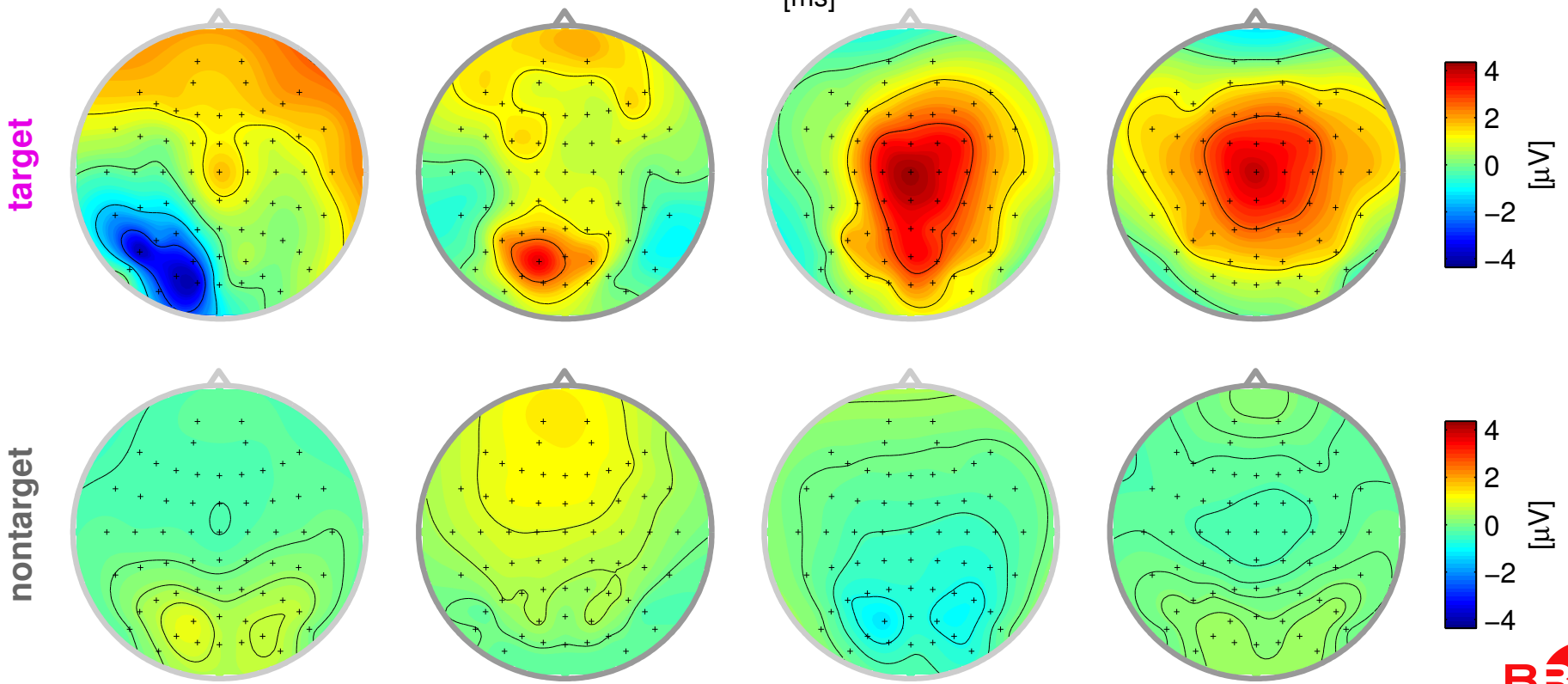


ERPs in a Grid Plot

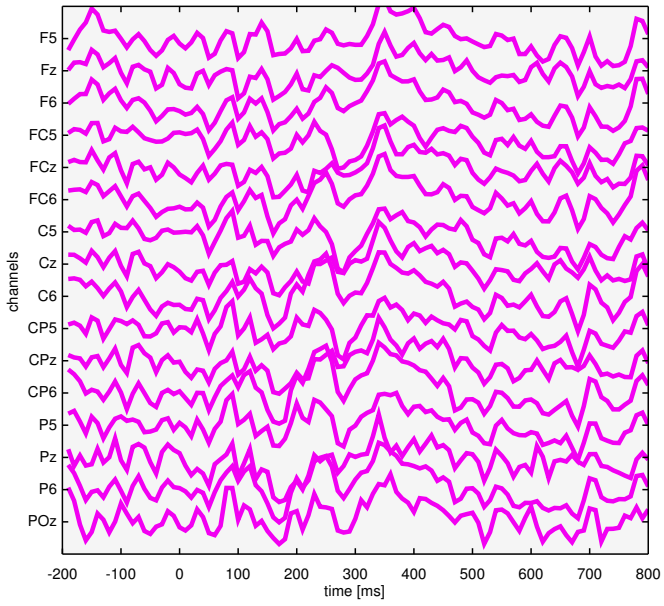


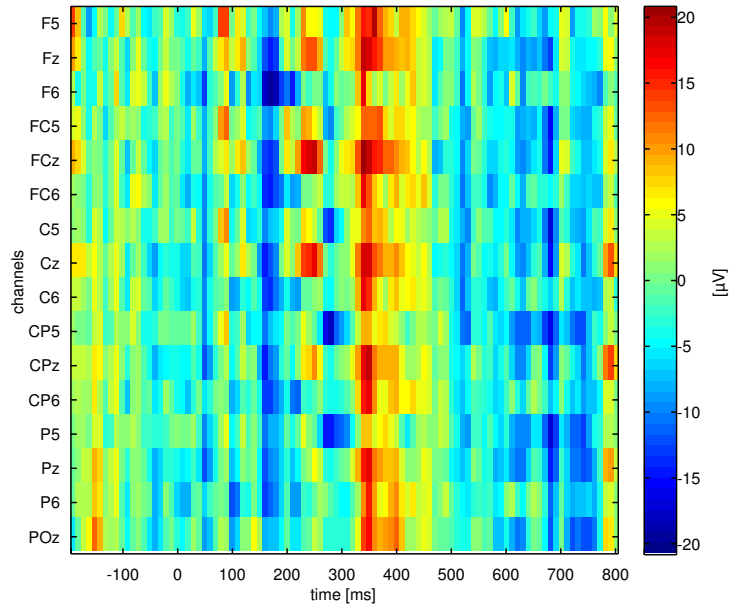
ERP Topographies

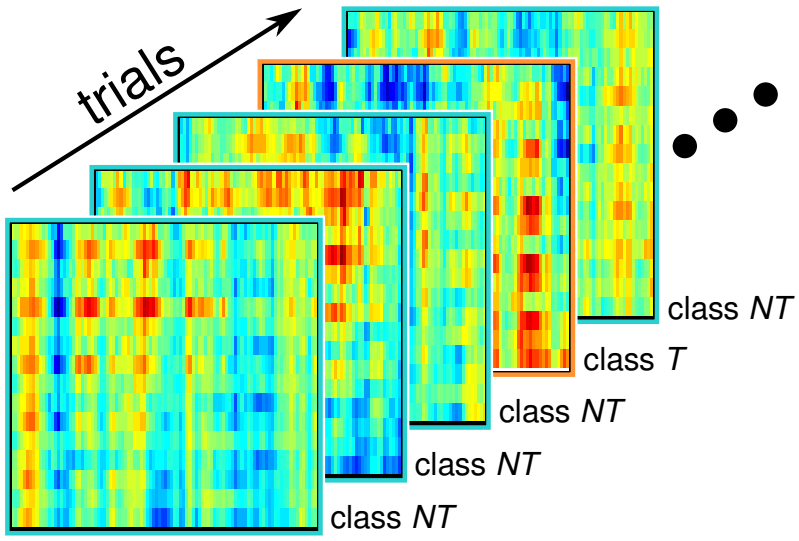




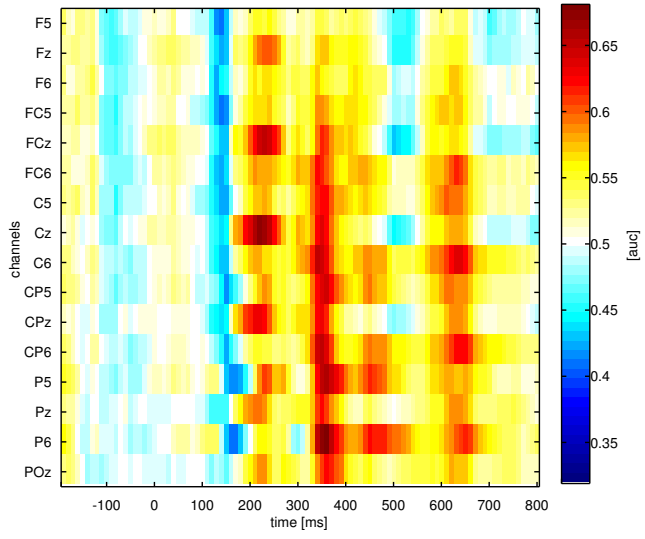
Interlude: Representation as Matrix







AUC across trials

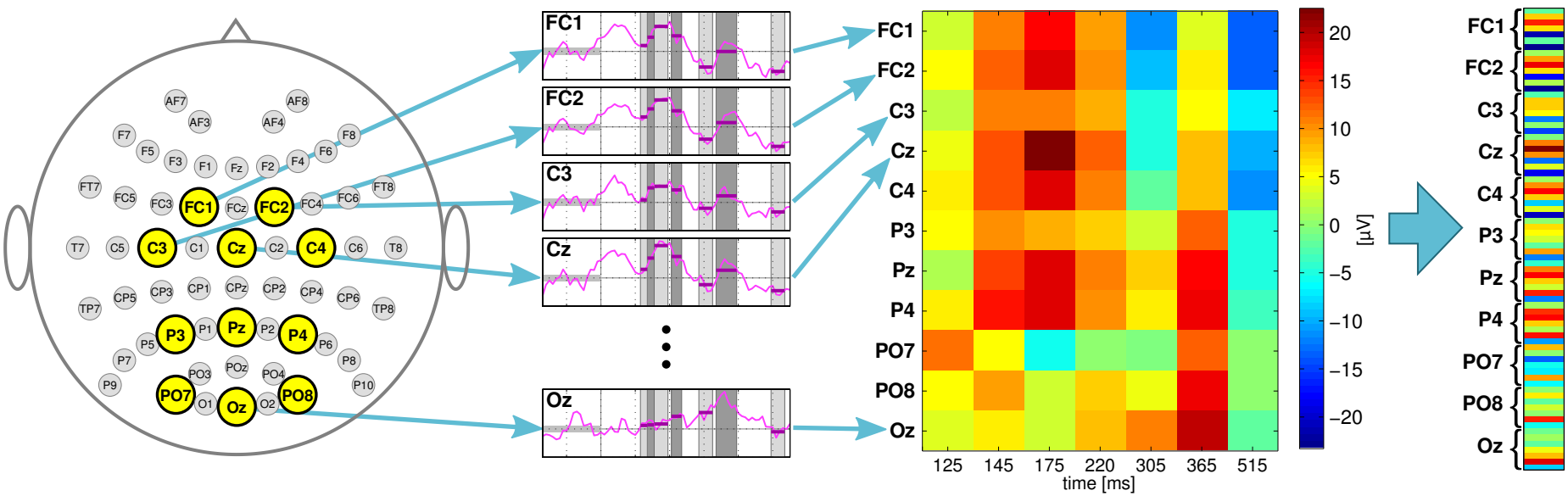


Extraction of Spatio-Temporal ERP Features

Given a set of channels $\mathcal{C}$ and time intervals $\langle \mathcal{T}_i \rangle_{i=1,\ldots,I}$, the *spatio-temporal* feature are defined as

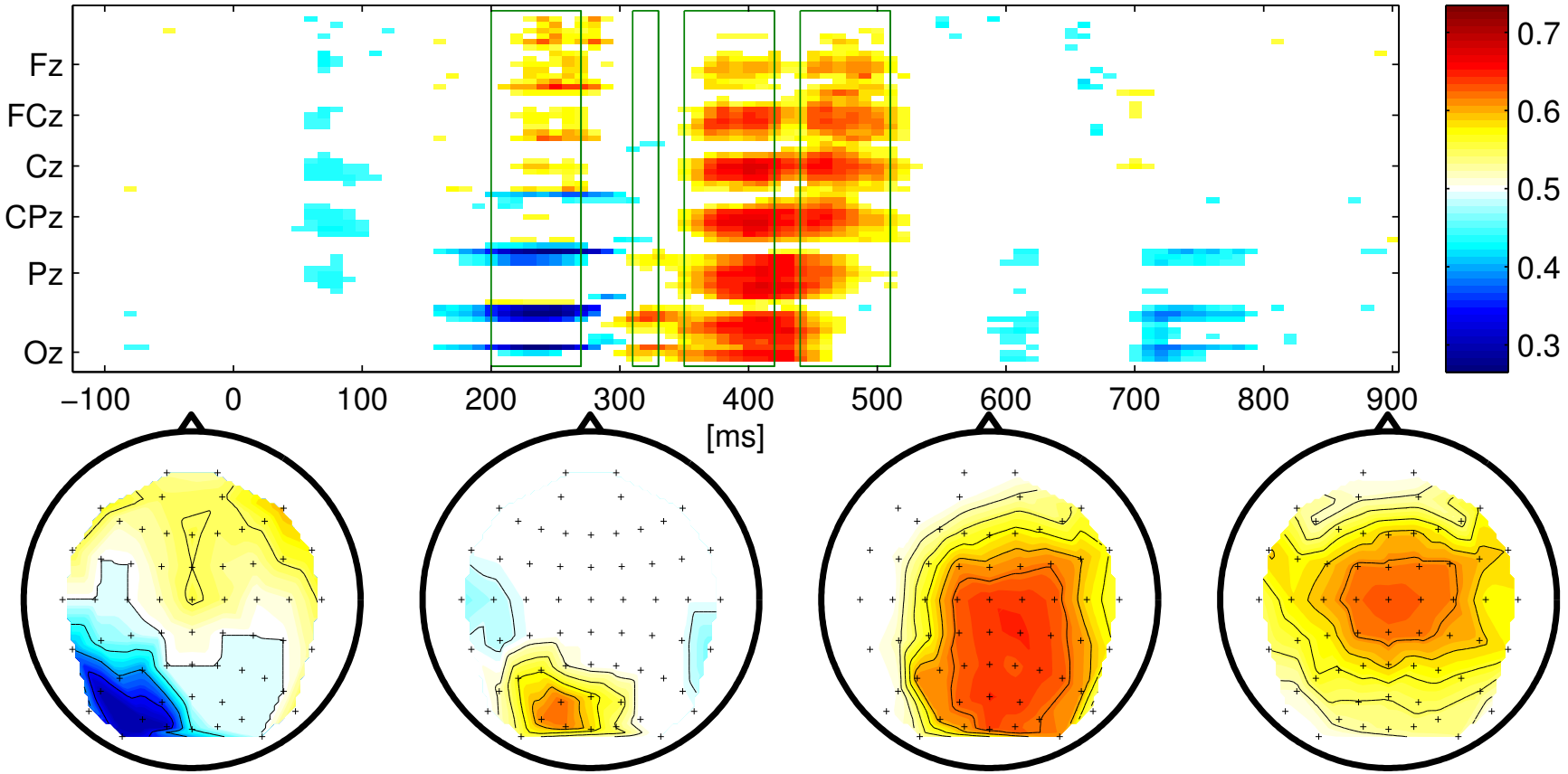
$$\mathbf{X}(\mathcal{C}, \mathcal{T}) = [\text{mean}\, \langle \mathbf{x}_{\mathcal{C}}(t) \rangle_{t \in \mathcal{T}_1}; \ldots; \text{mean}\, \langle \mathbf{x}_{\mathcal{C}}(t) \rangle_{t \in \mathcal{T}_I}].$$

Potentials are averaged within the time intervals, and then the averaged values are concatenated for all intervals and channels.



Dimensionality of features: # channels $\times$ # intervals

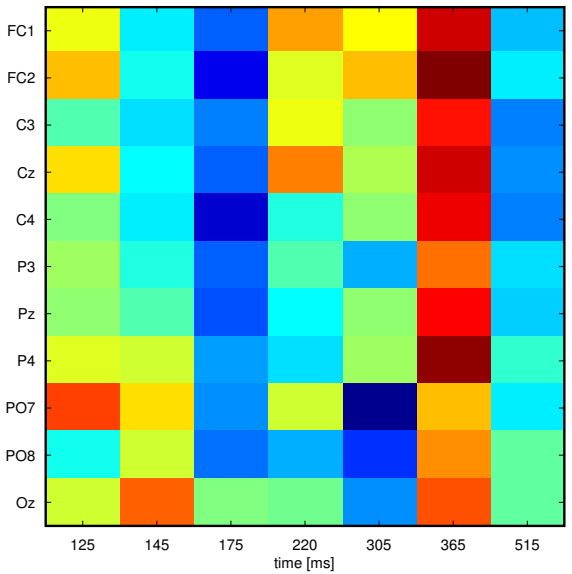
AUC Matrix: Selection of Channels and Time Intervals



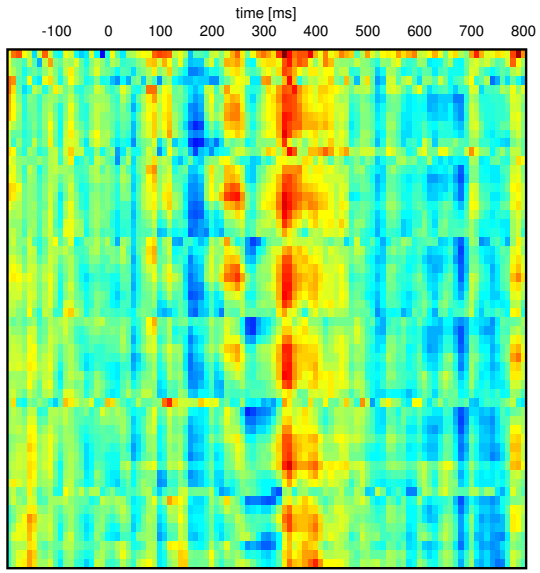
- Each cell in the matrix is one uni-variate feature.
- Multi-variate features: typically full set of **channels**
- **Time intervals** chosen by a heuristic based on AUC values.

Multivariate ERP Features

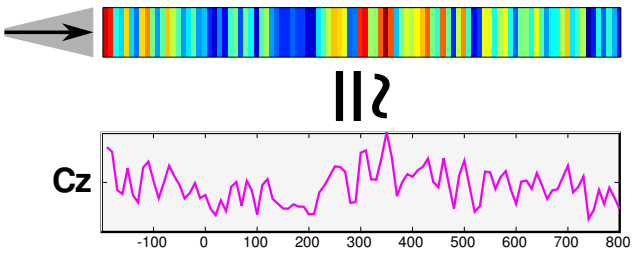
spatio-temporal ERP feature



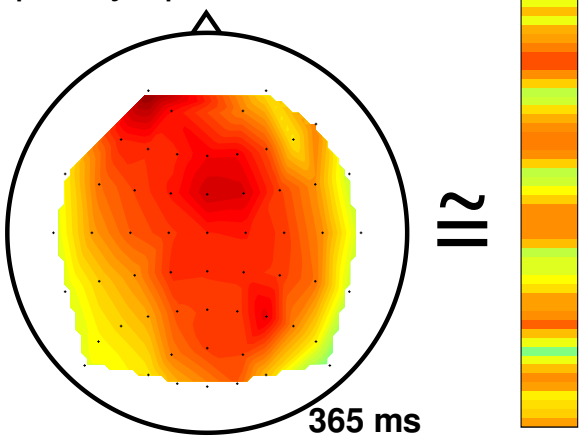
single-trial data matrix



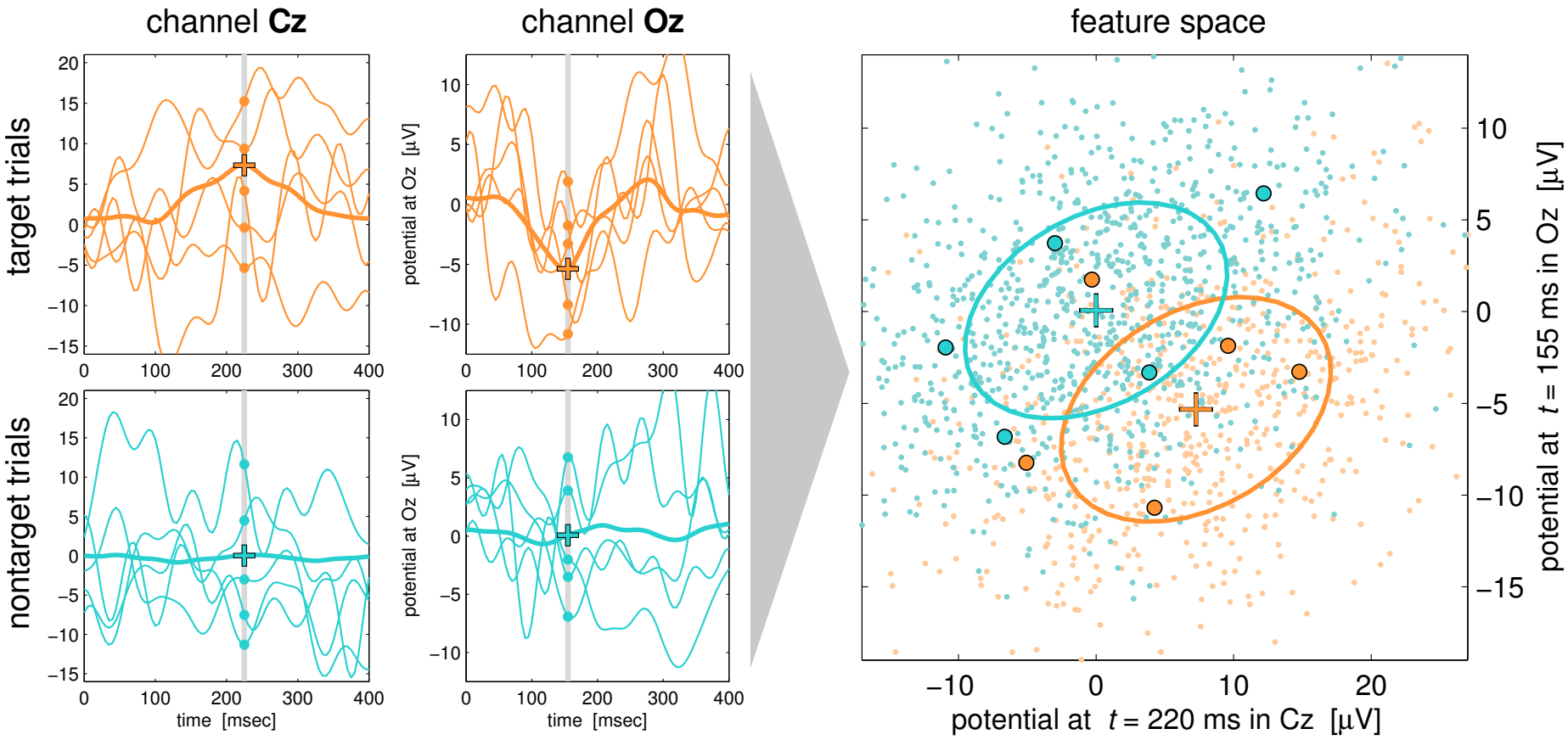
select one channel:
purely temporal feature



select one time point:
purely spatial feature



Summary: Representation of Multivariate Distributions



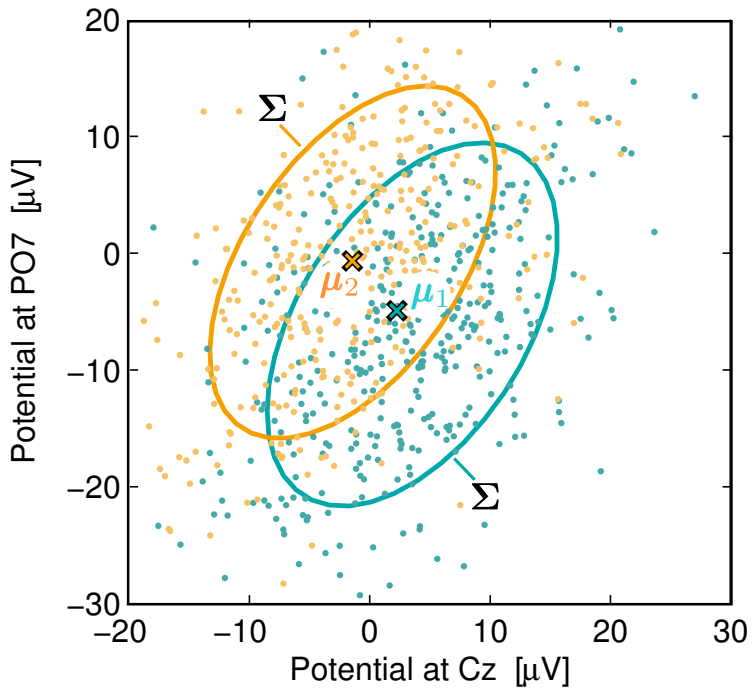
Distribution of ERP Features

For classification, we have to consider the distribution of those features. According to our model (ERPs are constant across trials):

$$\mathbf{x}^{(k)}(t) = \mathbf{p}_1(t) + \mathbf{m}^{(k)}(t) \qquad \text{for trials } k \text{ of condition 1}$$

$$\mathbf{x}^{(k)}(t) = \mathbf{p}_2(t) + \mathbf{m}^{(k)}(t) \qquad \text{for trials } k \text{ of condition 2}$$

with Gaussian noise: $\mathbf{m}^{(\cdot)}(t) \sim \mathcal{N}(0, \mathbf{\Sigma})$. Empirically, the Gaussian assumption seems justified:



For features of ERP data:

- μ_1: ERP of condition 1
- μ_2: ERP of condition 2
- $\mathbf{\Sigma}$: noise: non-phase-locked activity (independent of condition)

[Blankertz et al, NeuroImage 2011]

Interlude: Matrix Speller (Classical ERP-based Typewriter)

Classical example: *Matrix Speller*

- User concentrates on a symbol
- Rows and columns are intensified randomly
- Target rows and columns elicit specific ERPs (oddball)
- BCI detects target ERPs (averaged over few repetitions)

➤ Effective communication

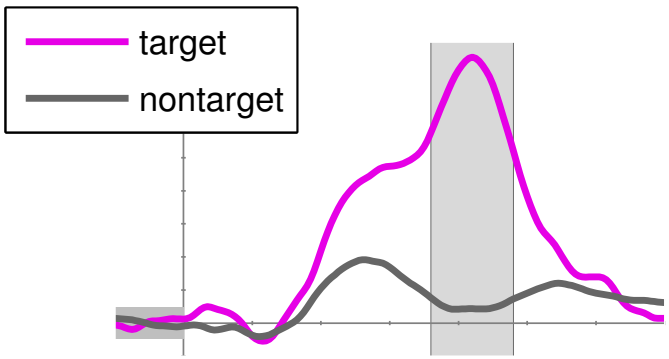
[Farwell & Donchin, 1988]

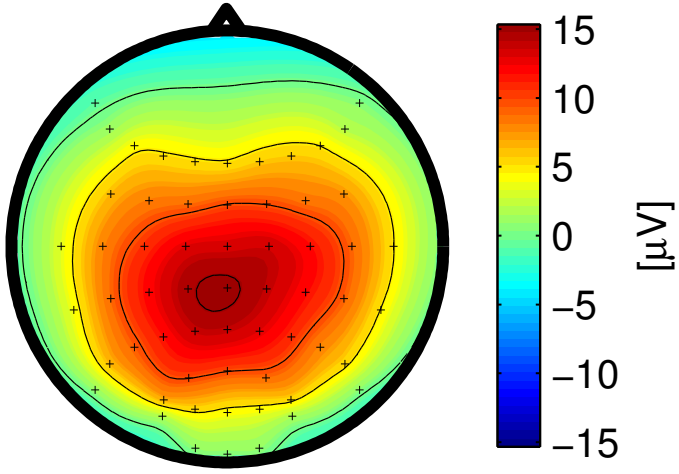
Interlude: Matrix Speller (Classical ERP-based Typewriter)

Classical example: *Matrix Speller*

- User concentrates on a symbol
- Rows and columns are intensified randomly
- Target rows and columns elicit specific ERPs (oddball)
- BCI detects target ERPs (averaged over few repetitions)
- ➤ Effective communication

[Farwell & Donchin, 1988]





Critical Questions

- Do the positive results of the Matrix Speller transfer to the target patient group?
- More specifically, how much does the performance depend on fixating the target symbol by gaze?

The latter question was addressed with a specific study (see subsequent slides. But getting aware of such issues is already possible when the original data are investigated thoroughly.

This illustrates that, it is important to be aware of where discriminative information in a BCI originates from, and in particular,

- ML methods should not be uesd as black box

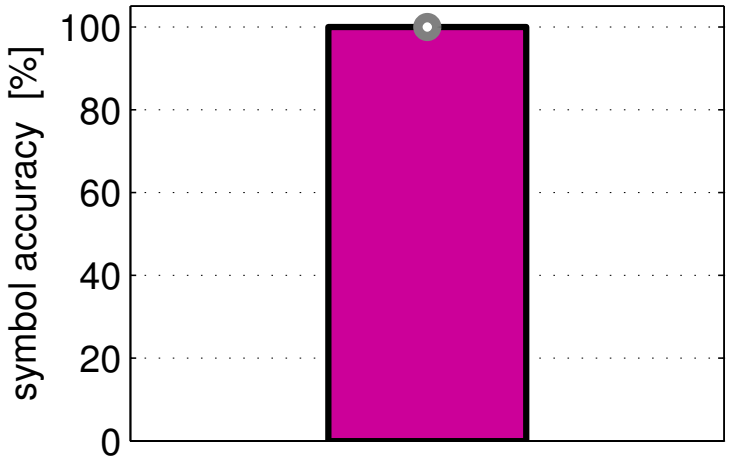
(i.e., just doing classification and looking at the results).

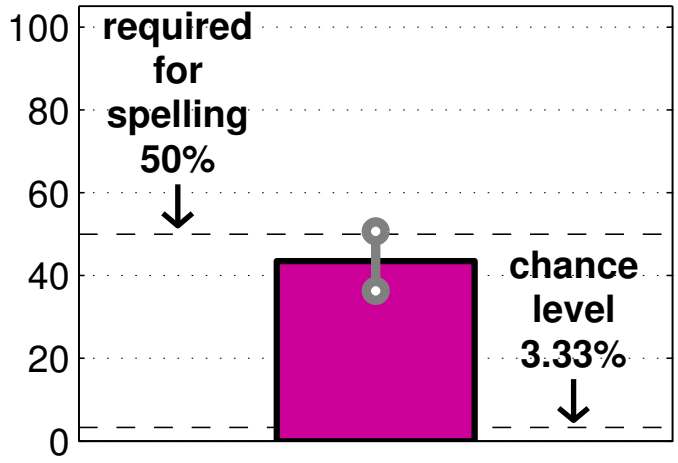
The Role of Gaze Control in the Matrix Speller

Matrix Speller - Overt

Matrix Speller - Covert

To select **B**: ○ = focus of gaze; ○ = focus of attention



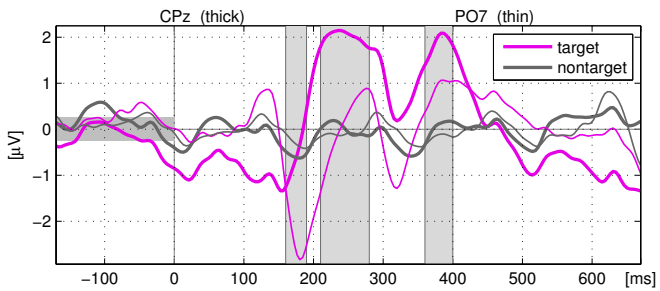


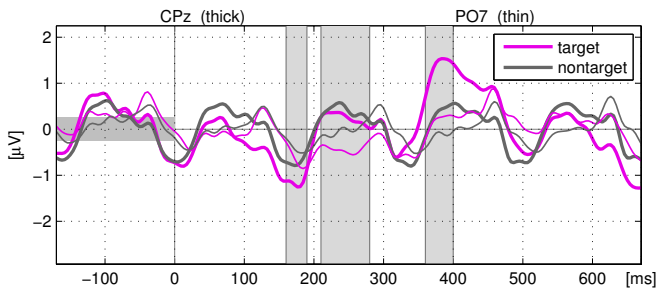
The Role of Gaze Control in the Matrix Speller

Matrix Speller - Overt

Matrix Speller - Covert

To select **B**: ○ = focus of gaze; ○ = focus of attention

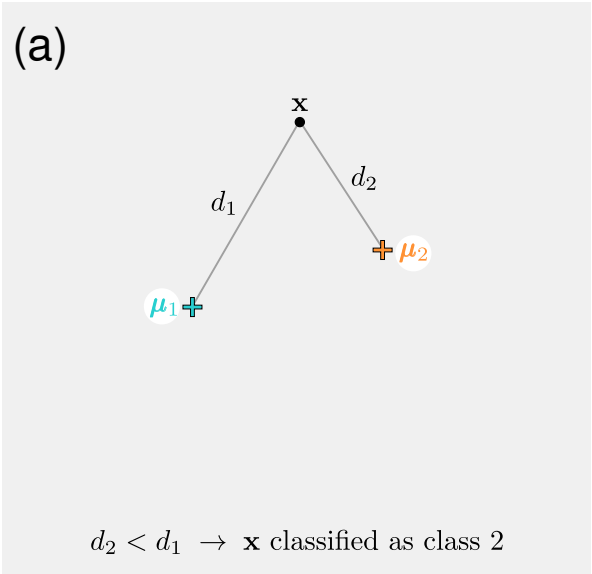


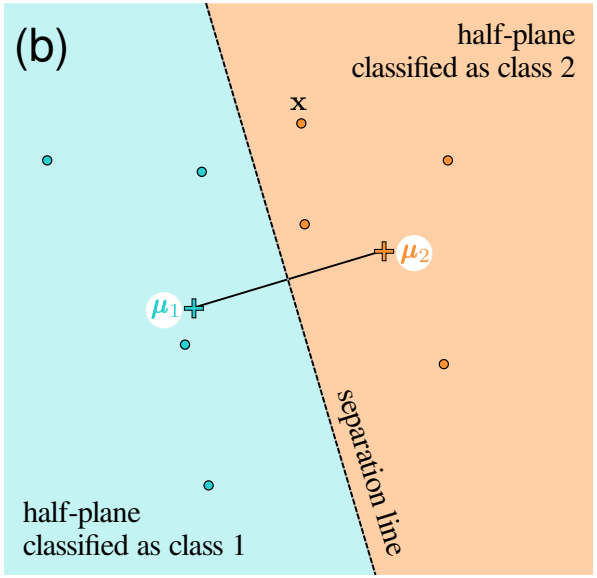


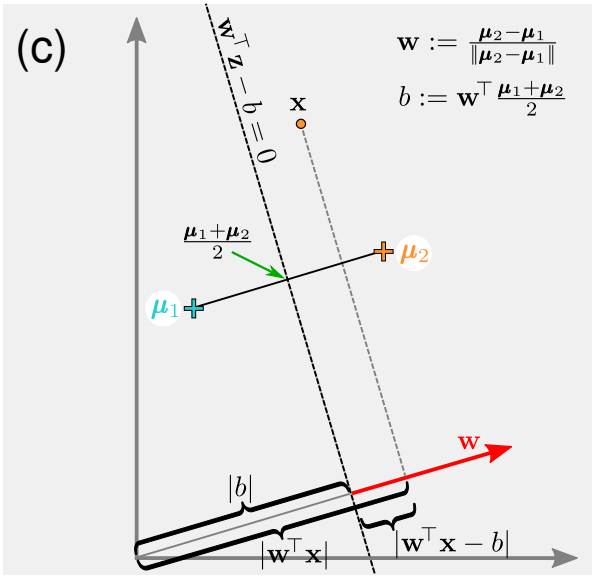
Part II: Classification of ERP Features

Nearest Centroid Classifier (NCC)

(a) Let us assume a simple setting of a classification problem with little information: Only the means (or centroids) $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of the two distributions are known.

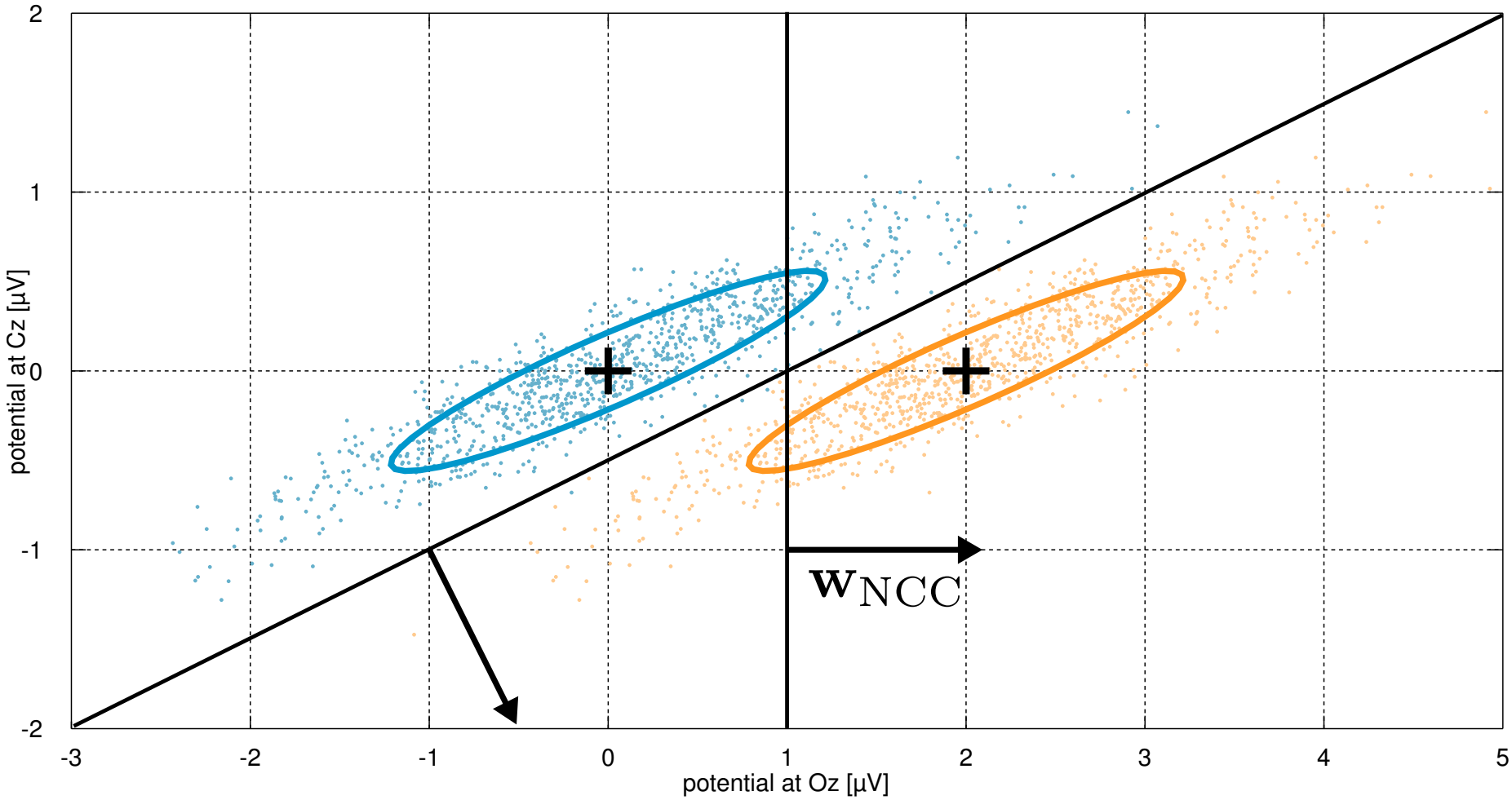






(b) This leads to a linear separation of the space with the separation line (or hyperplane in higher dimensions) intersecting perpendicularly the line connecting the centroids in the middle. **(c)** Mathematical formalism: Classify according to the sign of $\mathbf{w}^\top \mathbf{x} - b$ with $\mathbf{w} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

Can We Expect NCC to Perform Well for ERP Features?



Linear Discriminant Analysis (LDA)

Using probability theory, one can derive from the following three assumptions the optimal classifier for the given class distributions. *Optimality means that the classifier has the minimum risk of misclassification for new samples that are drawn from these class distributions.*

1. Features of each class are Gaussian distributed.
2. Gaussians of all classes have the same covariance matrix.
3. True class distributions are known.

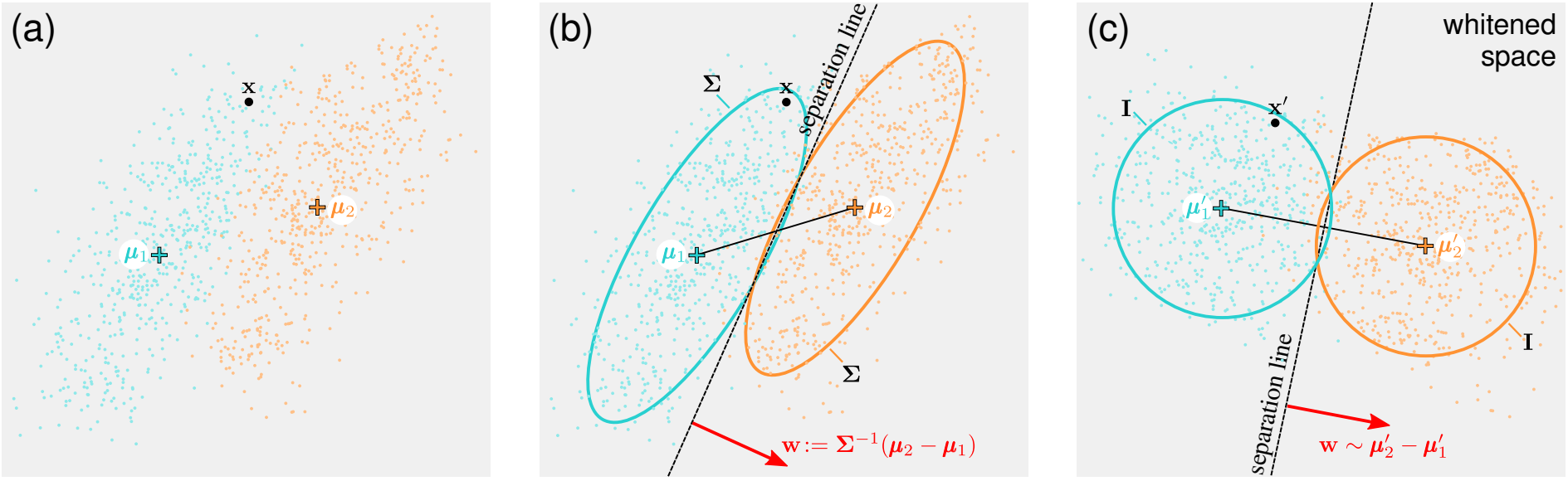
This optimal classifier is called *Linear Discriminant Analysis* (LDA) and it can be formalized in the following way: Given two Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, LDA is defined by the normal vector

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \quad \text{and bias} \quad b = \mathbf{w}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2. \tag{1}$$

First we will look at how the LDA classification looks like and later discuss the assumptions.

Linear Disciminant Analysis

(a) Means as in the NCC example, but specific distributions are shown.



(b) In Linear Discriminant Analysis, a common covariance matrix for both classes is estimated, which describes the (class-independent) noise. Note, that $\mathbf{x}$ is classified here differently with LDA than with NCC.

(c) Correspondence to NCC via whitening.

Knowing the noise (Σ) improves classification!

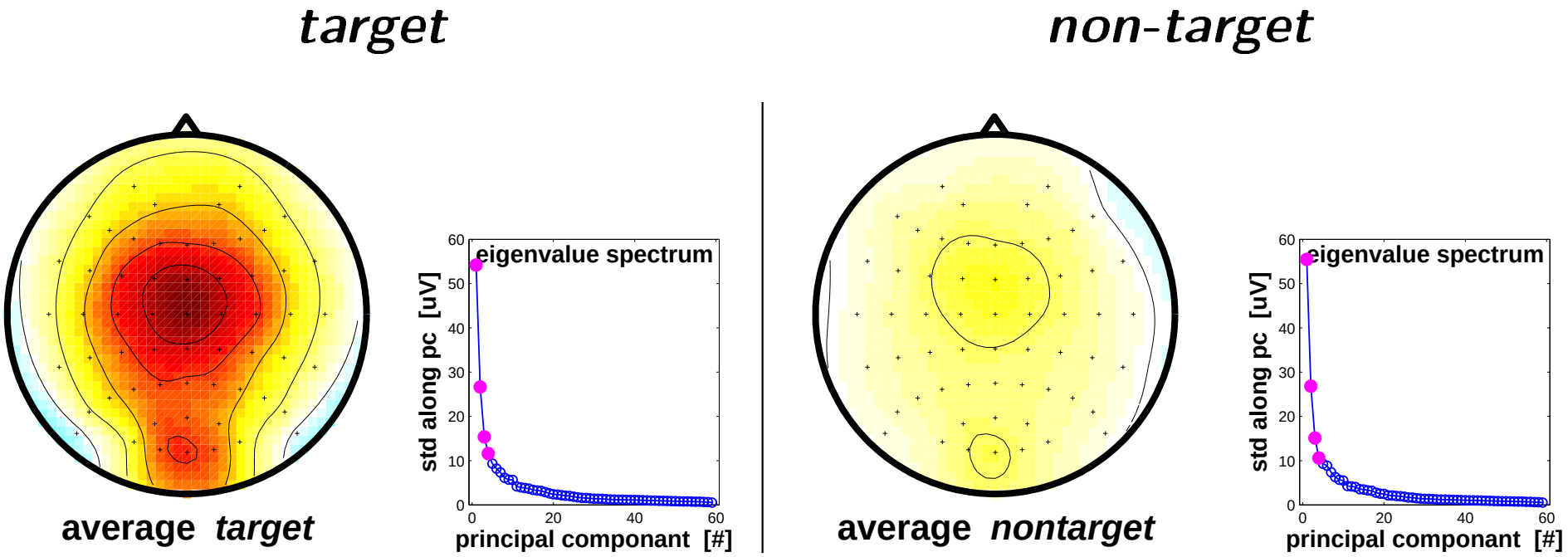
Linear Discriminant Analysis – Assumptions for Optimality

Now, we come back to the assumptions which are required to warrant optimality of the LDA classifier:

1. Features of each class are Gaussian distributed.
2. Gaussians of all classes have the same covariance matrix.
3. True class distributions are known.

We will verify the first two assumptions empirically by investigating one exemplary dataset. The last assumption will be discussed later in this lecture.

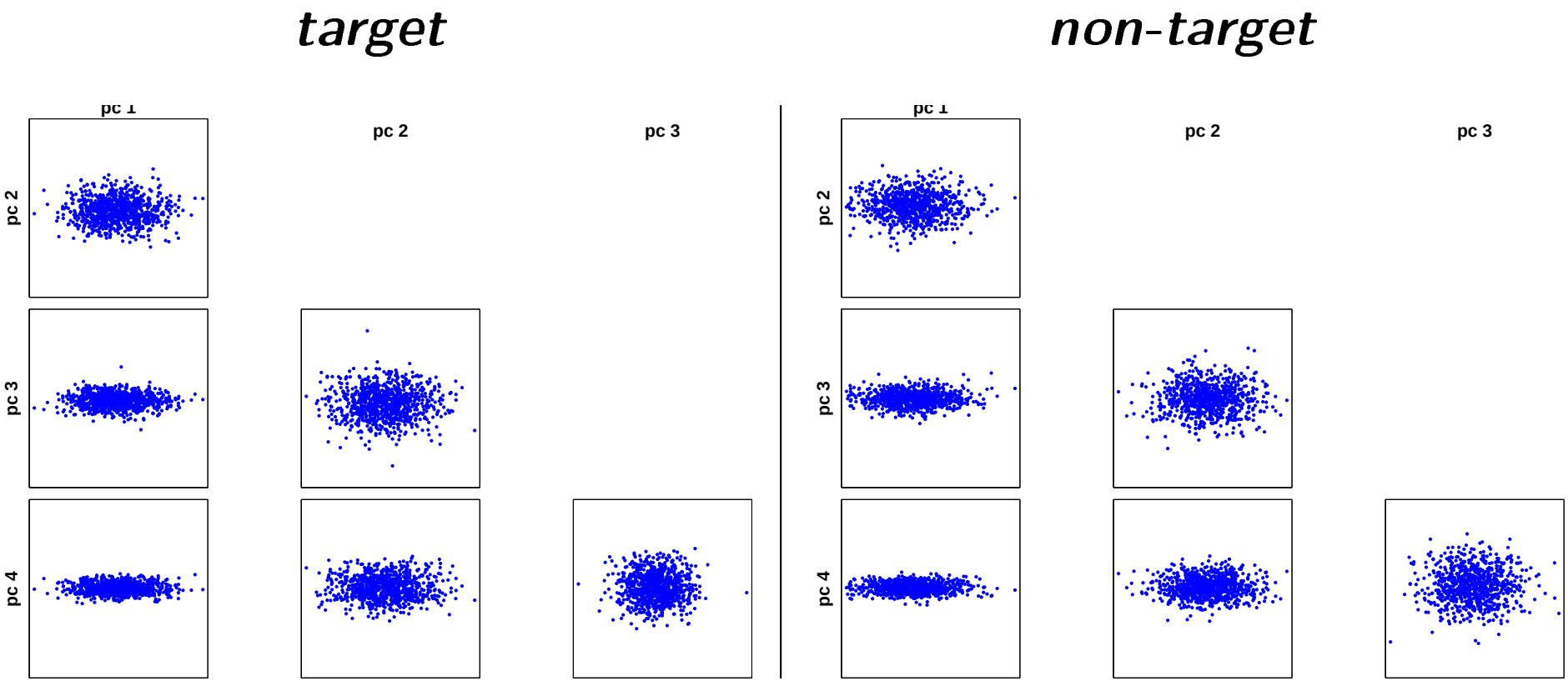
Mean and Eigenvalue Spectrum for a P300 Data Set



➤ In the following, we will look at the PCs (Eigenvectors) that correspond to the four largest Eigenvalues.

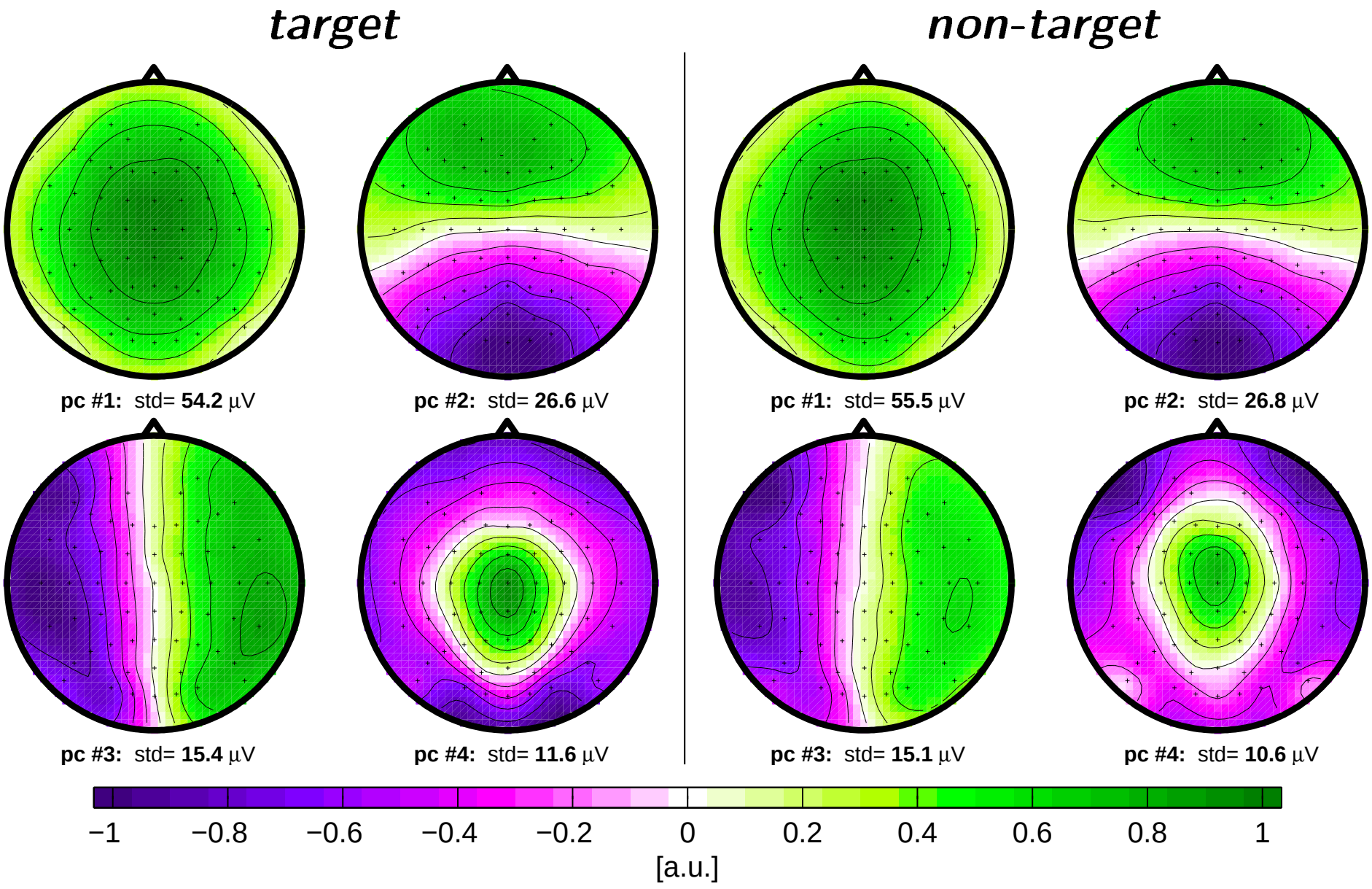
Distribution of the Noise

Scatter plots of projections on PCs (wrt. 4 largest Eigenvalues):



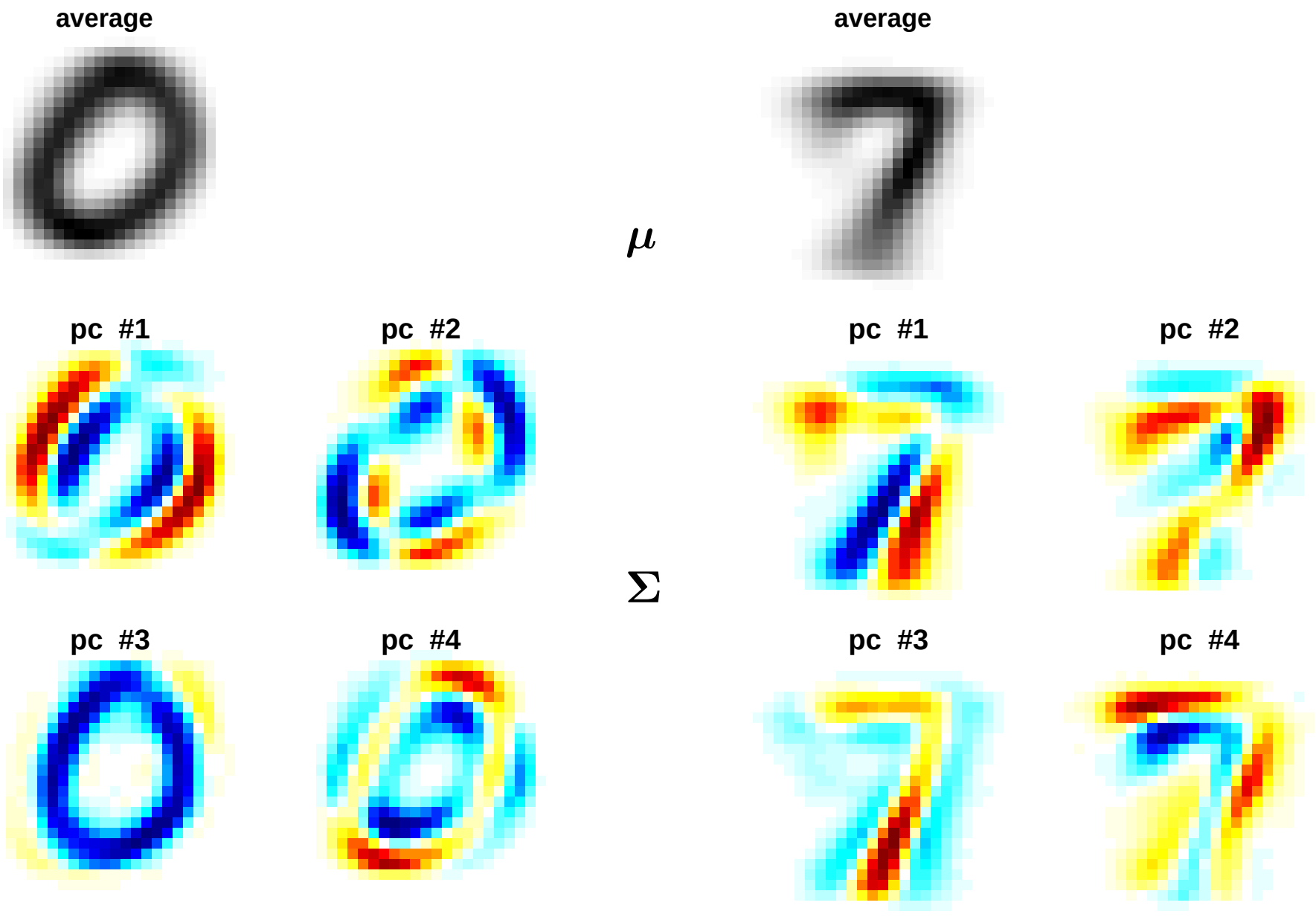
➤ These projections look very much Gaussian.

The Structure of the Noise



➤ Covariances of both classes look very similar.

For Comparison: Covariances in Handwritten Digits



➤ Here, covariances of both classes **do not** look similar.

Validation of Classification Procedures

To validate the performance of a classifier, one needs to have a

- **training set** on which all parameters of the model are estimated (weights of the classifier; selection of features etc.), and a
- **validation set** on which the performance is calculated.

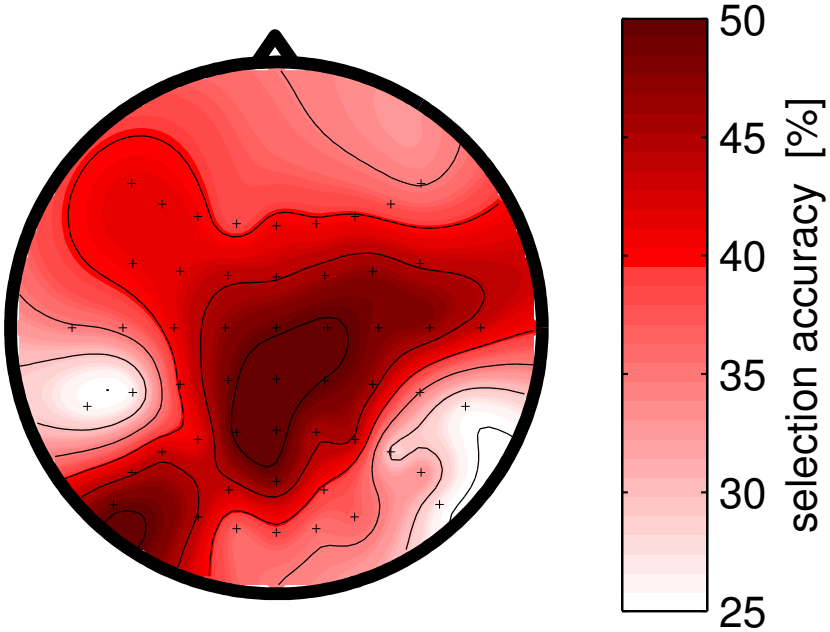
These sets of samples have to be disjoint and **INDEPENDENT**.

To that end, one can use a fixed training and validation set (e.g., first half / second half) or cross-validation.

See [Lemm et al, NeuroImage 2011] for details on validation.

Application of (Purely) Temporal Features

Single channel data does (in most cases) not contain sufficient information for a competitive classification. An application of *temporal features* can be used to investigate the spatial distribution of discriminative information:

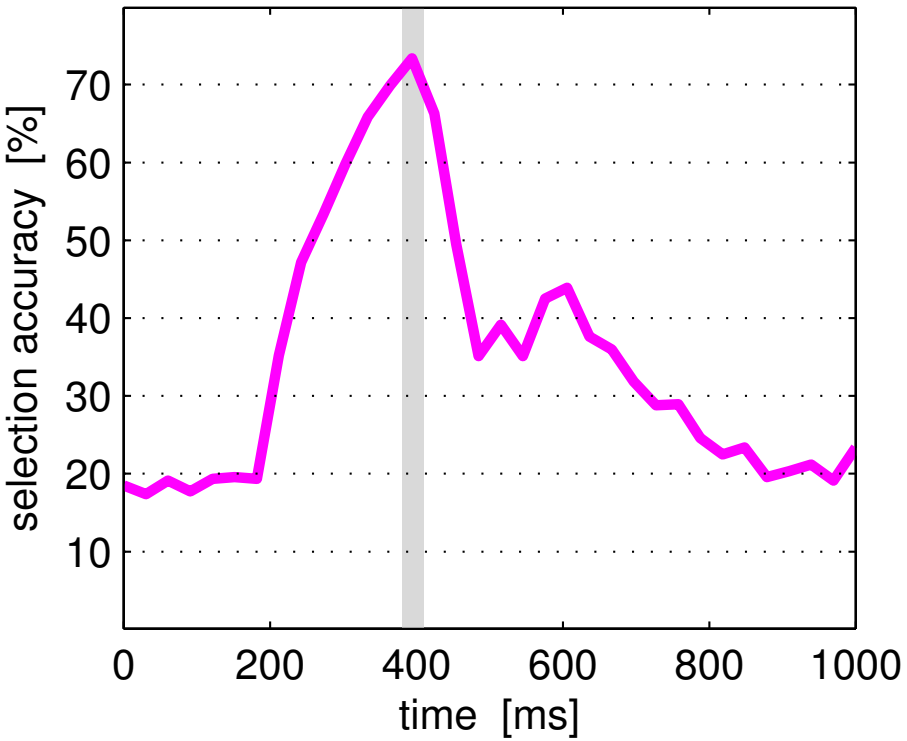


For each single channel the classification performance is determined for temporal features with LDA by cross validation. The resulting error values can be visualized as scalp topography.

Here, two foci are discernible, probably related to visual and cognitive areas.

Application of (Purely) Spatial Features

Spatial features can be used to investigate the distribution along time of discriminative information:

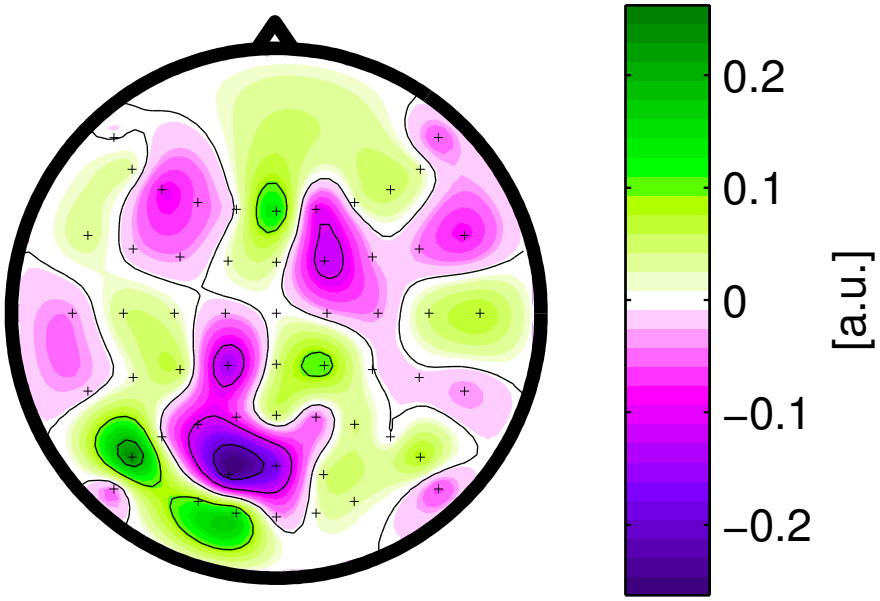


The classification error of spatial features was determined for each time interval of 30 ms duration, shifted from 0 to 1000 ms. (Here, chance level was 16.6%).

In some settings, classification of spatial feature may already yield powerful classification, given an appropriate selection of the time interval.

Application of (Purely) Spatial Features

Spatial features can be used to investigate the distribution along time of discriminative information:

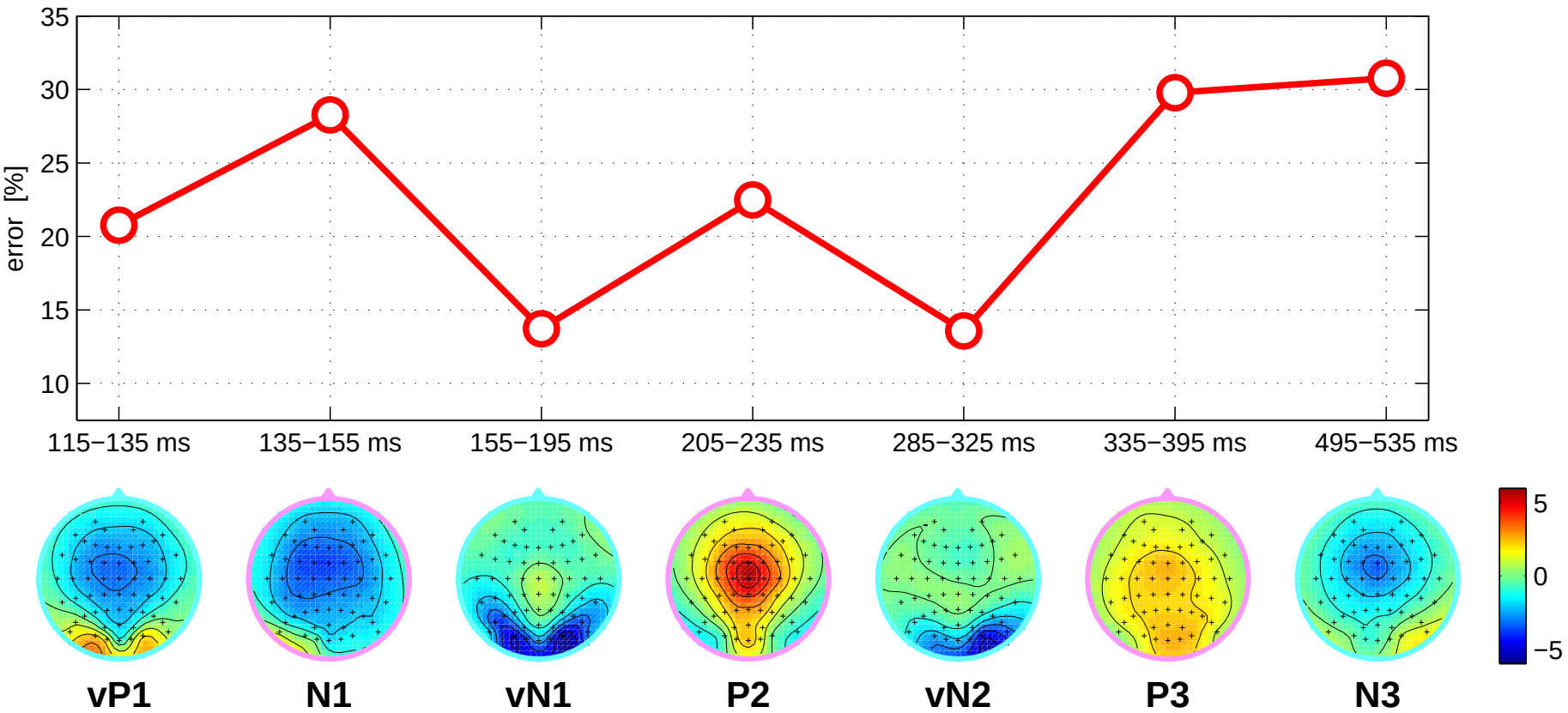


LDA trained on spatial features extracted from the time interval 380–410 ms. The resulting weight vector can be visualized as a topography and can be regarded as a spatial filter.

For the interpretation of spatial filters, you have to act with caution. This issue will be discussed in a later part of the lecture.

Results of Classifying Spatial Features

Classifying on various spatial features results in error rates between 14% and 31% in this example data set (visual speller):



Classification of Spatio-Temporal Features

Advancing from temporal or spatial features to *spatio-temporal* features means increasing the information.

Accordingly, a better classification performance is to be expected.

But in our example data set, the classification error **increases** from

- **14%** for the spatial feature at the best interval to
- **25%** for spatio-temporal features

when classifying with LDA.

Overfitting

Using a larger (more complex) function class allows a better fit with the training data. This is the case, e.g., if the dimensionality of the feature space is increased.

However, despite a low training error, the selected function might not describe the regularity in the data well. It may also be *overfitted* to the noise that is present in the particular set of samples that is available as training data.

This *overfitting* becomes apparent in a cross-validation when the error on the training data deviates substantially from the error on the test data.

Overfitting in LDA

When LDA was applied to high-dimensional (spatio-temporal) features, the performance broke down (result worse than on sub-features).

Given the optimality theorem, this should not happen. Or?

So far, we did not discuss the third assumption:
The true distributions are known.

- This assumptions is *always* violated in non-artificial problems.
- Distribution parameters have to be estimated from given data.
- Estimated (empirical) distribution parameters necessarily deviate from the true ones.
- How much this deviation deteriorates performance depends on various factors.

Bias in Estimating Covariance Matrices

For LDA we need estimates for the distribution parameters:

- $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$ **empirical mean**
- $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^\top$ **emp. covariance matrix**

But, if the number of samples n is not large relative to the dimension d, the estimation, in particular $\hat{\boldsymbol{\Sigma}}$, is error-prone.

This may affect classification with LDA badly.

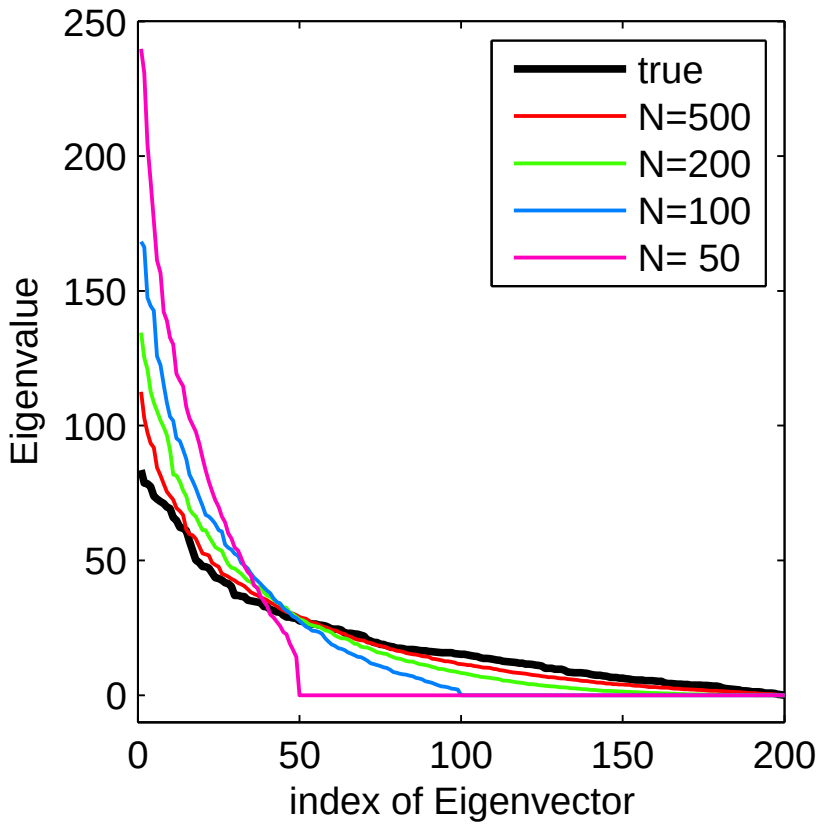
There is a systematical bias in the empirical covariance matrix:

- Large Eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are too large
- Small Eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are too small

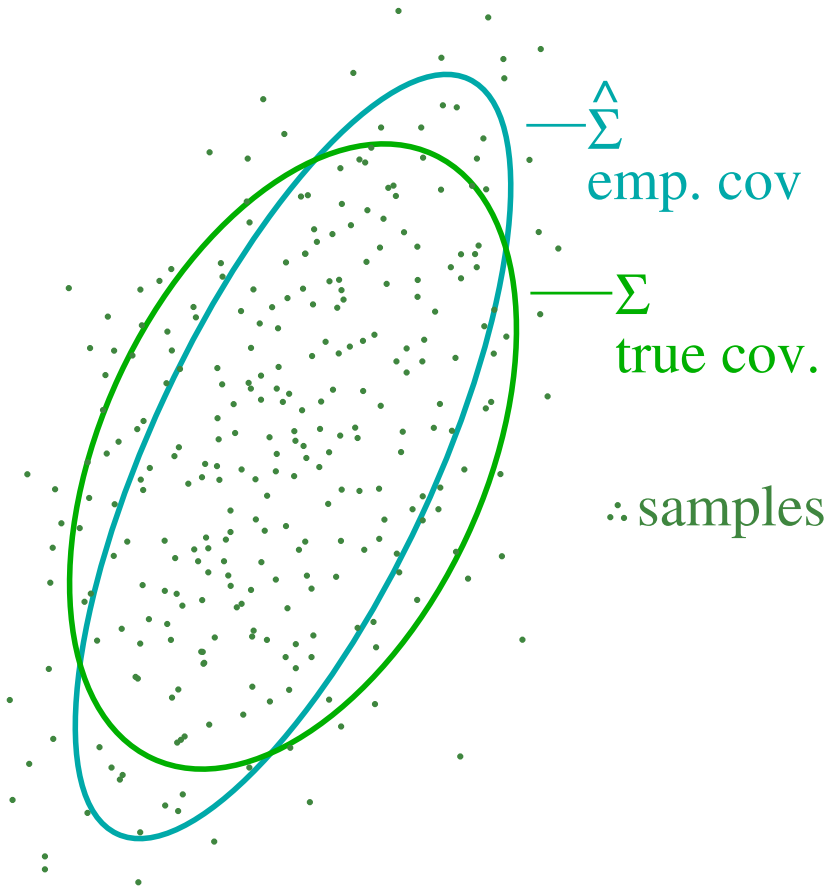
assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ are drawn from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Bias in Estimating Covariances (2)

Simulation for $d = 200$:



Cartoon in 2D:

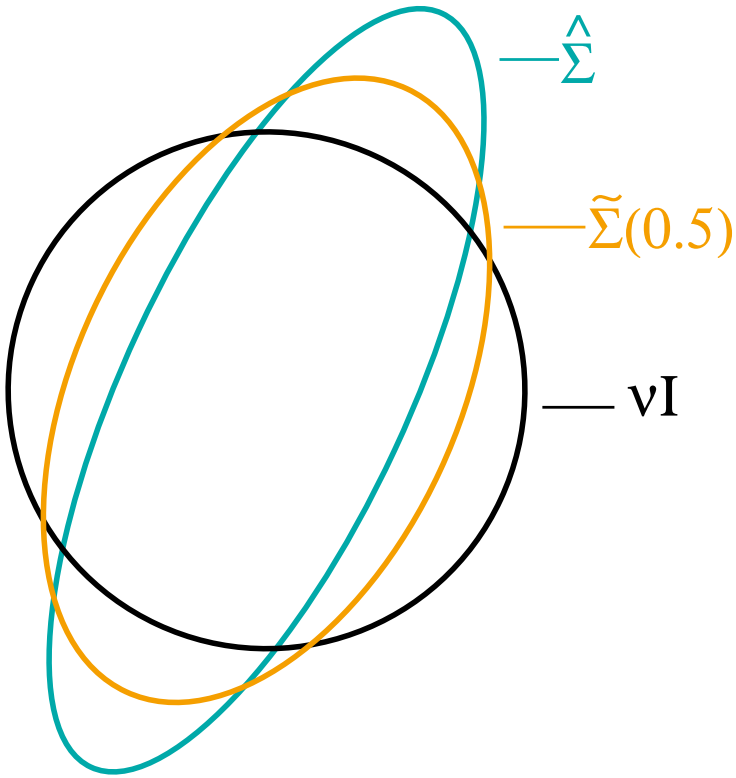


A Remedy for the Estimation Bias

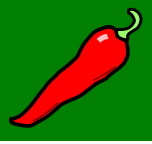
A simple way that counteracts the bias is **shrinkage**:
The empirical covariance matrix $\hat{\mathbf{\Sigma}}$ is modified to be more spherical:

$$\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$$

for a $\gamma \in [0, 1]$ and ν defined as average Eigenvalue $\mathrm{trace}(\hat{\mathbf{\Sigma}})/d$.



Next, we check that shrinkage serves the intended purpose. Covariance matrices are described by their Eigenvectors and Eigenvalues. So, we have to investigate, what happens to those, when we change over from the empirical covariance matrix $\hat{\mathbf{\Sigma}}$.



Properties of the Shrunk Covariance Matrix

From the Eigenvalue decomposition of the empirical covariance matrix $\hat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ with orthonormal $\mathbf{V}$ and diagonal $\mathbf{D}$, we get an Eigenvalue decomposition of $\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$ like this:

$$\begin{aligned}\tilde{\mathbf{\Sigma}}(\gamma) &= (1-\gamma)\mathbf{V}\mathbf{D}\mathbf{V}^\top + \gamma\nu\mathbf{I} \\ &= (1-\gamma)\mathbf{V}\mathbf{D}\mathbf{V}^\top + \gamma\nu\mathbf{V}\mathbf{I}\mathbf{V}^\top \\ &= \mathbf{V}\underbrace{\left((1-\gamma)\mathbf{D} + \gamma\nu\mathbf{I}\right)}_{\text{diagonal matrix}}\mathbf{V}^\top\end{aligned}$$

We see that

- $\hat{\mathbf{\Sigma}}$ and $\tilde{\mathbf{\Sigma}}(\gamma)$ have the same Eigenvectors (columns of $\mathbf{V}$)
- Extreme Eigenvalues (large/small) are shrunk/extended towards the average Eigenvalue ν as $d_i \mapsto (1-\gamma)d_i + \gamma\nu$
- $\gamma = 0$ means no shrinkage: $\tilde{\mathbf{\Sigma}}(0) = \hat{\mathbf{\Sigma}}$
- $\gamma = 1$ corresponds to spherical covariances matrices: $\tilde{\mathbf{\Sigma}}(1) = \nu\mathbf{I}$

Regularized Linear Discriminant Analysis

This technique can be used to enhance LDA to work better in the case of a low number-of-samples to dimensionality ratio. The empirical covariance matrix $\hat{\mathbf{\Sigma}}$ is replaced by a shrunk covariance matrix $\tilde{\mathbf{\Sigma}}(\gamma)$:

$$\mathbf{w}_\gamma := \tilde{\mathbf{\Sigma}}(\gamma)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

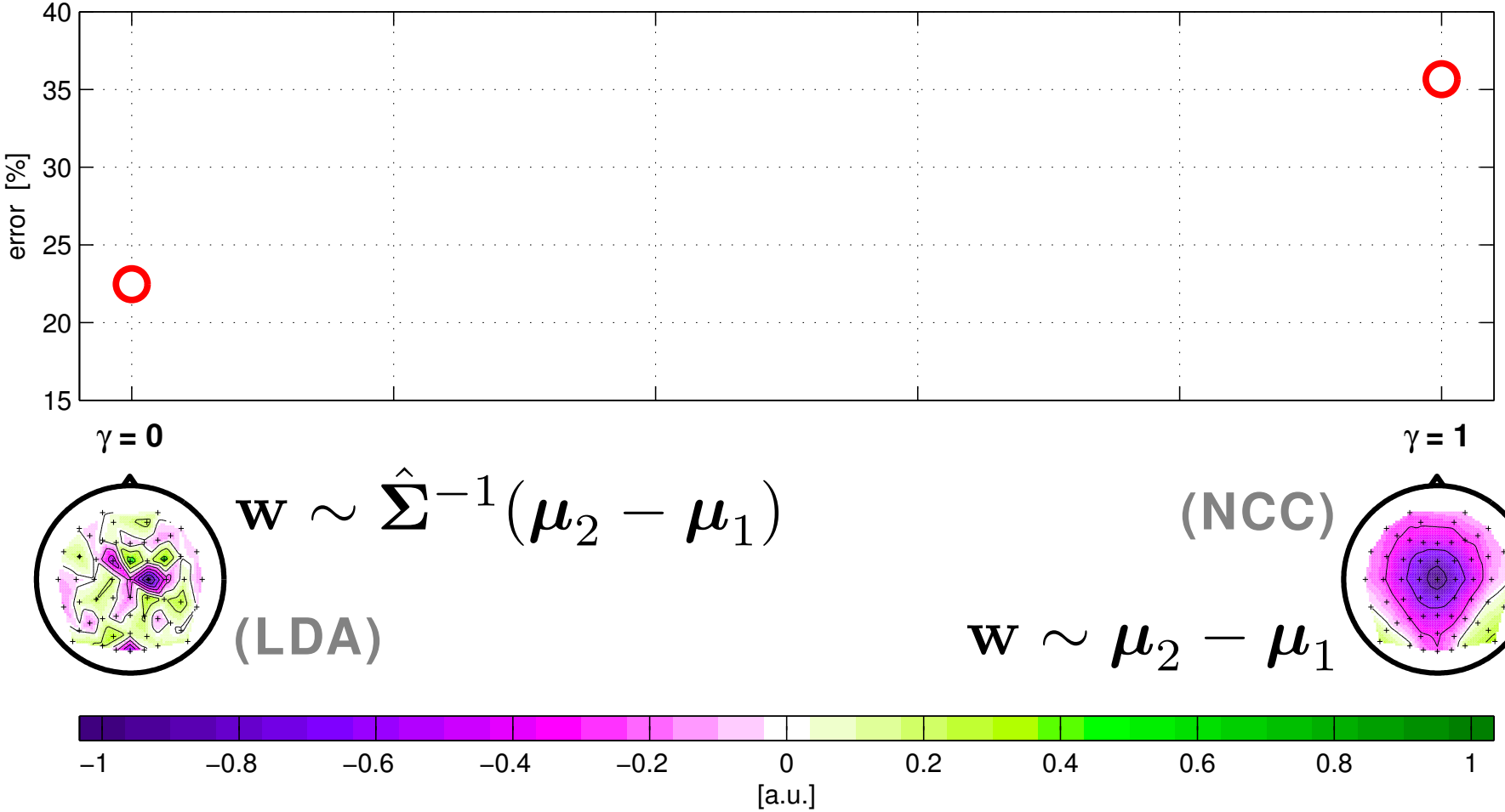
Here, γ is a hyper parameter that has to be selected between 0 and 1.

- $\gamma = 0$ yields $\mathbf{w}_0 = \hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, i.e. unregularized LDA
- $\gamma = 1$ yields $\mathbf{w}_1 = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, i.e. NCC

Before addressing the choice of γ, let us look at the impact of the shrinkage parameter.

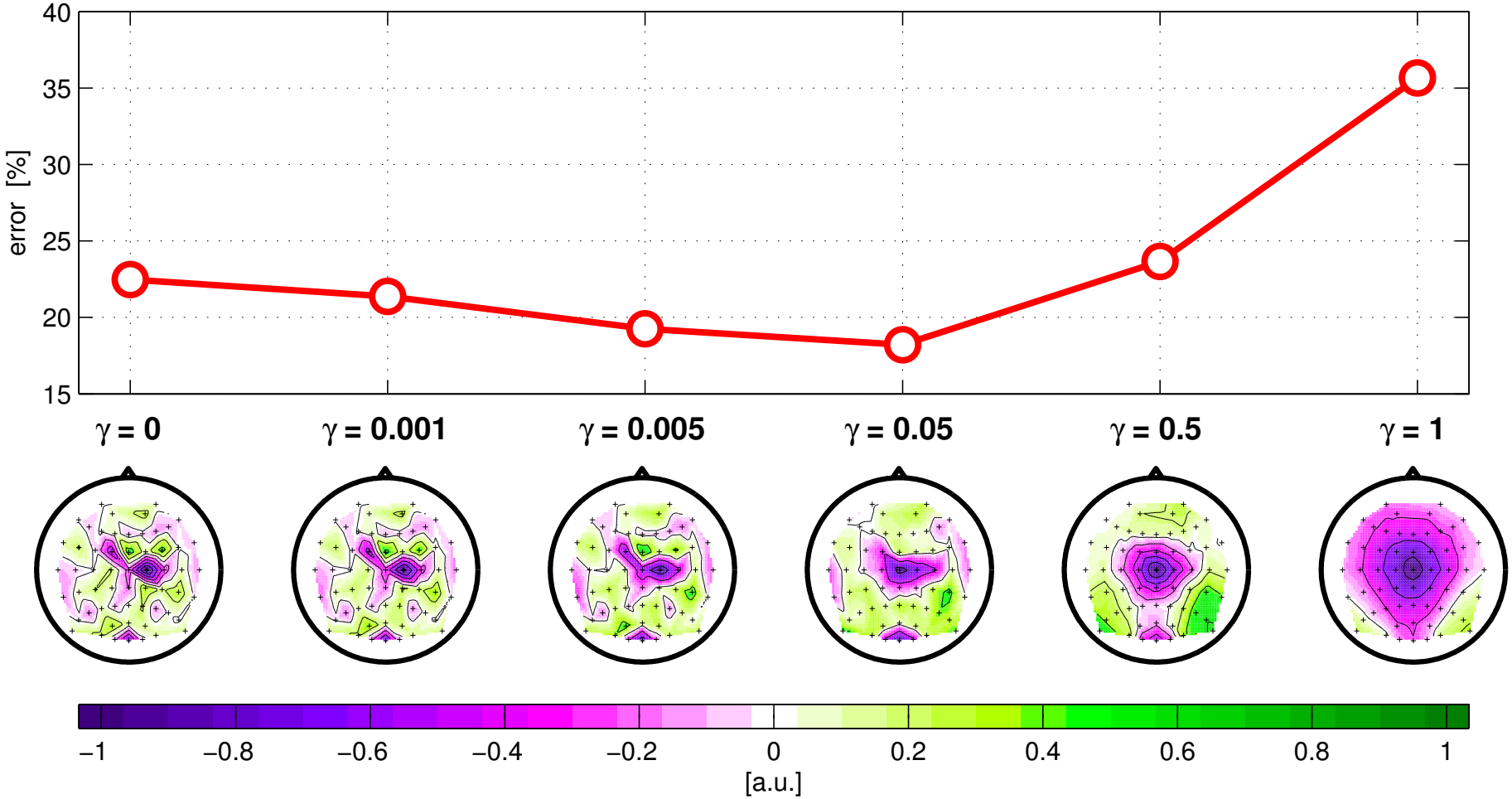
Impact of Shrinkage as Trade-off

LDA with shrinkage: $\mathbf{w} = \tilde{\mathbf{\Sigma}}(\gamma)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$;
$\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$



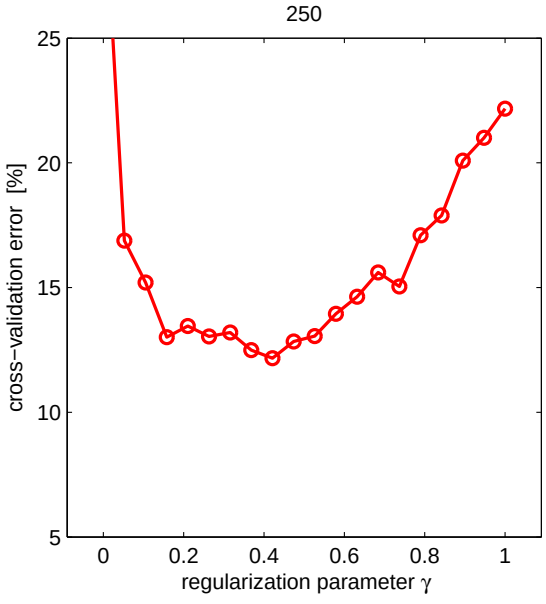
Impact of Shrinkage as Trade-off

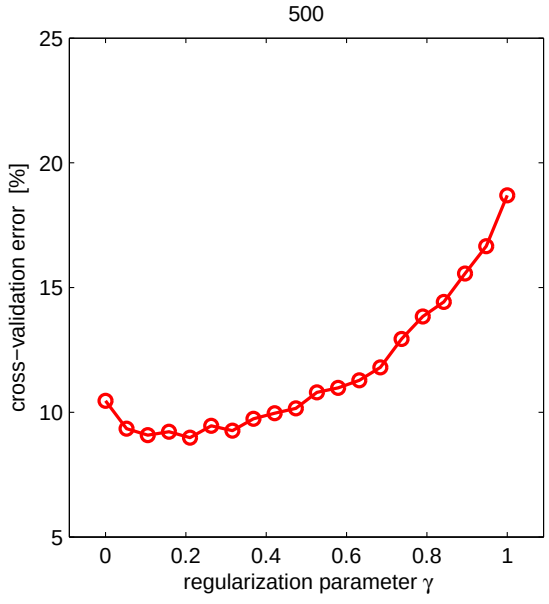
With increasing shrinkage, the spatial filters (classifier) look smoother, but classification may degrade with too much shrinkage.

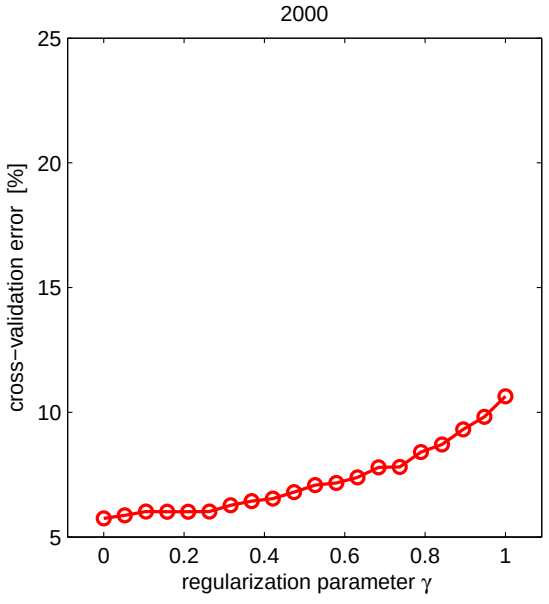


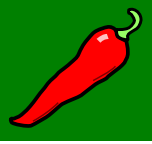
LDA with Different Shrinkage Parameters

Cross-validation results for different sizes of training data (250, 500, 2000) for different values of the shrinkage parameter γ (x-axis). Features vectors have 250 dimensions.









Optimal Selection of Shrinkage Parameter

As a (relatively) novel method for selecting the free parameter (γ) other than cross-validation, there is an analytical method.

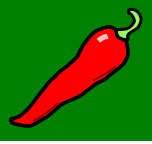
Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be n feature vectors and let $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$ be the empirical mean.

Aim: get a better estimate of the true covariance matrix $\boldsymbol{\Sigma}$ (especially in case $n < d$) than the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^\top$$

by selecting a γ in

$$\tilde{\boldsymbol{\Sigma}}(\gamma) := (1-\gamma)\hat{\boldsymbol{\Sigma}} + \gamma\nu\mathbf{I}.$$



Optimal Selection of Shrinkage Parameter

The approach of [Ledoit & Wolf, J Multivar Anal, 2004] is to minimize

$$\|\tilde{\mathbf{\Sigma}}(\gamma) - \mathbf{\Sigma}\|_F^2 \qquad \text{with } \|\cdot\|_F^2 \text{ being the Frobenius norm.}$$

We denote by $(\mathbf{x}_k)_i$ resp. $(\hat{\boldsymbol{\mu}})_i$ the i-th element of the vector $\mathbf{x}_k$ resp. $\hat{\boldsymbol{\mu}}$ and define the covariance of feature i and j in trial k:

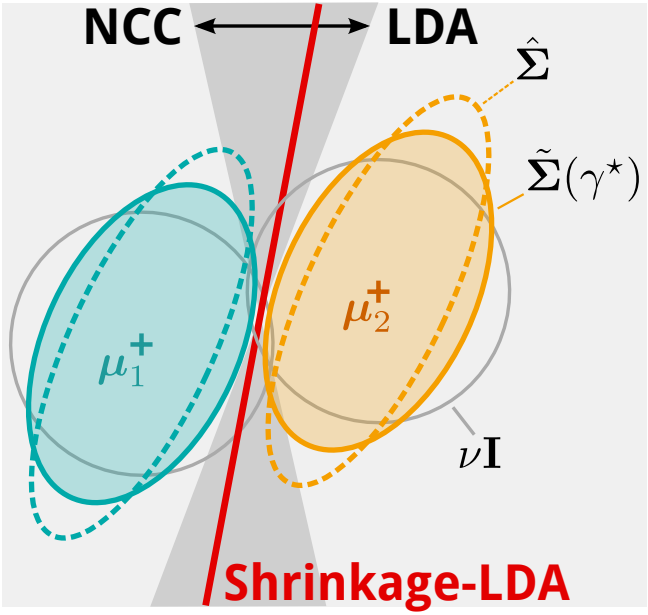
$$z_{ij}(k) = ((\mathbf{x}_k)_i - (\hat{\boldsymbol{\mu}})_i)\ ((\mathbf{x}_k)_j - (\hat{\boldsymbol{\mu}})_j)^\top$$

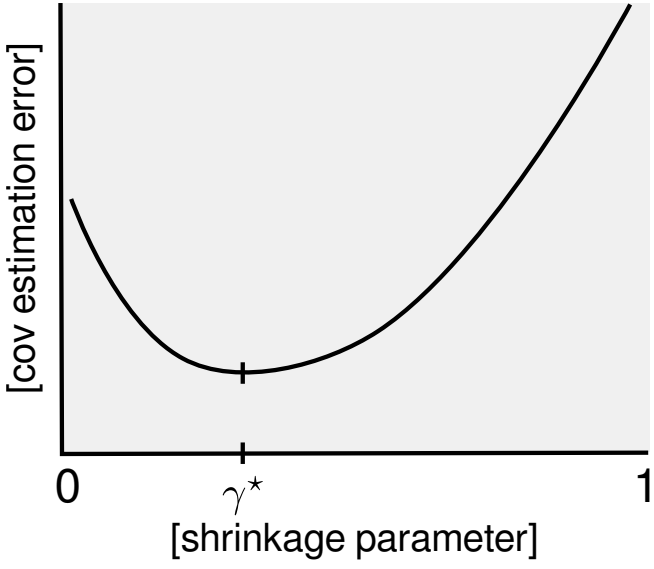
Denoting by s_{ij} the element in the i-th row and j-th column of the matrix $\hat{\mathbf{\Sigma}} - \nu\mathbf{I}$, the optimal shrinkage parameter $\gamma^\star = \text{argmin}_\gamma \|\tilde{\mathbf{\Sigma}}(\gamma) - \mathbf{\Sigma}\|_F^2$ can be analytically calculated as [Schäfer & Strimmer 2005]

$$\gamma^\star = \frac{n}{(n-1)^2} \frac{\sum_{i,j=1}^d \text{var}_k(z_{ij}(k))}{\sum_{i,j=1}^d s_{ij}^2}.$$

Shrinkage-LDA: use $\tilde{\mathbf{\Sigma}}(\gamma^\star)$ instead of $\hat{\mathbf{\Sigma}}$.

Classification with Shrinkage-LDA at a Glance





Shrinkage-LDA hyperplane is defined by:

$$\mathbf{w} := \tilde{\mathbf{\Sigma}}(\gamma^\star)^{-1}\,(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

$$\tilde{\mathbf{\Sigma}}(\gamma) := (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$$

Calculate optimal $\gamma^\star$ analytically:

$$\gamma^\star = \operatorname*{argmin}_{\gamma} \|\tilde{\mathbf{\Sigma}}(\gamma) - \mathbf{\Sigma}\|_F^2$$

$$= \frac{n}{(n-1)^2}\,\frac{\sum_{i,j=1}^{d} \operatorname{var}_k(z_{ij}(k))}{\sum_{i,j=1}^{d} s_{ij}^2} \quad \text{with}$$

$$z_{ij}(k) := ((\mathbf{x}_k)_i - (\hat{\boldsymbol{\mu}})_i)\,((\mathbf{x}_k)_j - (\hat{\boldsymbol{\mu}})_j)^\top$$

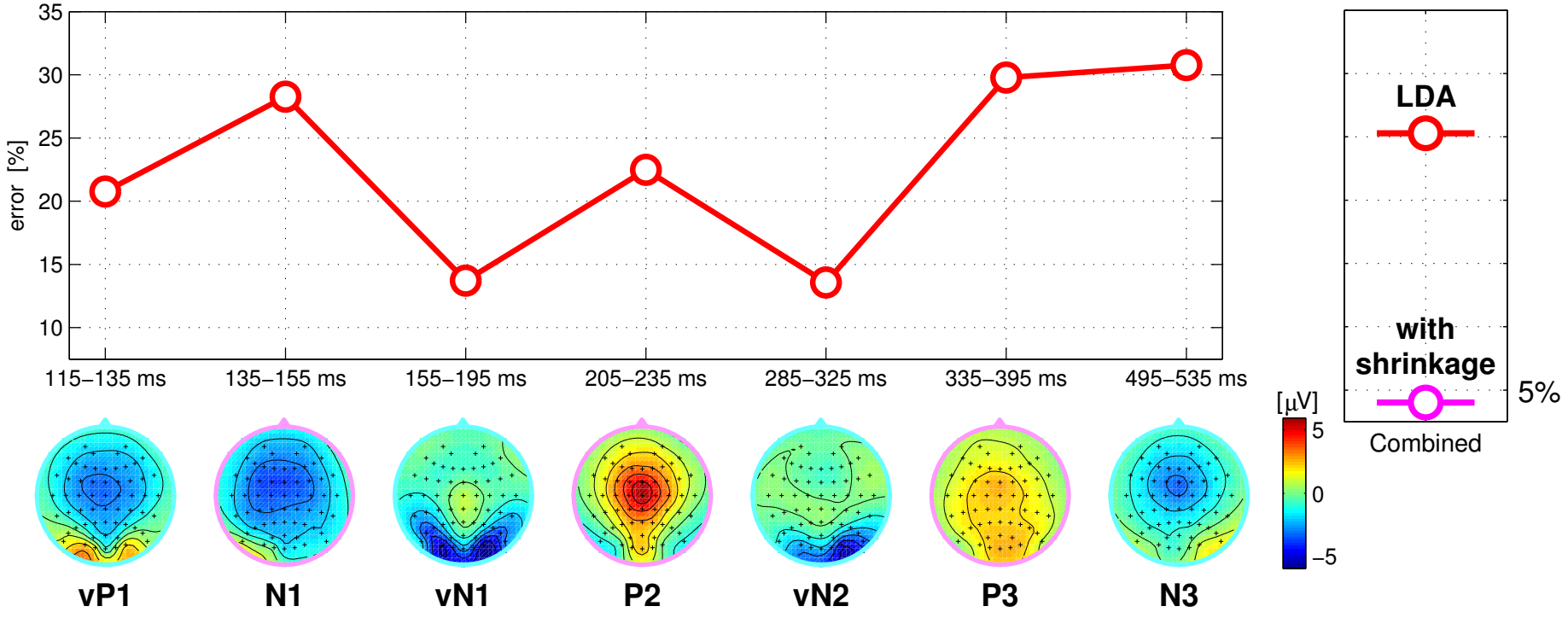
Selection of shrinkage parameter γ:

[Ledoit & Wolf 2004], [Schäfer & Strimmer 2005]

Tutorial on ERP classification:

[Blankertz et al, NeuroImage 2011]

Classification on Single Components and Combined



Classification (with $N = 750$ training samples) on seven different single components ($d = 55$) yields errors between **14%** and 31%.

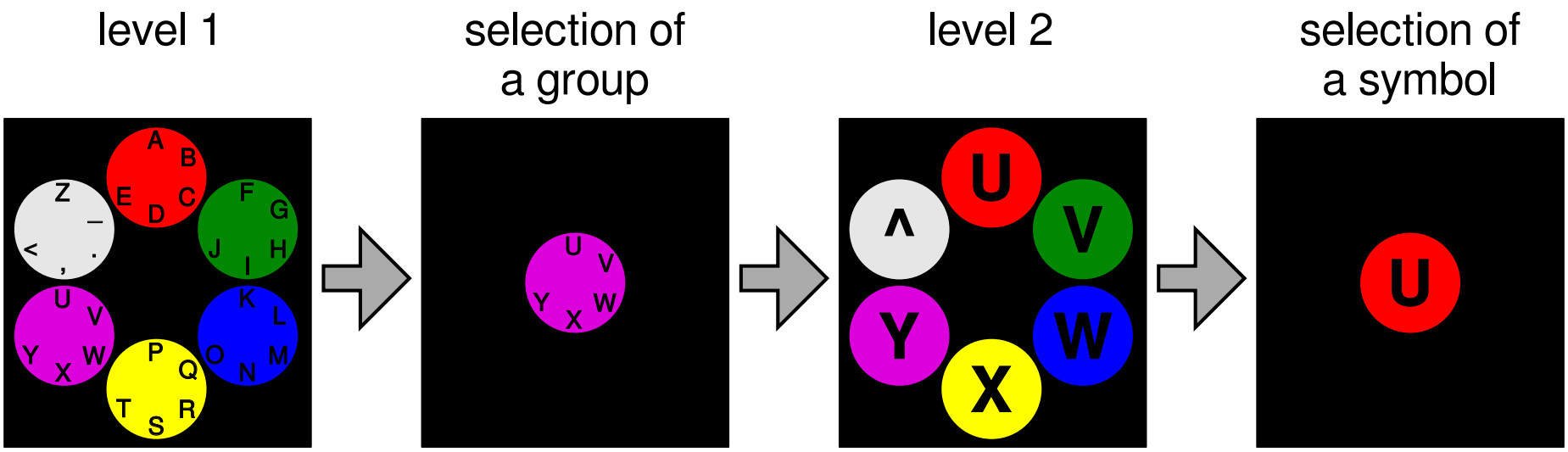
LDA on the concatenated feature ($d = 7 \cdot 55 = 385$) performs with **25%** worse, although information is added: *overfitting*.

Shrinkage-LDA: only **4%** error. [Blankertz et al, NeuroImage 2011]

Center Speller

To counteract the problems in visual layout of the Matrix Speller, a new paradigm for gaze-independent spelling was established:

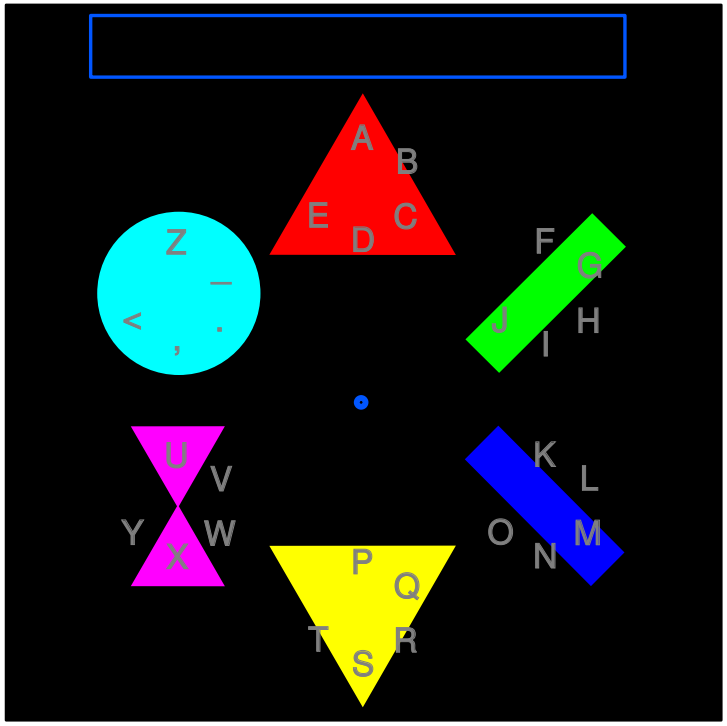
➤ The Hex-o-Spell selection principle:



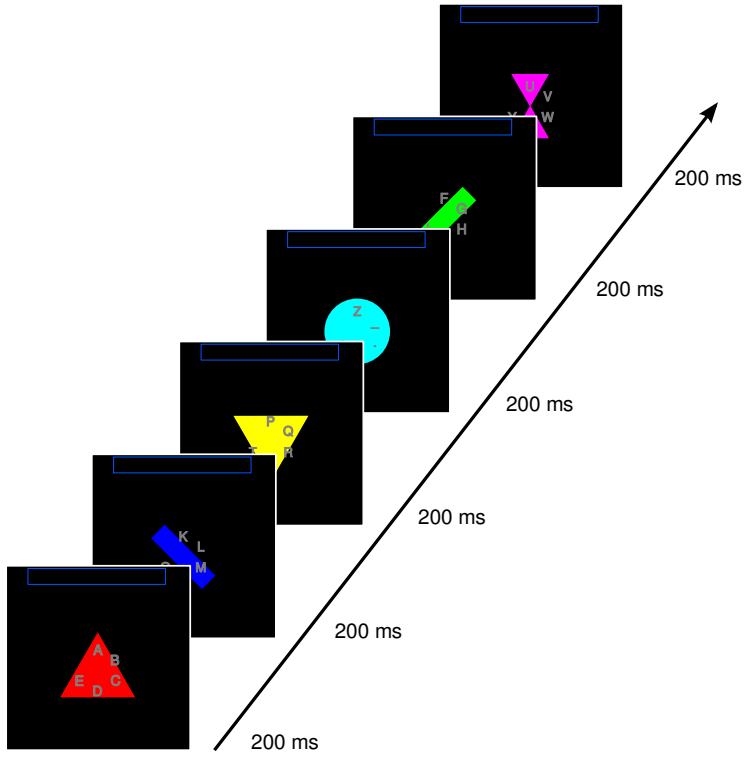
[Matthias Treder & Blankertz, Behav Brain Funct 2010]

Gaze Independent ERP-Speller

Center Speller



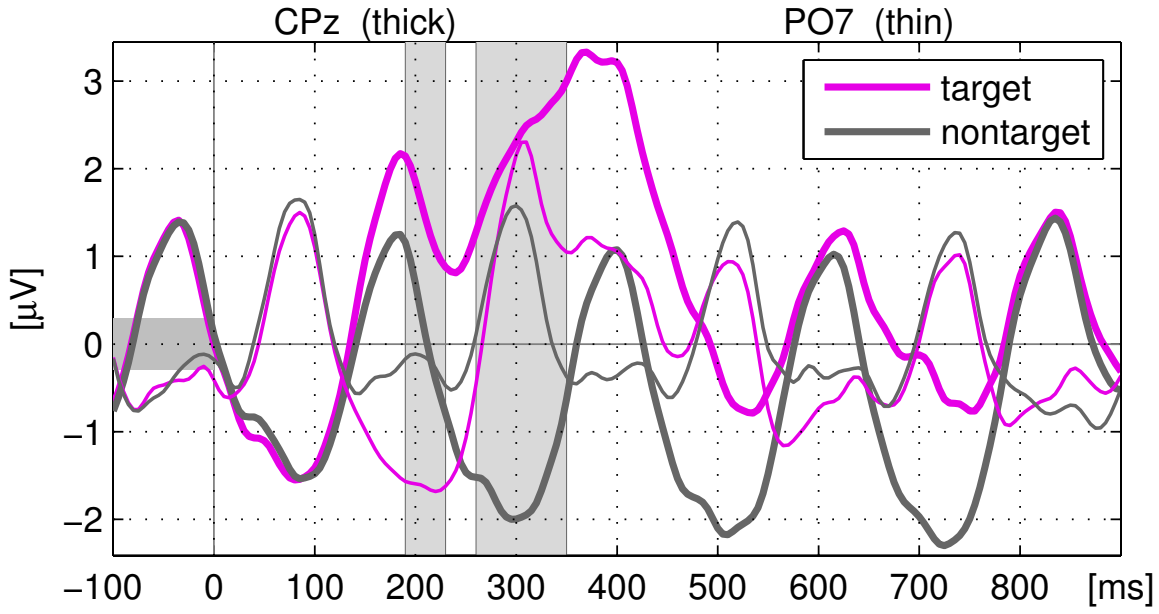
Experimental flow

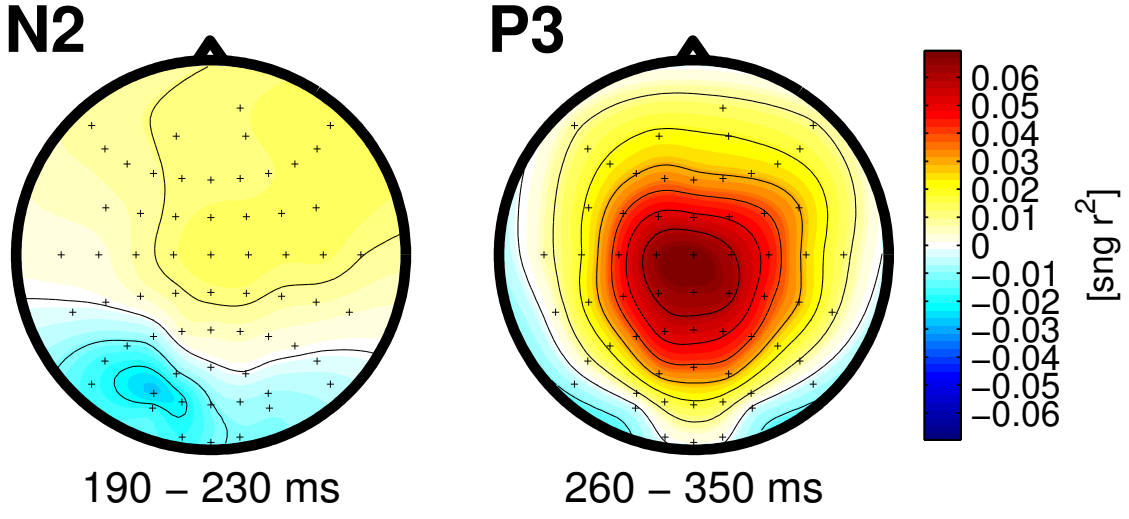


- Two level selection process
- Feature attention (form)
- Feature attention (color)

[Matthias Treder, Schmidt & Blankertz, J Neural Eng 2011]

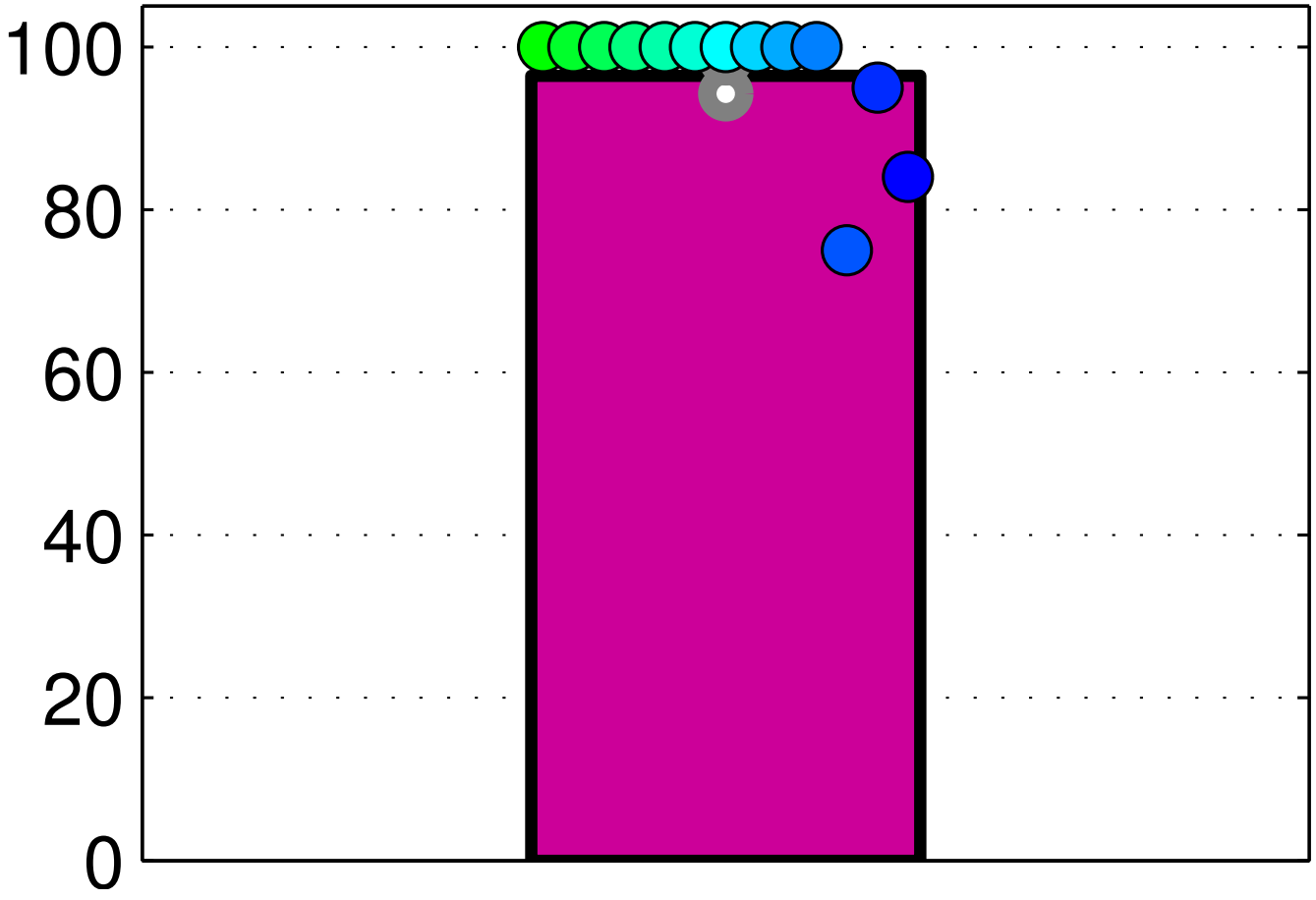
Center Speller: Results - ERPs



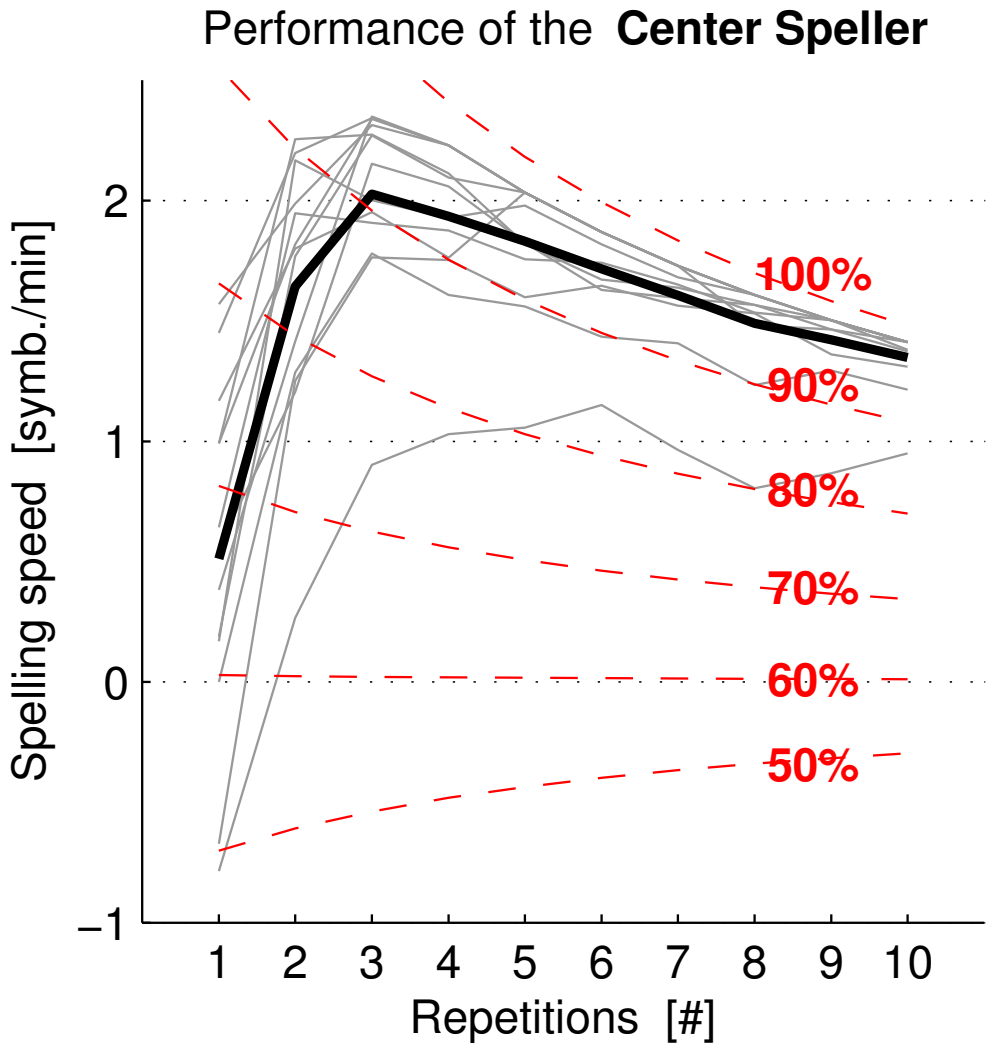


Results - Online Spelling Performance

Online symbol selection accuracy was 100% for 10 out of 13 participants.



Results - Simulated Spelling Performance



Notes performance measures for spellers: [Blankertz et al, in prep.]

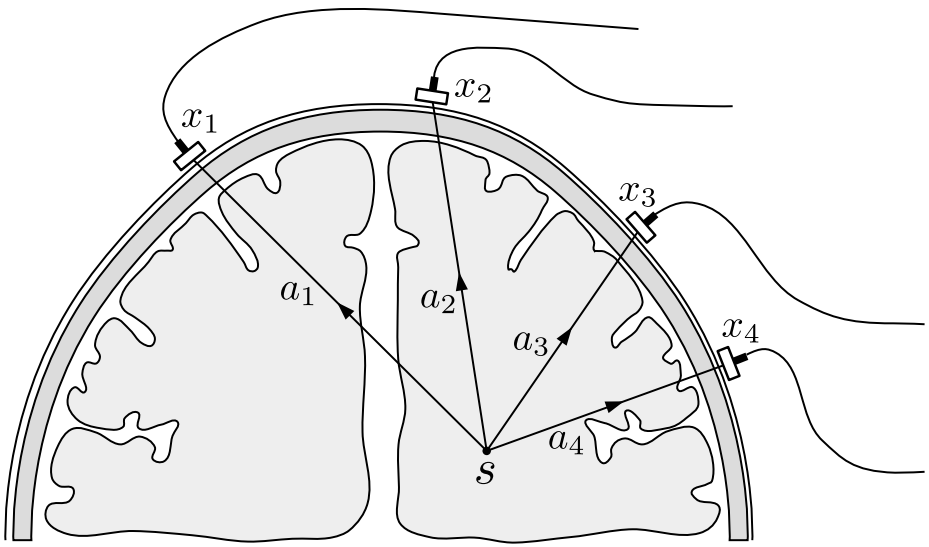
Part III: The Linear Model of EEG / Spatial Patterns and Spatial Filters

Linear Model of EEG: Forward Model

- **Assumption:** The contribution of a current source $s(t)$ to the scalp potentials $\mathbf{x}(t) = [x_1, \ldots, x_k]^\top$ is linear in $s(t)$:

$$\mathbf{x}(t) = [a_1 s(t), \ldots, a_k s(t)]^\top = \mathbf{a}\, s(t)$$

- The proportionality factors in vector $\mathbf{a}$ are typically unknown and depend on the spatial distribution and orientation of the current source and the conductivity distribution of the anatomy.



Linear Model of EEG: Forward Model (2)

- Now, we consider several sources with distribution vectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$.
- Potentials are additive. Defining the matrix $\mathbf{A}$ as being composed of the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$ (i.e., $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]$), the **Forward Model** is

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t) = \mathbf{a}_1\, s_1(t) + \mathbf{a}_2\, s_2(t) + \ldots \mathbf{a}_k\, s_p(t)$$

- Contributions not captured by this model are considered as noise, $\mathbf{n}(t)$, typically assumed to be Gaussian distributed.
- This gives a simple linear model representing the electrophysics of EEG:

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t) + \mathbf{n}(t)$$

Linear Model of EEG: Backward Model

Recovering of sources is formalized in the **backward model**:

$$\hat{\mathbf{s}}(t) = \mathbf{W}^\top \mathbf{x}(t)$$

Given a forward model $\mathbf{A}$, taking $\mathbf{W}^\top$ as $\mathbf{A}^\# = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top$, the pseudo inverse of $\mathbf{A}$, is the least mean squares estimator:

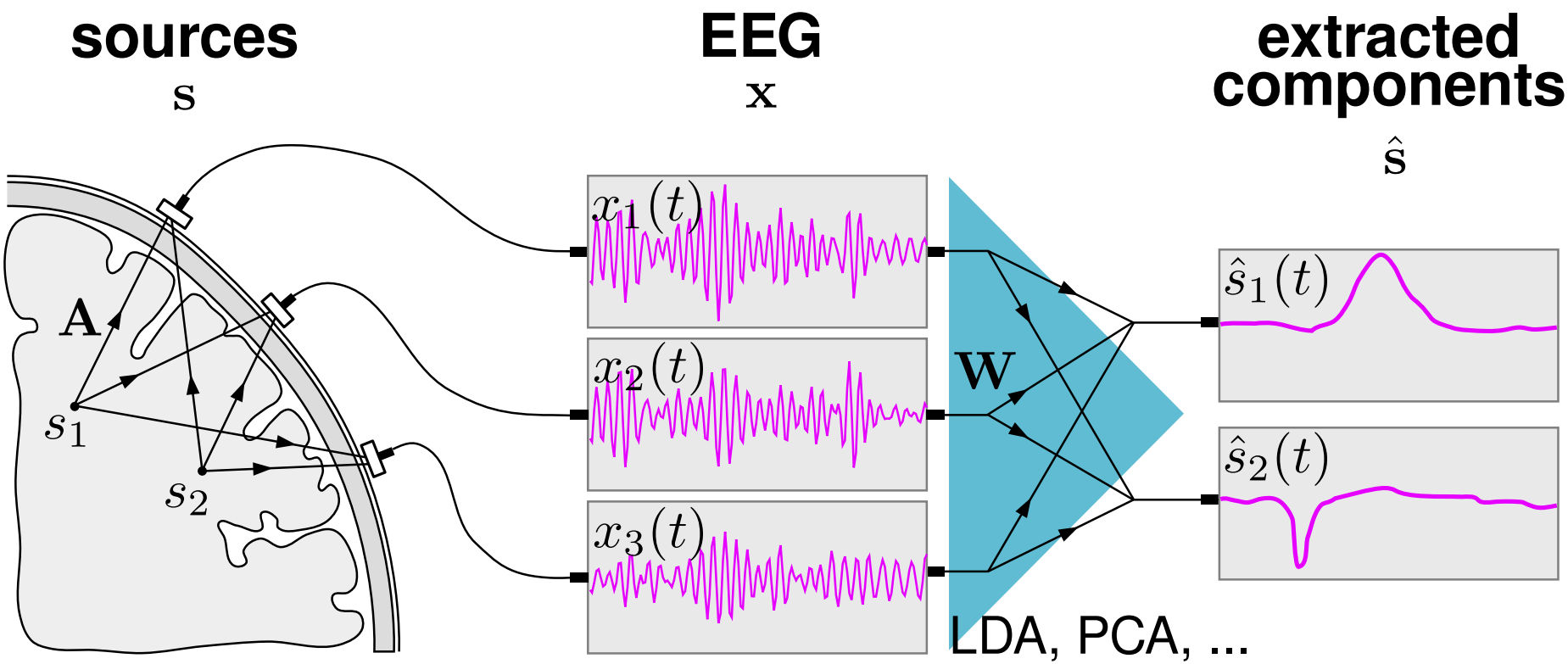
$$\arg\min_{\mathbf{V}} \sum_t \|\mathbf{V}^\top \mathbf{A}\mathbf{s}(t) - \mathbf{s}(t)\|^2 = \mathbf{A}^\#$$

Note that, even for invertible $\mathbf{A}$ a backward model captures also the portion of the noise that is collinear with the source estimates.

$$\hat{\mathbf{s}}(t) = \mathbf{s}(t) + \mathbf{W}^\top \mathbf{n}(t).$$

Anyway, $\mathbf{A}$ is unknown and difficult to estimate (*inverse problem*).

Linear Model of EEG



$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t)$$

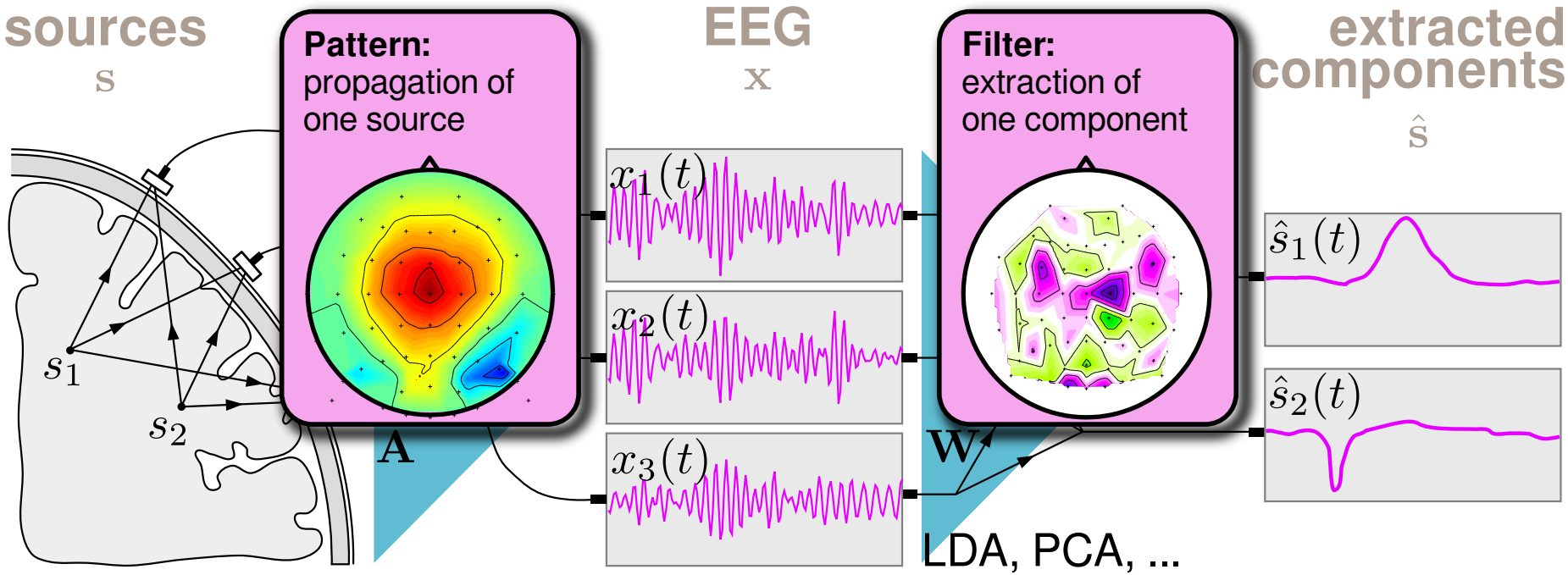
forward model

$$\hat{\mathbf{s}}(t) = \mathbf{W}^\top \mathbf{x}(t)$$

backward model

Each column of $\mathbf{A}$ is a spatial *pattern*: propagation of a source to sensors
Each row of $\mathbf{W}^\top$ is a spatial *filter*: weighting of EEG channels.

Patterns and Filters in the Linear Model of EEG



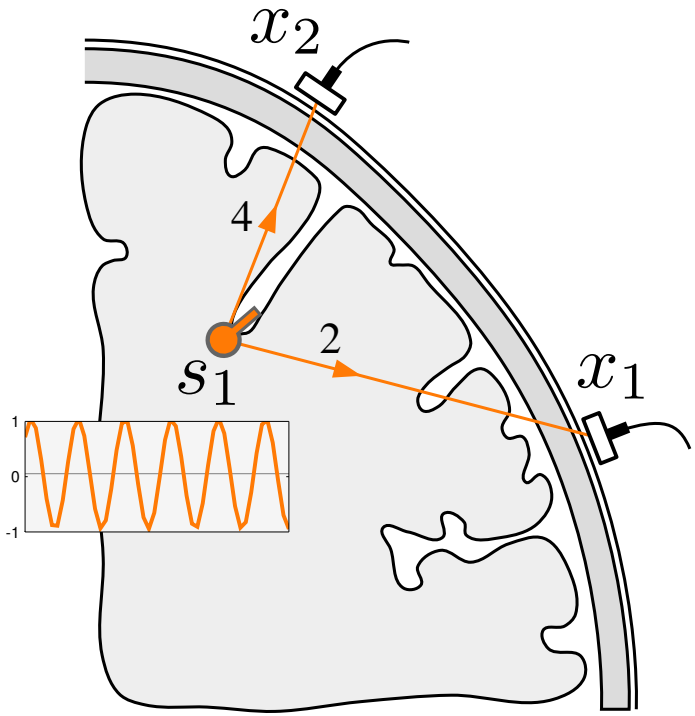
$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t)$$

forward model

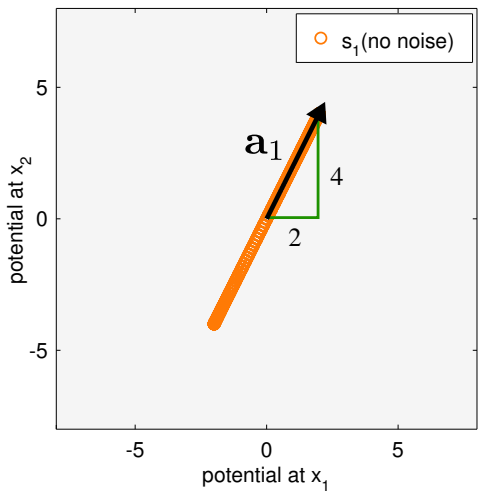
$$\hat{\mathbf{s}}(t) = \mathbf{W}^\top \mathbf{x}(t)$$

backward model

Explaining Spatial Patterns and Spatial Filters

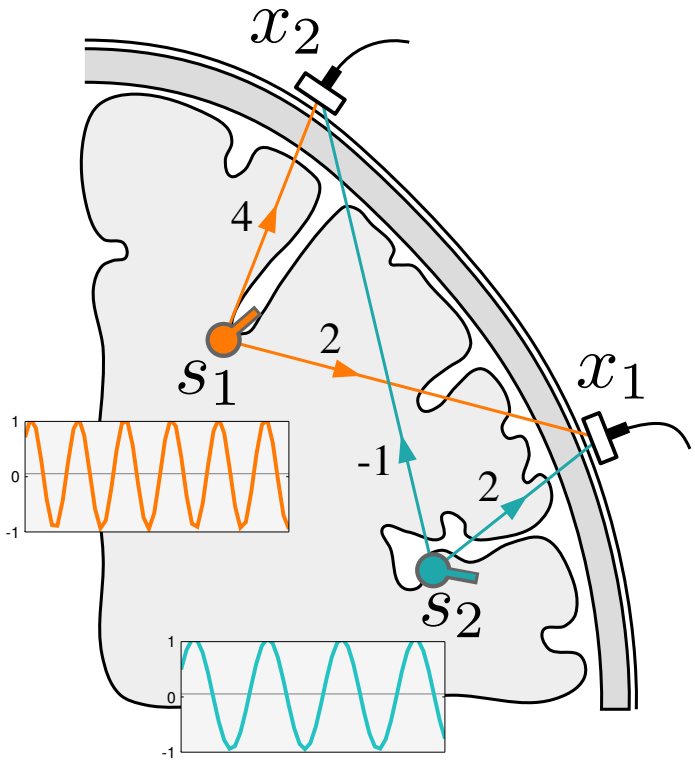


$$\mathbf{x}(t) = \mathbf{a}_1 s_1(t)$$



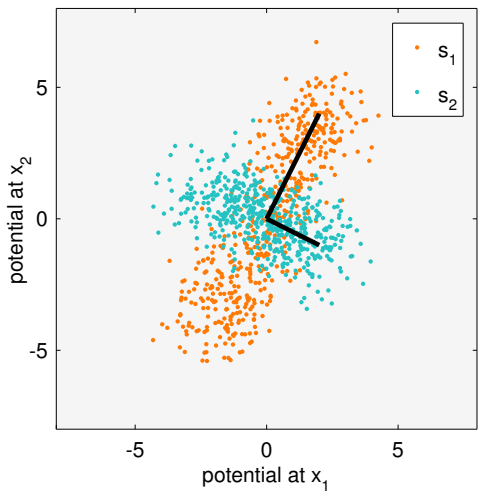
$$\mathbf{a}_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

Explaining Spatial Patterns and Spatial Filters

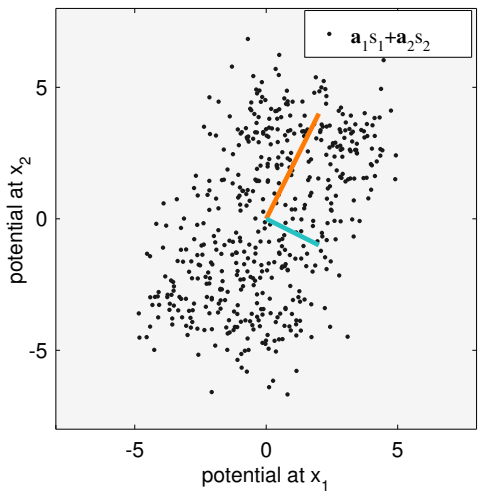


$$\mathbf{a}_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad \mathbf{a}_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

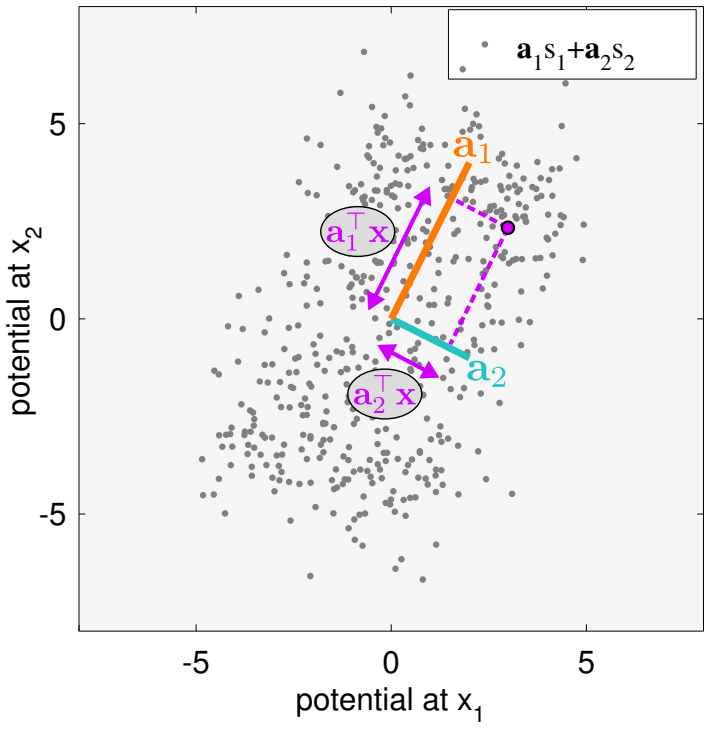
$$\mathbf{x}(t) = \mathbf{a}_1 s_1(t) + \mathbf{n}(t)$$
$$\mathbf{x}(t) = \mathbf{a}_2 s_2(t) + \mathbf{n}(t)$$



$$\mathbf{x}(t) = \mathbf{a}_1 s_1(t) + \mathbf{a}_2 s_2(t) + \mathbf{n}(t)$$

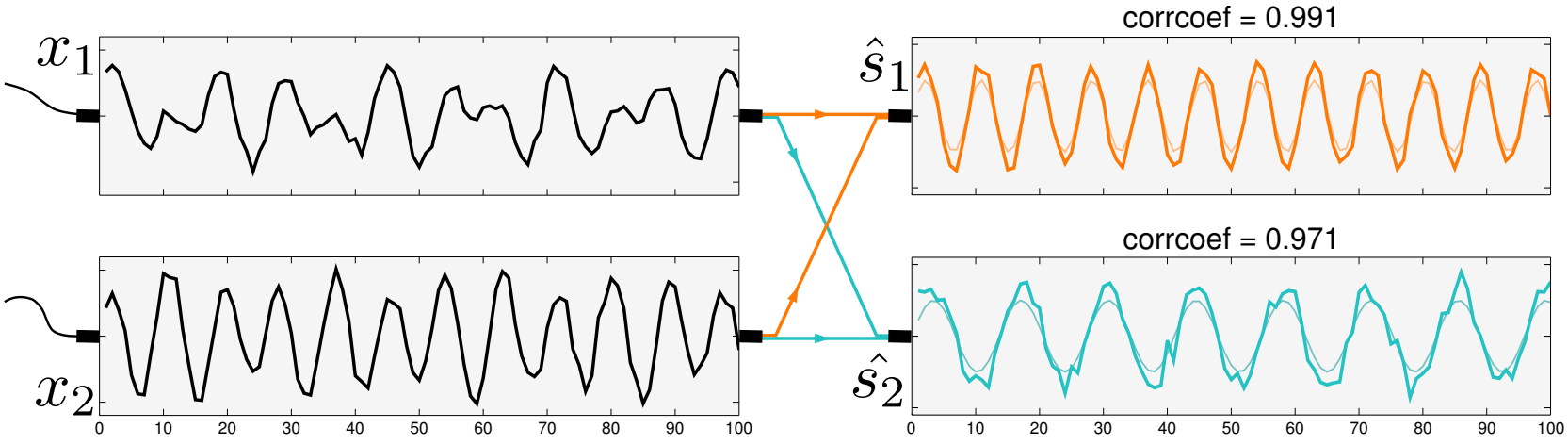


Explaining Spatial Patterns and Spatial Filters

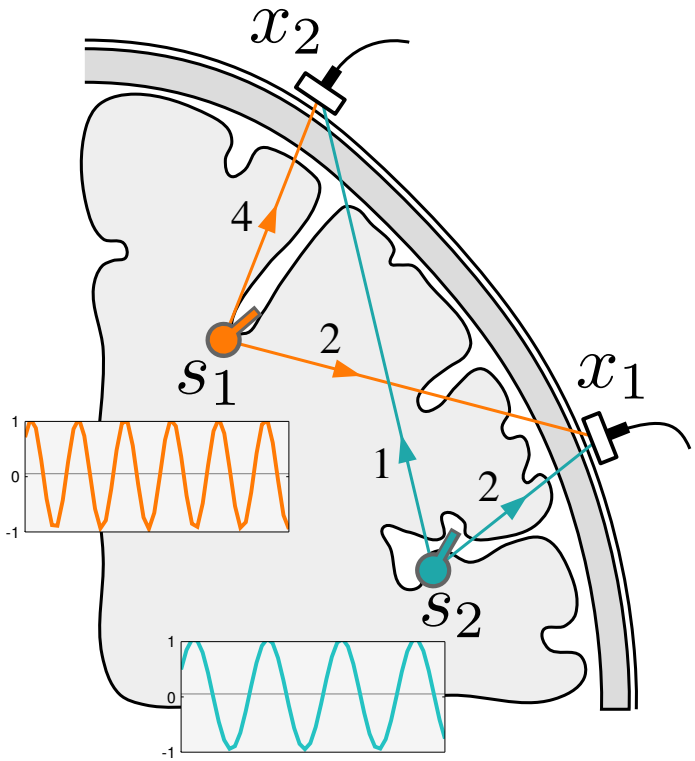


$$\hat{s}_1 = \mathbf{a}_1^\top \mathbf{x}$$
$$= \mathbf{a}_1^\top \mathbf{a}_1 s_1 + \underbrace{\mathbf{a}_1^\top \mathbf{a}_2}_{=0} s_2 + \mathbf{a}_1^\top \mathbf{n}$$
$$\sim s_1$$

$$\hat{s}_2 \sim s_2 \quad \text{(as above)}$$

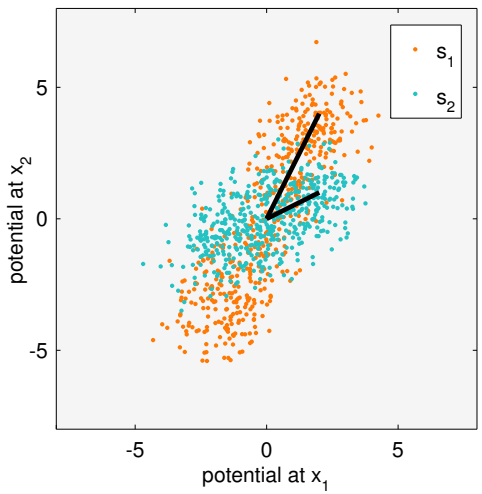


Explaining Spatial Patterns and Spatial Filters

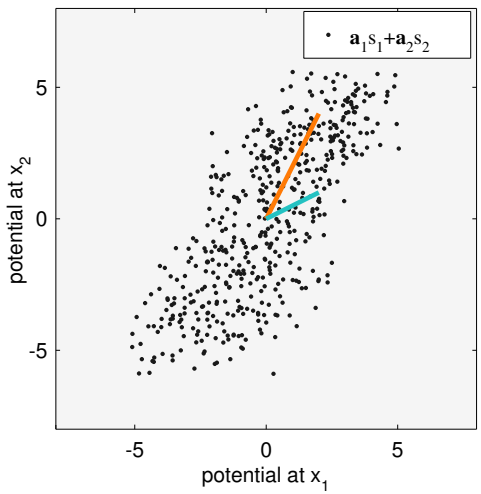


$$\mathbf{a}_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad \tilde{\mathbf{a}}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

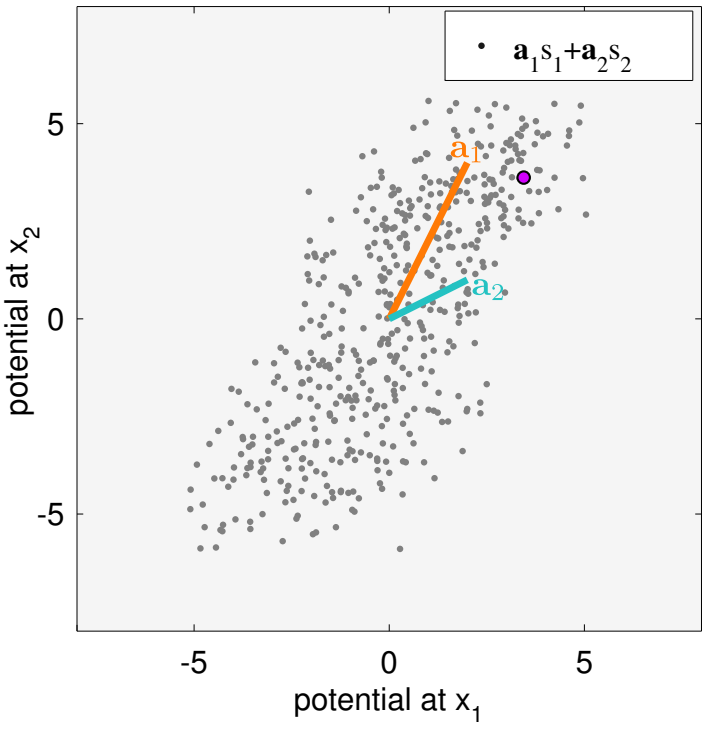
$$\mathbf{x}(t) = \mathbf{a}_1 s_1(t) + \mathbf{n}(t)$$
$$\mathbf{x}(t) = \tilde{\mathbf{a}}_2 s_2(t) + \mathbf{n}(t)$$



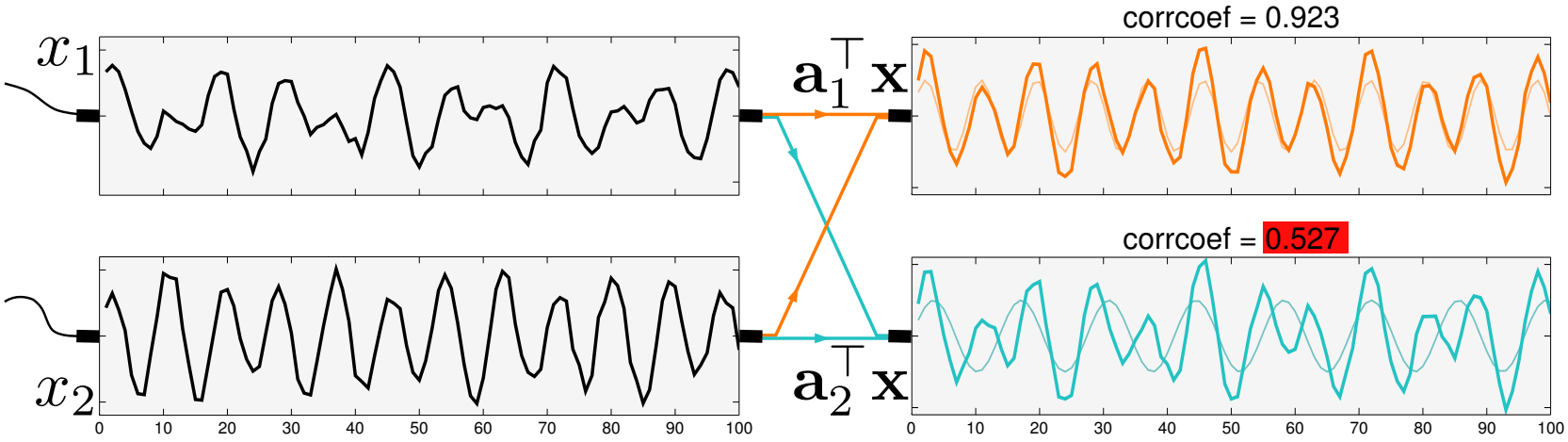
$$\mathbf{x}(t) = \mathbf{a}_1 s_1(t) + \tilde{\mathbf{a}}_2 s_2(t) + \mathbf{n}(t)$$



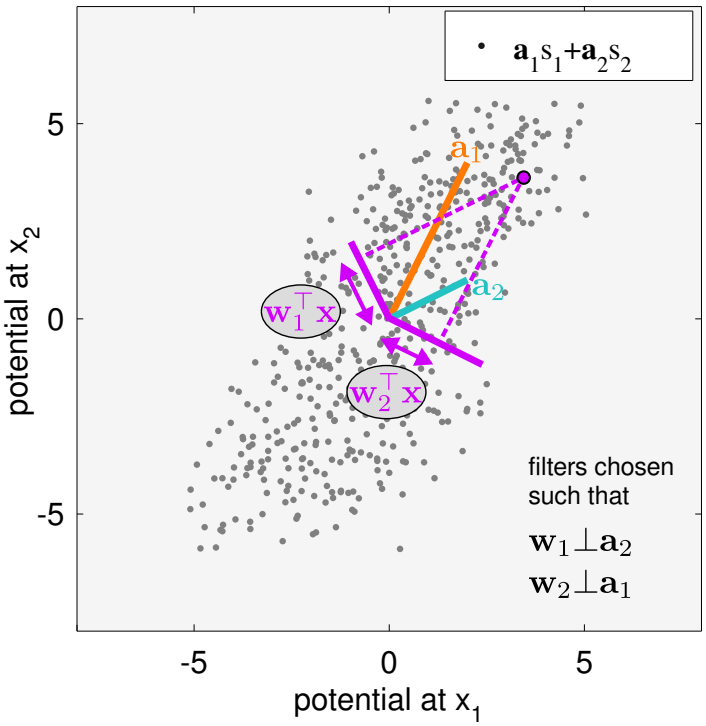
Explaining Spatial Patterns and Spatial Filters



$$\hat{s}_1 = \mathbf{a}_1^\top \mathbf{x}$$
$$= \mathbf{a}_1^\top \mathbf{a}_1 s_1 + \underbrace{\mathbf{a}_1^\top \mathbf{a}_2 s_2}_{\text{does not vanish}} + \mathbf{a}_1^\top \mathbf{n}$$

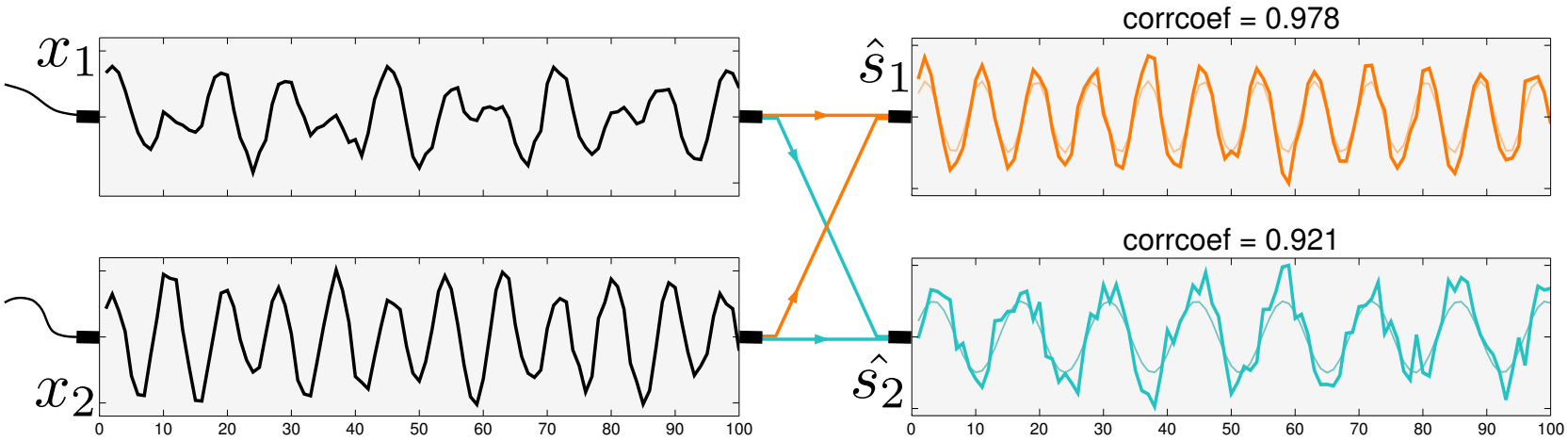


Explaining Spatial Patterns and Spatial Filters

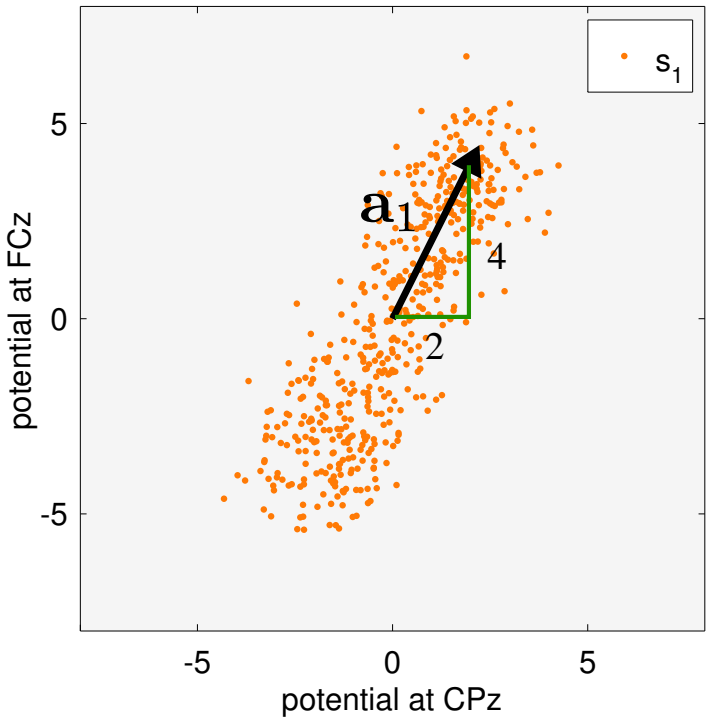


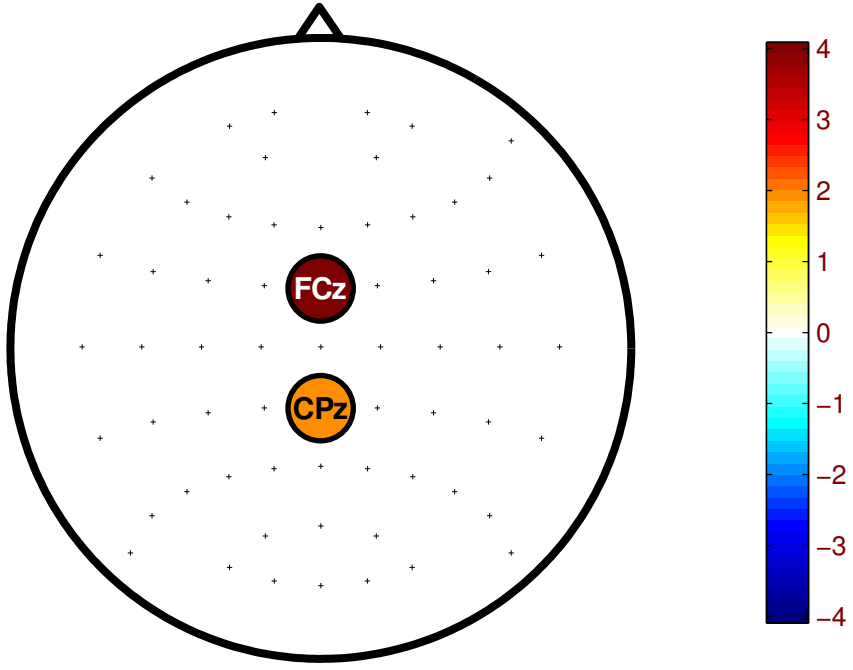
$$\hat{s}_1 = \mathbf{w}_1^\top \mathbf{x}$$
$$= \mathbf{w}_1^\top \mathbf{a}_1 s_1 + \underbrace{\mathbf{w}_1^\top \mathbf{a}_2}_{=0} s_2 + \mathbf{w}_1^\top \mathbf{n}$$
$$\sim s_1$$

$$\hat{s}_2 \sim s_2$$

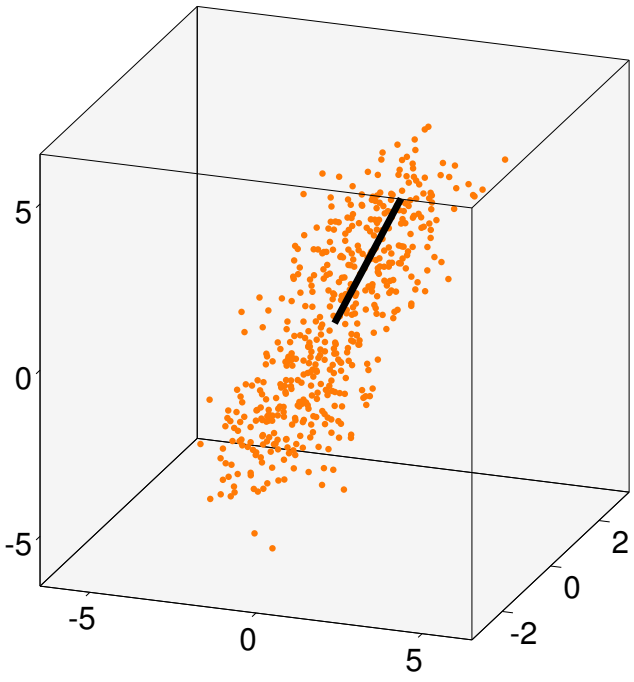


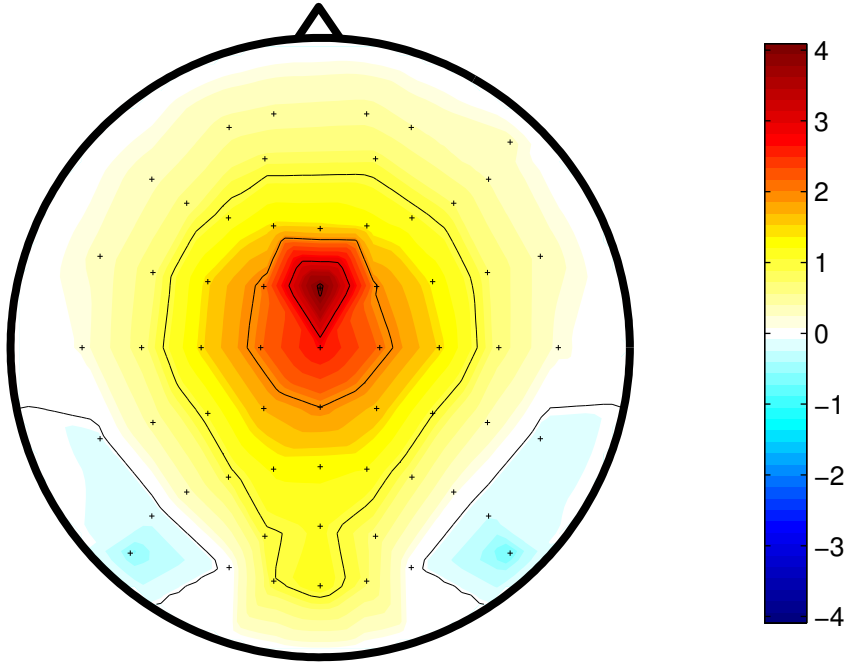
Correspondence of Vectors in Feature Space and Patterns





Correspondence of Vectors in Feature Space and Patterns





Special Case: The Linear Model According to PCA

Using PCA (Eigenvalue decomposition) one gets a linear model in which the 'sources' (not meant to approximate the real ones) are uncorrelated. Given signals $\mathbf{x}(t)$, the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ is decomposed as

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

with diagonal $\mathbf{D}$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$.

The sources in the PCA model are

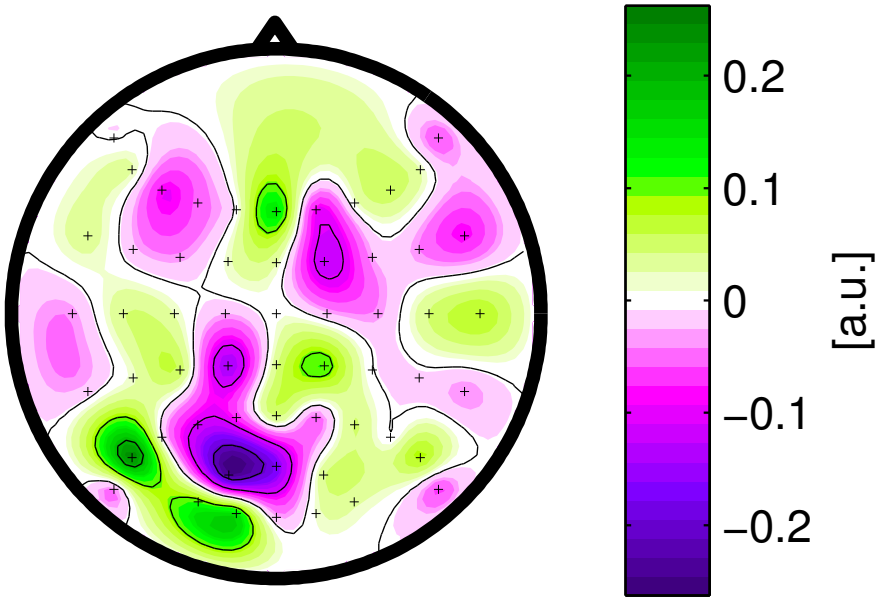
$$\hat{\mathbf{s}}(t) = \mathbf{V}^\top \mathbf{x}(t) \qquad \text{(backward model)}$$

and they are projected to the sensors with the same matrix:

$$\mathbf{x}(t) = \mathbf{V}^\top \hat{\mathbf{s}}(t) \qquad \text{(forward model)}$$

$\Rightarrow$ In this model, **patterns and filters coincide!**

Recap: Classification of (Purely) Spatial Features



The weight vector of an LDA trained on spatial features can be visualized as a topography and can be regarded as a spatial filter.

For the interpretation of spatial filters there is a caveat, that we will discuss next.

Interpretation of Spatial Filters (Recapitulation)

Let's assume we have a mixture of two sources (ignoring the noise here)

$$\mathbf{x} = \mathbf{a}_1 s_1 + \mathbf{a}_2 s_2$$

and the task is to find a spatial filter $\mathbf{w}$ to recover s_1. Applying the filter to $\mathbf{x}$ yields

$$\mathbf{w}^\top \mathbf{x} = \mathbf{w}^\top \mathbf{a}_1 s_1 + \mathbf{w}^\top \mathbf{a}_2 s_2$$

Case 1: $\mathbf{a}_1^\top \mathbf{a}_2 = 0$ (untypical). Then $\mathbf{w} = \mathbf{a}_1$ does the job: For orthorgonal propagation vectors, the best filter corresponds to the propagation direction of the source, i.e., a pattern.

Case 2: $\mathbf{a}_1^\top \mathbf{a}_2 \neq 0$ (typical). To recover s_1, the filter $\mathbf{w}$ needs to be chosen such that $\mathbf{w}^\top \mathbf{a}_2 = 0$, i.e., the filter $\mathbf{w}$ is orthogonal to $\mathbf{a}_2$.

Interpretation of Spatial Filters (Conclusion)

In the typical case ($\mathbf{a}_1^\top \mathbf{a}_2 \neq 0$), the best filter $\mathbf{w}$ to recover source s_1 also depends on the interfering source s_2, as it must be orthogonal to its propagation vector $\mathbf{a}_2$.

Example. We would like to extract

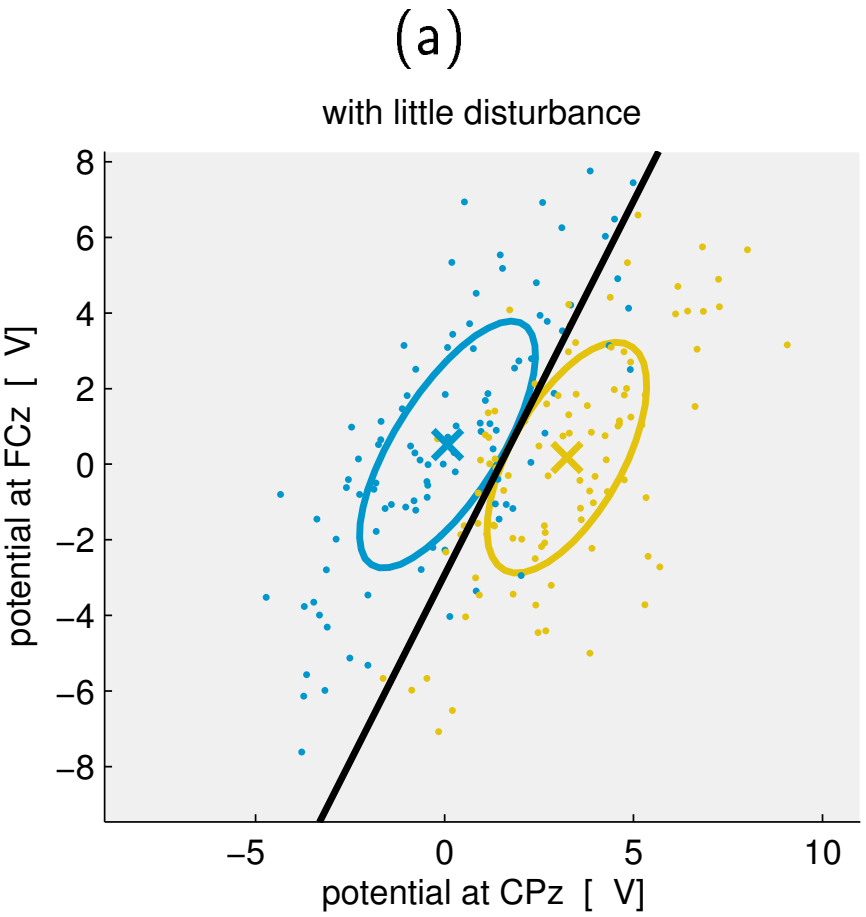
- s_1, the cognitive P300 component

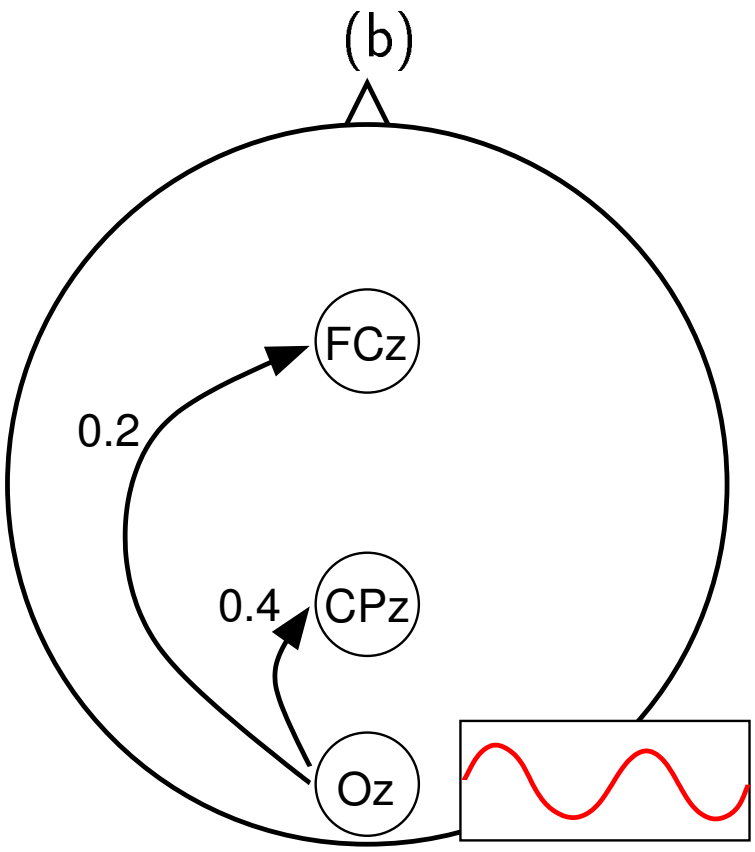
but there is interference from

- s_2, the visual area.

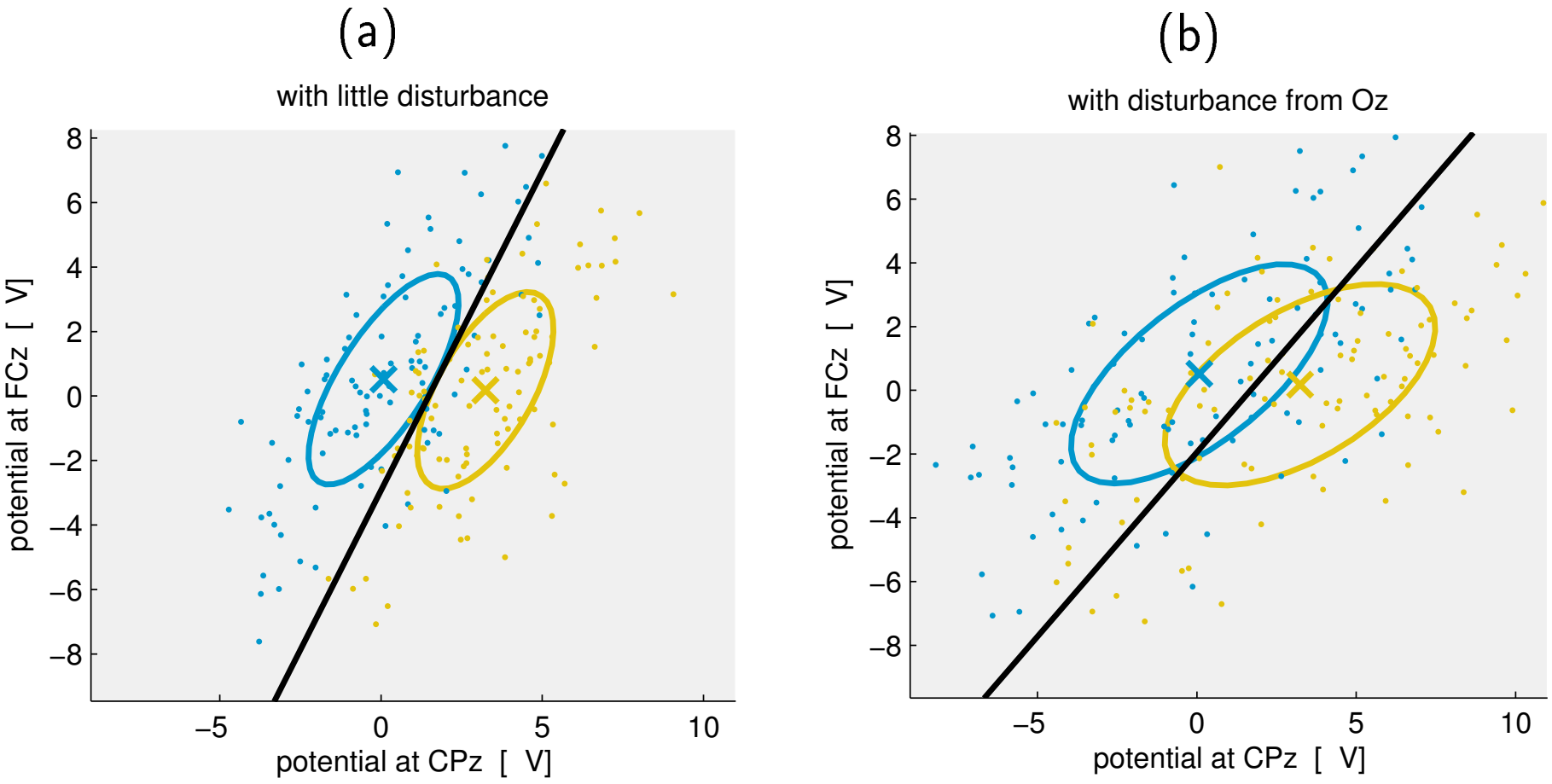
The best filter to recover the P300 component (s_1) depends also on the interfering source of the visual area (s_2). In particular, the spatial map of the filter probably shows strong weights over occipital area, although the P300 component originates from the central region.

Understanding Spatial Filters





Understanding Spatial Filters

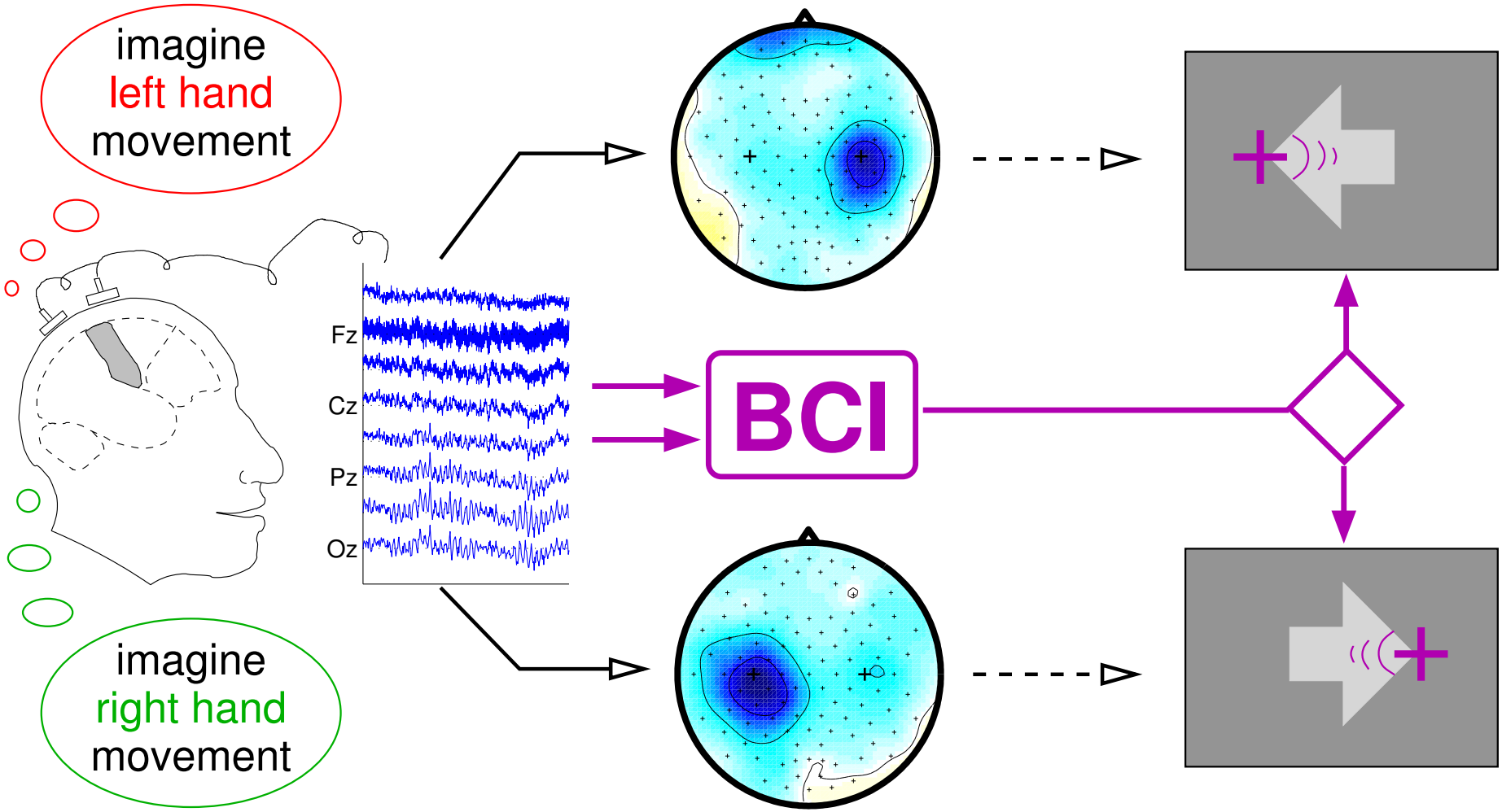


Two channel classification of (a): 15% error, (b): 37% error

When disturbing channel Oz is added to the data (3D): 16% error. Here, channel Oz is required for good classification although itself is not discriminative.

Part IV: Feature Extraction and Classification for Modulations of Oscillatory Brain Activity

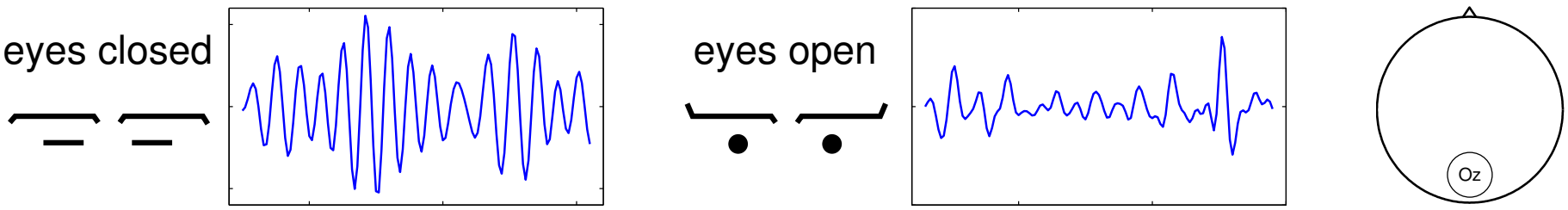
SMR-based BCI Systems



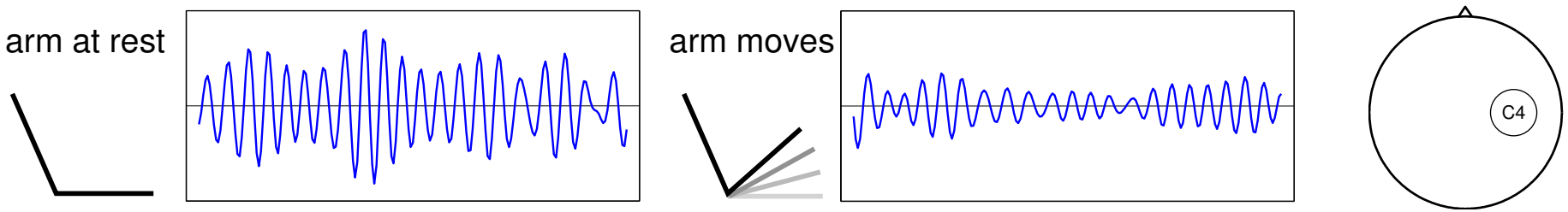
Modulation of Brain Rhythms

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

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

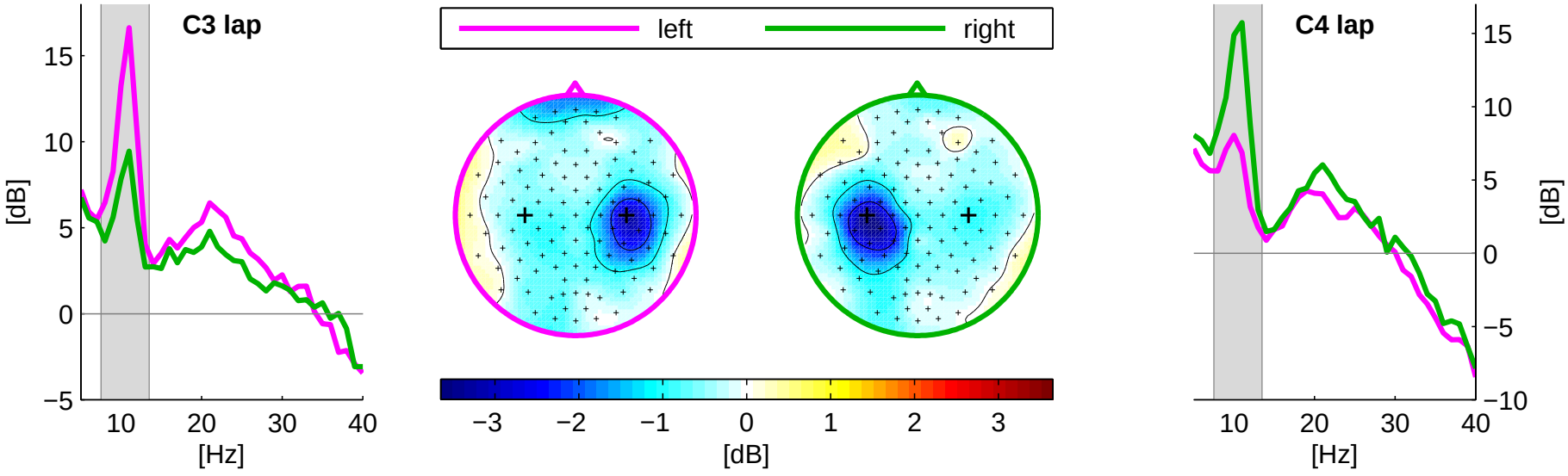


- μ-rhythm (around 10 Hz) in motor and sensory cortex:



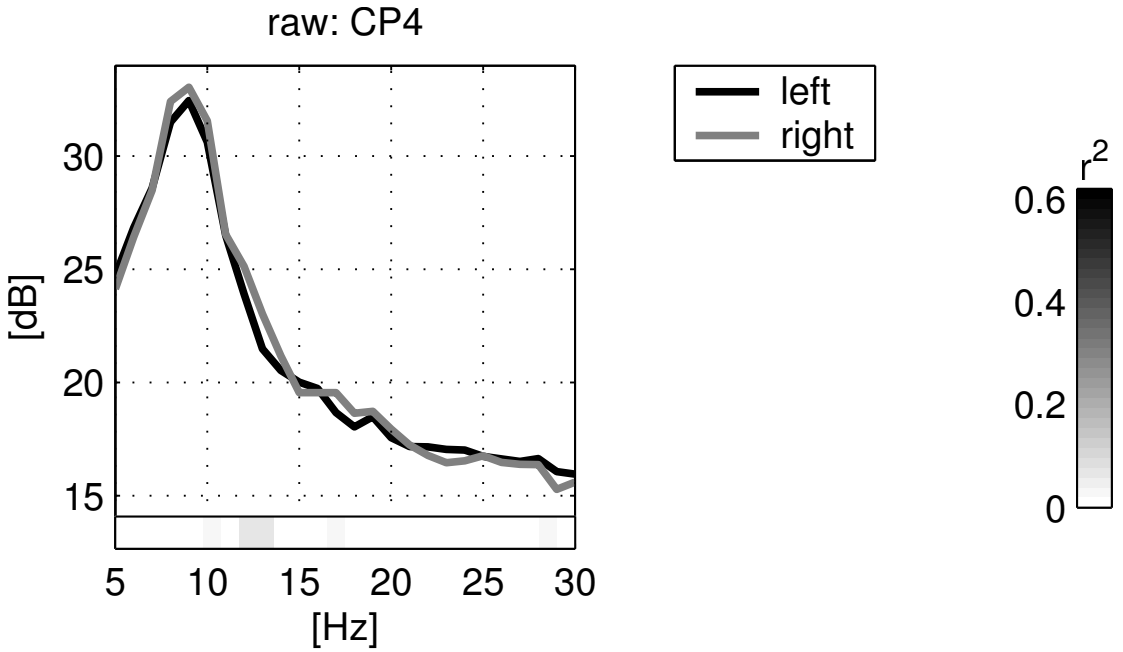
Modulation of SMRs in Motor Imagery

Data from an individual with very clear and prototypical patterns:



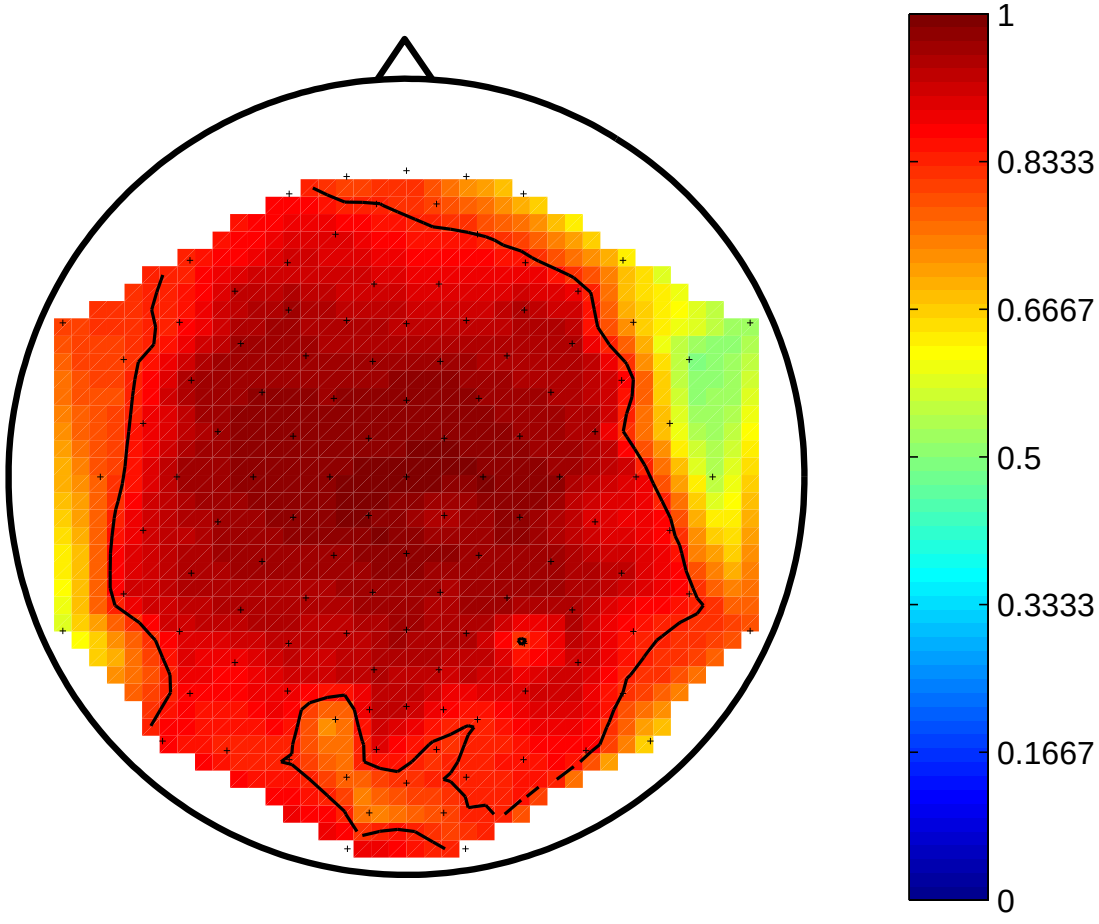
The Effect of Spatial Filtering

In practice, the difference might not show up so pronounced:



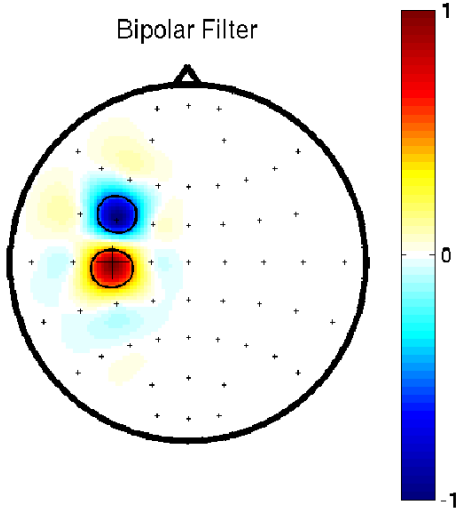
Reminder: Prominent Problem in EEG is Spatial Smearing

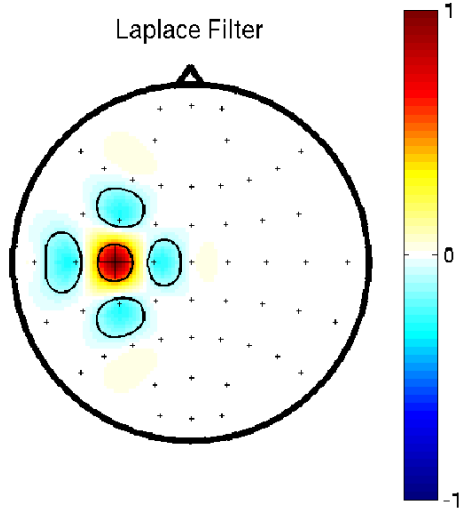
- Raw EEG scalp potentials are known to be associated with a large spatial scale owing to volumne conduction.
- In this typical example data set, most of the channels are highly correlated:

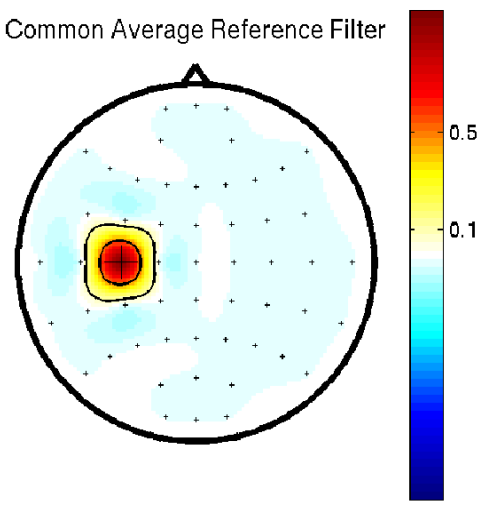


This May Help: Spatial Filters

- **Bipolar:** Subtract values from two electrode positions, e.g.: $\text{Bip}_{\text{C3,FC3}} = \text{C3} - \text{FC3}$
- **Common Average Reference (CAR):** Subtract the average of all EEG electrodes ($\mathcal{C} = \{\text{F3,Fz,F4,C3,Cz,C4,} \dots\}$) from the given electrode: $\text{C3}_{\text{CAR}} = \text{C3} - \frac{1}{|\mathcal{C}|} \sum_{C \in \mathcal{C}} C$
- **Laplace (Lap):** Subtract from each channel the average of its immediate neighbours: $\text{C3}_{\text{Lap}} = \text{C3} - {}^1\!/_4(\text{FC3} + \text{C1} + \text{CP3} + \text{C5})$



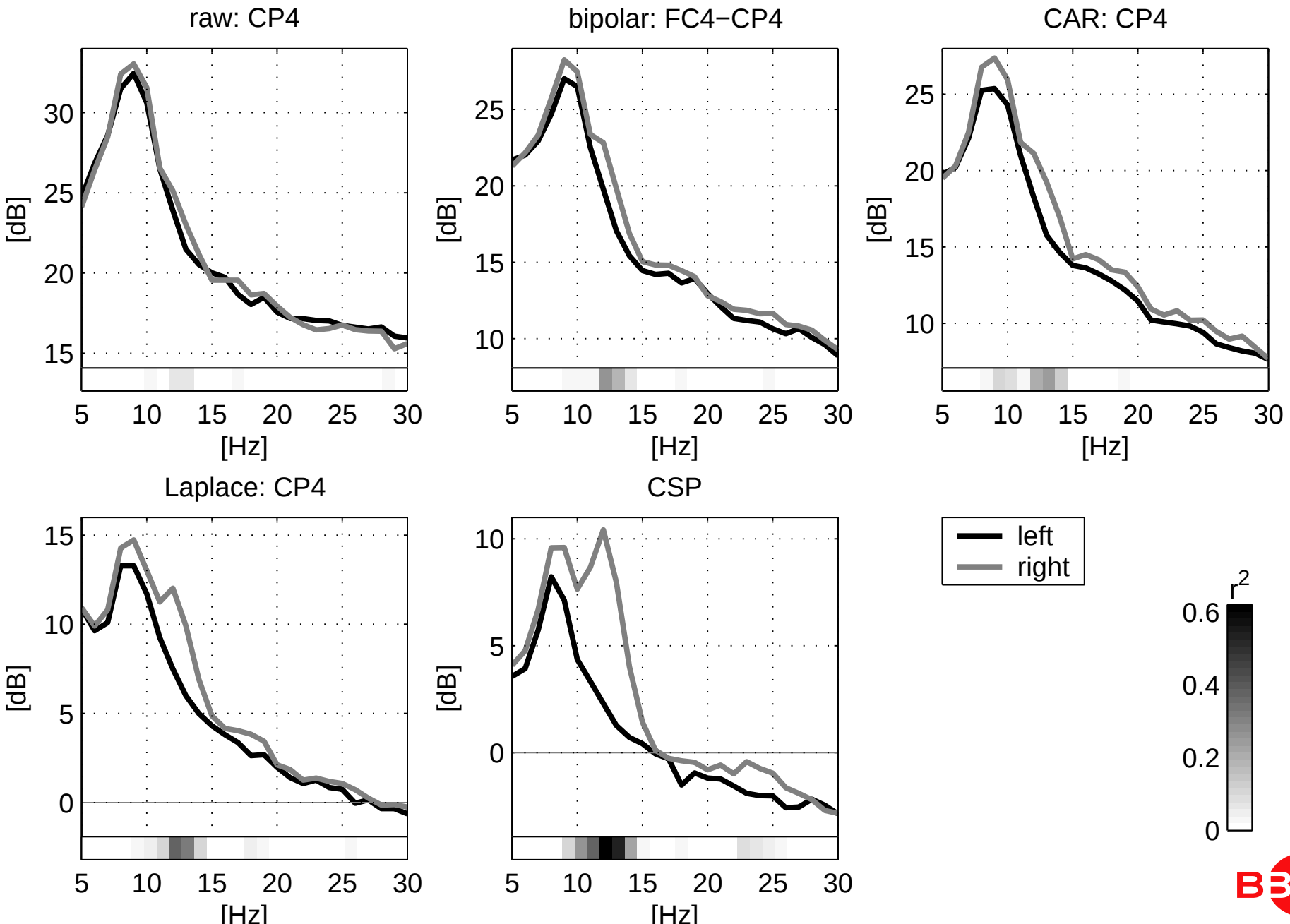




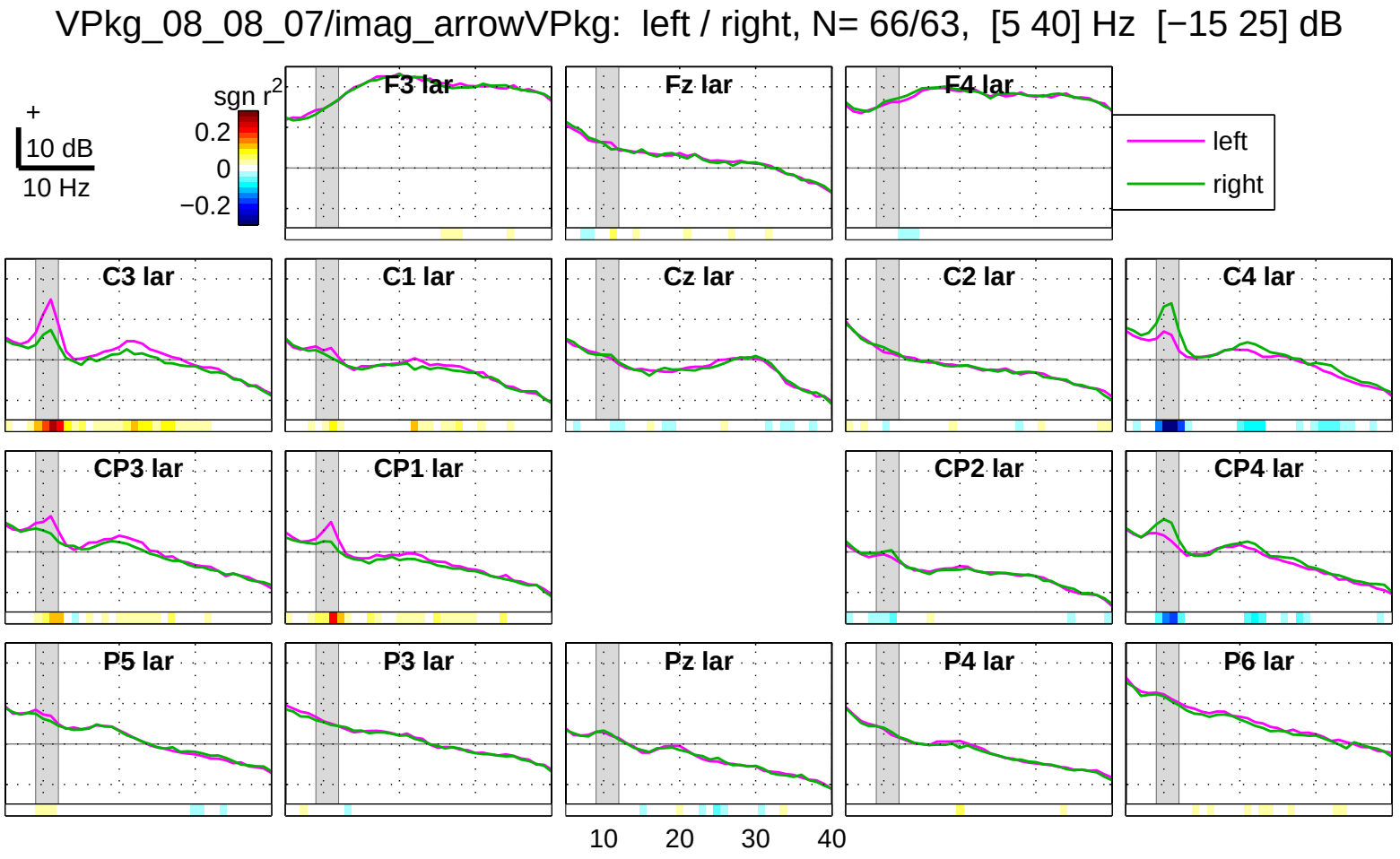
Data Driven Spatial Filters

- **Principal Component Analysis (PCA):** A data-driven method that can be used, e.g., to extract components of most variance in the data, see first lecture.
- **Independent Component Analysis (ICA):** Data-driven methods that extract components of independent activity. If succesful, these components correspond to sources in the brain.
- **Common Spatial Patterns (CSP):** A data-driven method that can be used to find optimized filters that reflect amplitude modulations of brain rythms (topic of today).

The Effect of Spatial Filtering



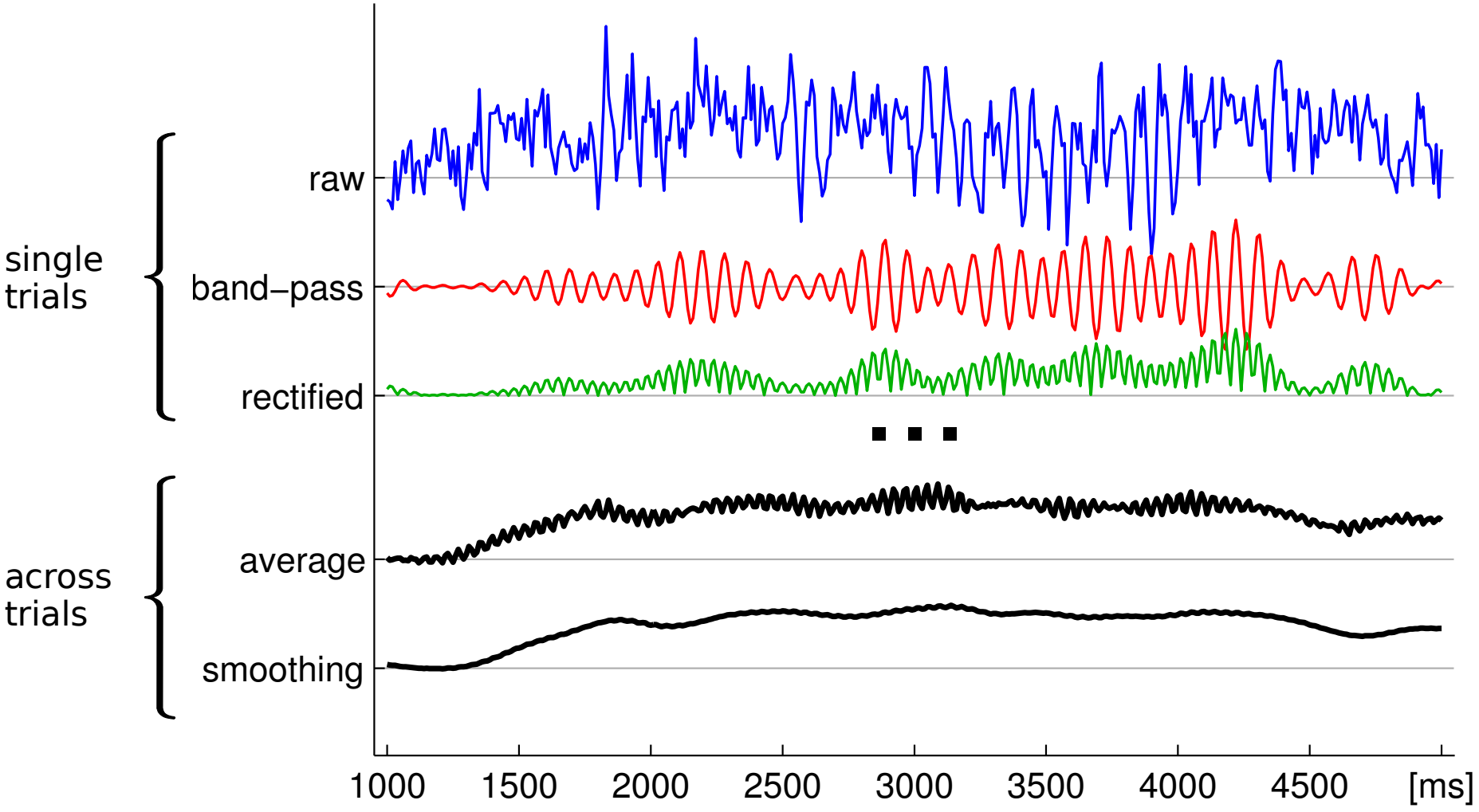
Analysis of Motor Imagery Conditions: Spectra



First step: determine a suitable frequency band that shows good discrimination between the conditions.

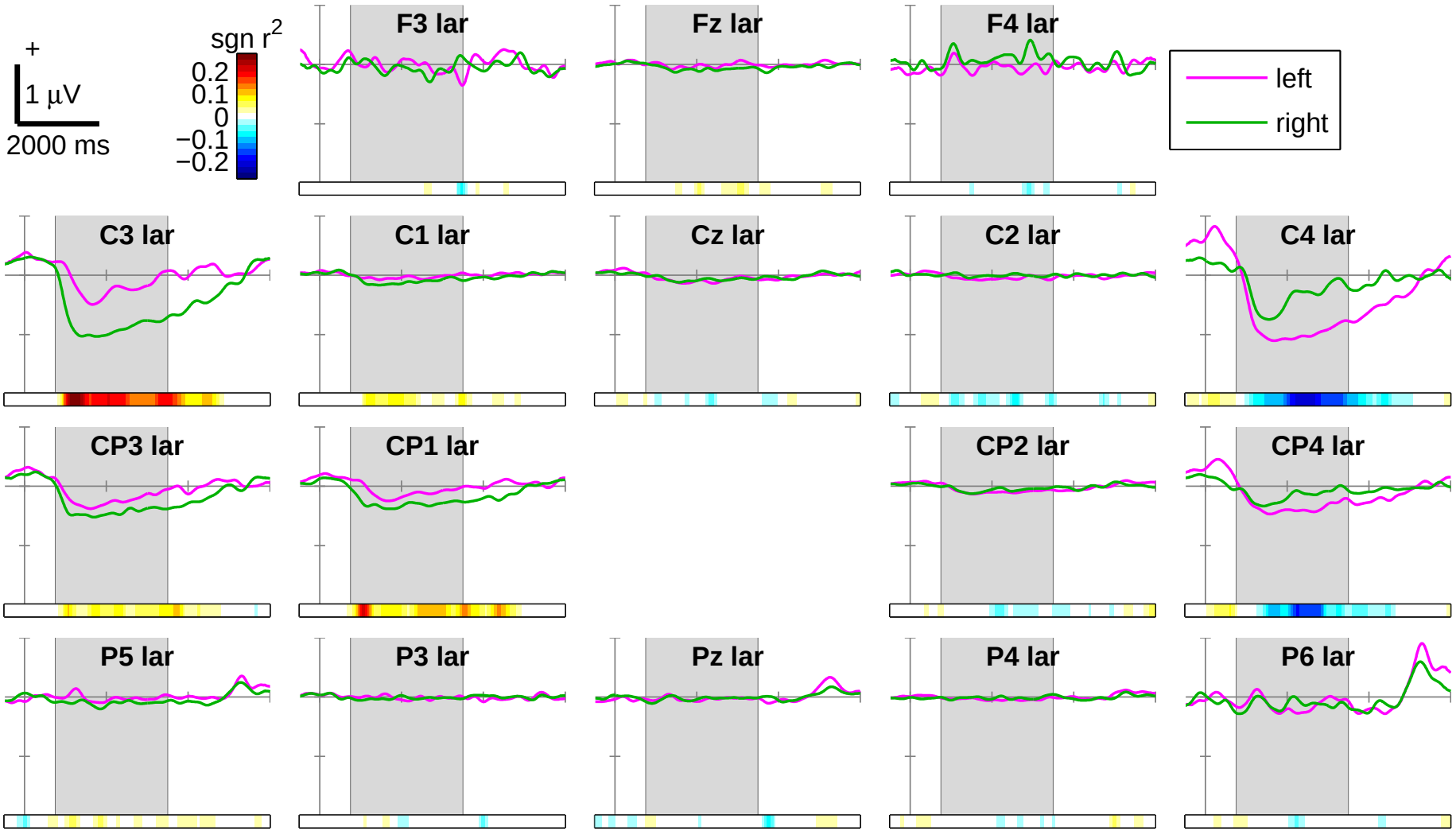
Next, we investigate the time course of band power.

Calculation of ERD/ERS Curves

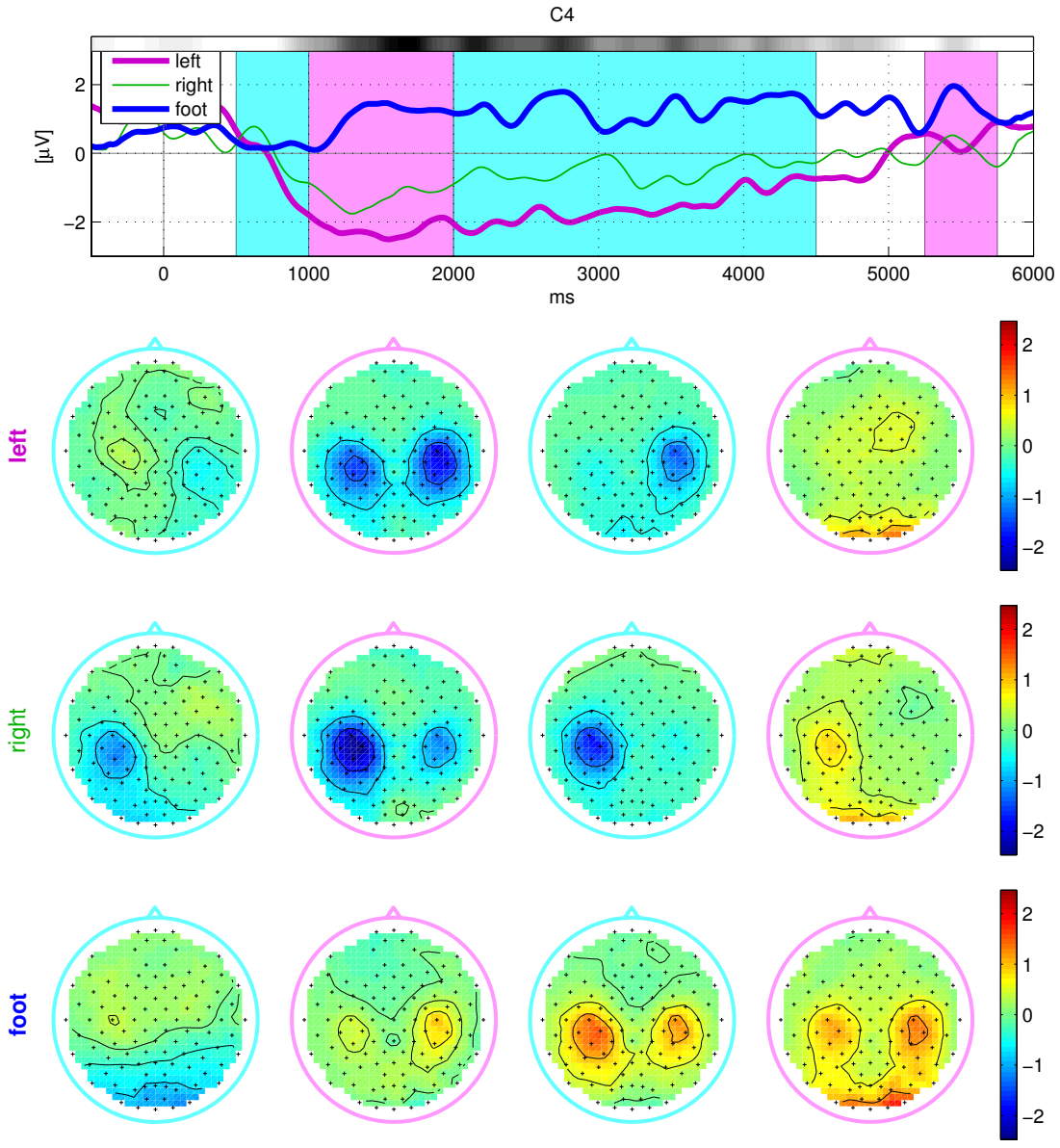


ERD/ERS: Event-Related (De)Synchronization. These curves display time courses of band-power.

ERD/ERS Curves of Motor Imagery



Topography of ERD Curves of Motor Imagery



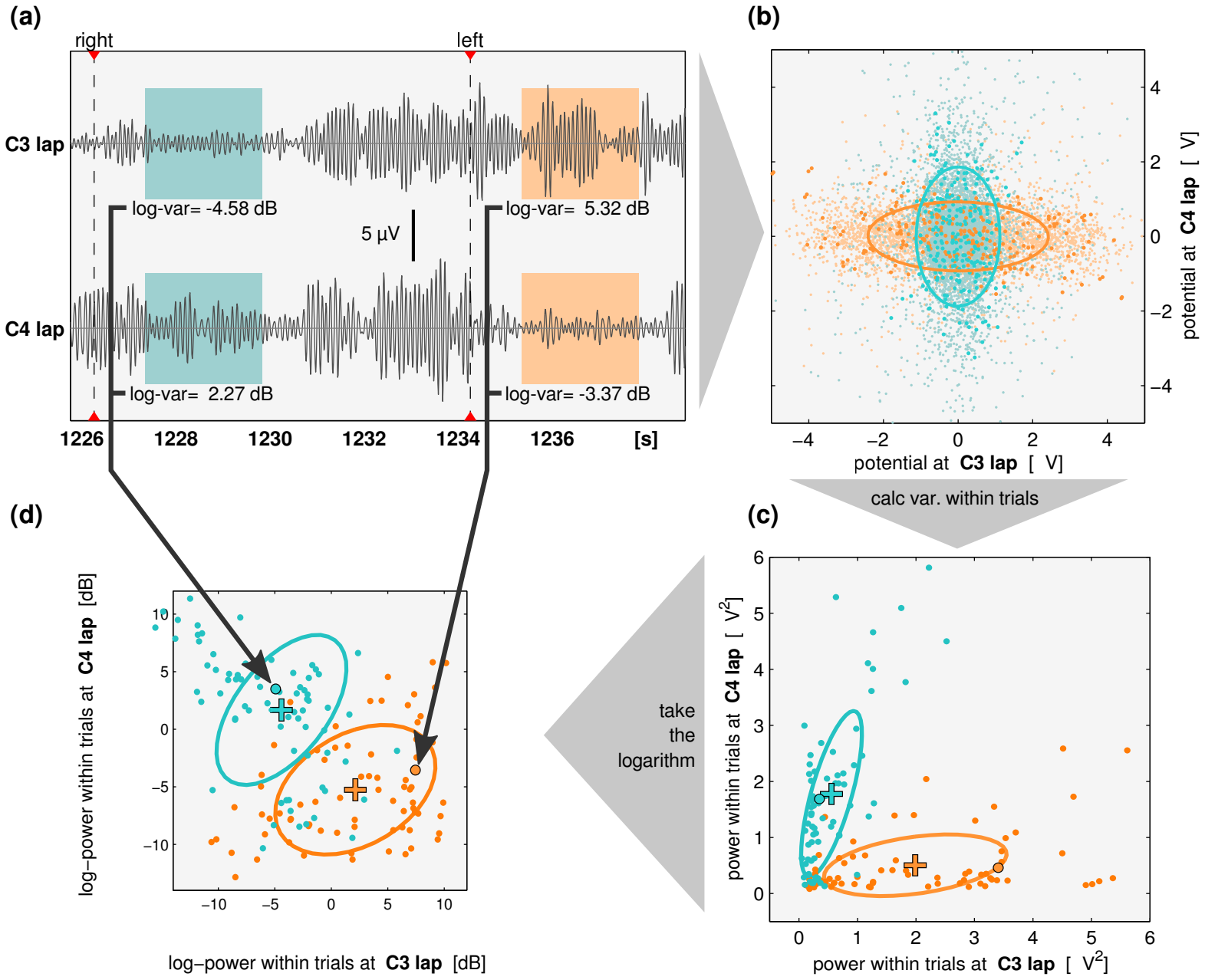
Remark

Using the Hilbert transform, ERD/ERS curves of single-trial can be determined. Those could in principle be classified in the same way as ERPs.

However, this does mostly result in a weak classification performance. The reason for that will become appearant later (e.g., slide “Demixing has to be [...]”).

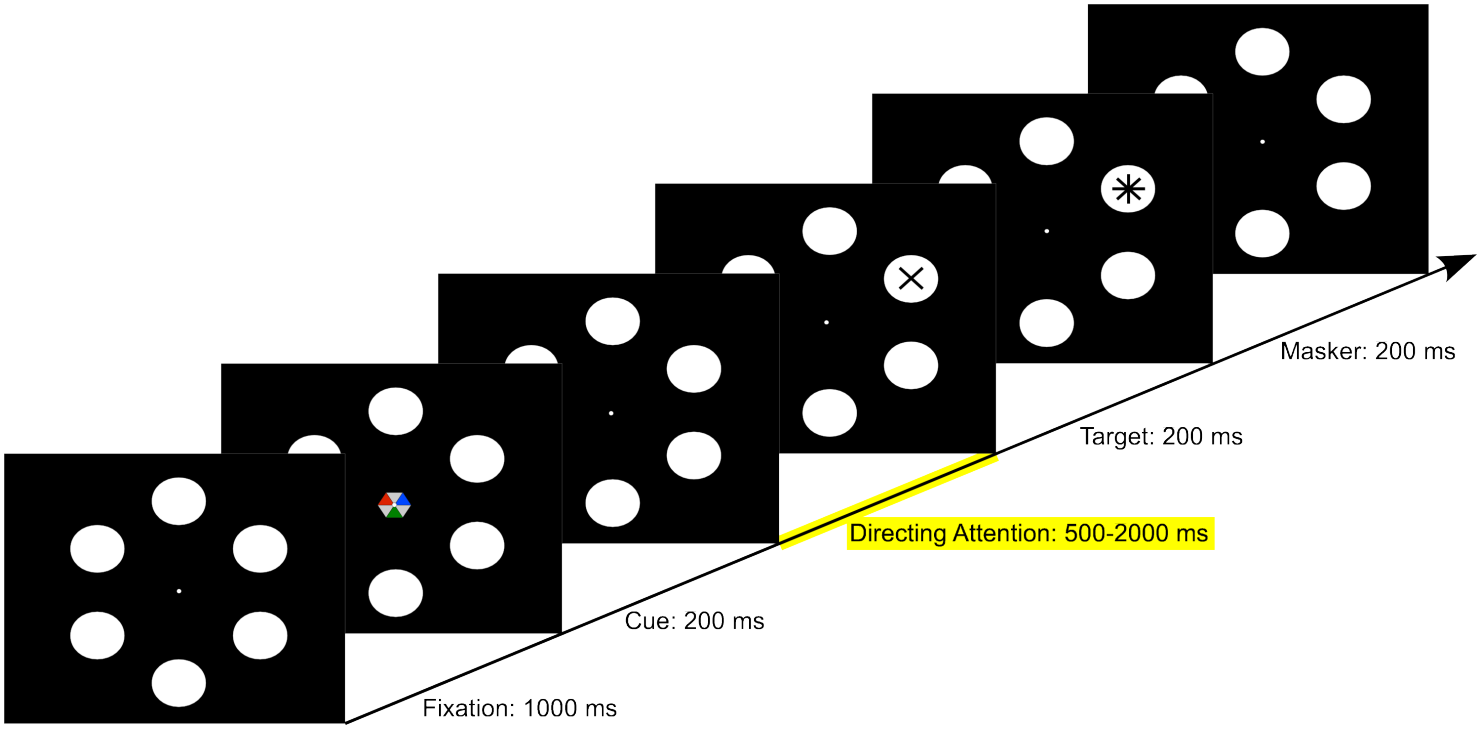
Next, we discuss first the extraction of band-power features disregarding the spatial mixing of sources, and then take care of spatial filtering.

Extraction of log Band-Power Features



Interlude: BCI based on Covert Shits of Attention

Following the MEG experiment of [van Gerven & Jensen, J Neurosci Methods 2009]:



- Study with $N = 8$ participants [Nico Schmidt et al, IEEE SMC 2010].
- After 1000 ms fixation, a cue indicated the direction in which the participant had to covertly shift attention.
- After a variable amount of time (500–2000 ms), a target ('+' or '×') appeared that had to be detected.

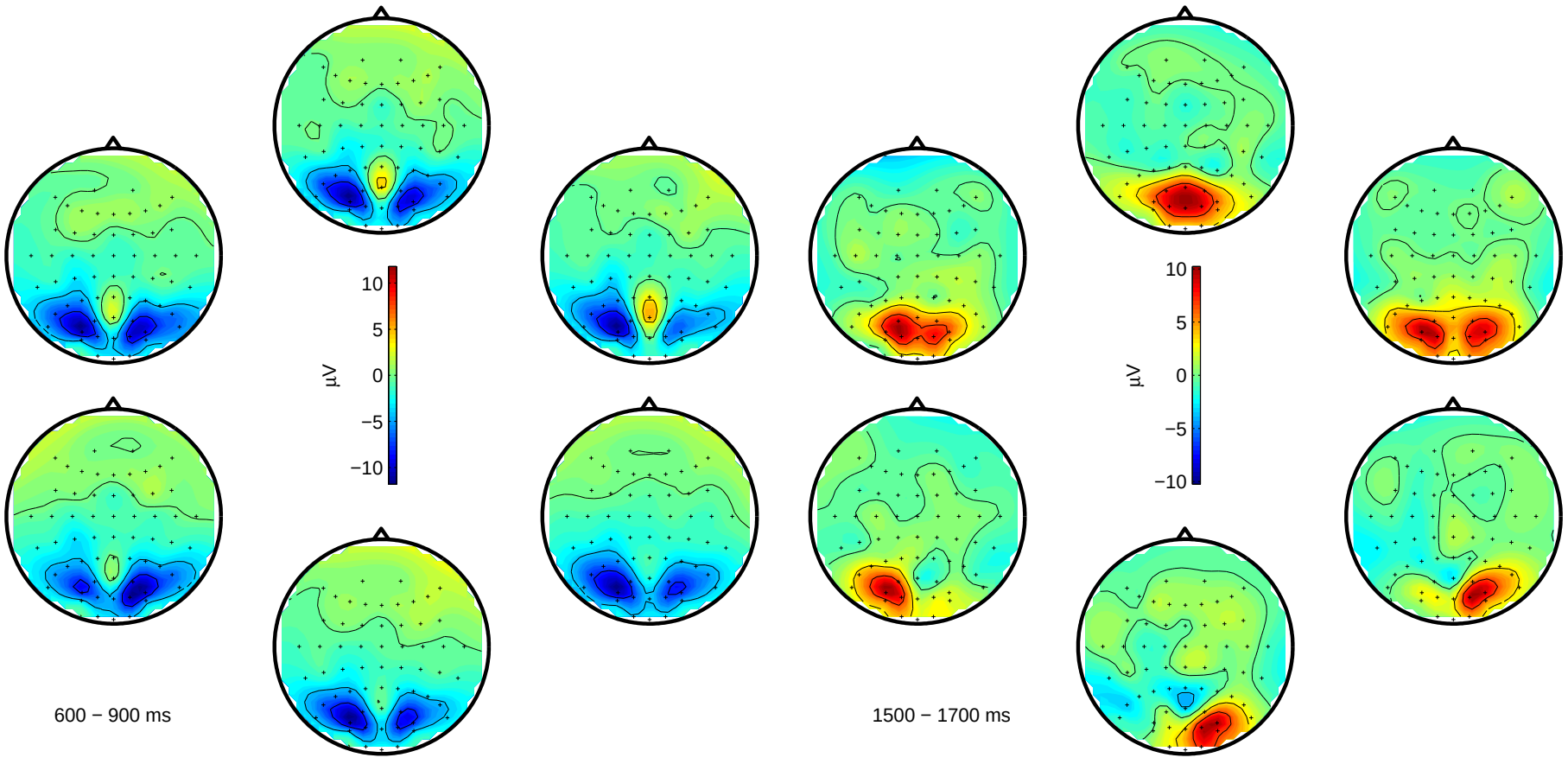
Important Details of Experimental Design

- Fixation of central dot: prevent eye movements.
- Cue-to-target interval was
  - 2000 ms in **50%** of the trials: Long period of focused attention
  - 500 ms in **30%** of the trails: Force quick and time locked shifts of attention
  - >500ms and <2000ms in **20%** of the trials: Sustain attention of the whole time interval
- Cue target direction without a direction specific cue: Each participant had a target color blue, red, or green, which indicated the direction in with s/he had to covertly shift attention.
- The target was replaced after 200 ms by a masker ('$*$') to prevent an afterimage to increase task difficulty (i.e., require more focused attention).
- Response: Indicate type of target ('+' or '$\times$') by button press with the right or left hand: to validate task compliance, in combination with the next point.
- There were **80%** *valid* and **20%** *invalid* trials, in which the target appear at a location different from the cued one to validate task

Behavioral Results

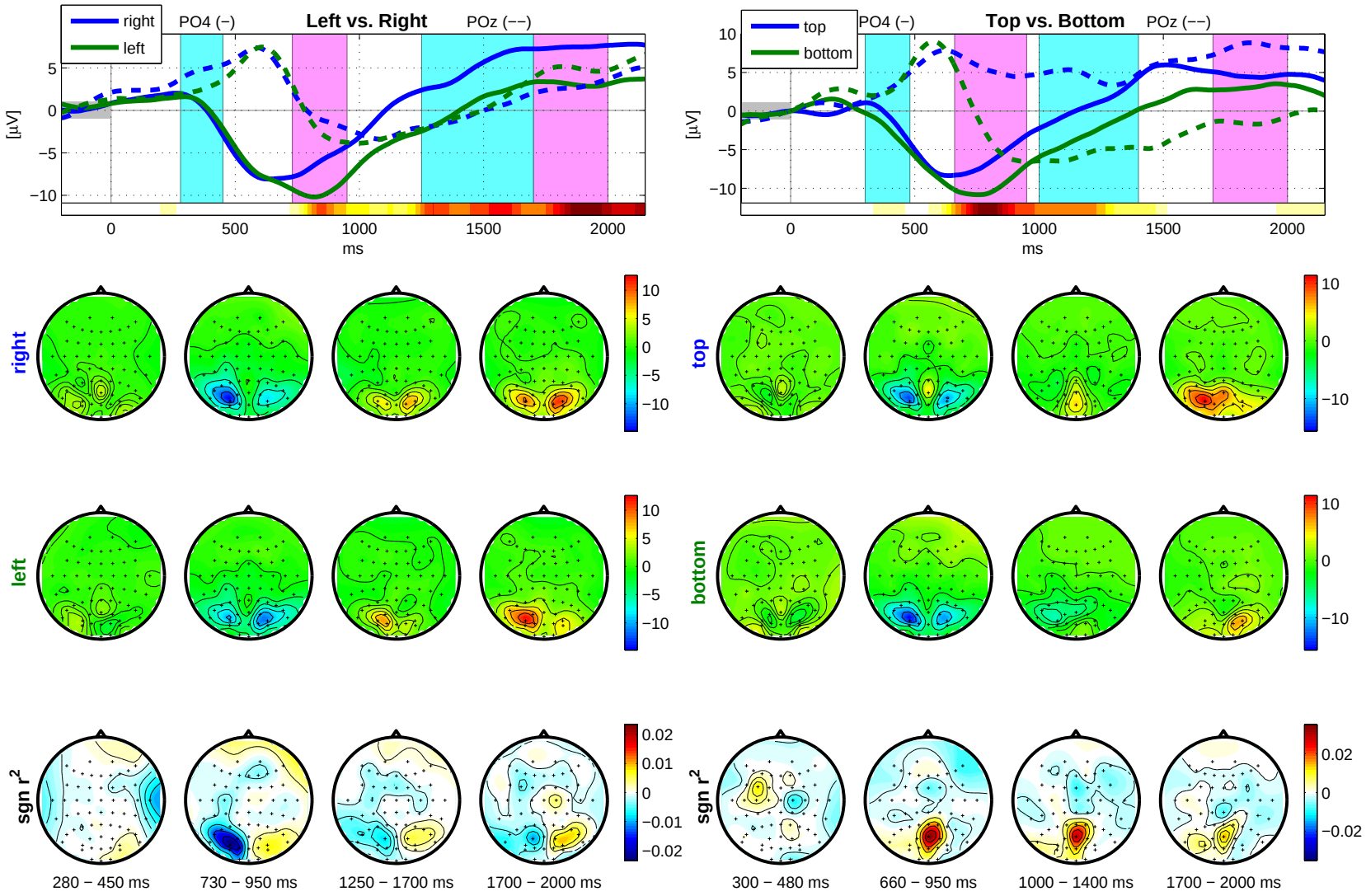
- Overall response accuracy (RACC) was $86.62\% \pm 8.46\%$ SEM.
- Paired t-test: RACC was not significantly different in *valid* vs. *invalid* condition ($p = .199$).
- Geometric means of reaction times were significantly smaller in the *valid* condition than in the *invalid* one ($t = 4.49$, $p < .01$):
  valid: 719 ms $\pm$ 51 ms SEM; *invalid*: 881 ms $\pm$ 76 ms SEM.
- $\Rightarrow$ participants attended correctly the cued positions.

Topographies of Alpha Modulation



- *Left:* Contralateral Alpha ERD 600–900 ms after cue onset
- *Right:* Ipsilateral Alpha ERS 1700–1900 ms after cue onset

Differential Alpha Modulation (L vs R; top vs bottom)

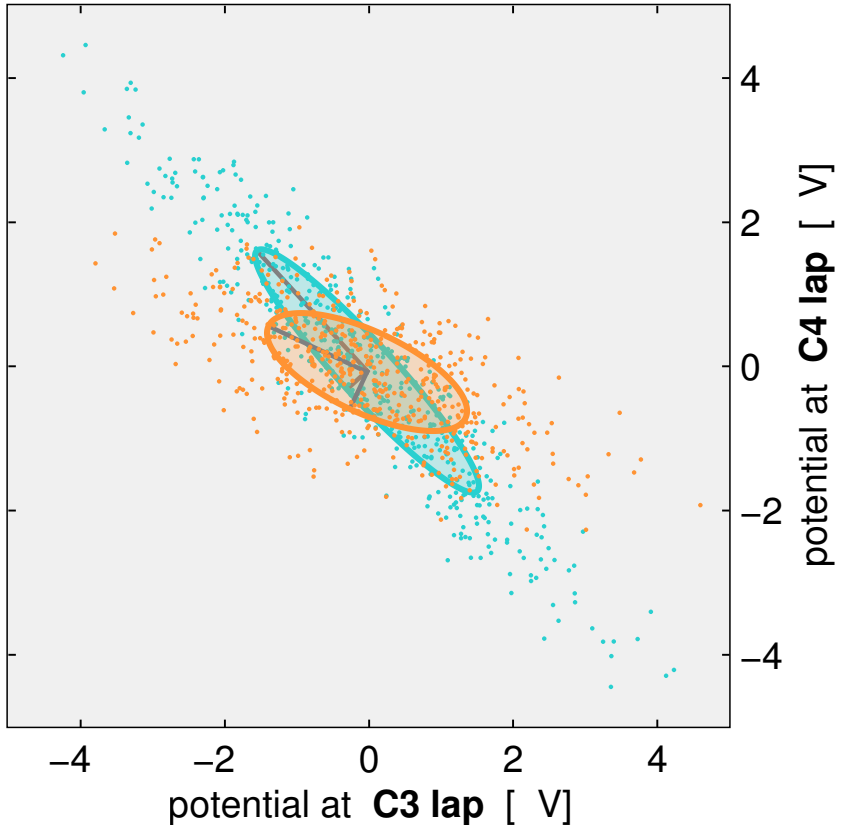


However, even best the binary classification is only 66 to 87%.

[Matthias Treder et al, J Neuroeng Rehabil 2011]

Demixing has to be Performed Before Feature Extraction

Typically, the information is more mixed across channels than in the previous figure (even after Laplace filtering):



Calculating log-variance in those raw channels would make this mixing of information irreversible for subsequent classification.

Preprocessing for Band-Power Features

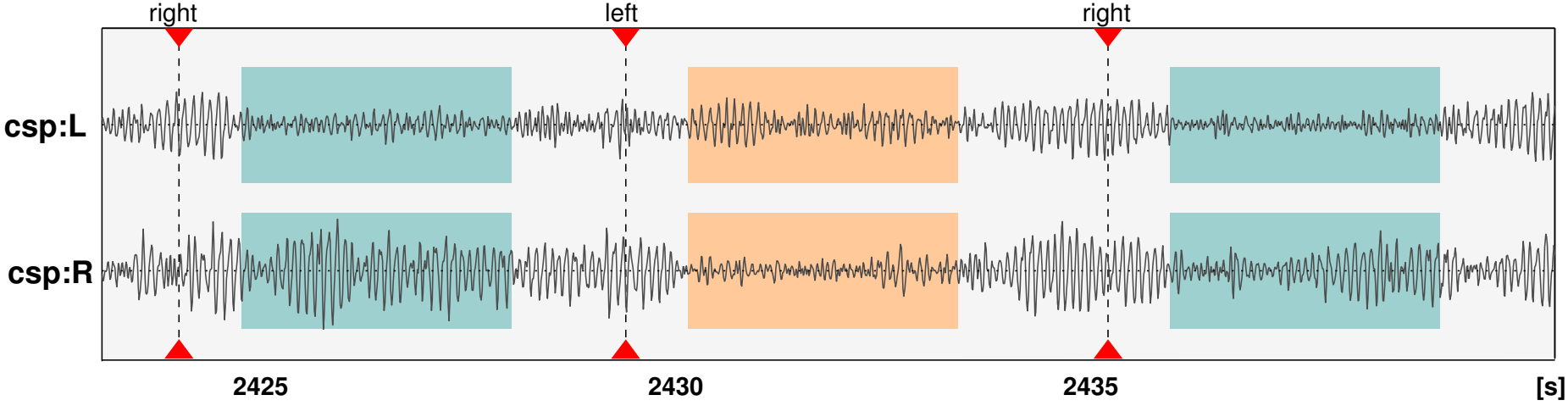
In order to obtain good band-power features, we typically need to apply some spatial filtering **before** calculating *log band-power*.

Note, that for ERP features, spatial filtering was done implicitly by the classifier. The extraction of band-power features involves non-linear processing. In that case, spatial filters have to be applied in advance.

To this end, we will use Common Spatial Patterns (CSP) Analysis [Fukanaga 1990]. The goal of CSP is to determine spatial filters that optimally capture modulations of brain rhythms.

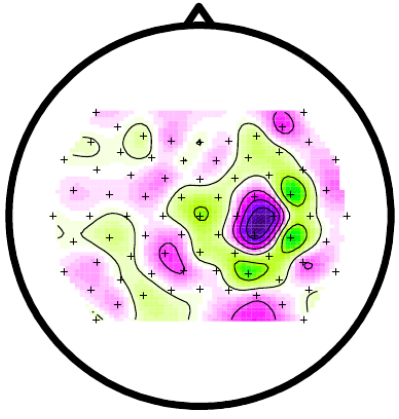
The Goal of Common Spatial Pattern (CSP) Analysis

The goal of CSP: Determine spatial filters such that (log-) variance in an epoch of each filtered signal is indicative of the class.

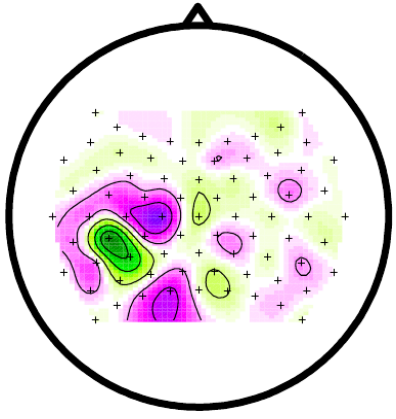


CSP analysis yields spatial filters that can be visualized:

CSP filter **'left'**



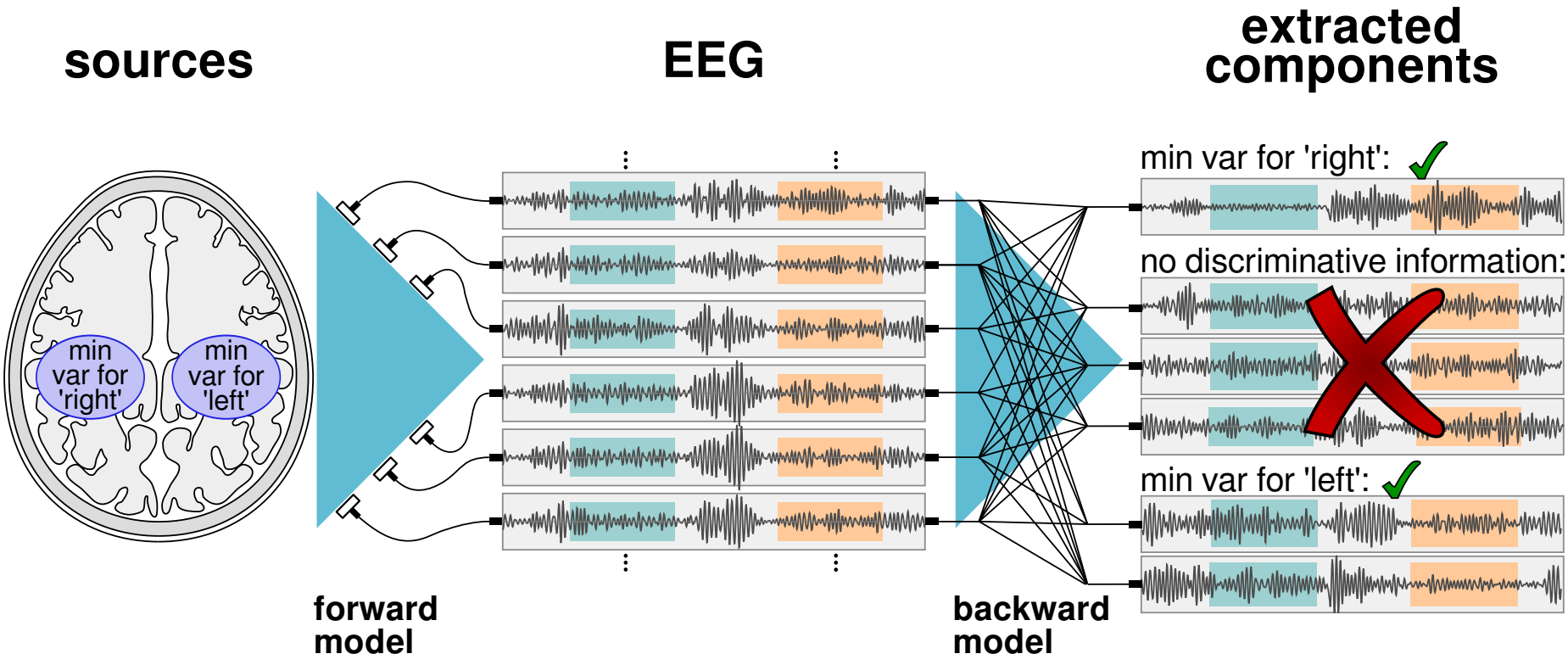
CSP filter **'right'**



But the caveat for the interpreation of filters also applies here.

CSP Analysis Interpretation as Linear Source Model

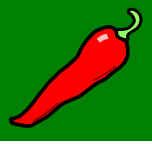
CSP analysis can be interpreted in terms of our linear model of the EEG:



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

Illustration of CSP in 2D

Left: The blue and orange ellipsoids refer to the two class conditional covariance matrices, while the covariance sum is depicted in white. (*Means of the distributions are shifted away from the origin for better illustation.*) **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 blue class, while it is minimal for the orange class and vice versa along the vertical direction.



Generalized Eigenvalue Decomposition

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

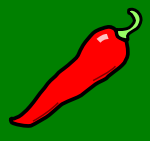
Given $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}$ *symmetric and pos. definite (satisfied for covariance matrices), there exists an invertible matrix* $\mathbf{W} \in \mathbb{R}^{p \times p}$ *and a diagonal matrix* $\mathbf{D} \in \mathit{Diag}(p)$*, such that*

$$\mathbf{AW} = \mathbf{BWD} \quad \& \quad \mathbf{W}^\top \mathbf{BW} = \mathbf{I} \tag{2}$$

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

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

$$\mathbf{W}^\top \mathbf{AW} = \mathbf{W}^\top \mathbf{BWD} = \mathbf{D} \tag{3}$$



CSP with Generalized Eigenvalue Decomposition

Thus, we obtain CSP analysis by performing a generalized Eigenvalue decomposition wrt. the matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$, cf. eqns (2) and (3):

$$\mathbf{W}^\top \mathbf{\Sigma}_1 \mathbf{W} = \mathbf{D} \quad \& \quad \mathbf{W}^\top (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2) \mathbf{W} = \mathbf{I} \tag{4}$$

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

$$\begin{aligned} \mathbf{W}^\top \mathbf{\Sigma}_1 \mathbf{W} &= \mathbf{\Lambda}_1, \qquad \text{with } \mathbf{\Lambda}_1 := \mathbf{D} \\ \mathbf{W}^\top \mathbf{\Sigma}_2 \mathbf{W} &= \mathbf{\Lambda}_2, \qquad \text{with } \mathbf{\Lambda}_2 := \mathbf{I} - \mathbf{D} \end{aligned} \tag{5}$$

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

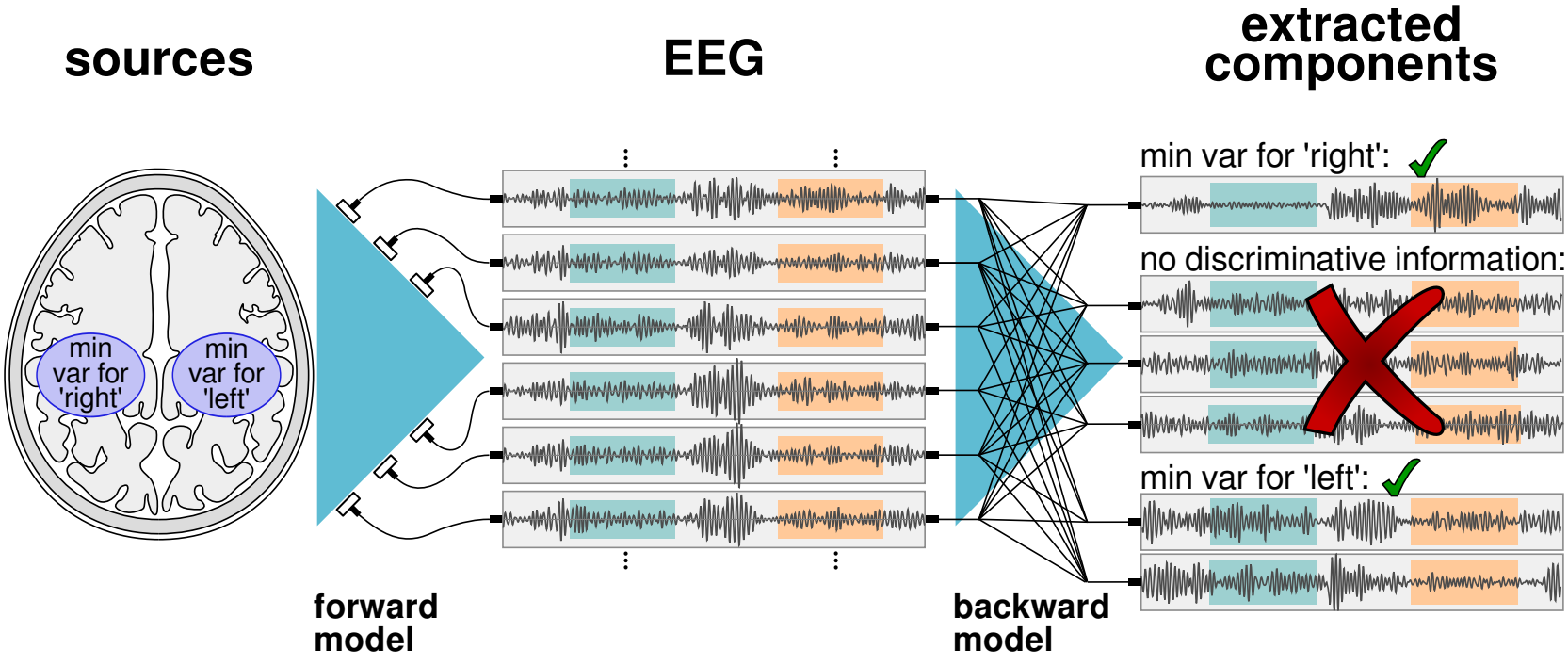
In Matlab this can be done by

```
» [V,D]= eig(Sigma1, Sigma1+Sigma2).
```

CSP Analysis Supported by the Linear Source Model

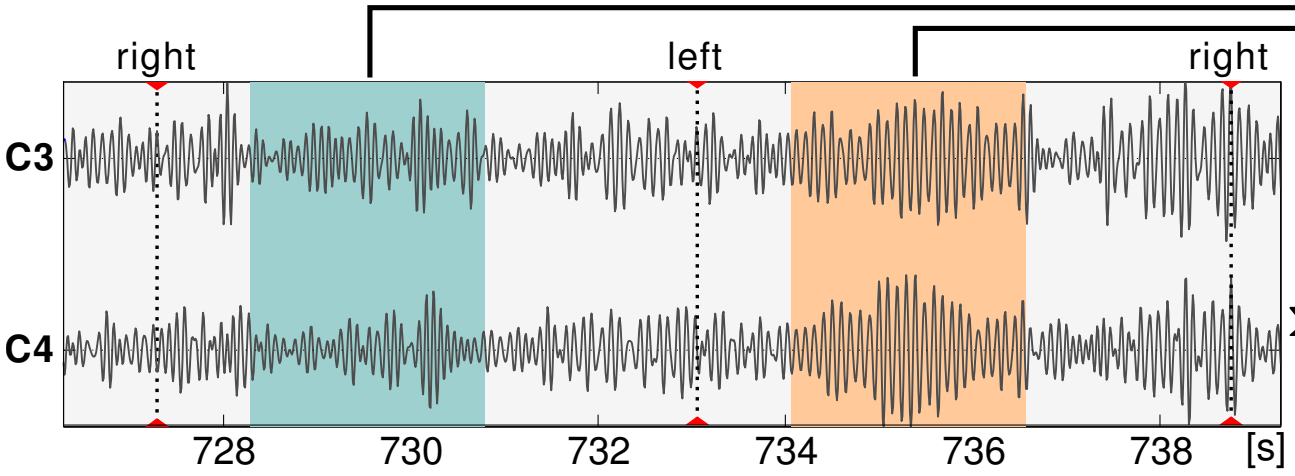
But: Can we hope that there exists a *useful* solution? (If all generalized Eigenvalues are near 0.5, there is no gain of discriminative information.)

According to neurophysiology (see above), the sources in the sensorimotor areas have **low band-power** (small variance in the band-pass filtered signals) during motor imagery of the contralateral hand and **high band-power** for the ipsilateral hand. Accordingly, appropriate CSP filters should exist as backward model.



Practical Wrap-Up of CSP

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



cl #1 (left)
$\mathbf{X}_1$
(chans × T_1)

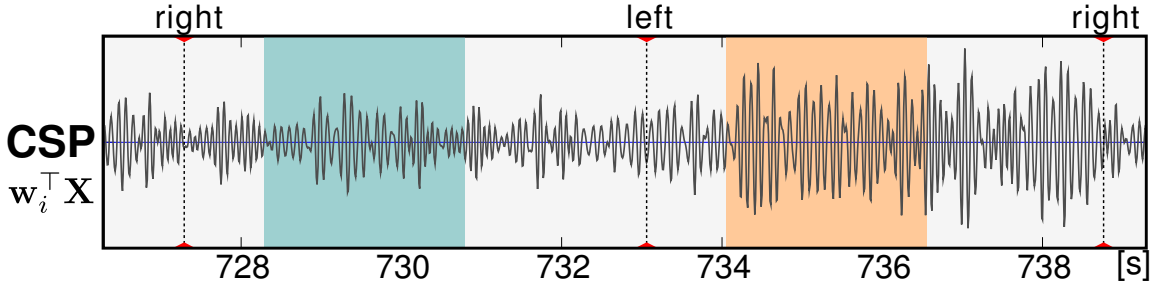
$\mathbf{\Sigma}_1 = \frac{1}{T_1}\mathbf{X}_1\mathbf{X}_1^\top$

cl #2 (right)
$\mathbf{X}_2$
(chans × T_2)

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

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

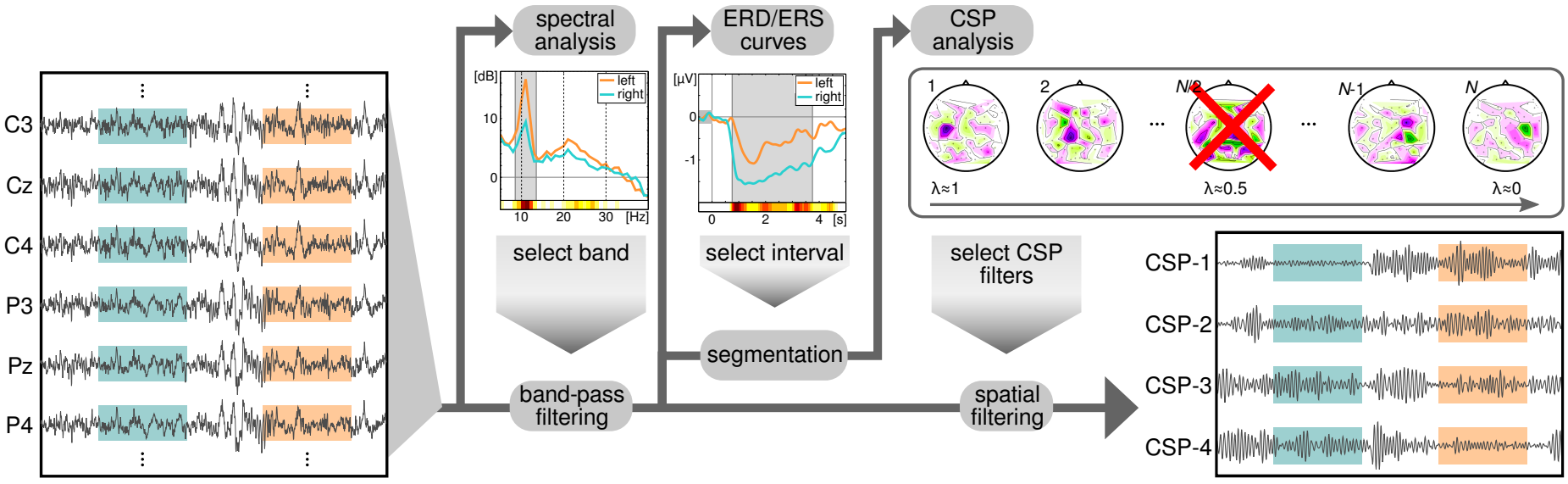
1) choose eigenvector $\mathbf{w}_i$ from $\mathbf{W}$ having a **large** eigenvalue d_i w.r.t. $\mathbf{\Sigma}_1$.



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

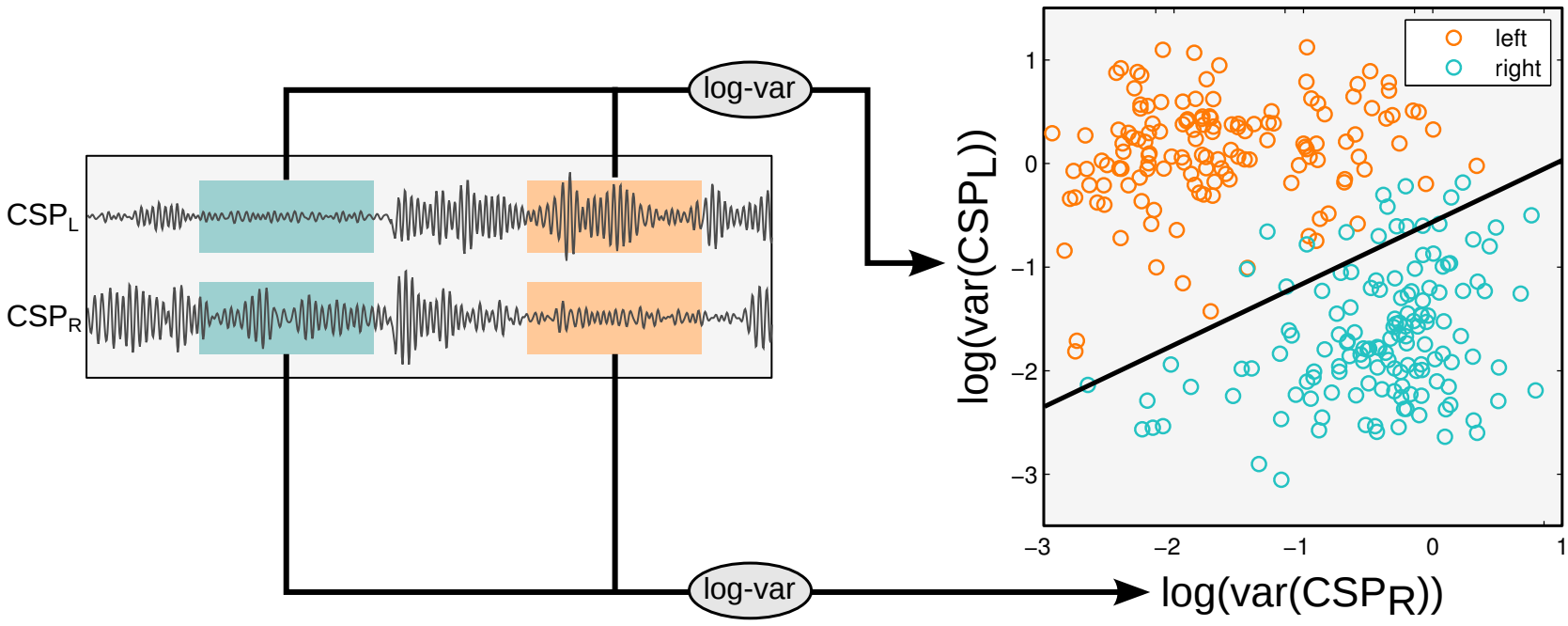
$\mathrm{var}(\mathbf{w}_i^\top\mathbf{X}_2) = 1 - d_i$ small

Work Flow of the Whole CSP Filtering Procedure



Training a Classifier on CSP-based Features

To obtain features from the CSP filtered EEG, in each channel and trial, the variance across time is calculated and the logarithm is applied. On the right there is a scatter plot of the resulting CSP features:



Here, only two dimensions are shown. Note, that applying the logarithm to the band power features makes the distribution more Gaussian and therefore enhances linear separability.

Training of CSP-based Classification

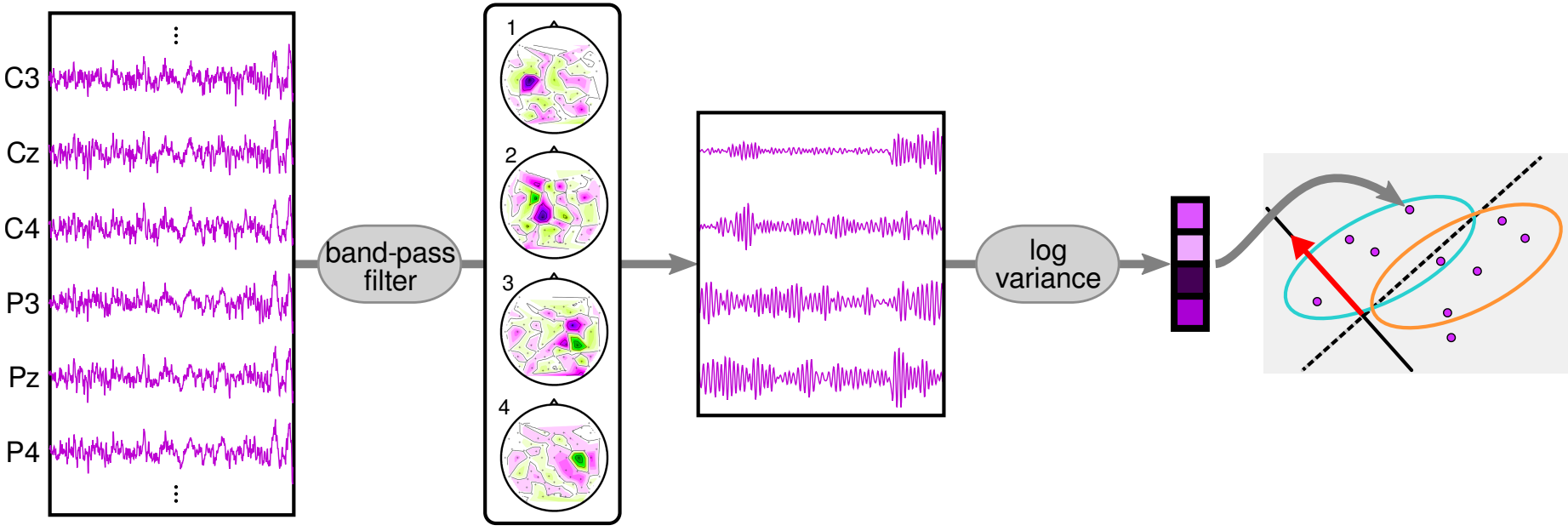
- Determine most discriminative frequency band,
- band-pass filter EEG in that band,
- extract single trials using the time interval in which ERD/ERS is expected,
- calculate and select CSP filters,
- and apply them to EEG single trials,
- calculate the log variance within trials.

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.
  (Since these features are low-dimensional, shrinkage is typically not necessary.)

Applying CSP-based Classification

- Project band-pass filtered EEG with spatial CSP filters,
- calculate the variance in short windows (e.g. last 500 ms),
- take the logarithm,
- and apply the classifier weighting.



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

Caveats in Validation

When machine learning techniques are used for classification of EEG single-trials, the expected performance of a method has to be evaluated carefully, and there are several possible pitfalls.

The estimation of generalization performance requires a training and a test set. The estimation is only proper

- if the test set was not used in any way to determine parameters of the method, and
- if the samples in the test set are independent from the samples in the training set.

Although these principles are quite obvious, it happens that they are violated (mostly unintentionally).

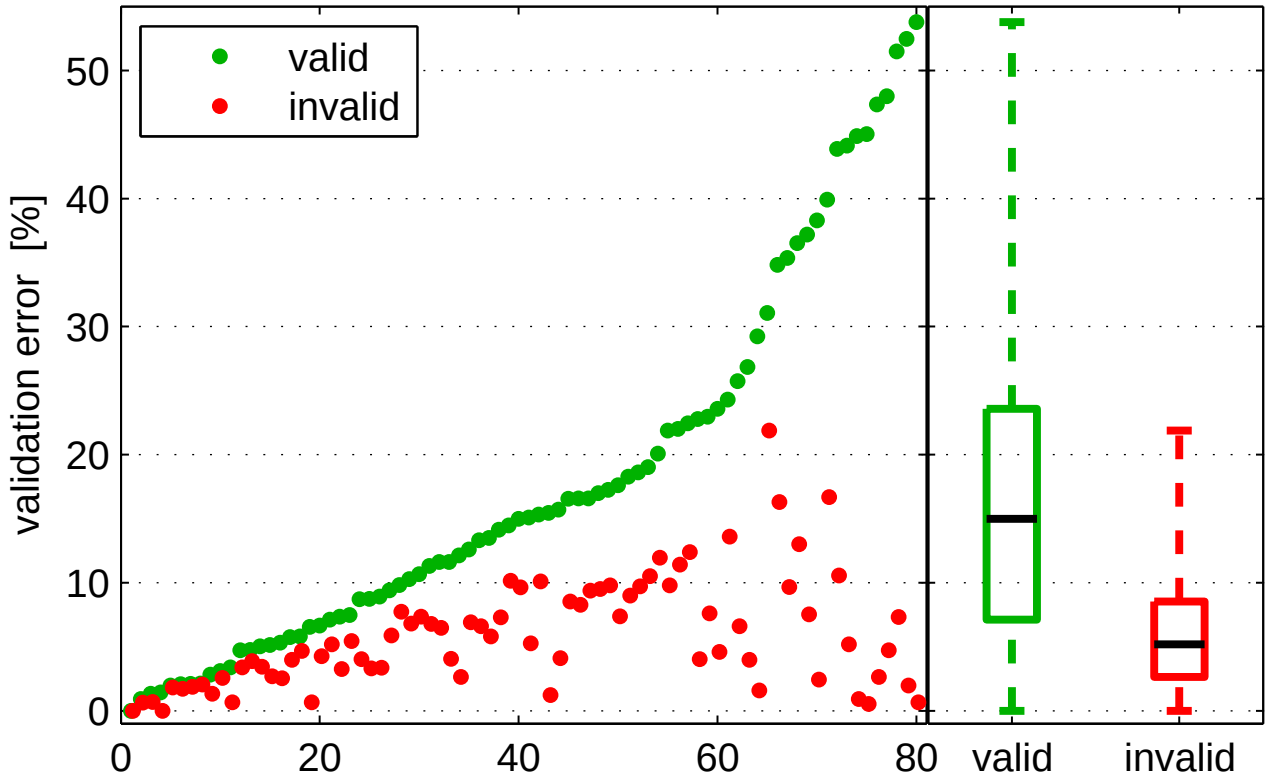
Validating a CSP-based Classification Method

Crucial measure: **generalization performance**, i.e., the accuracy obtained when the classifier is applied to new data, which have not been used in any way before.

Note: When a preprocessing step (like CSP) uses the class labels, it needs to be performed on the training sets only! Take the spatial filter obtained by CSP on the 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.

Demo: Faulty 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.

References I

- Blankertz, B., Lemm, S., Treder, M. S., Haufe, S., and Müller, K.-R. (2011).
  Single-trial analysis and classification of ERP components – a tutorial.
  Neuroimage, 56:814–825.

- 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.

- Fukunaga, K. (1990).
  Introduction to statistical pattern recognition.
  Academic Press, Boston, 2nd edition edition.

- Ledoit, O. and Wolf, M. (2004).
  A well-conditioned estimator for large-dimensional covariance matrices.
  J Multivar Anal, 88:365–411.

- Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011).
  Introduction to machine learning for brain imaging.
  Neuroimage, 56:387–399.

- Neuper, C. and Klimesch, W., editors (2006).
  Event-related Dynamics of Brain Oscillations.
  Elsevier.

- Odom, J., Bach, M., Barber, C., Brigell, M., Marmor, M., Tormene, A., Holder, G., and Vaegan (2004).
  Visual evoked potentials standard (2004).
  Doc Ophthalmol, 108(2):115–123.

References II

- Pfurtscheller, G. and da Silva, F. H. L. (1999).
  Event-related EEG/MEG synchronization and desynchronization: basic principles.
  Clin Neurophysiol, 110(11):1842–1857.

- Schäfer, J. and Strimmer, K. (2005).
  A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.
  Stat Appl Genet Mol Biol, 4:Article32.

- Schmidt, N. M., Blankertz, B., and Treder, M. S. (2010).
  Alpha-modulation induced by covert attention shifts as a new input modality for EEG-based BCIs.
  In *Proceedings of the 2010 IEEE Conference on Systems, Man and Cybernetics (SMC2010)*, pages 481–487.

- Treder, M. S., Bahramisharif, A., Schmidt, N. M., van Gerven, M., and Blankertz, B. (2011a).
  Brain-computer interfacing using modulations of alpha activity induced by covert shifts of attention.
  J Neuroeng Rehabil, 8:24.

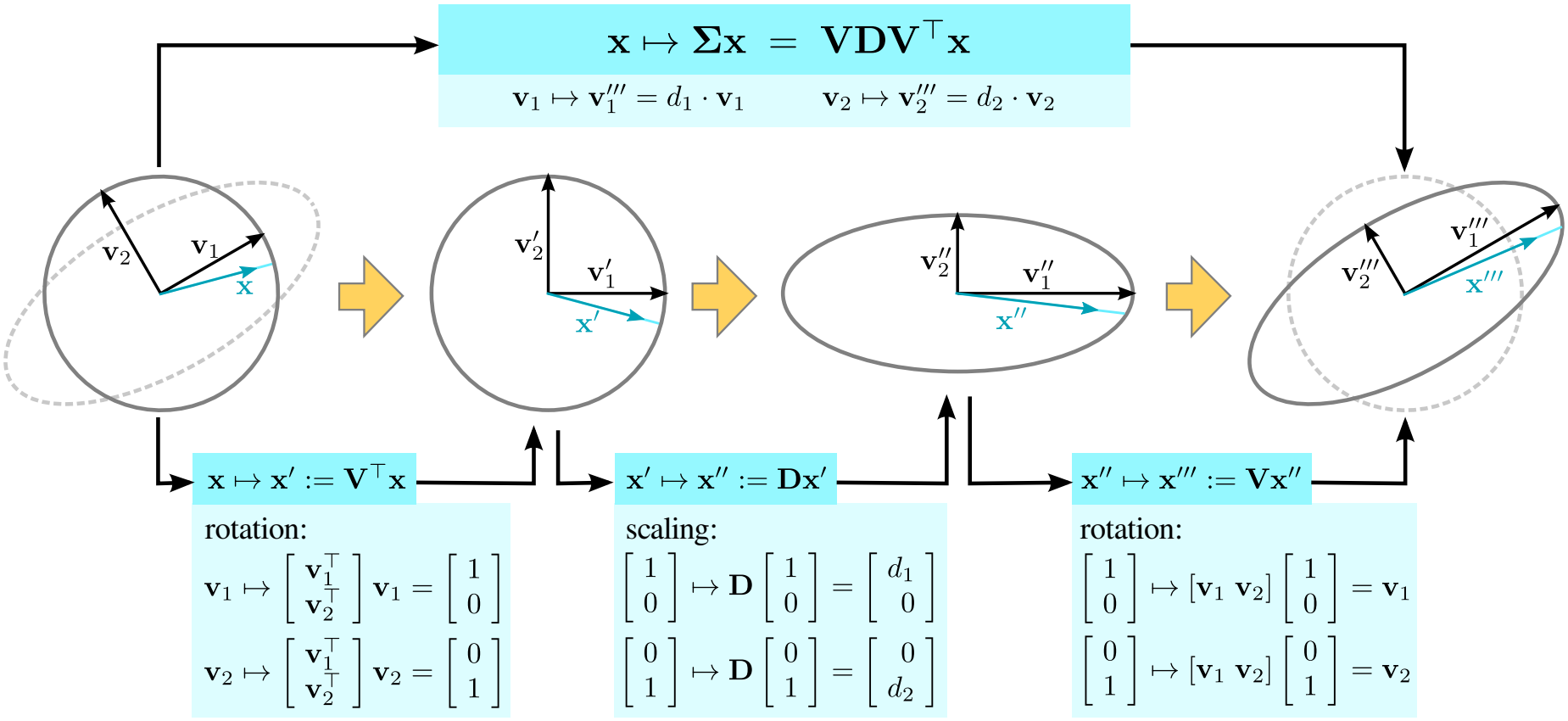
- Treder, M. S. and Blankertz, B. (2010).
  (C)overt attention and visual speller design in an ERP-based brain-computer interface.
  Behav Brain Funct, 6:28.

- Treder, M. S., Schmidt, N. M., and Blankertz, B. (2011b).
  Gaze-independent brain-computer interfaces based on covert attention and feature attention.
  J Neural Eng, 8(6):066003. Open Access.

- van Gerven, M. and Jensen, O. (2009).
  Attention modulations of posterior alpha as a control signal for two-dimensional brain-computer interfaces.
  J Neurosci Methods, 179:78–84.

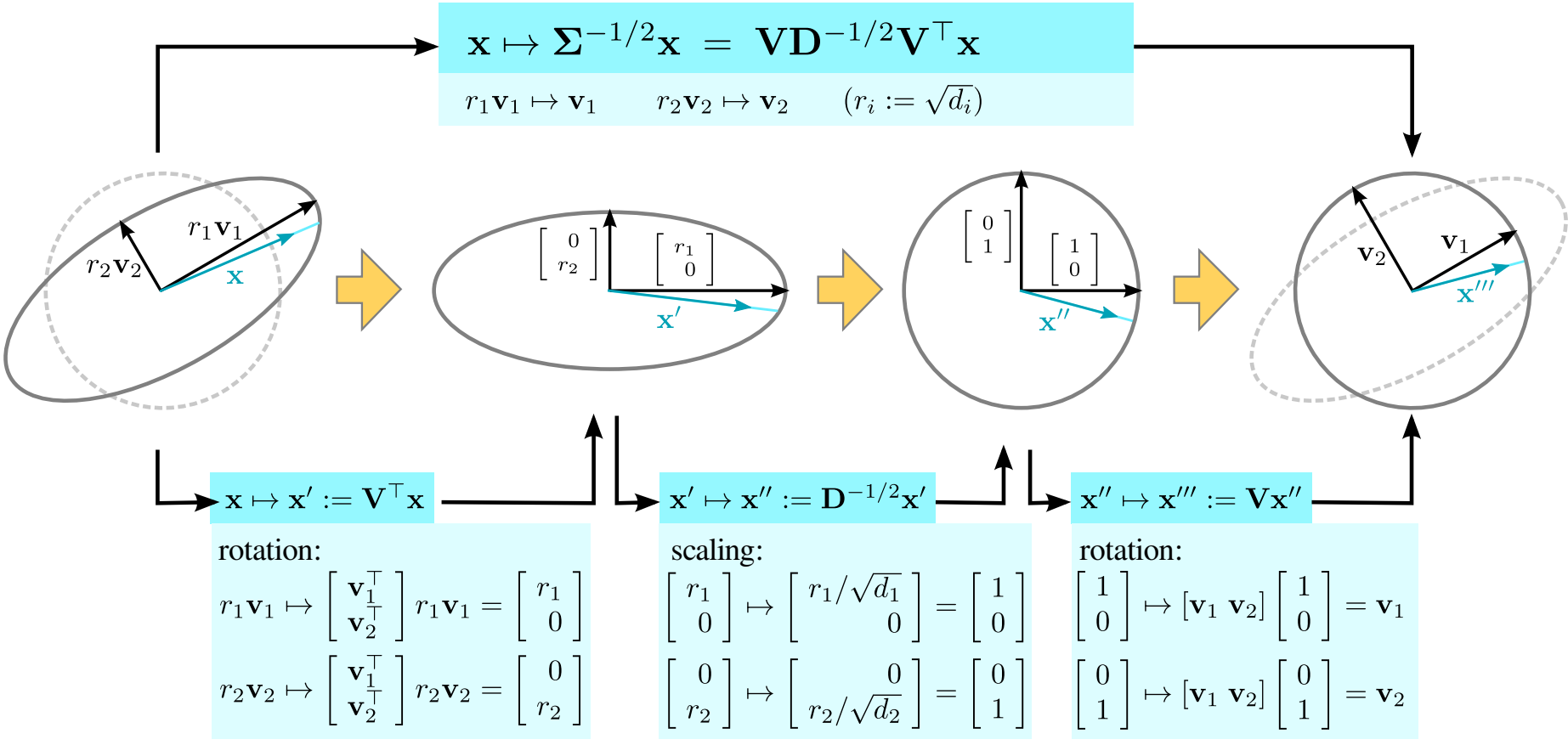
Appendix

Illustration of Multiplication by Covariance Matrix

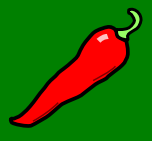


1. Step. The multiplication of a vector with the orthonormal matrix $\mathbf{V}^\top$ is a rotation. The calculation shows, that the rotation is defined by mapping the Eingevectors $\mathbf{v}_i$ to the coordinate axes. *2. Step.* The multiplication of a vector with the diagonal matrix $\mathbf{D}$ is a scaling along the coordinate axes. *3. Step.* The multiplication with $\mathbf{V}$ is the inverse rotation to the multiplication with $\mathbf{V}^\top$ (due to orthonormality). This means the coordinate axes are mapped 'back' to the Eigenvectors.

Illustration of Whitening Transform



The whitening transform maps the space such that a Gaussian distribution with the given covariance matrix becomes a stanard normal distribution, i.e., the variance in all directions is one. It maps the ellipsoid given by the standard isodensity line of the Gaussian distribution to the unit sphere.



Another view: CSP as Optimization Problem

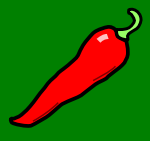
Let $\mathbf{X}_1 \in \mathbb{R}^{C \times T_1}$ be the concatenation of all band-pass filtered trials of class 1 along the time dimension (T_1 is the total number of time points of all trials of class 1, and C being the number of channels), and let $\mathbf{X}_2$ de defined analogously for class 2.

$\mathbf{\Sigma}_i = \frac{1}{T_i}\mathbf{X}_i\mathbf{X}_i^\top \in \mathbb{R}^{C \times C}$ 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:

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

This optimization is solved on the next slide.



CSP with the Rayleigh Coefficient

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

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

In order to obtain CSP filters, we need to find the min (resp. max) of R.

*The **Min-Max Theorem** states: $\lambda_1 \leq R_{\mathbf{A},\mathbf{B}}(\mathbf{w}) \leq \lambda_C$, if $\lambda_1 \leq \cdots \leq \lambda_C$ are the **generalized Eigenvalues** of $\mathbf{A}$ and $\mathbf{B}$.*

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

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

Accordingly, the min (max) of R is attained for $\mathbf{w}_1$ (for $\mathbf{w}_C$). However, practically CSP is determined by Eigenvalue decomposition as above.