Gentle Introduction to Signal Processing and Classification for Single-Trial EEG Analysis

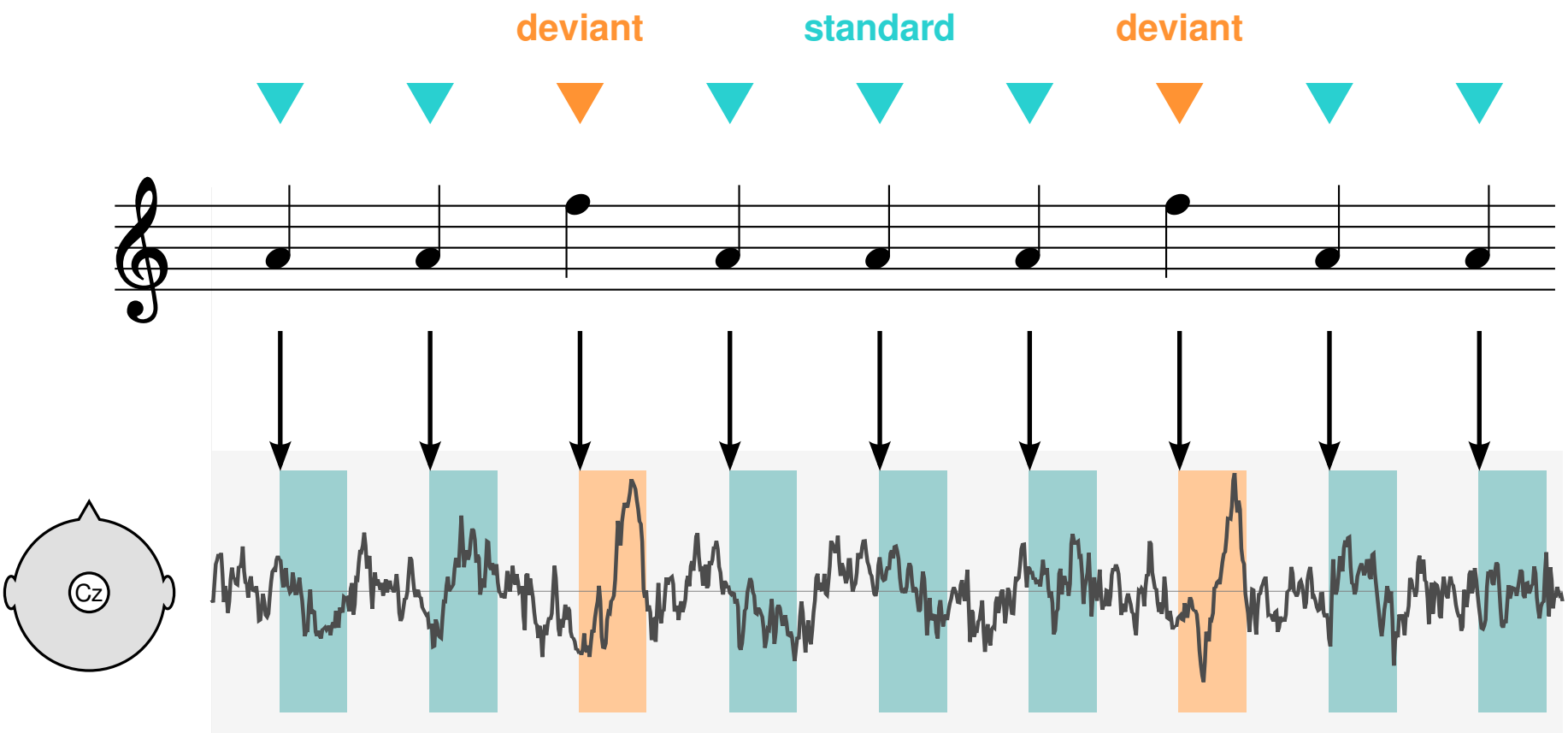
Benjamin Blankertz

Neurotechnology Group, Berlin Institute of Technology

benjamin.blankertz@tu-berlin.de

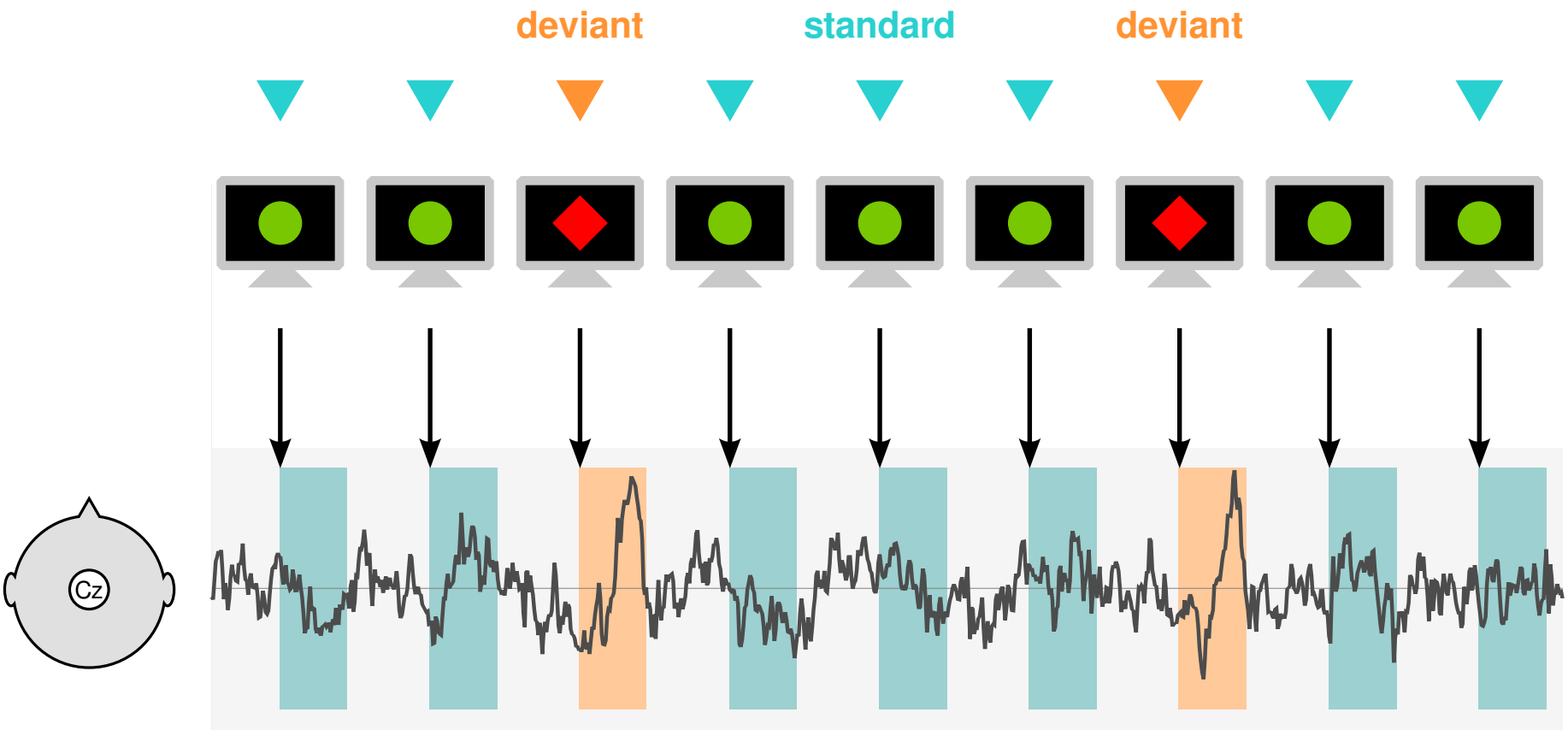
20|Sep|2012

Basics: Oddbal Paradigm, P300, BCI Speller

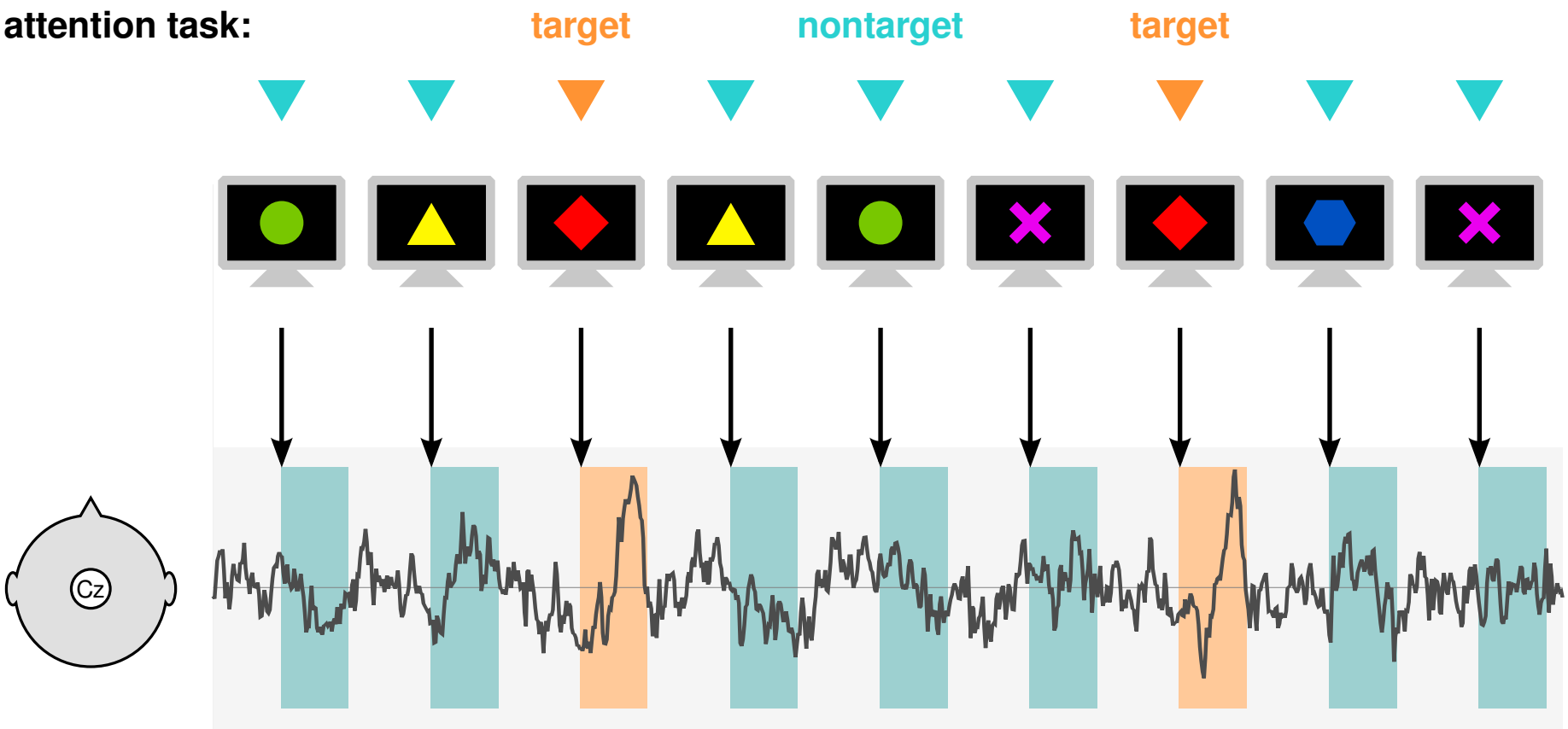


- Segments of the signals are called *epochs* or *single-trials*.

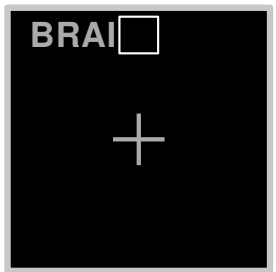
Basics: Oddbal Paradigm, P300, BCI Speller

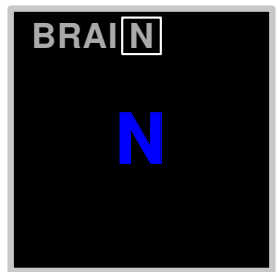


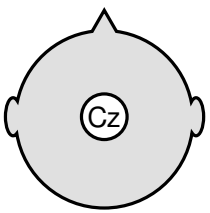
Basics: Oddbal Paradigm, P300, BCI Speller

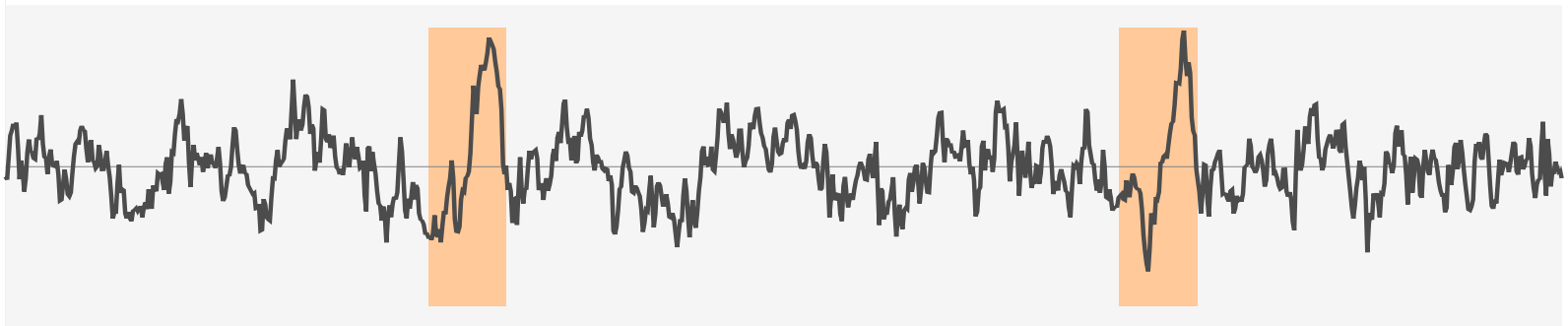


Basics: Oddbal Paradigm, P300, BCI Speller

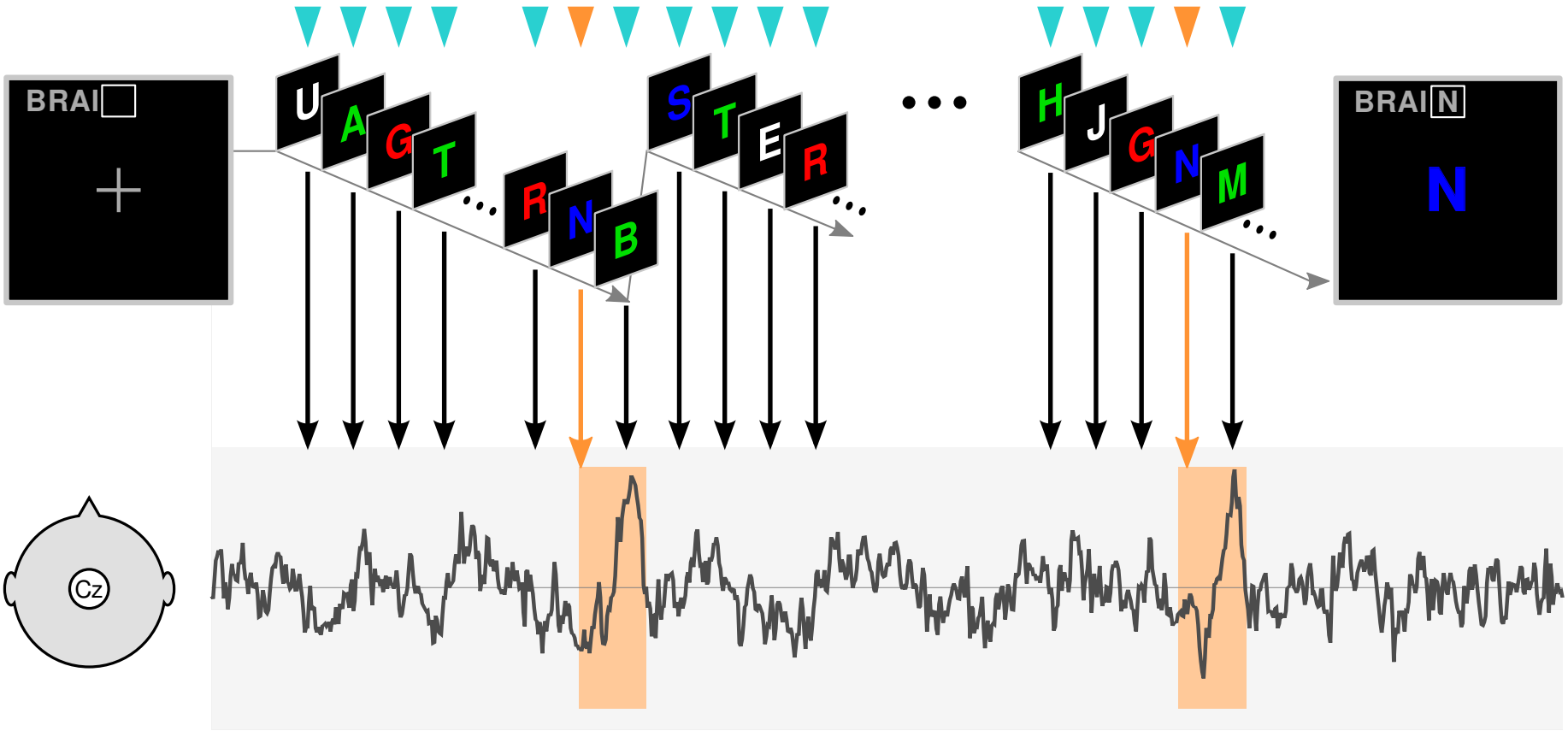






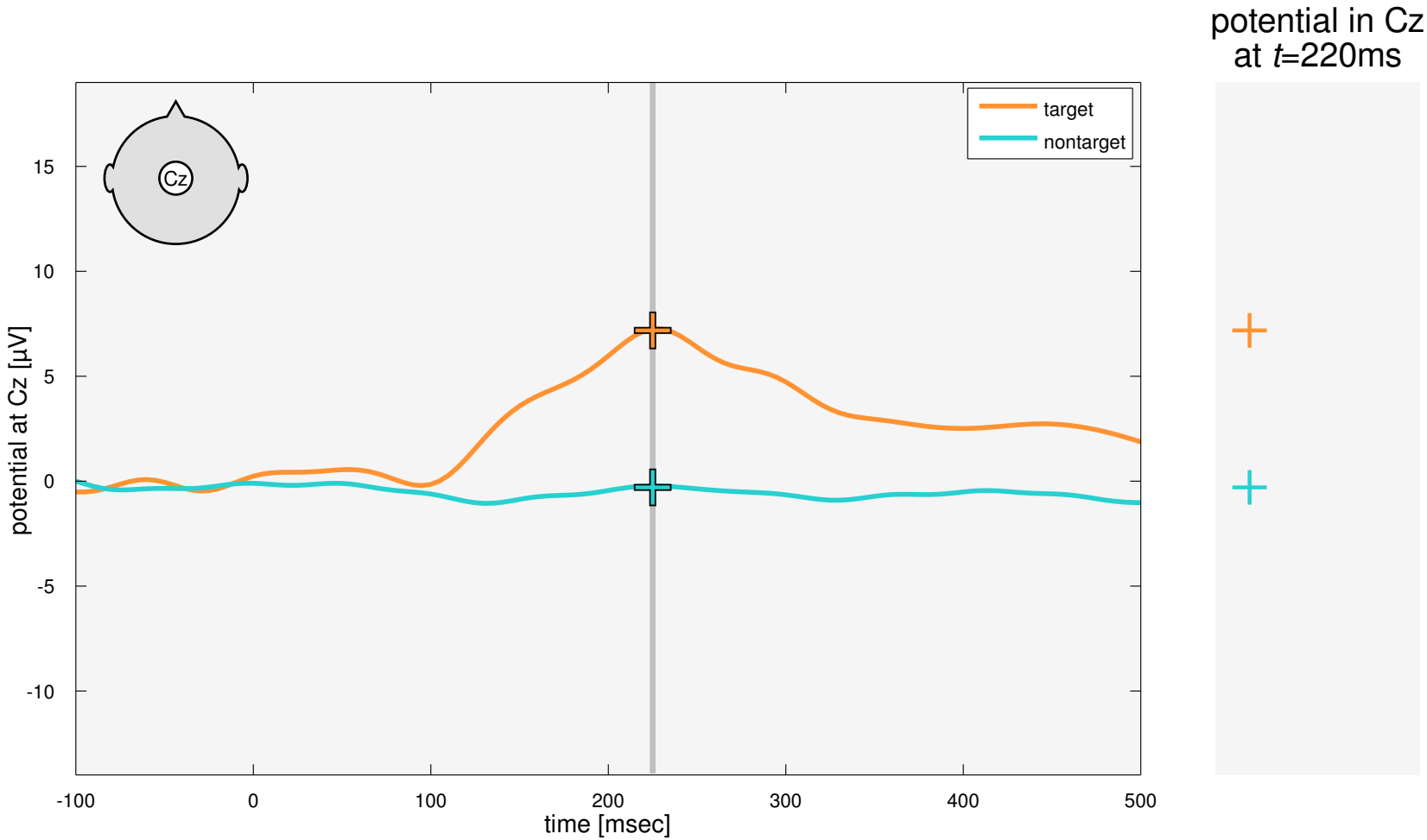


Basics: Oddbal Paradigm, P300, BCI Speller



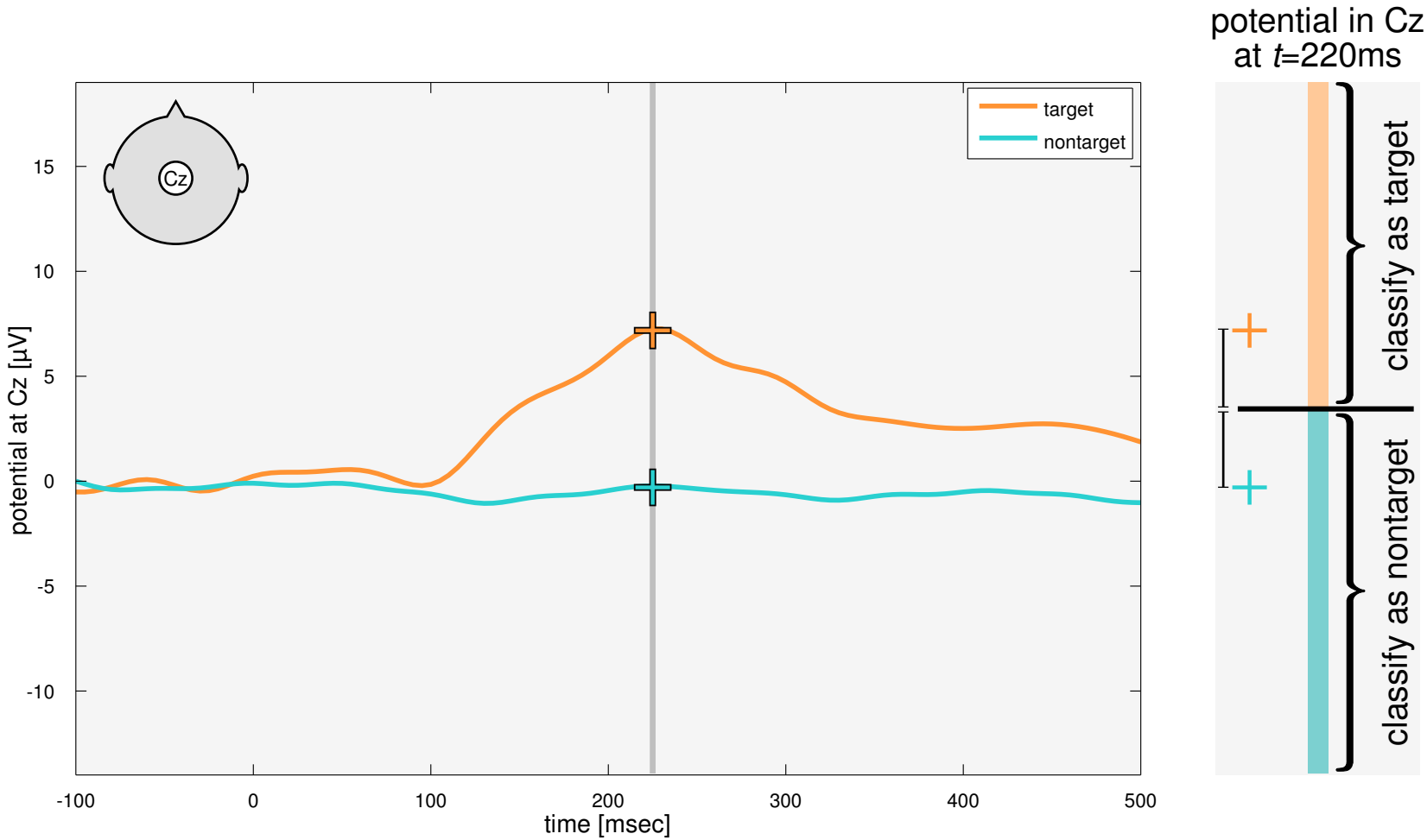
- In BCI epochs are typically strongly overlapping. (Non-target epochs are not shaded in this figure.)

Univariate Features: Averages and Single-Trials



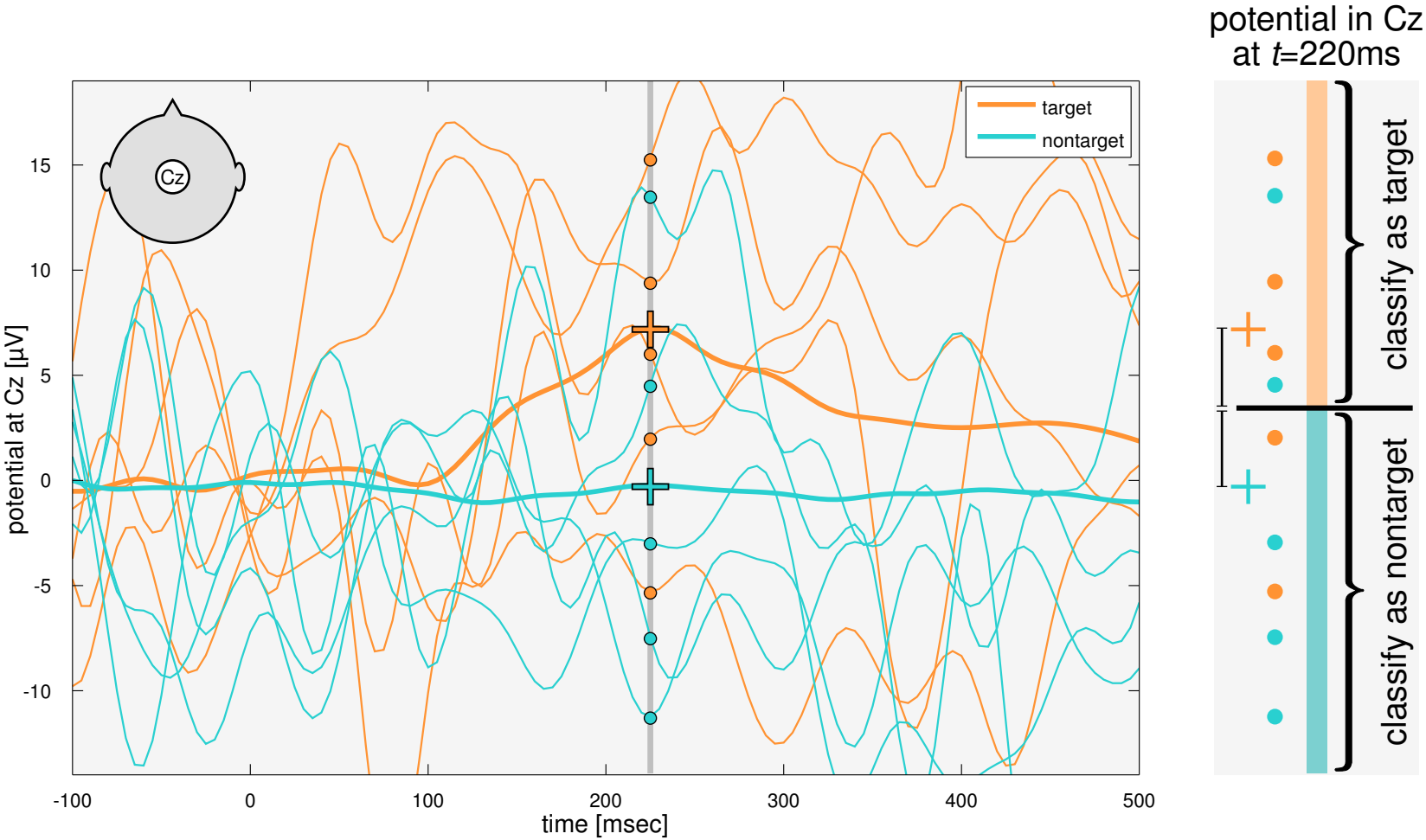
- ERPs can be voluntarily modulated according to an experimental condition, here selective attention to certain *target* stimuli.
- The potential measured 220ms post-stimulus at **Cz** is a one-dimensional observation variable: a *univariate* feature.

Univariate Features: Averages and Single-Trials



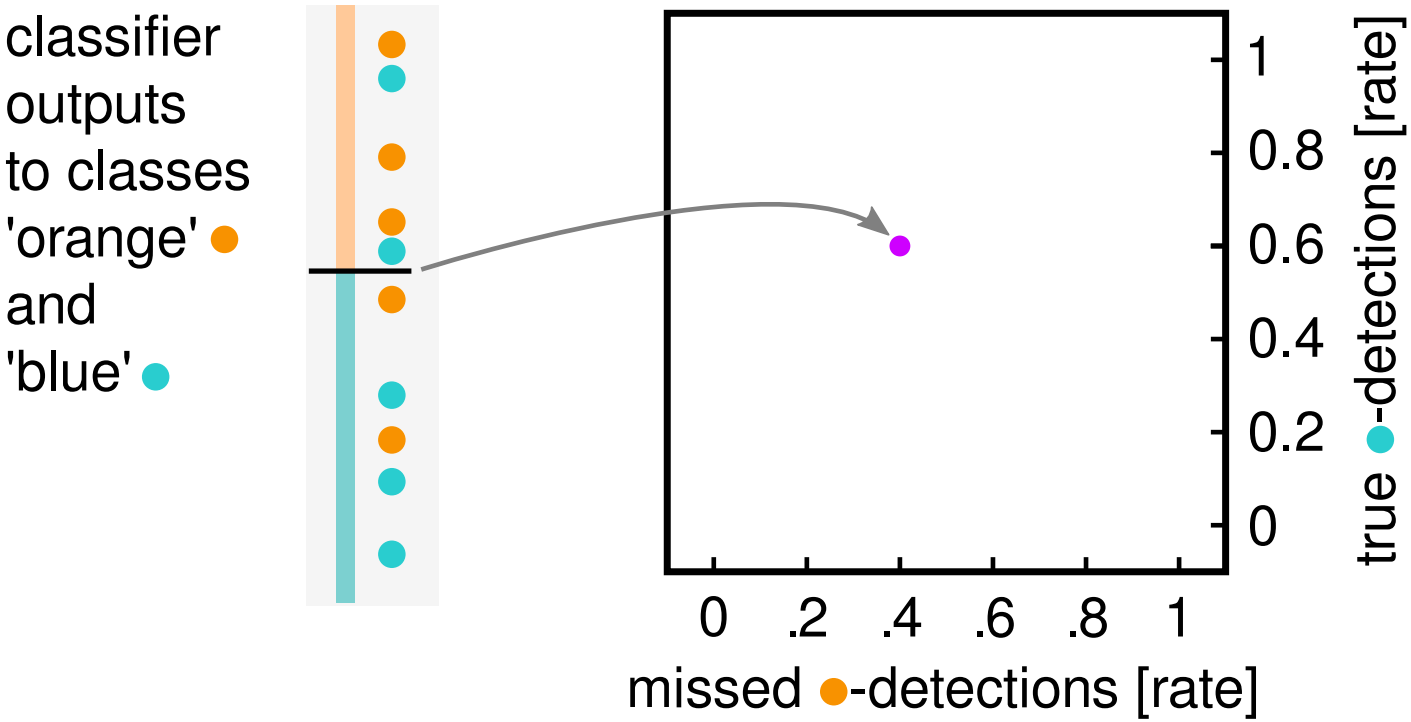
- ERPs can be voluntarily modulated according to an experimental condition, here selective attention to certain *target* stimuli.
- The potential measured 220ms post-stimulus at **Cz** is a one-dimensional observation variable: a *univariate* feature.

Univariate Features: Averages and Single-Trials

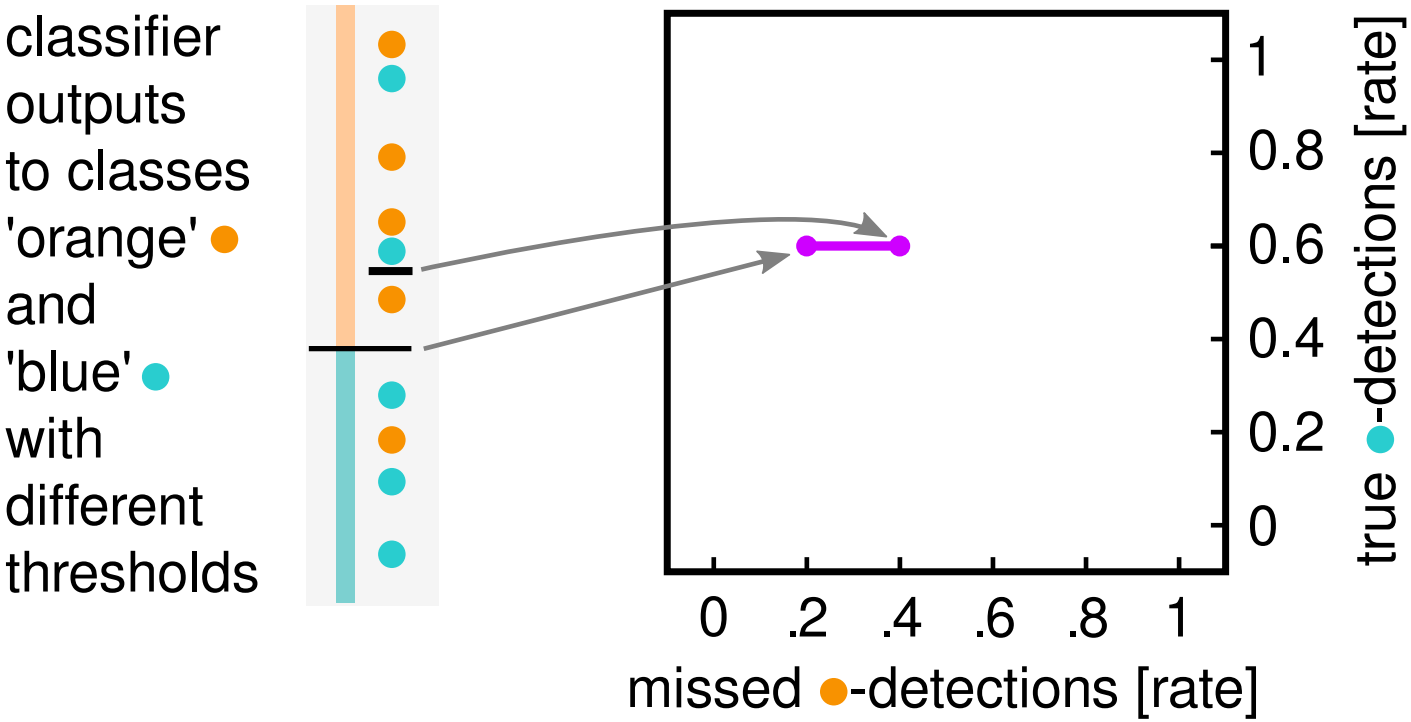


- ERPs can be voluntarily modulated according to an experimental condition, here selective attention to certain *target* stimuli.
- The potential measured 220ms post-stimulus at **Cz** is a one-dimensional observation variable: a *univariate* feature.

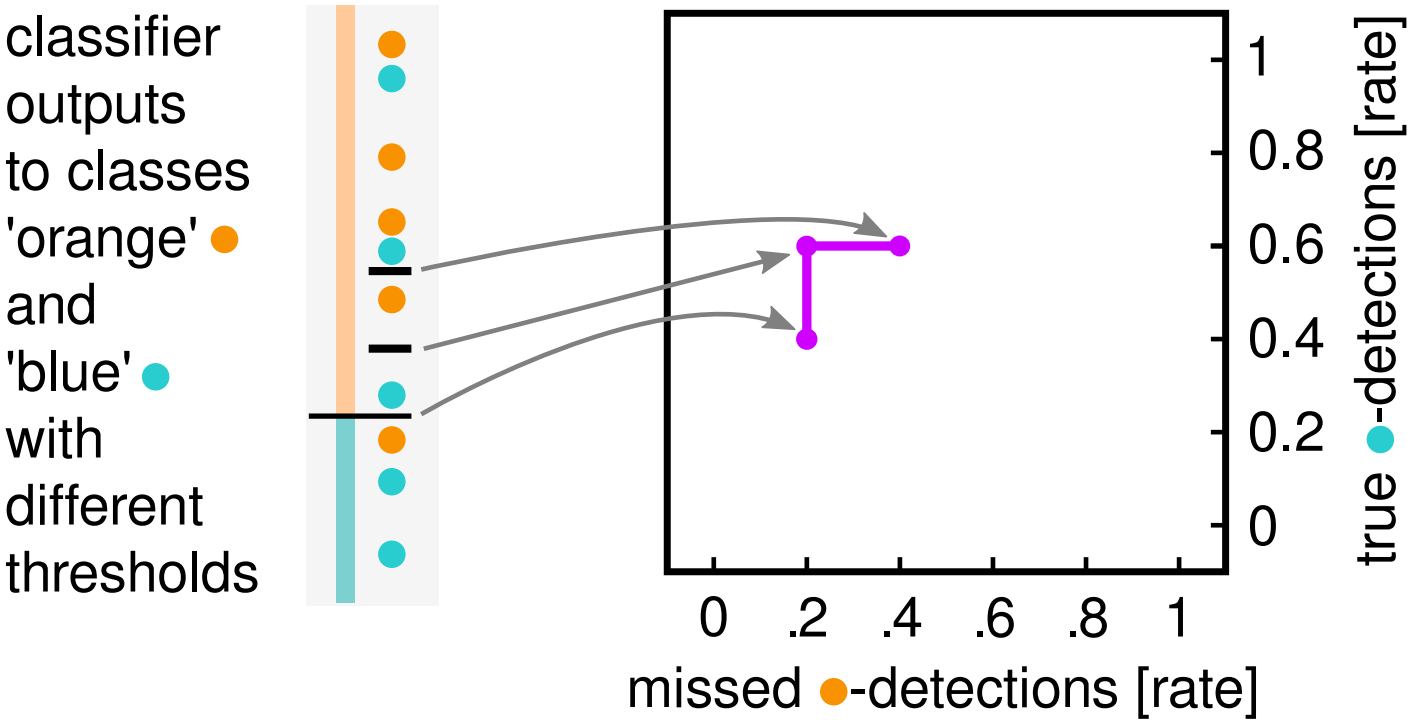
Receiver Operator Characteristics (ROC) and AUC



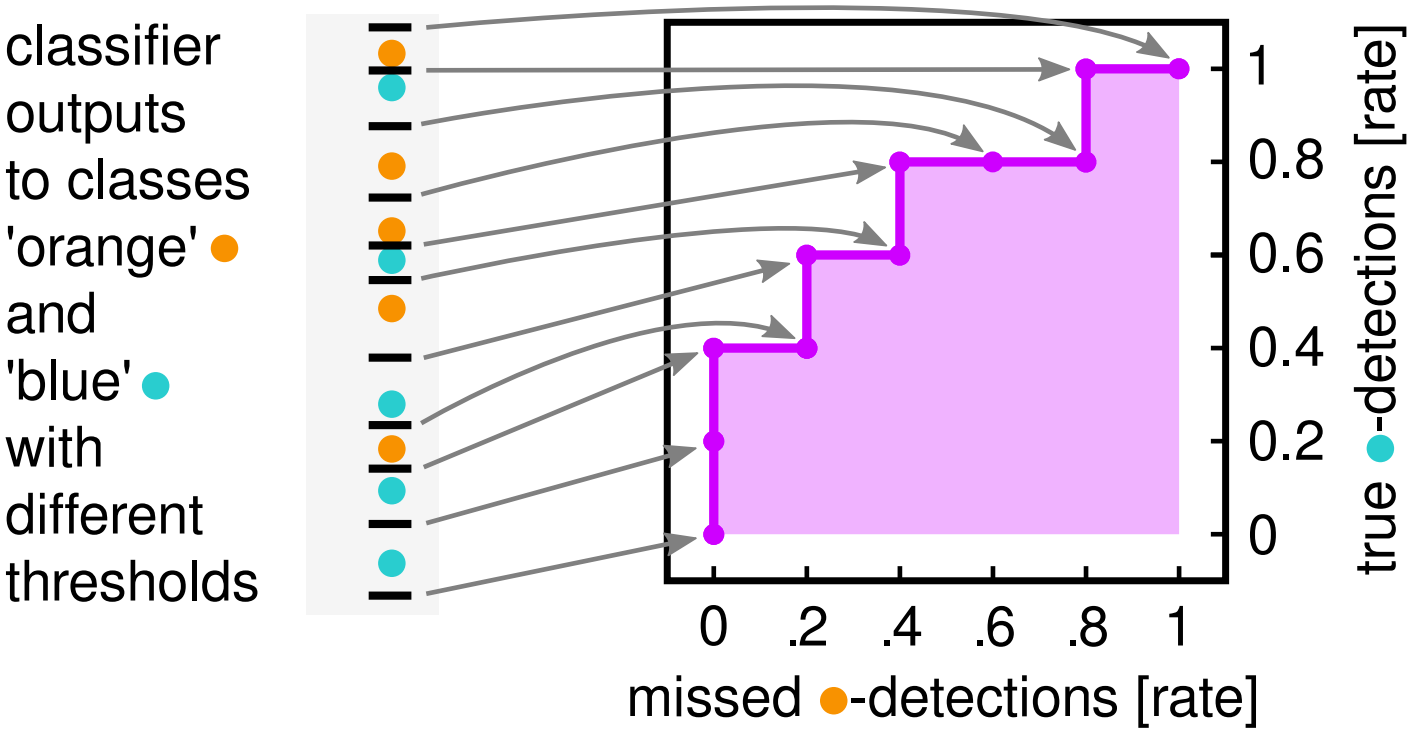
Receiver Operator Characteristics (ROC) and AUC



Receiver Operator Characteristics (ROC) and AUC

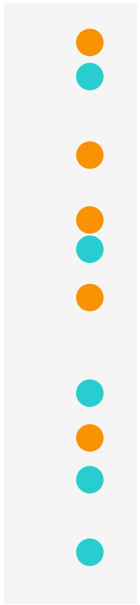


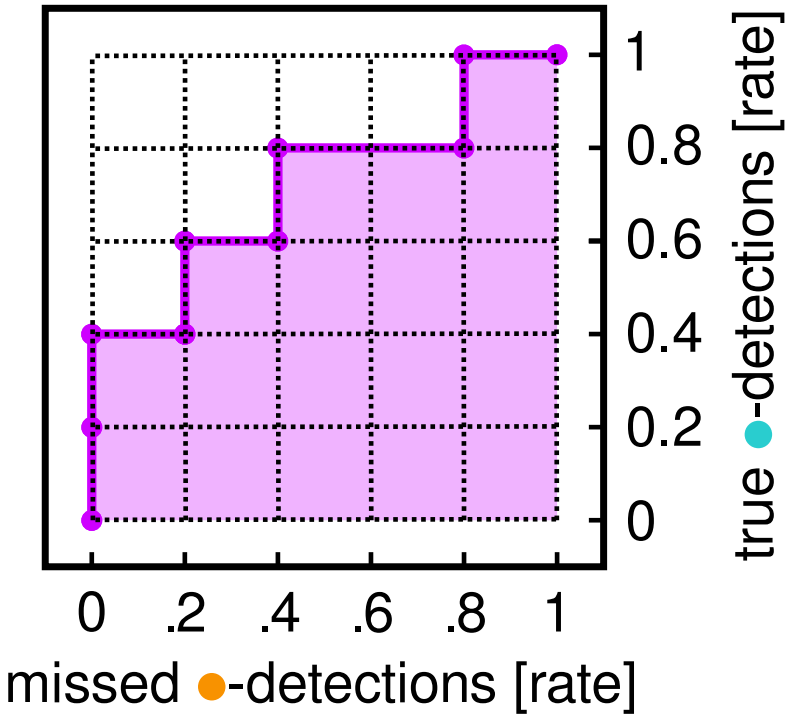
Receiver Operator Characteristics (ROC) and AUC



Receiver Operator Characteristics (ROC) and AUC

classifier outputs to classes 'orange' ● and 'blue' ● with different thresholds





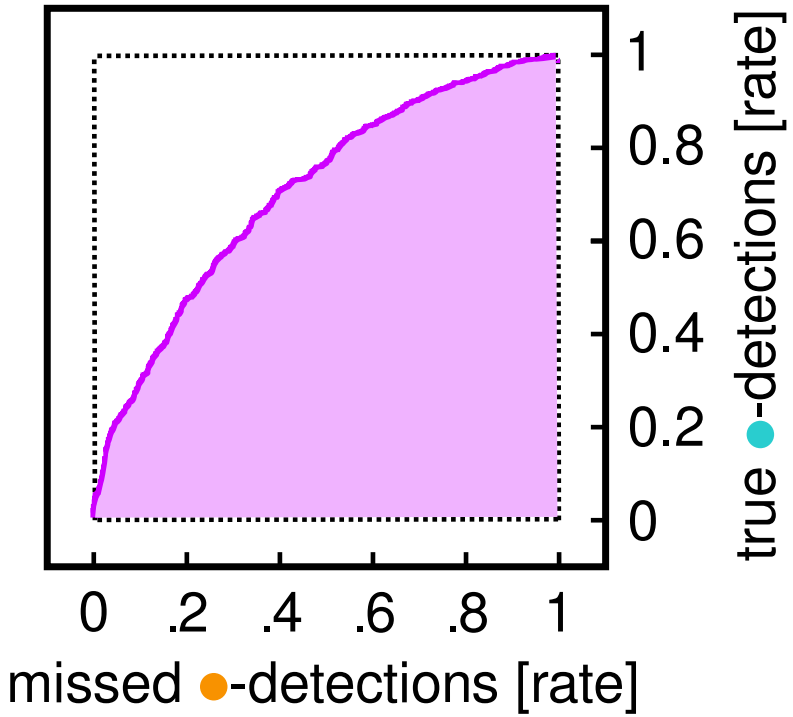
area under the ROC curve:

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

- 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.

Receiver Operator Characteristics (ROC) and AUC

classifier outputs to classes 'orange' ● and 'blue' ● with different thresholds



area under the ROC curve:

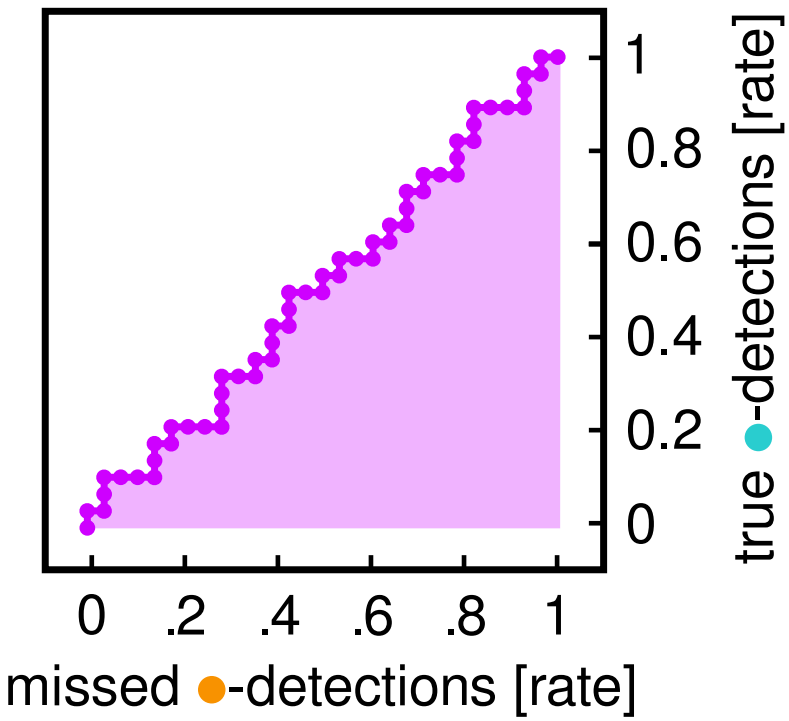
AUC ≈ 0.7

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

- With all trials of our example data set, the AUC is ≈ 0.7.

Receiver Operator Characteristics (ROC) and AUC

classifier outputs to classes 'orange' ● and 'blue' ● with different thresholds



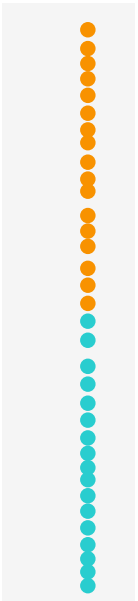
area under the ROC curve:

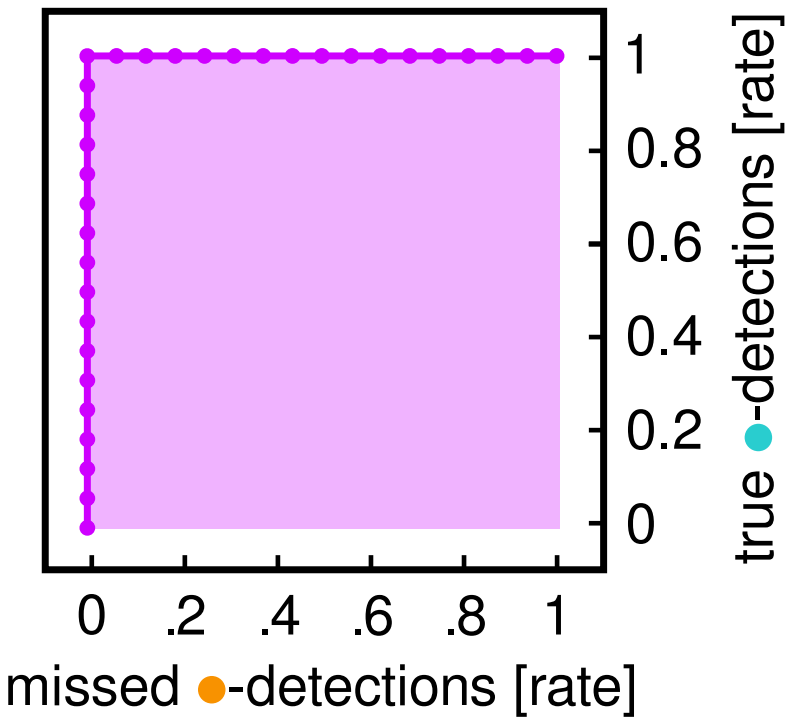
AUC ≈ 0.5

- For random values, the AUC is about 0.5.

Receiver Operator Characteristics (ROC) and AUC

classifier outputs to classes 'orange' ● and 'blue' ● with different thresholds





area under the ROC curve:

$\text{AUC} = 1$

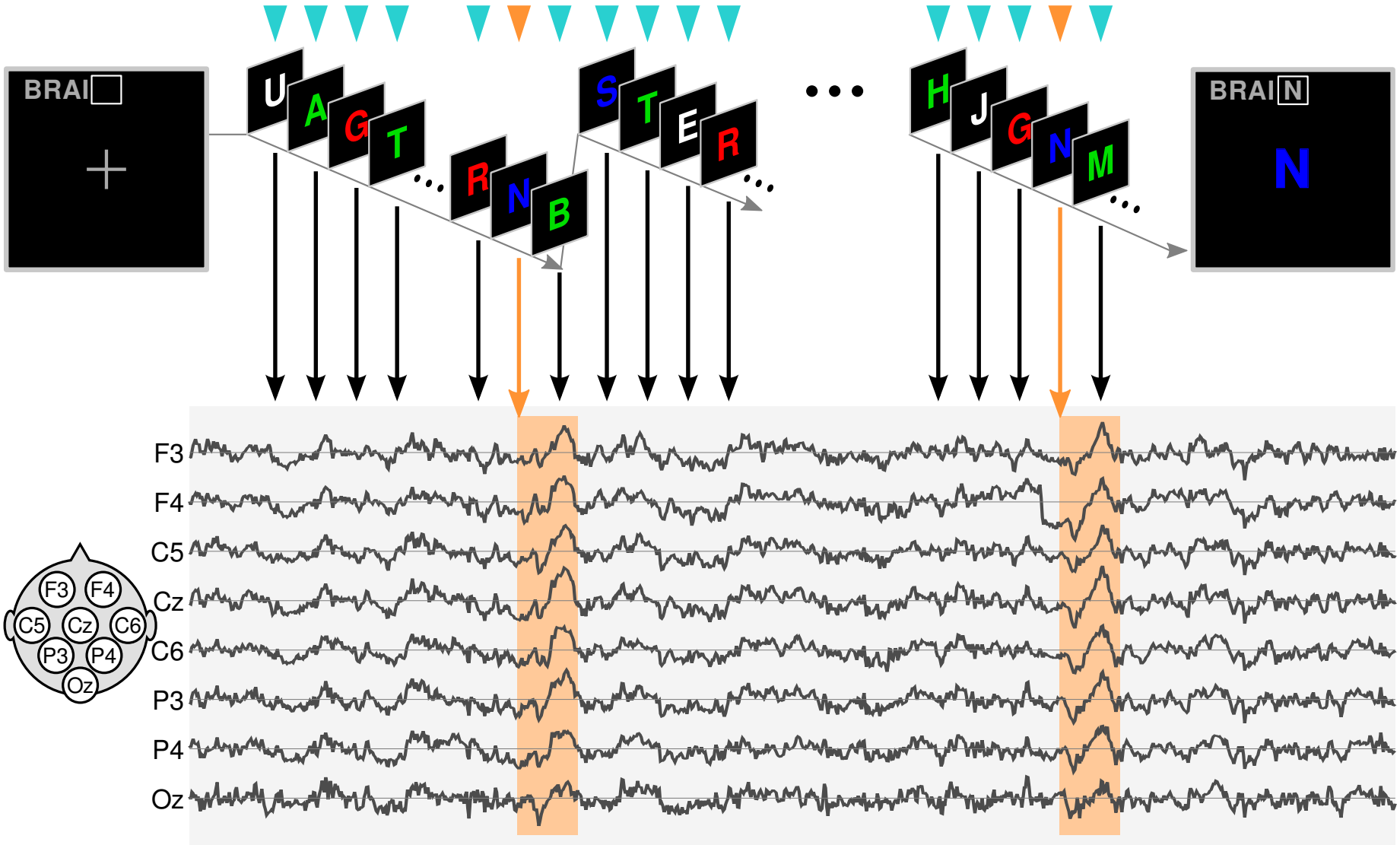
- If the classes are perfectly separated, the AUC is 1 (or 0 if the sign is reversed).

From Uni- to Multivariate Features

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

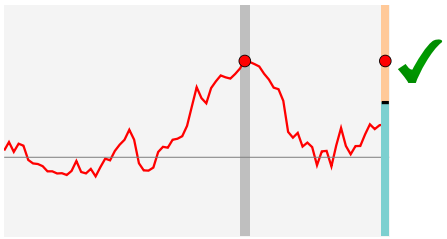
- sample ERP signals at *multiple* time points/intervals
  $\rightarrow$ *temporal feature*
- join signals from *multiple* channels
  $\rightarrow$ *spatial feature*
- do both things
  $\rightarrow$ *spatio-temporal feature*

Multi-channel Epochs

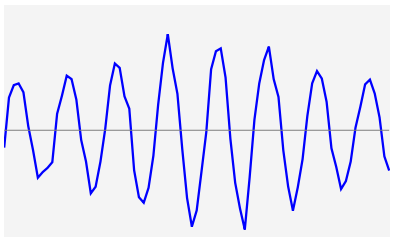


The Virtue of Multivariate Spatial Features

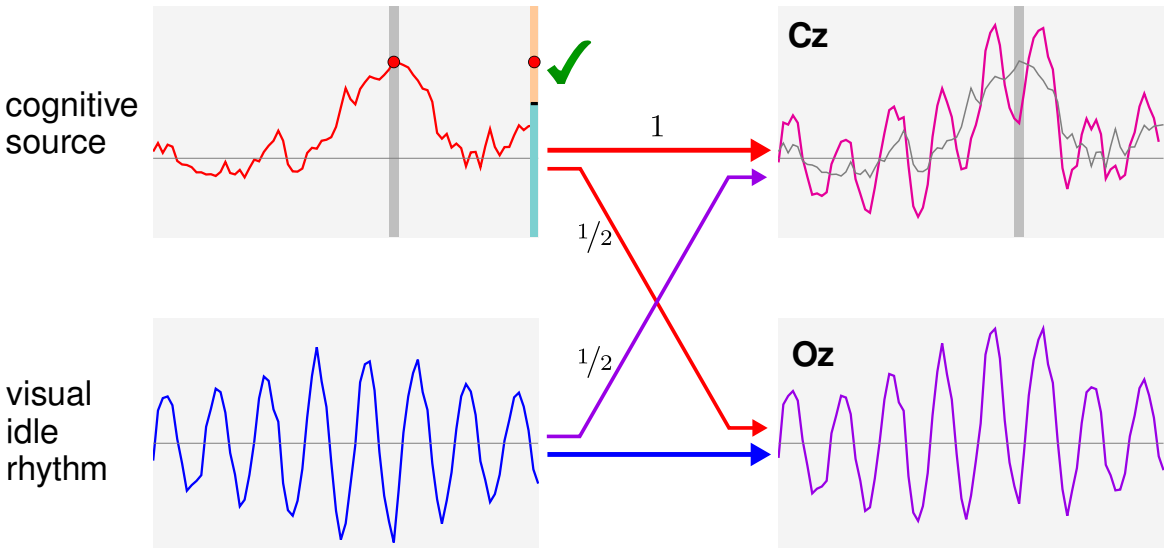
cognitive source



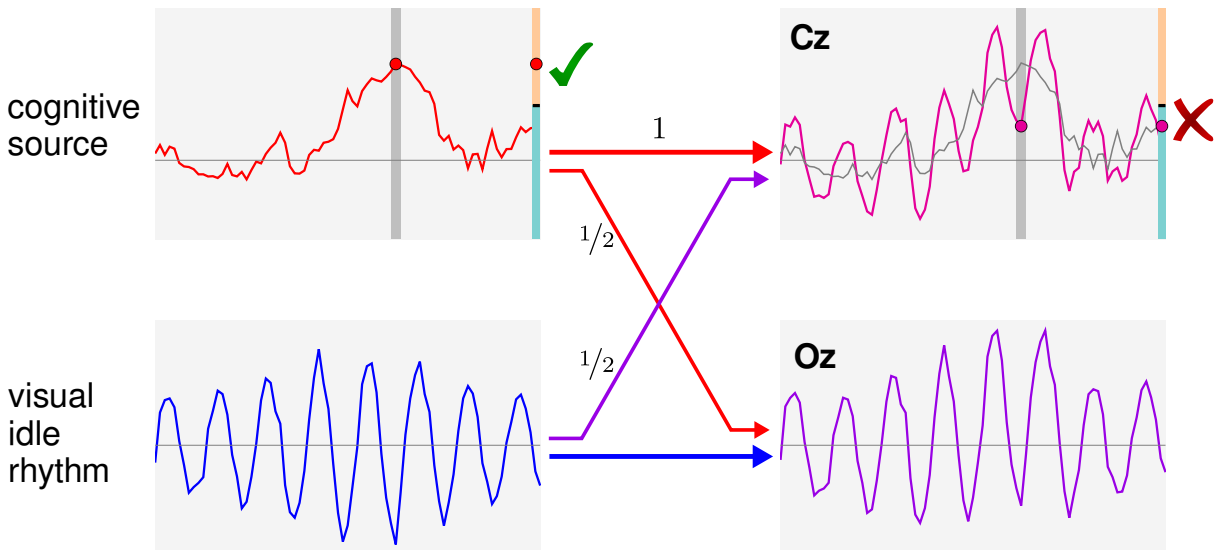
visual idle rhythm



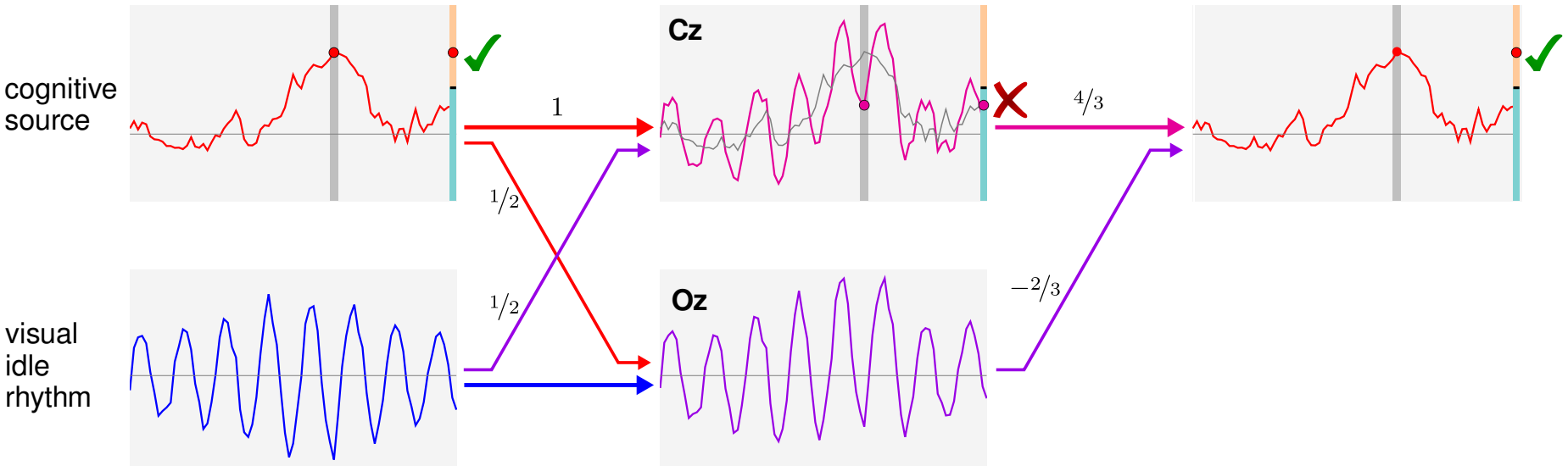
The Virtue of Multivariate Spatial Features



The Virtue of Multivariate Spatial Features

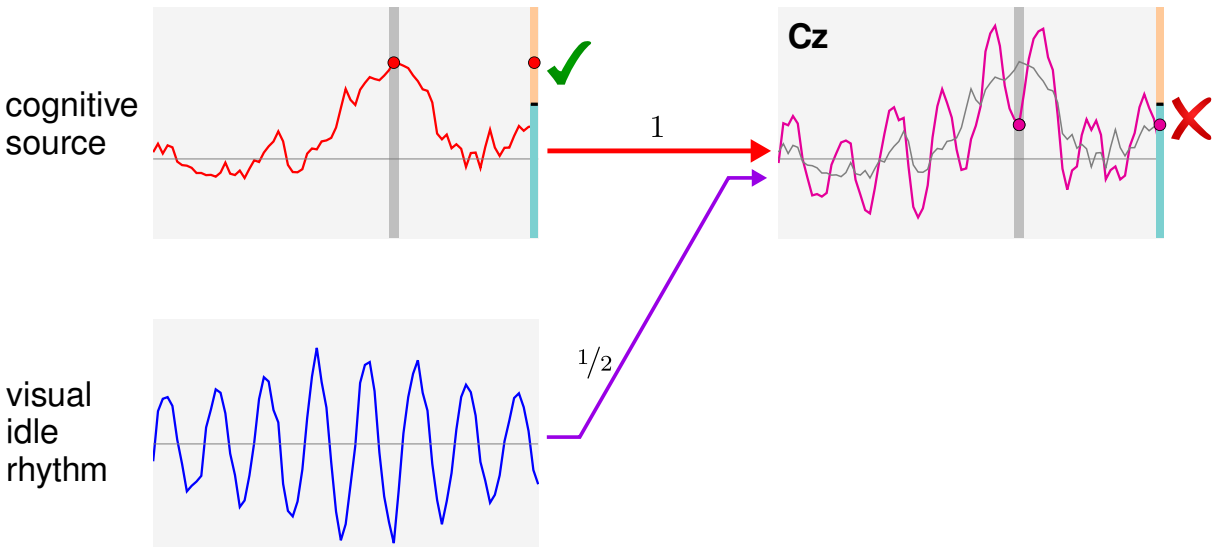


The Virtue of Multivariate Spatial Features

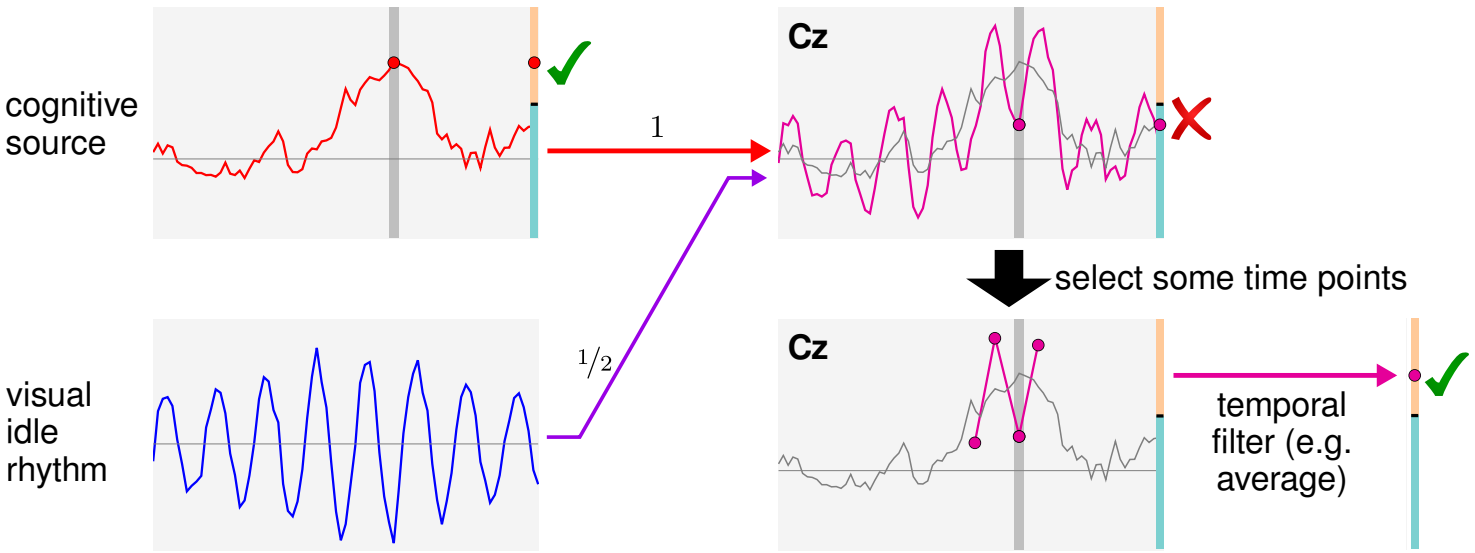


- Here, $\mathbf{w} = [4/3 \quad -2/3]^\top$ is a simple spatial filter.

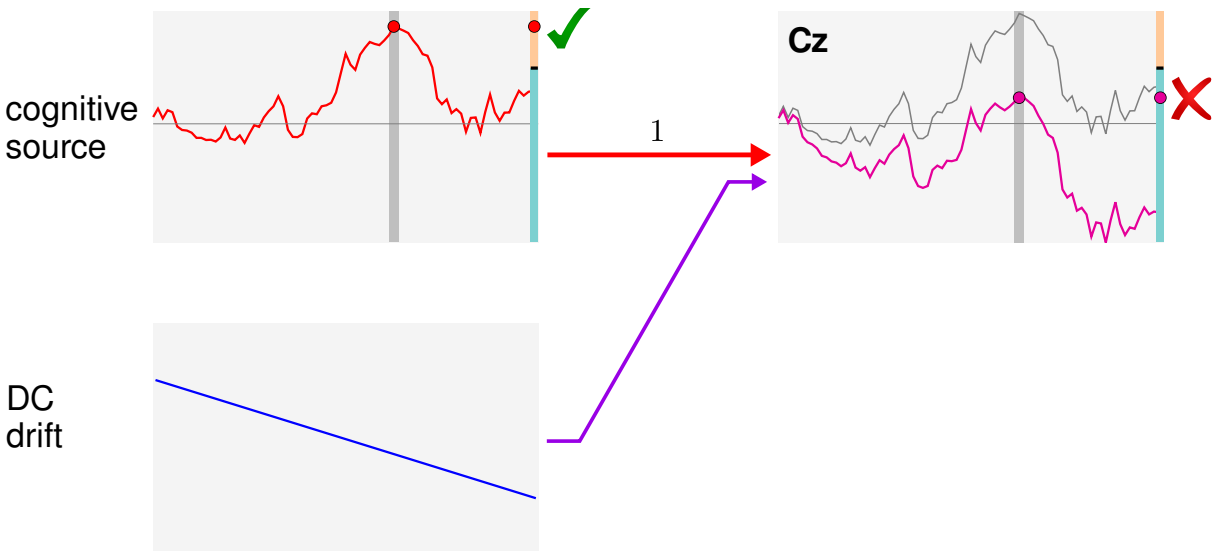
The Virtue of Multivariate Temporal Features



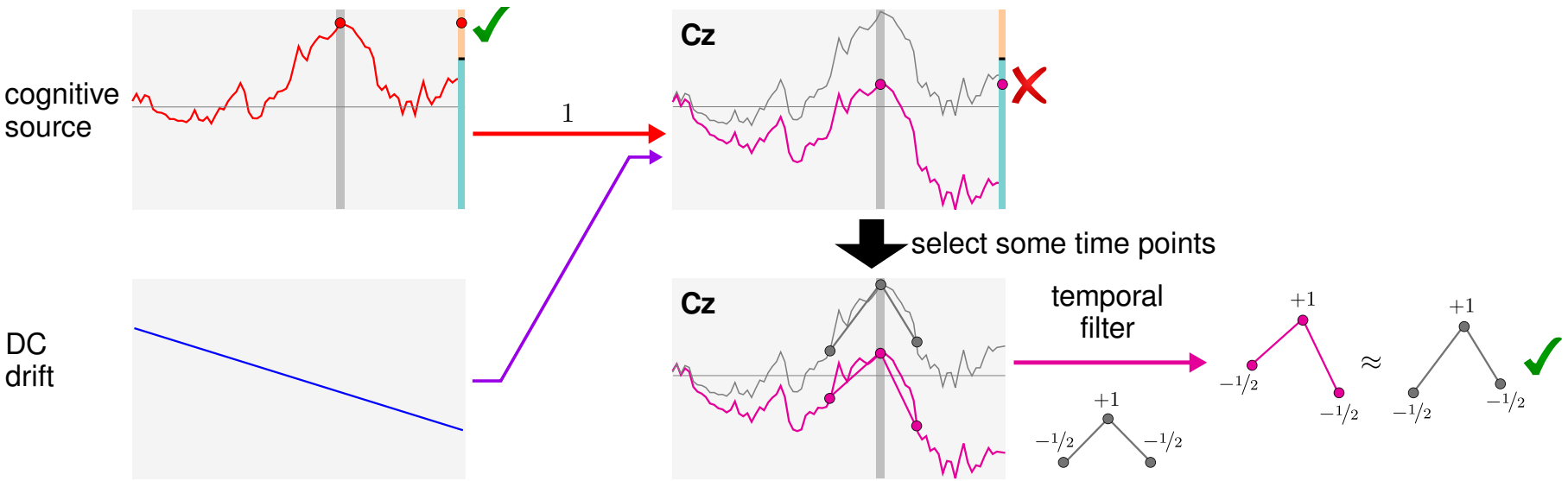
The Virtue of Multivariate Temporal Features



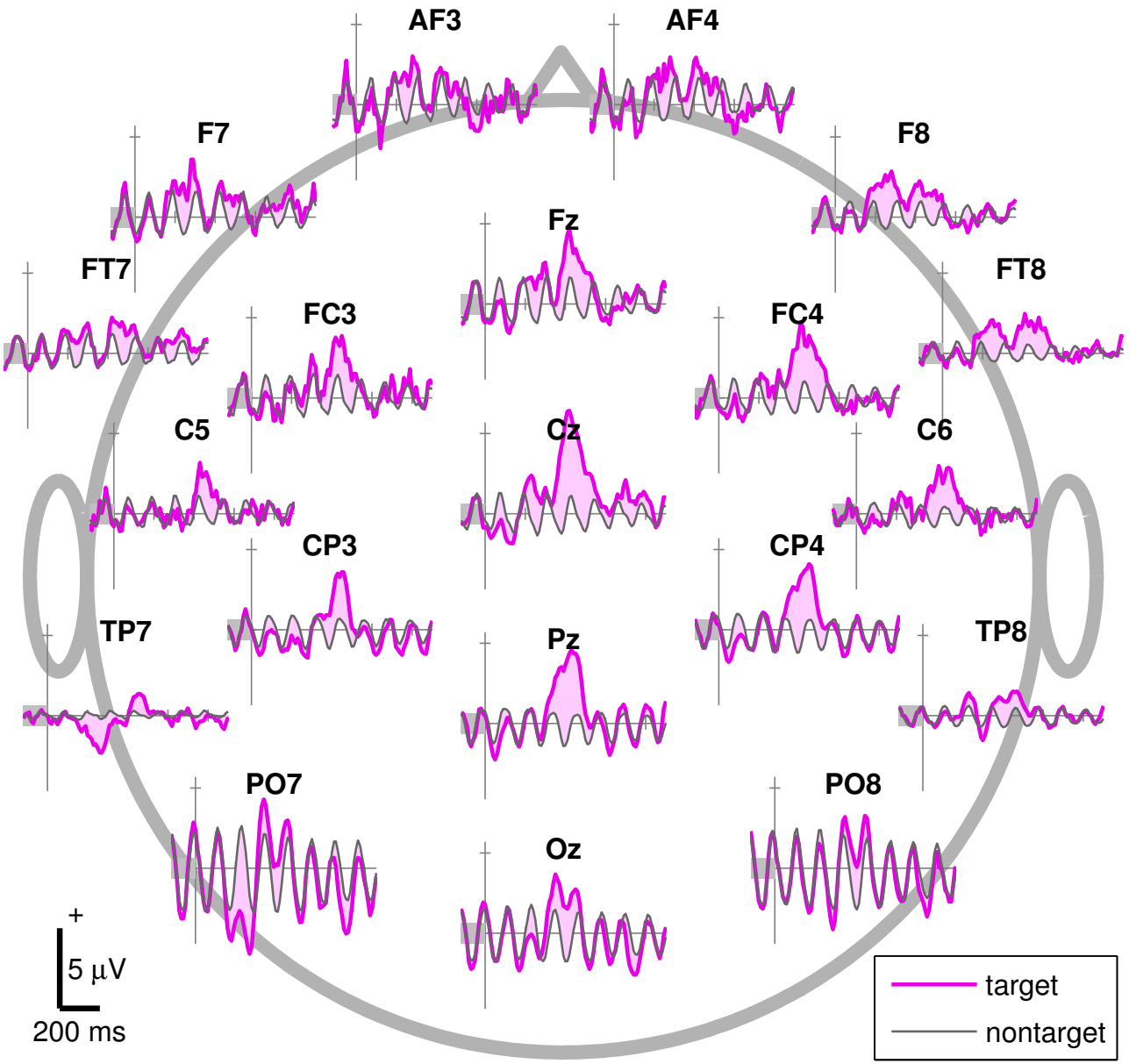
The Virtue of Multivariate Temporal Features



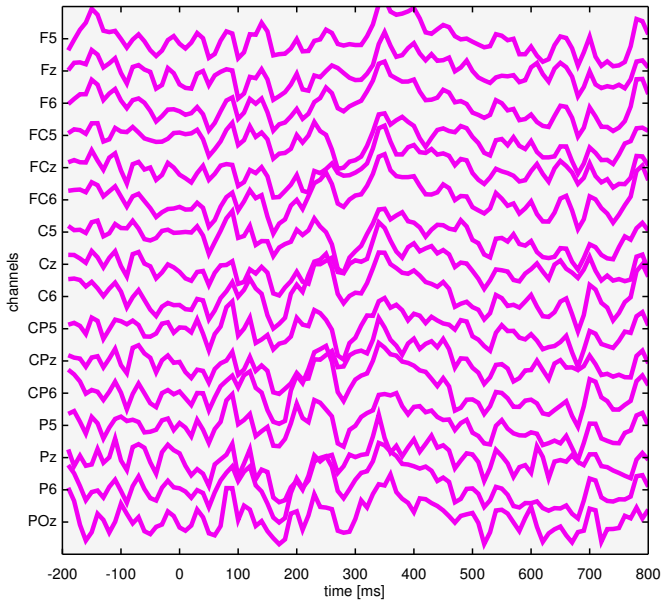
The Virtue of Multivariate Temporal Features

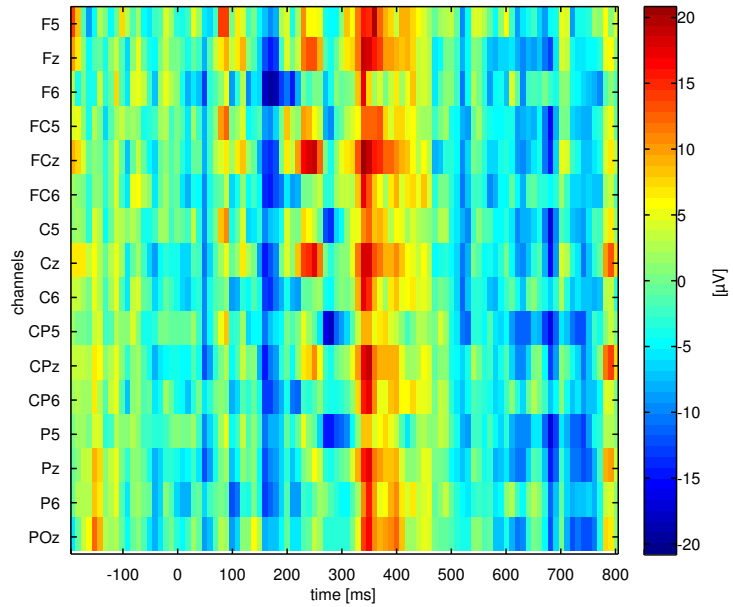


ERPs in a Head Plot

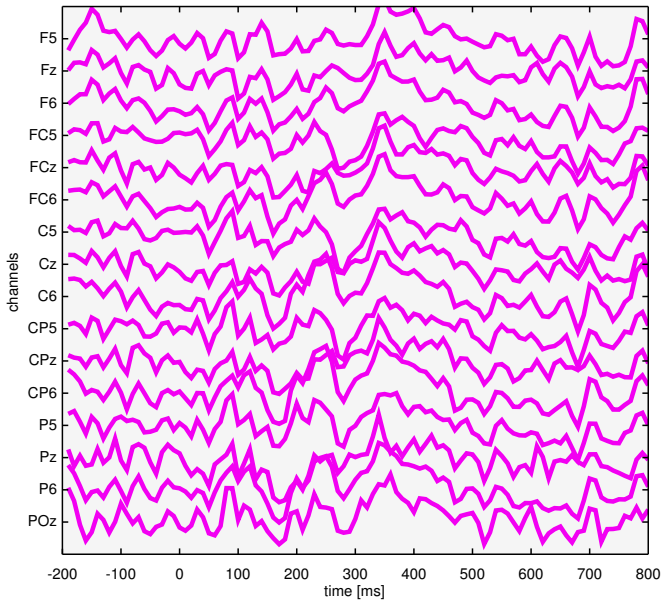


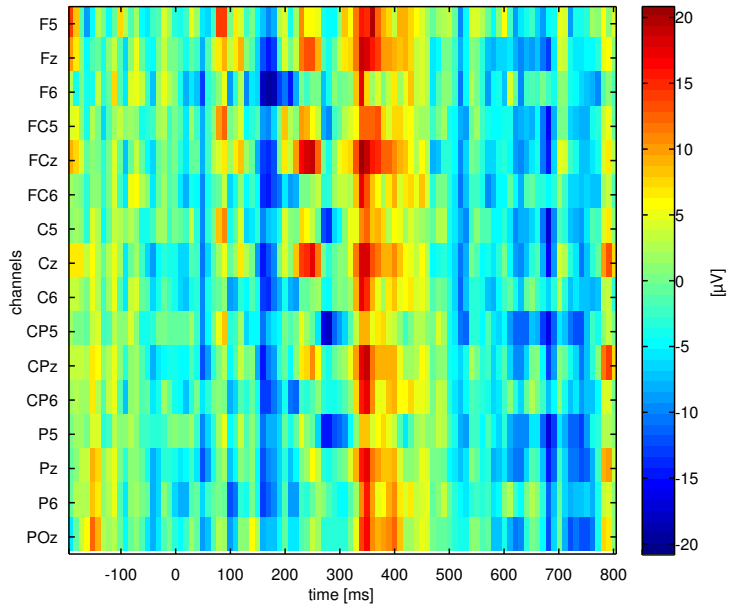
Interlude: Representation as Matrix

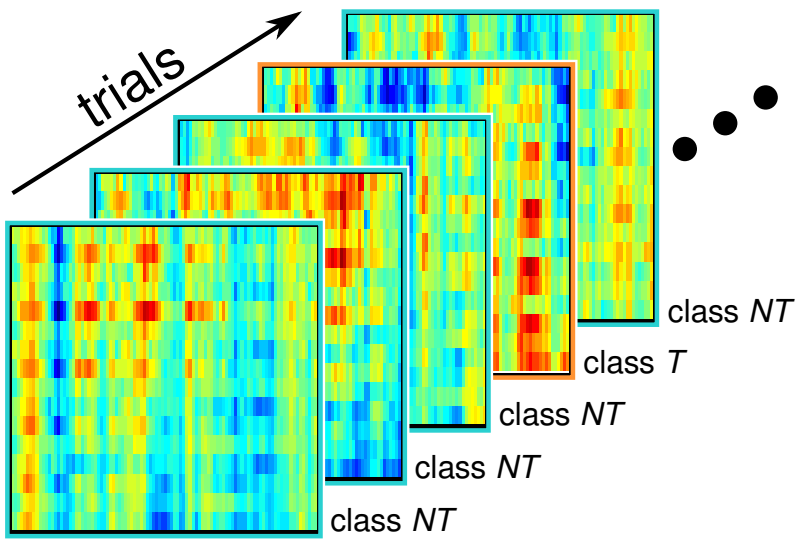




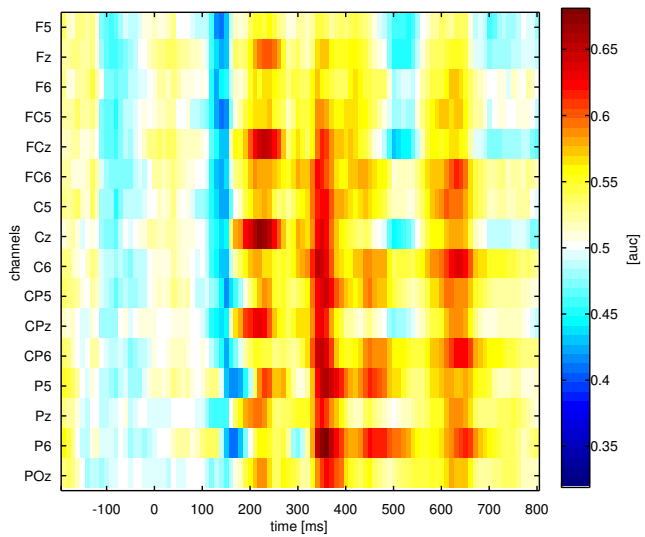
Interlude: Representation as Matrix



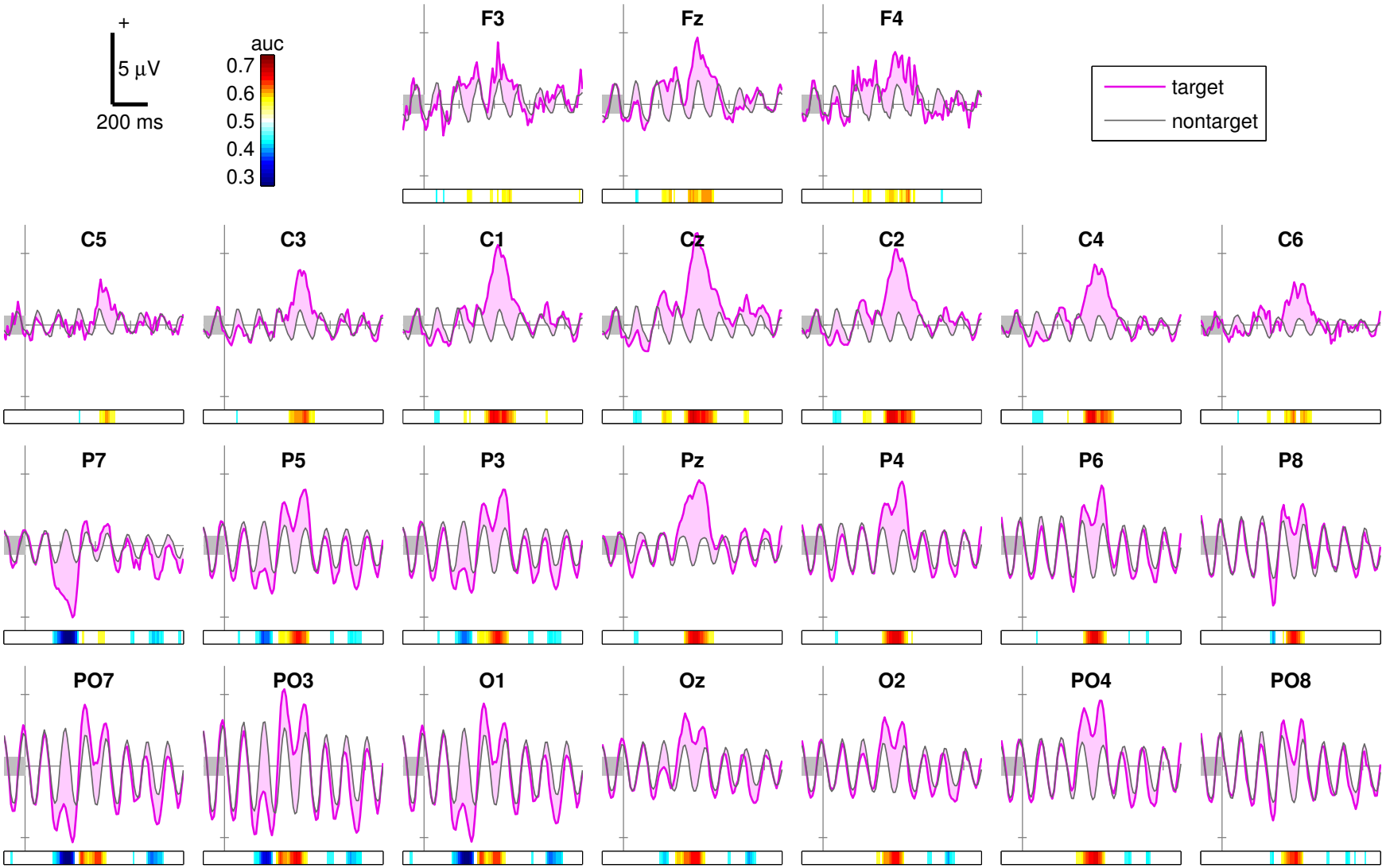




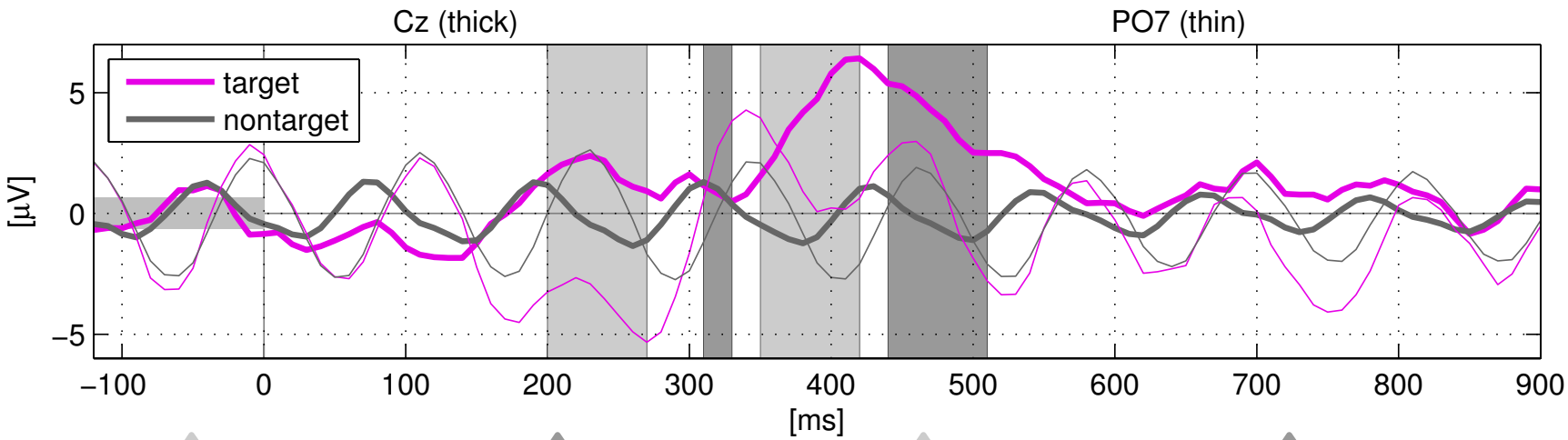
AUC across trials

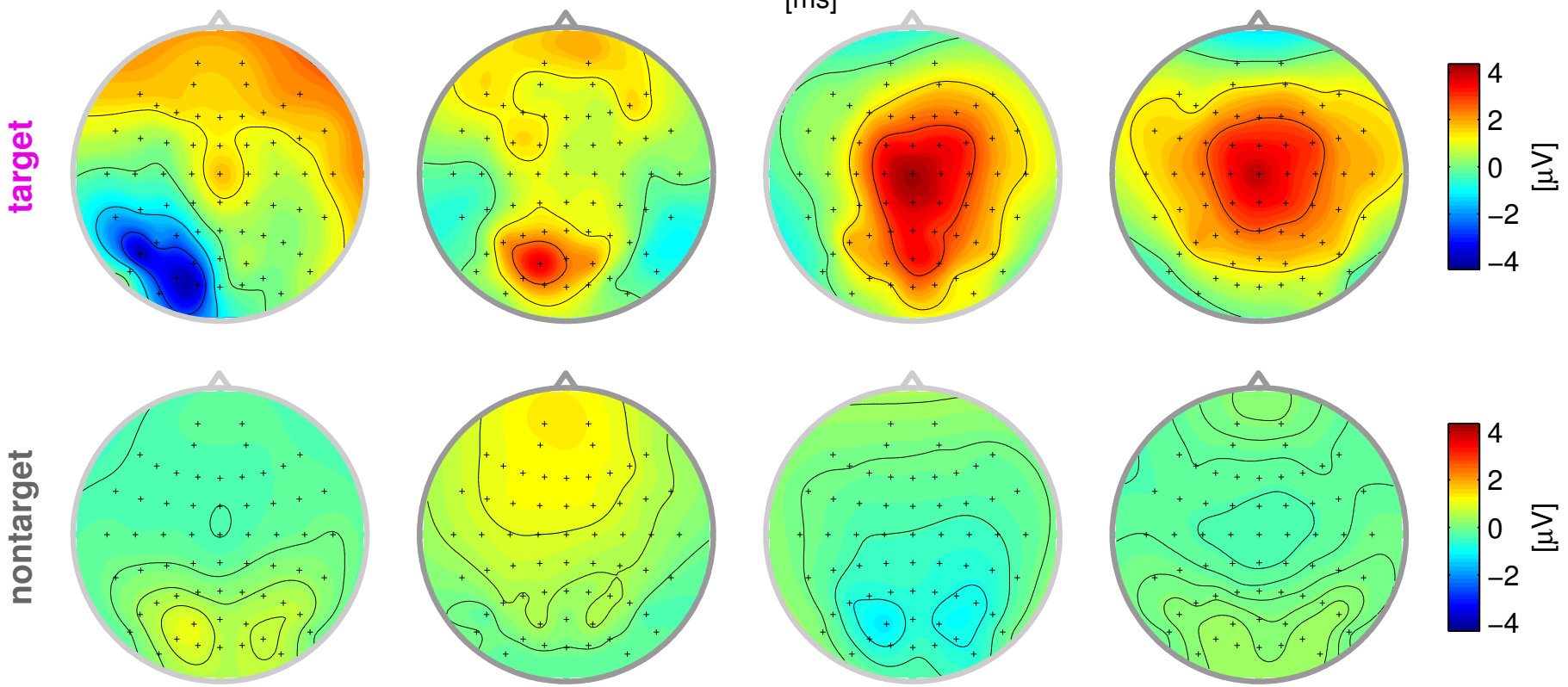


ERPs in a Grid Plot

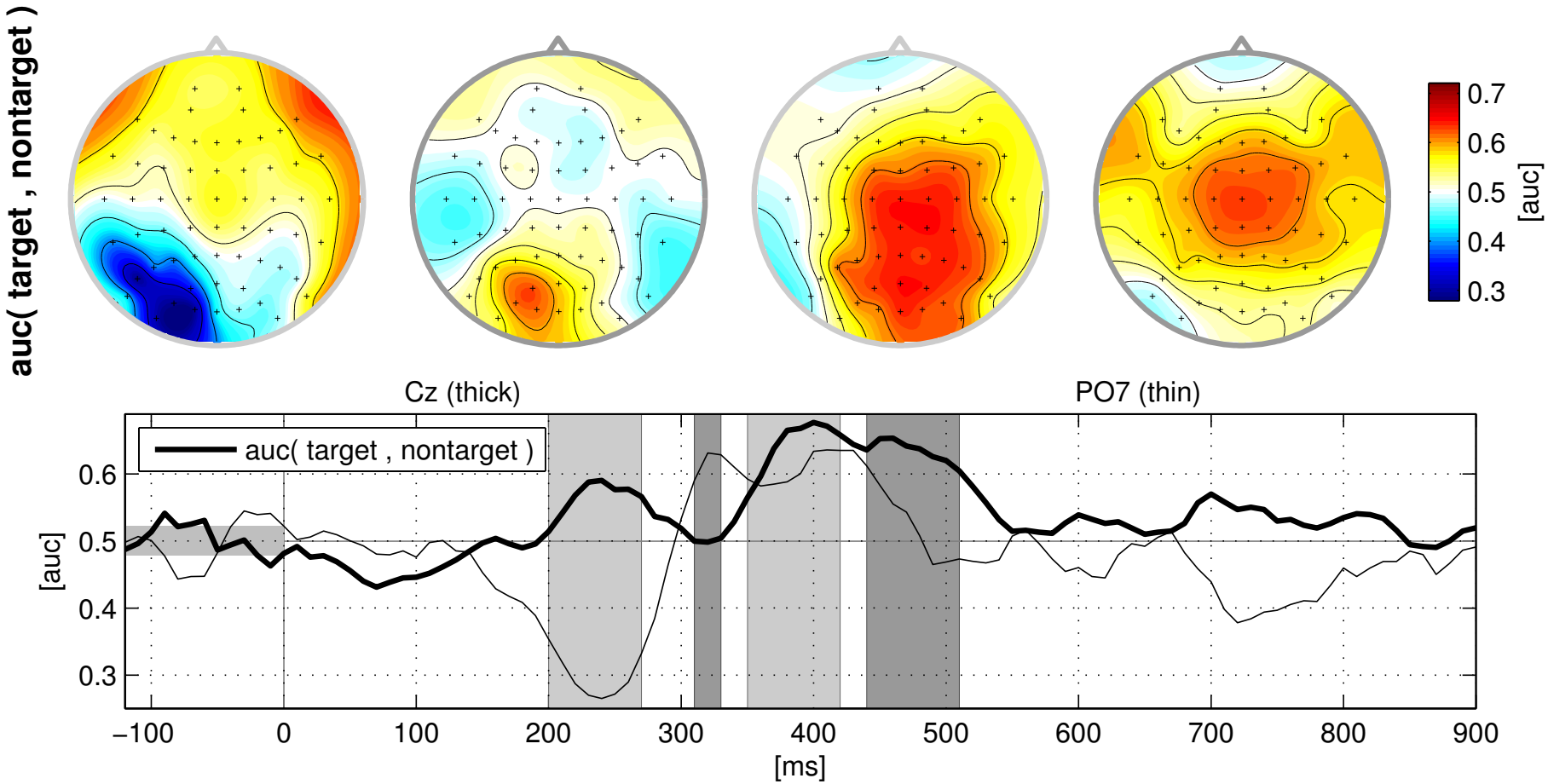


ERP Topographies

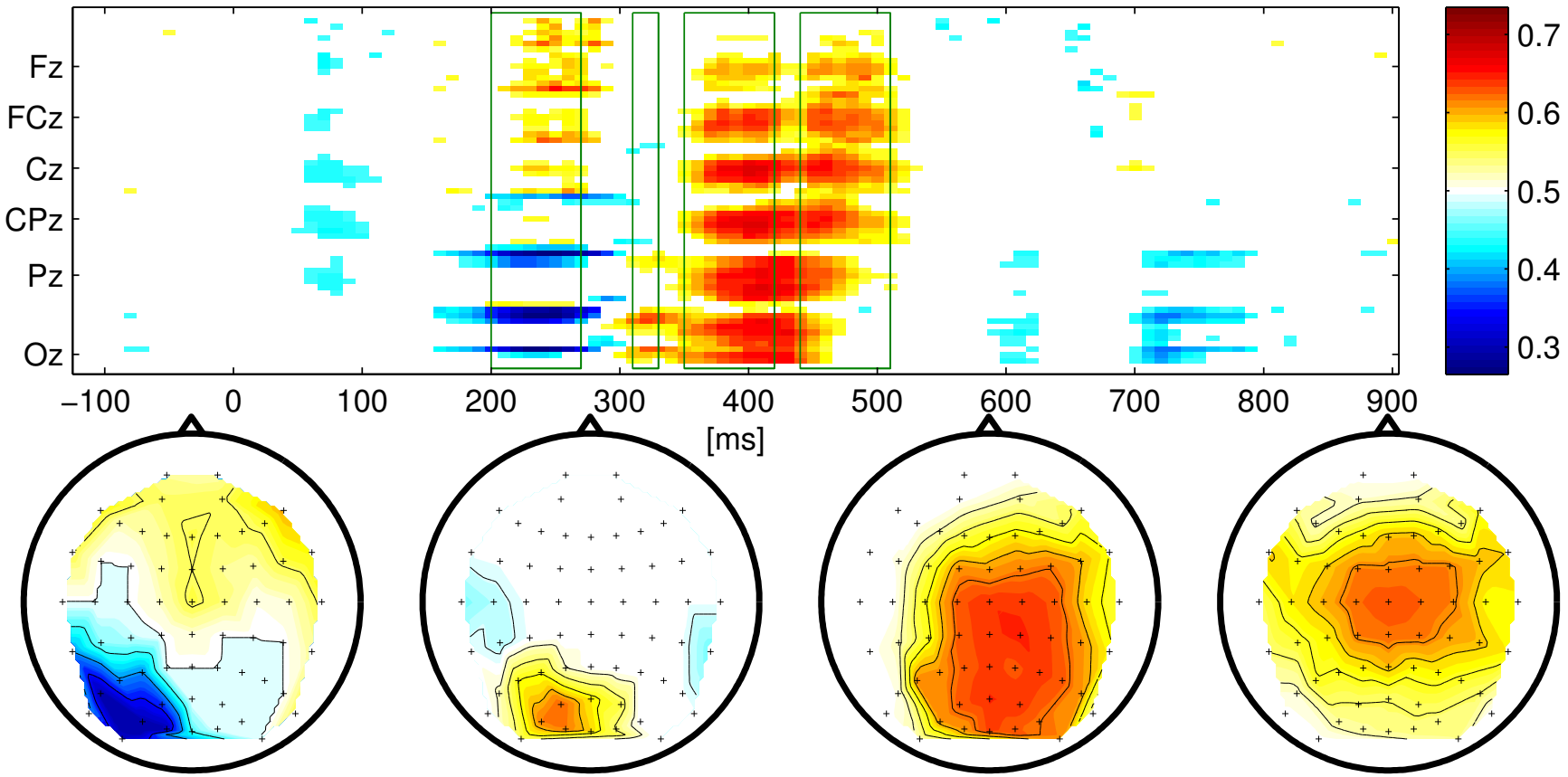




ERP Topographies

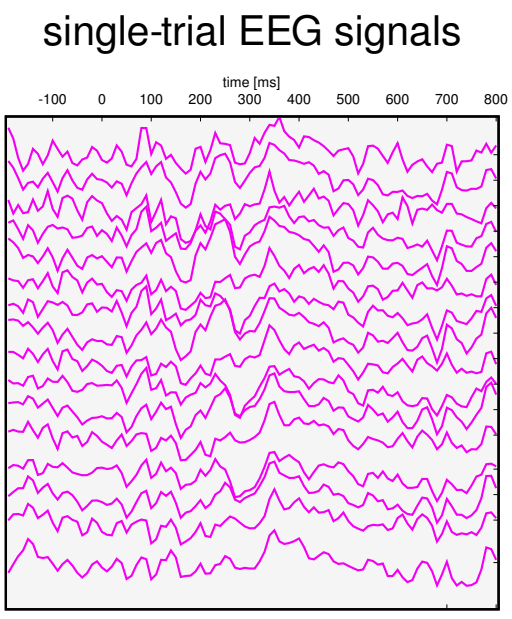


AUC Matrix: Selection of Channels and Time Intervals



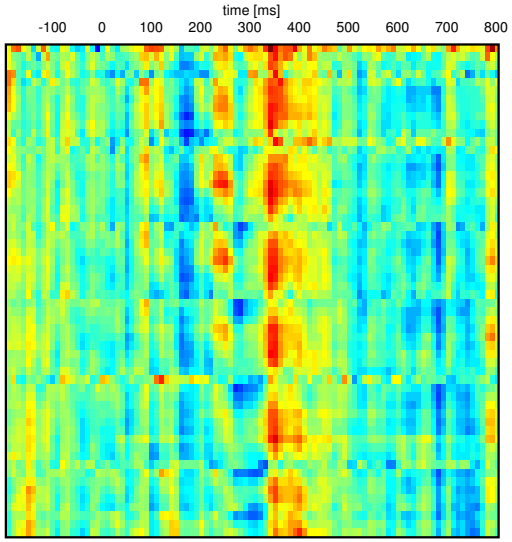
- Each cell in the matrix is one uni-variate feature.
- Let's combine them to multi-variate features!

Multivariate ERP Features

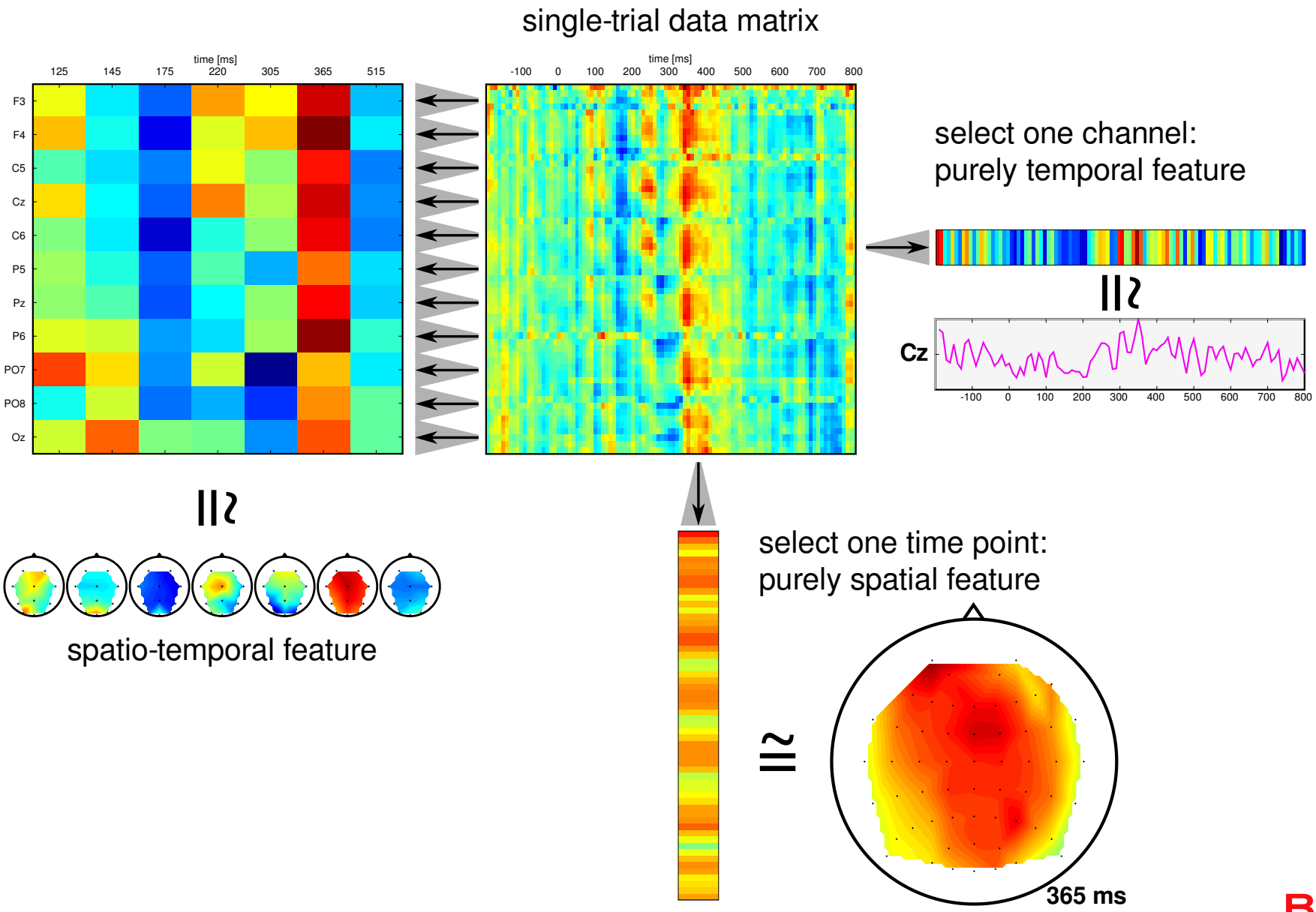


Multivariate ERP Features

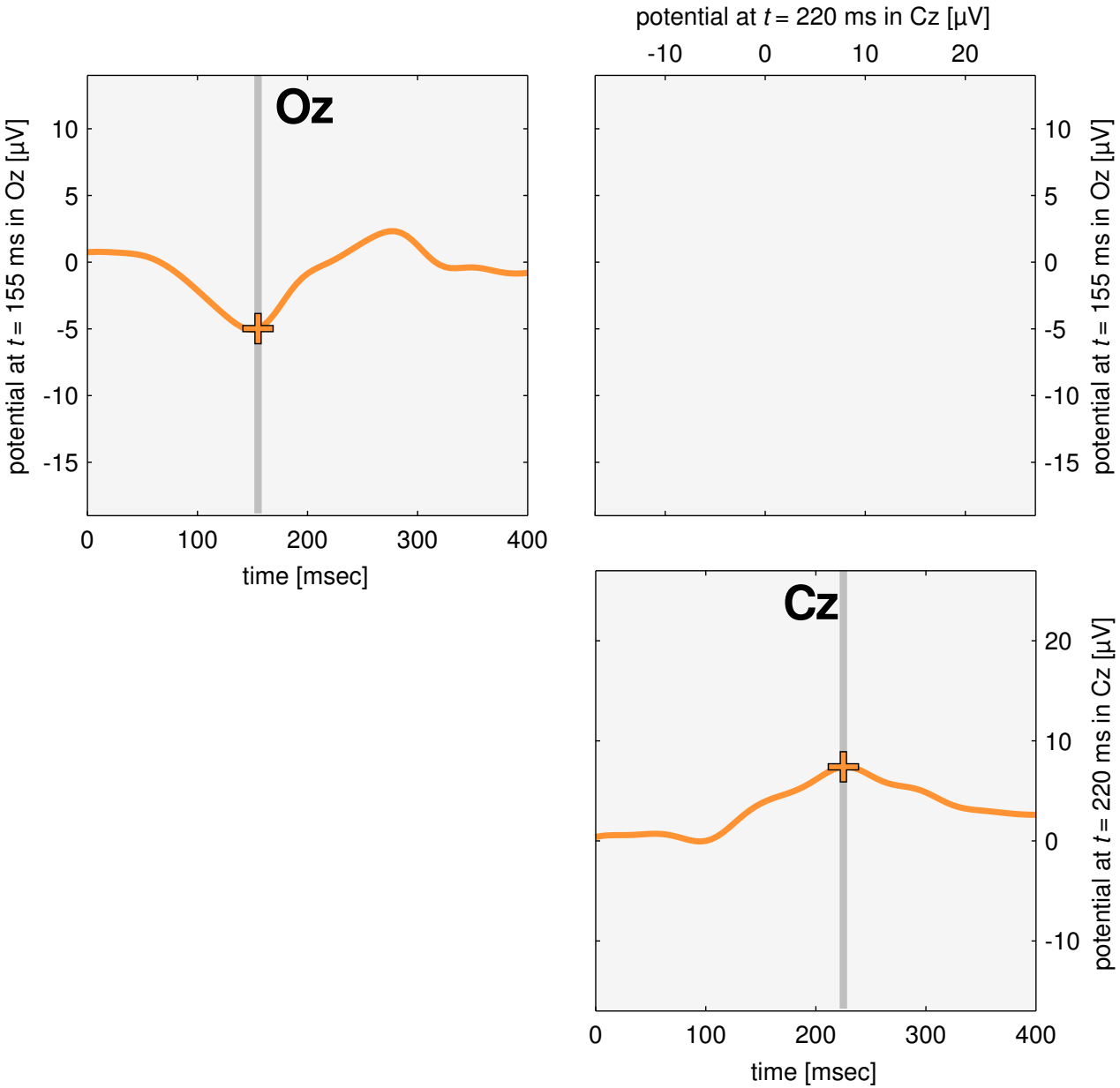
single-trial data matrix



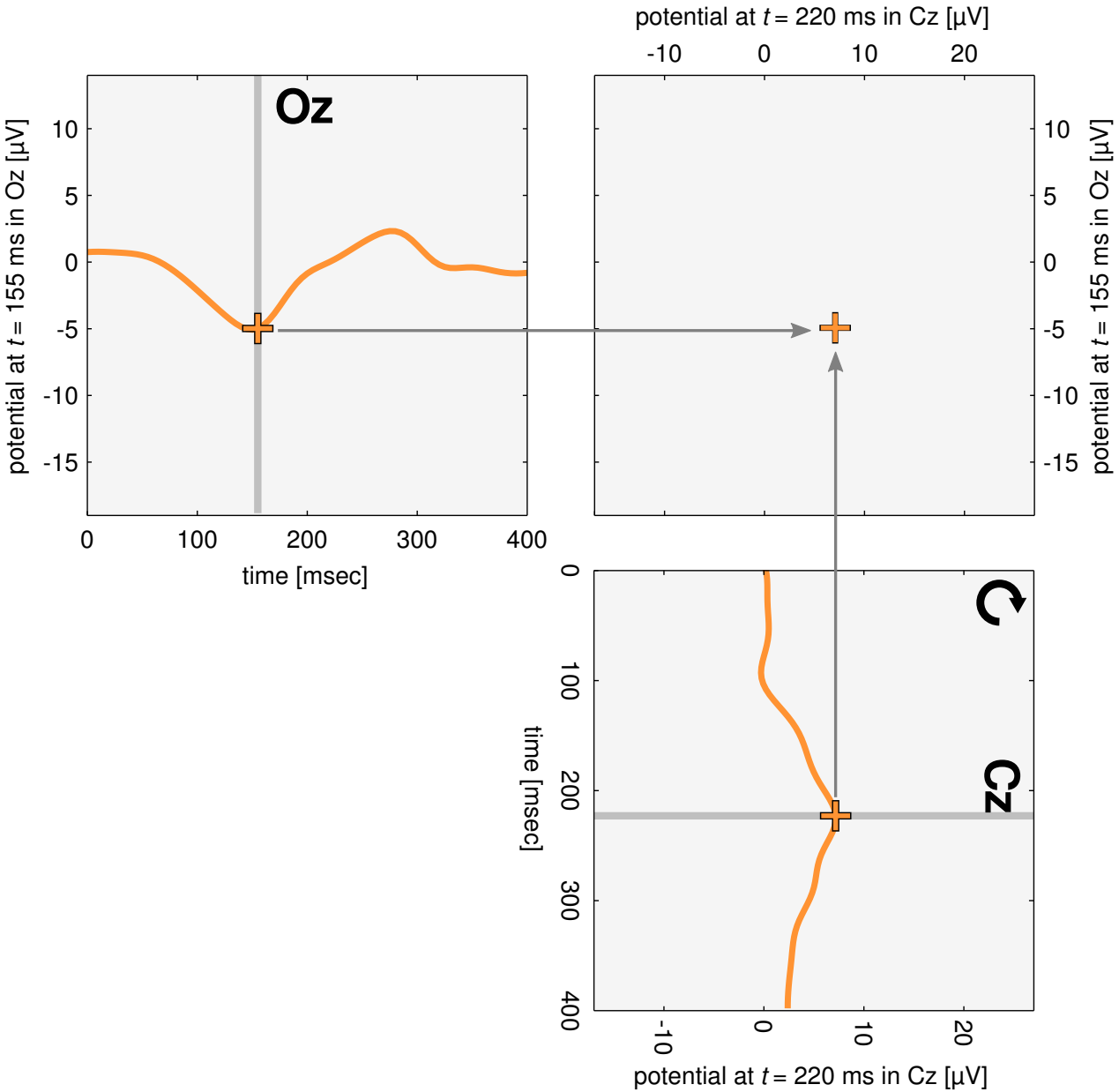
Multivariate ERP Features



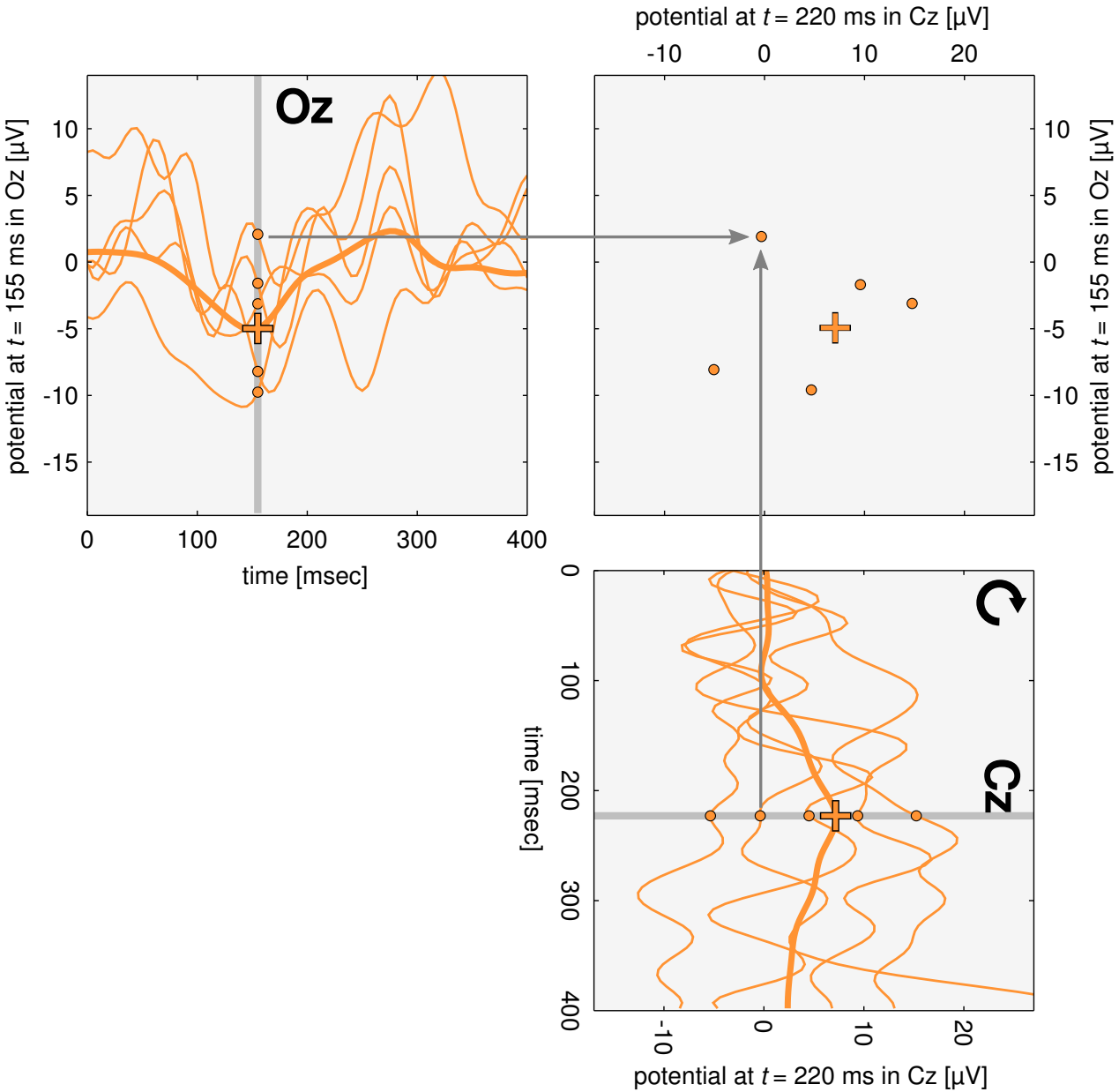
Representation of Multivariate Distributions: Scatter Plot



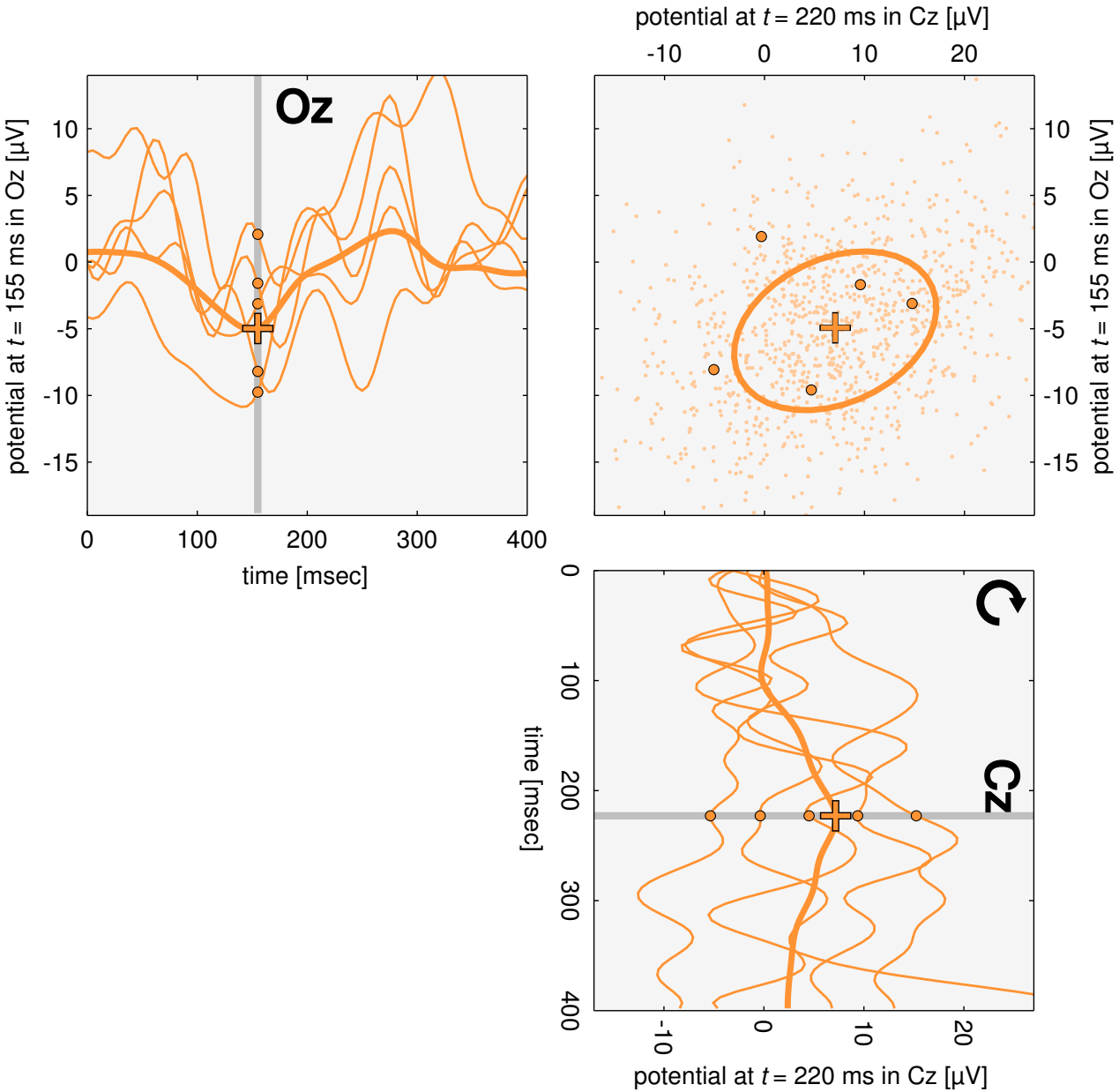
Representation of Multivariate Distributions: Scatter Plot



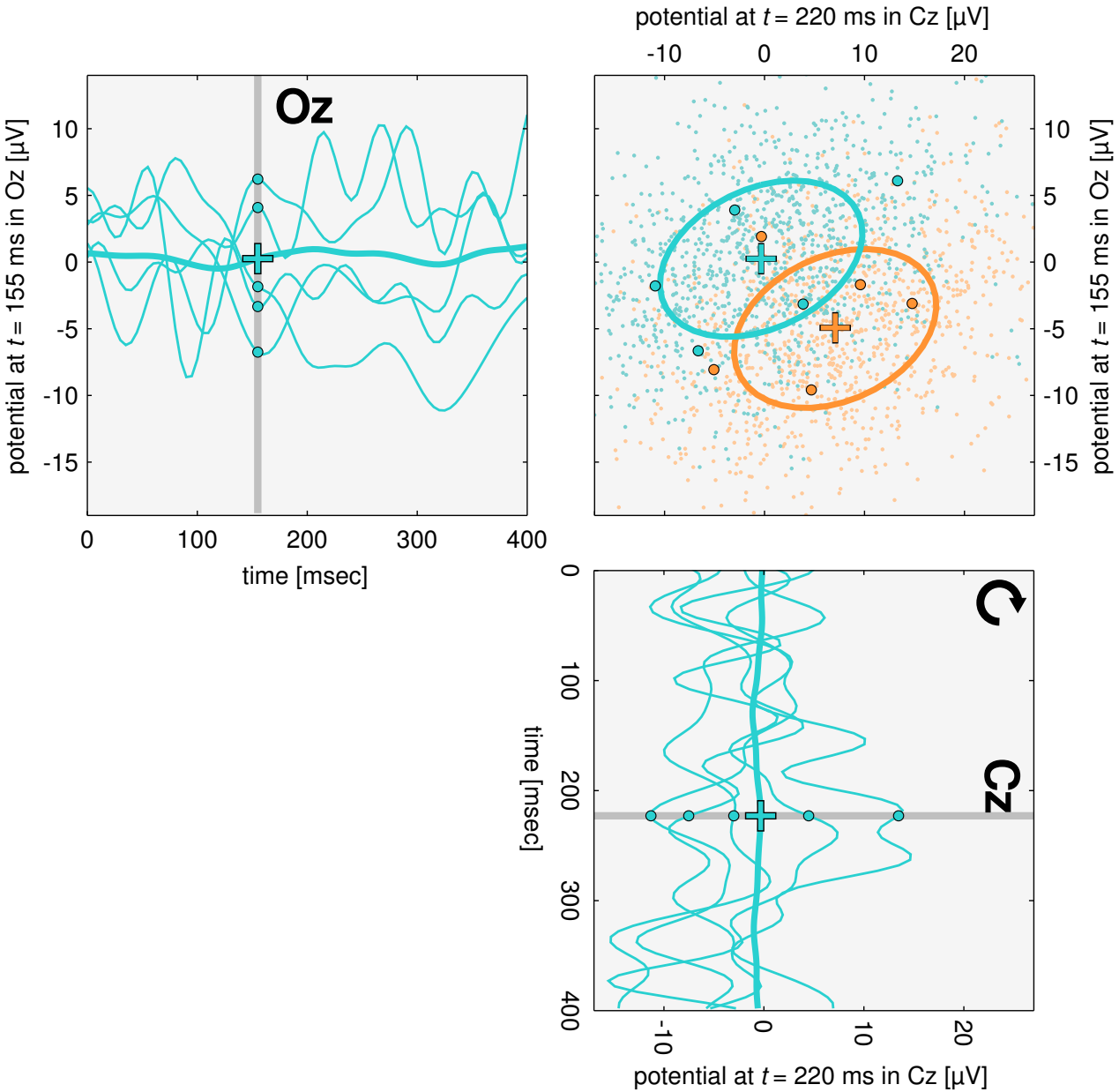
Representation of Multivariate Distributions: Scatter Plot



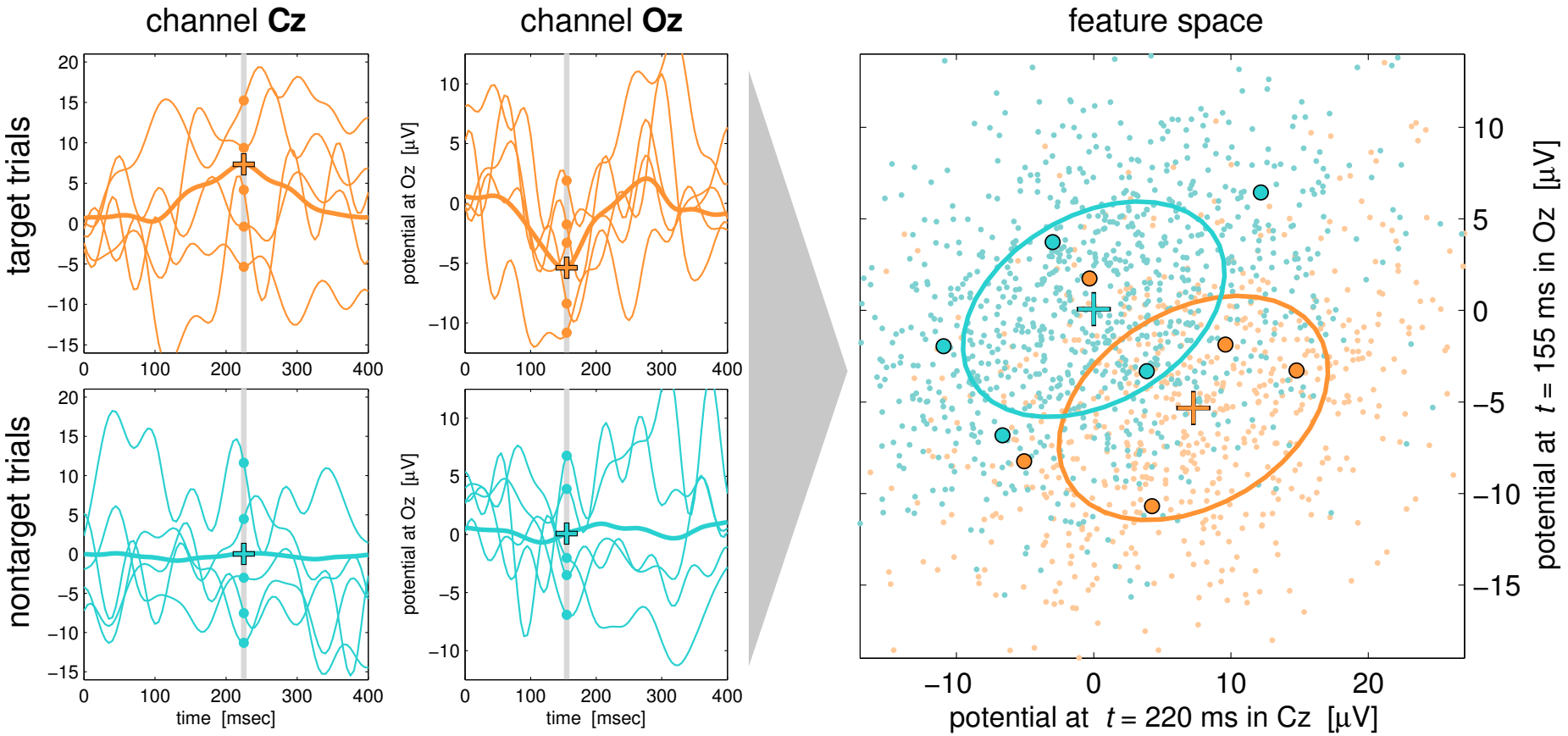
Representation of Multivariate Distributions: Scatter Plot



Representation of Multivariate Distributions: Scatter Plot

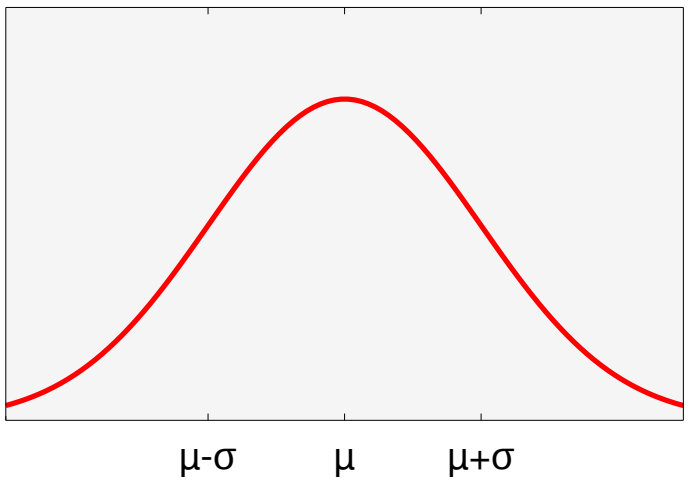


Representation of Multivariate Distributions (2)

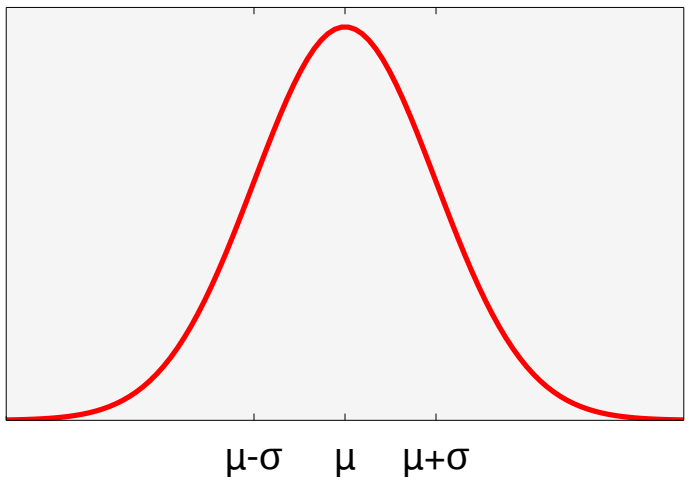


Two Univariate Gaussian Distributions

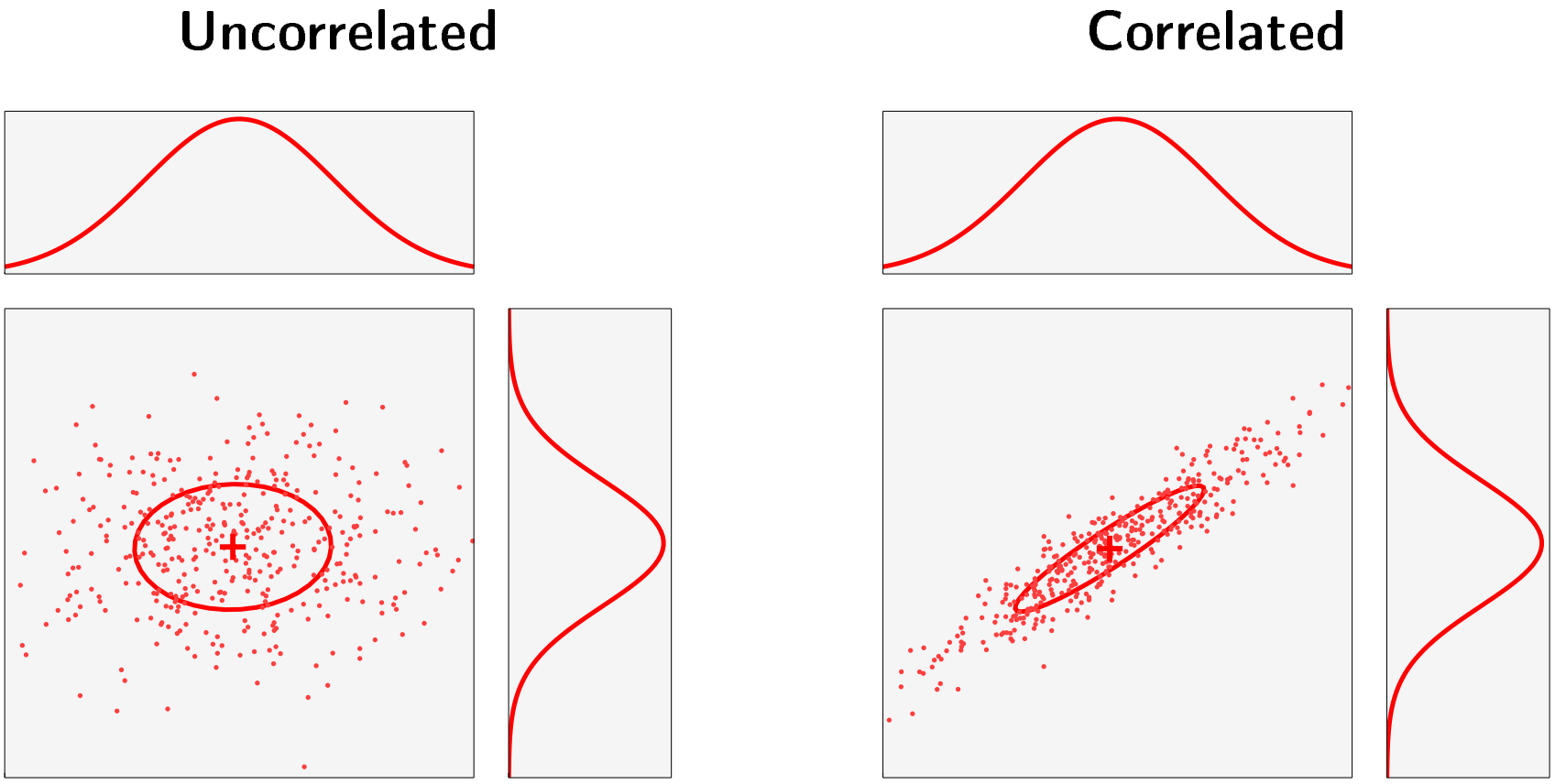
Component #1



Component #2



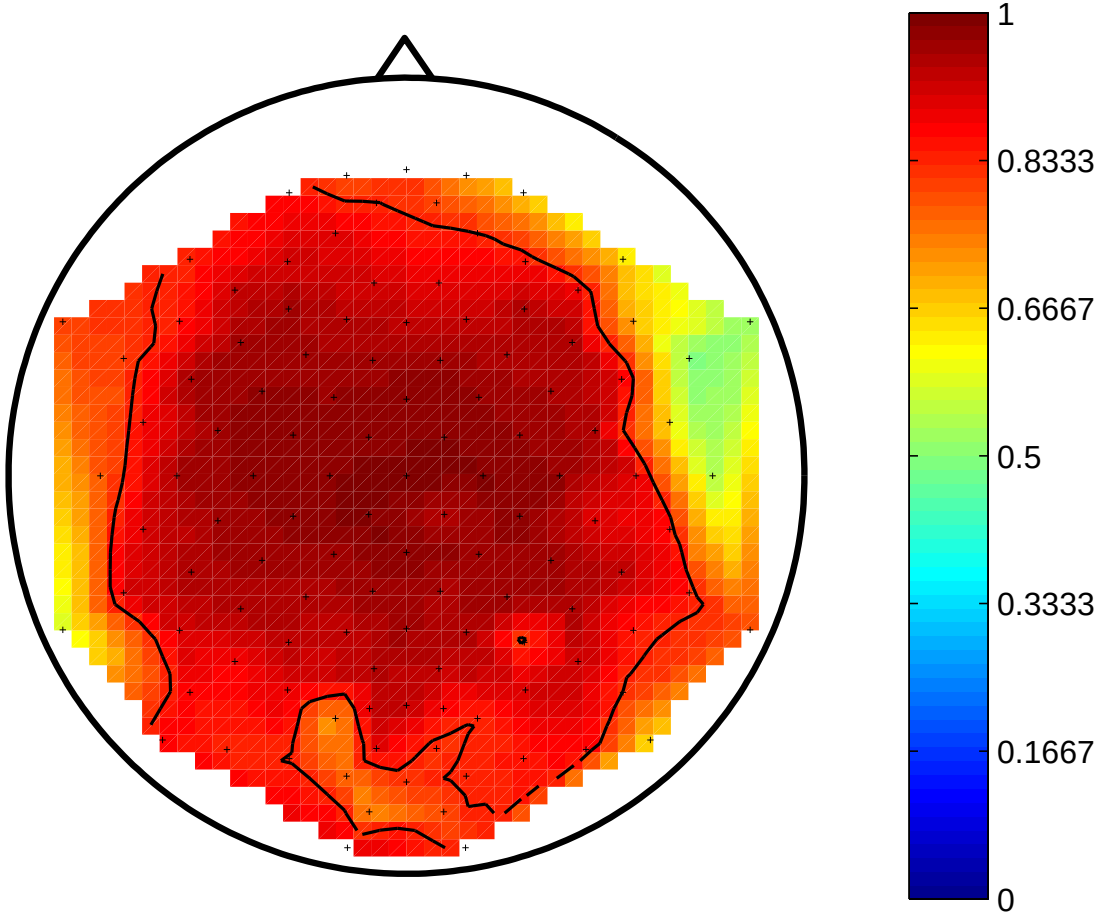
Two-Dimensional Gaussians - Correlated or Uncorrelated



- 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).

Correlated or Uncorrelated? Mind Spatial Smearing!

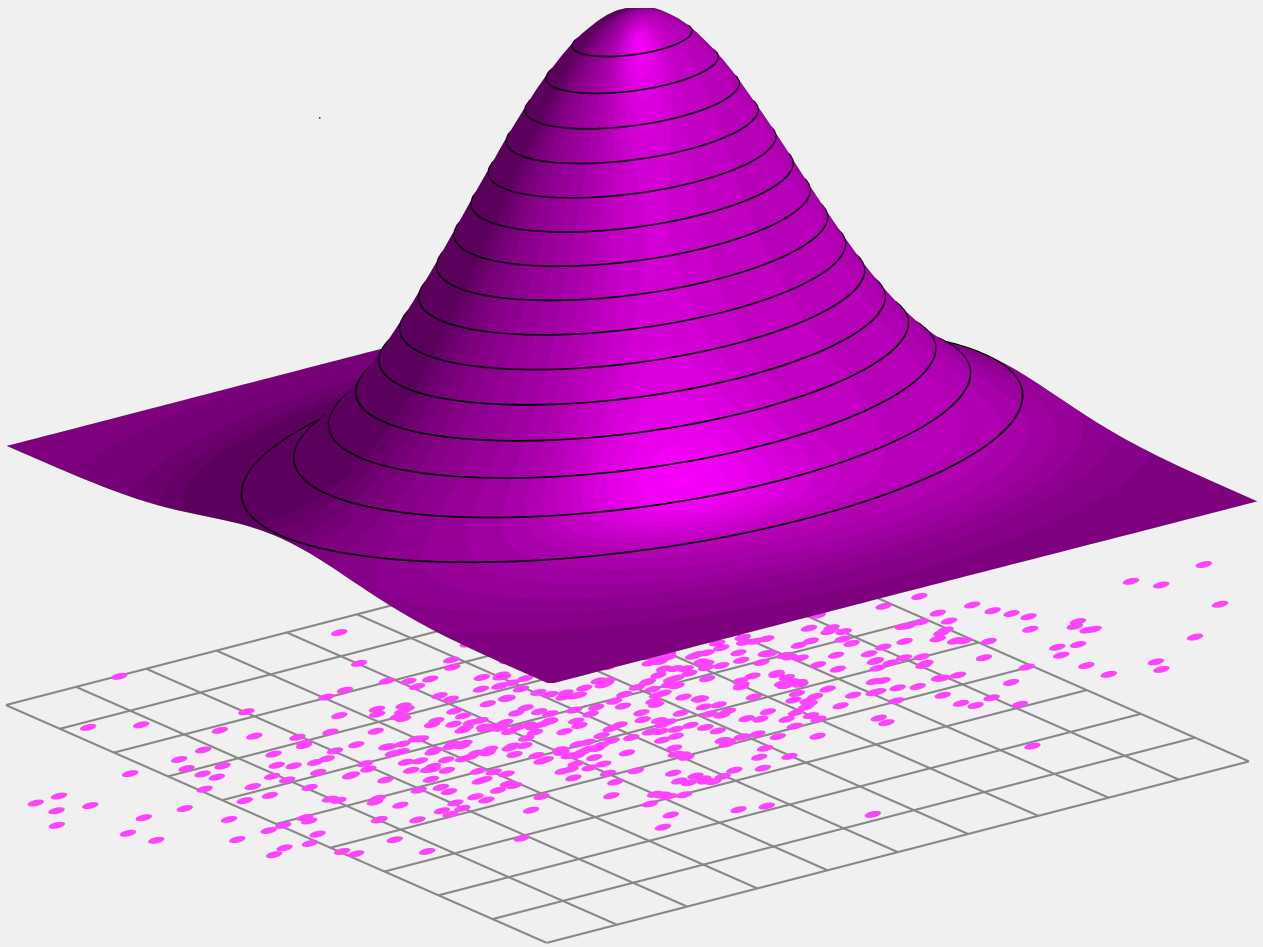
- Raw EEG scalp potentials are known to be associated with a large spatial scale owing to volumne conduction.
- In a simulation of Nunez et al [1] only half the contribution to one scalp electrode comes from sources within a 3 cm radius.



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

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

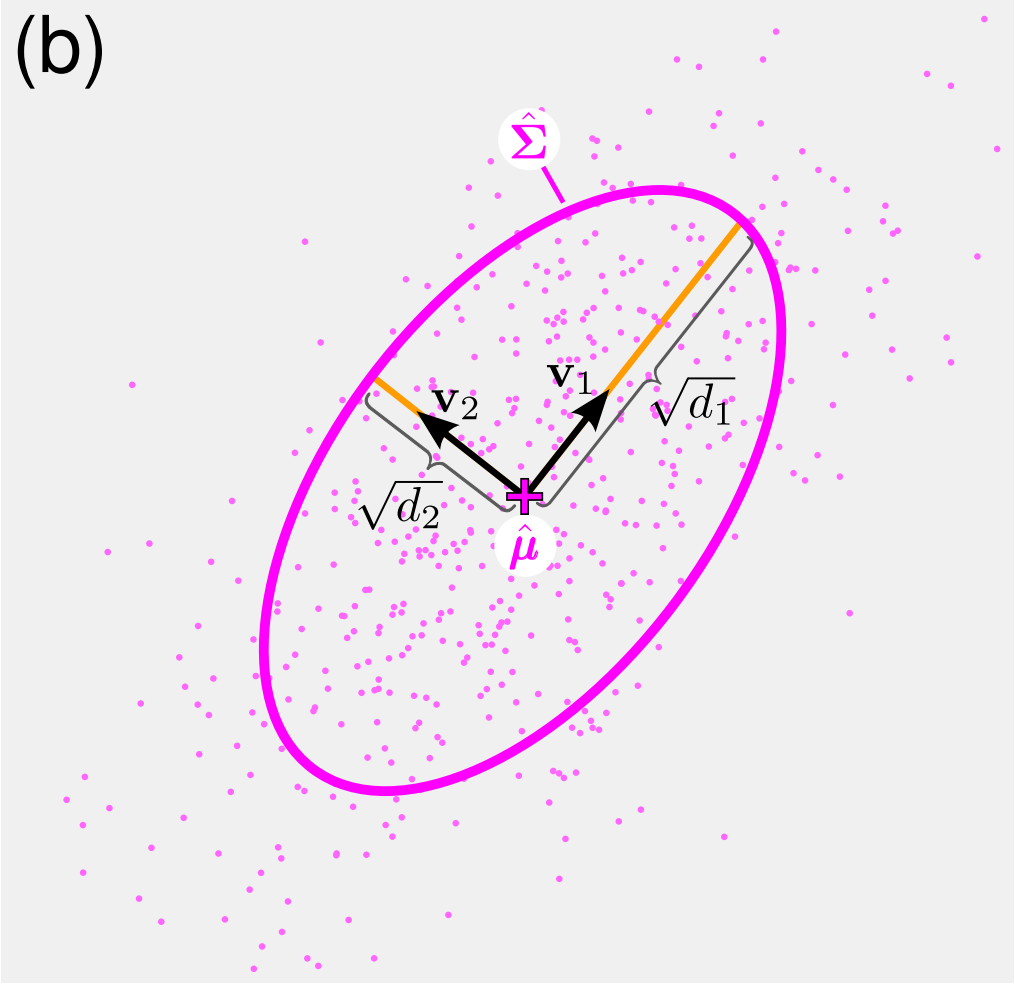
$$\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$$

Characterization of Gaussian Distributions

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

Eigenvalue decomposition of the empirical covariance matrix:

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



- 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 and $\boldsymbol{\Sigma}$ noise.
- The vector of Eigenvalues is called *Eigenvalue spectrum*

BBCI

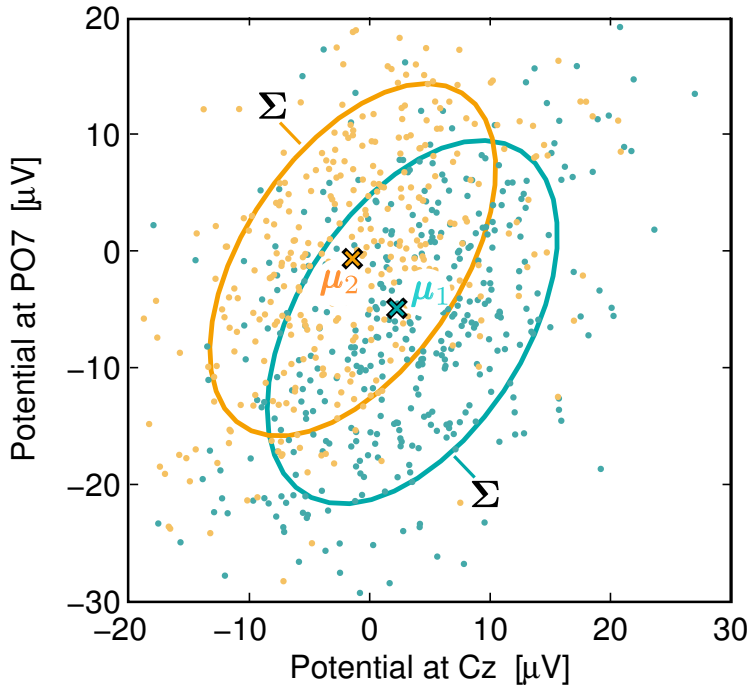
Distribution of ERP Features

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

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

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

with Gaussian noise: $\mathbf{n}^{(\cdot)}(t) \sim \mathcal{N}(0, \mathbf{\Sigma})$.



For features of ERP data:

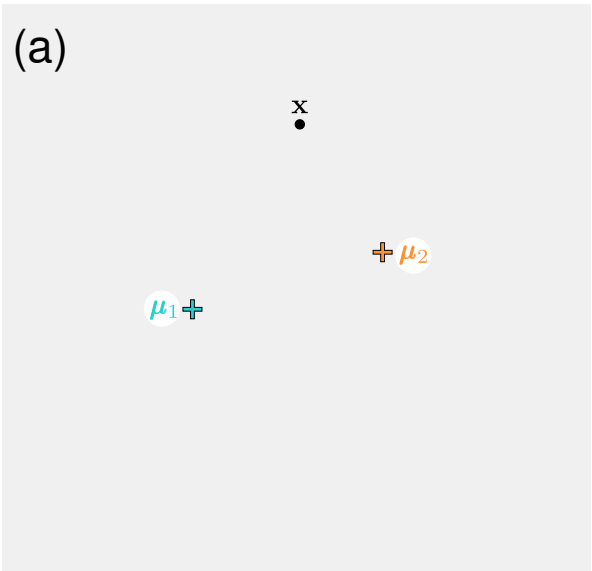
- $\boldsymbol{\mu}_1$: ERP of condition 1
- $\boldsymbol{\mu}_2$: ERP of condition 2
- $\mathbf{\Sigma}$: noise: non-phase-locked activity (independent of condition)

[Blankertz et al, NeuroImage 2011]

BBCI

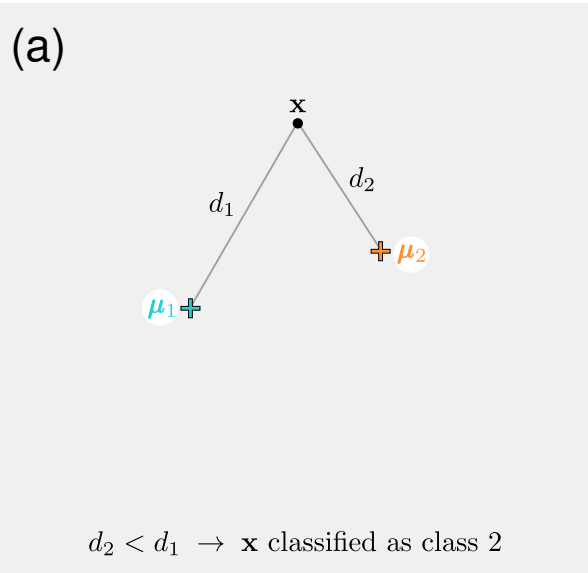
Nearest Centroid Classifier (NCC)

(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.



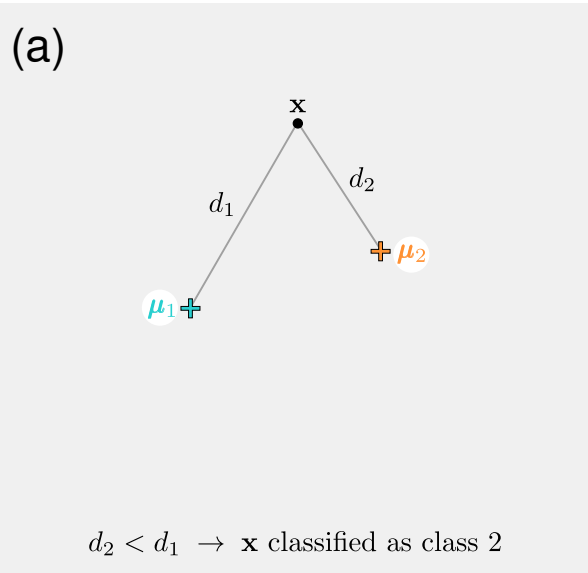
Nearest Centroid Classifier (NCC)

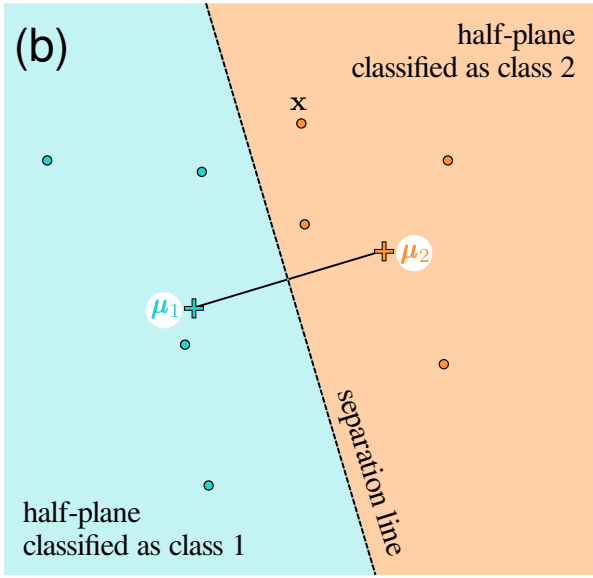
(a) Let us assume a simple setting of a classification problem with little information: Only the means (or centroids) $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of the two distributions are known.



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.

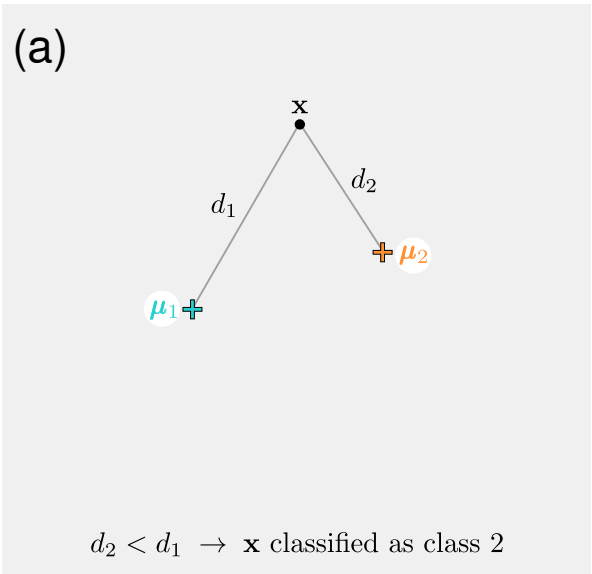


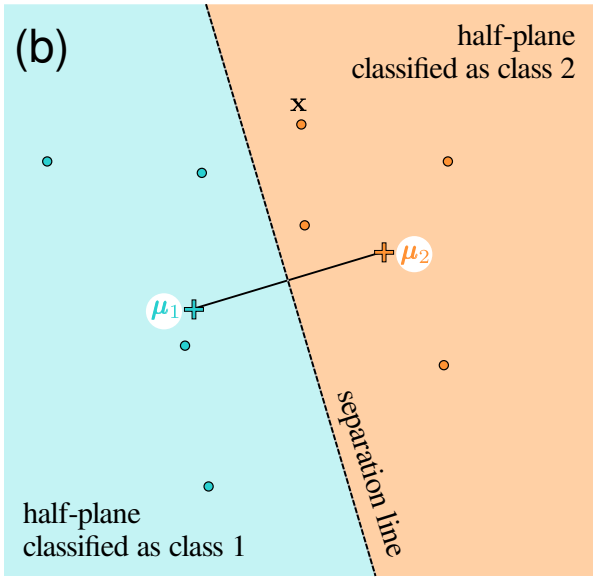


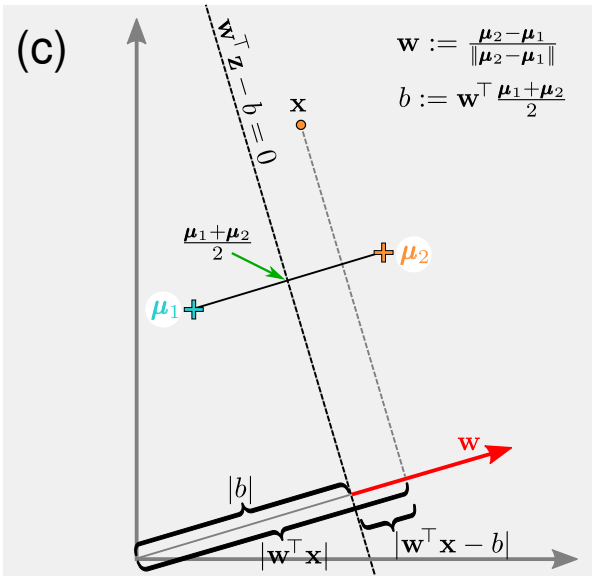
(b) This leads to a linear separation of the space with the separation line (or hyperplane in higher dimensions) intersecting perpendicularly the line connecting the centroids in the middle.

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.



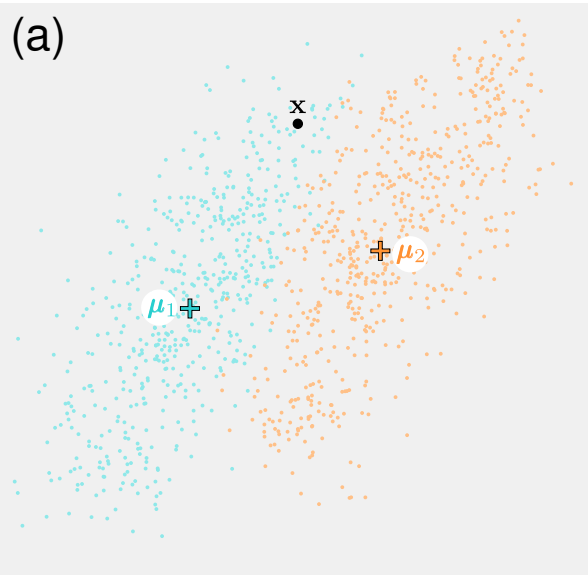




(b) This leads to a linear separation of the space with the separation line (or hyperplane in higher dimensions) intersecting perpendicularly the line connecting the centroids in the middle. **(c)** Mathematical formalism.

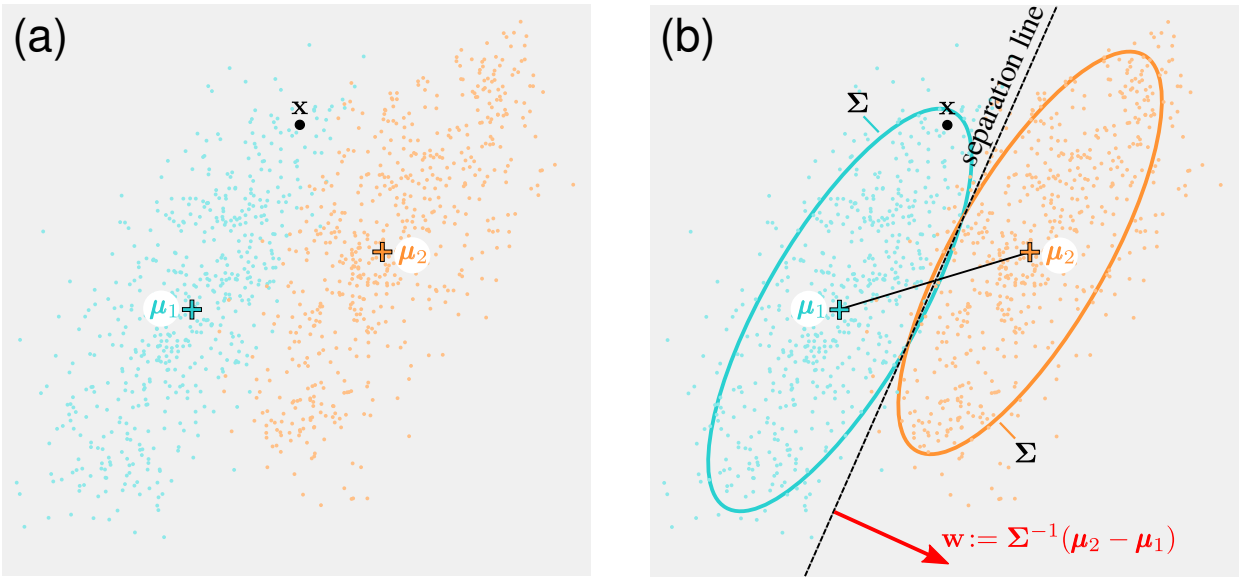
Linear Disciminant Analysis

(a) Means as before, but distributions according to real EEG data.



Linear Disciminant Analysis

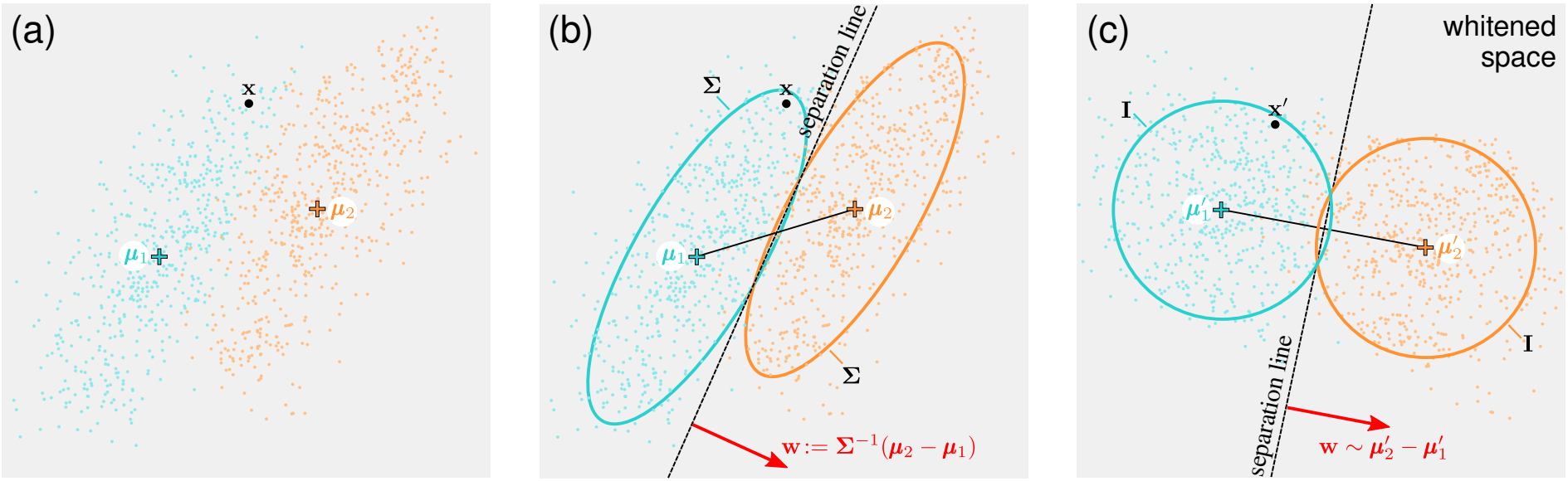
(a) Means as before, but distributions according to real EEG data.



(b) In Linear Discriminant Analysis, a common covariance matrix for both classes is estimated, which describes the (class-independent) noise.

Linear Disciminant Analysis

(a) Means as before, but distributions according to real EEG data.



(b) In Linear Discriminant Analysis, a common covariance matrix for both classes is estimated, which describes the (class-independent) noise.
(c) Correspondence to NCC.

Linear Discriminant Analysis

Linear Discriminant Analysis is based on the following assumptions:

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

Based on probability theory, the optimal classifier under these conditions can be derived:

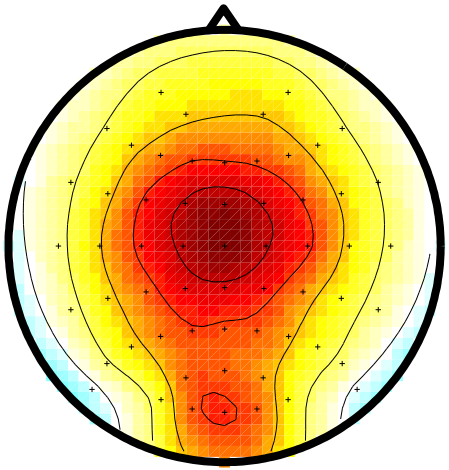
Given two Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, LDA is defined by the normal vector

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \quad \text{and bias} \quad b = \mathbf{w}^\top(\mu_1 + \mu_2)/2. \tag{1}$$

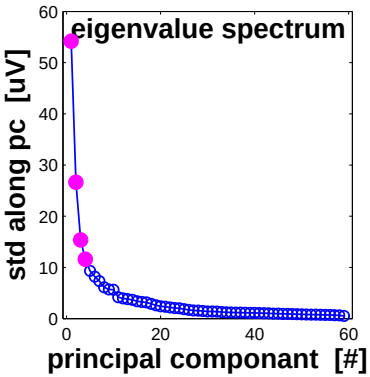
On the subsequent slides, we discuss the assumptions.

Mean and Eigenvalue Spectrum for a P300 Data Set

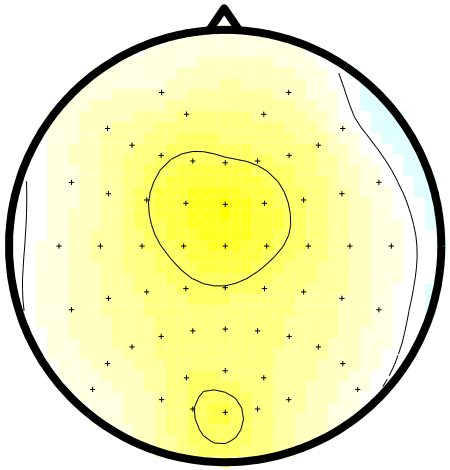
target



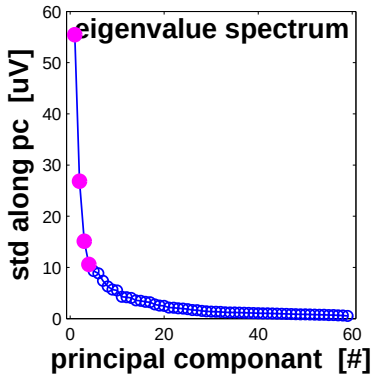
average *target*



non-target

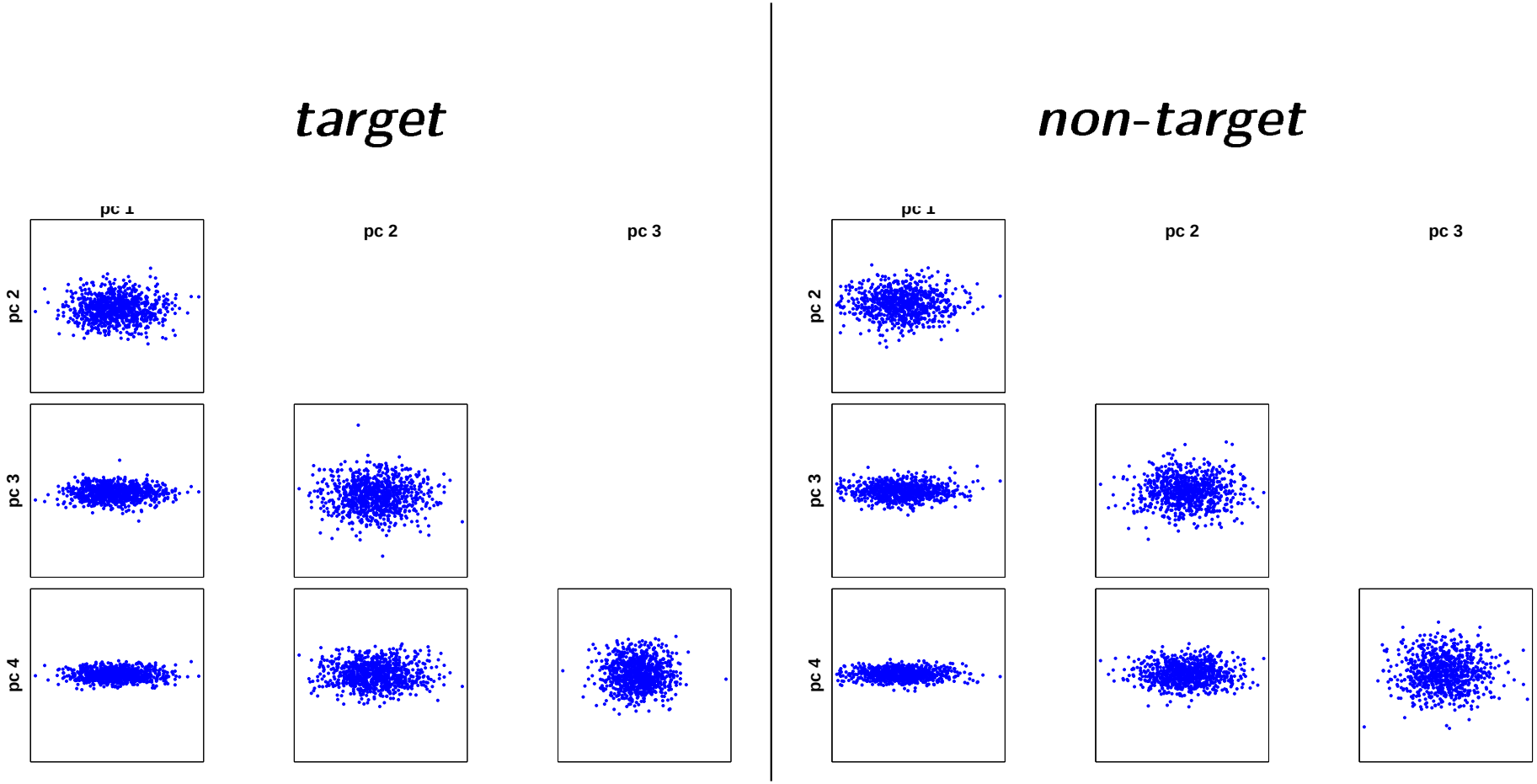


average *nontarget*

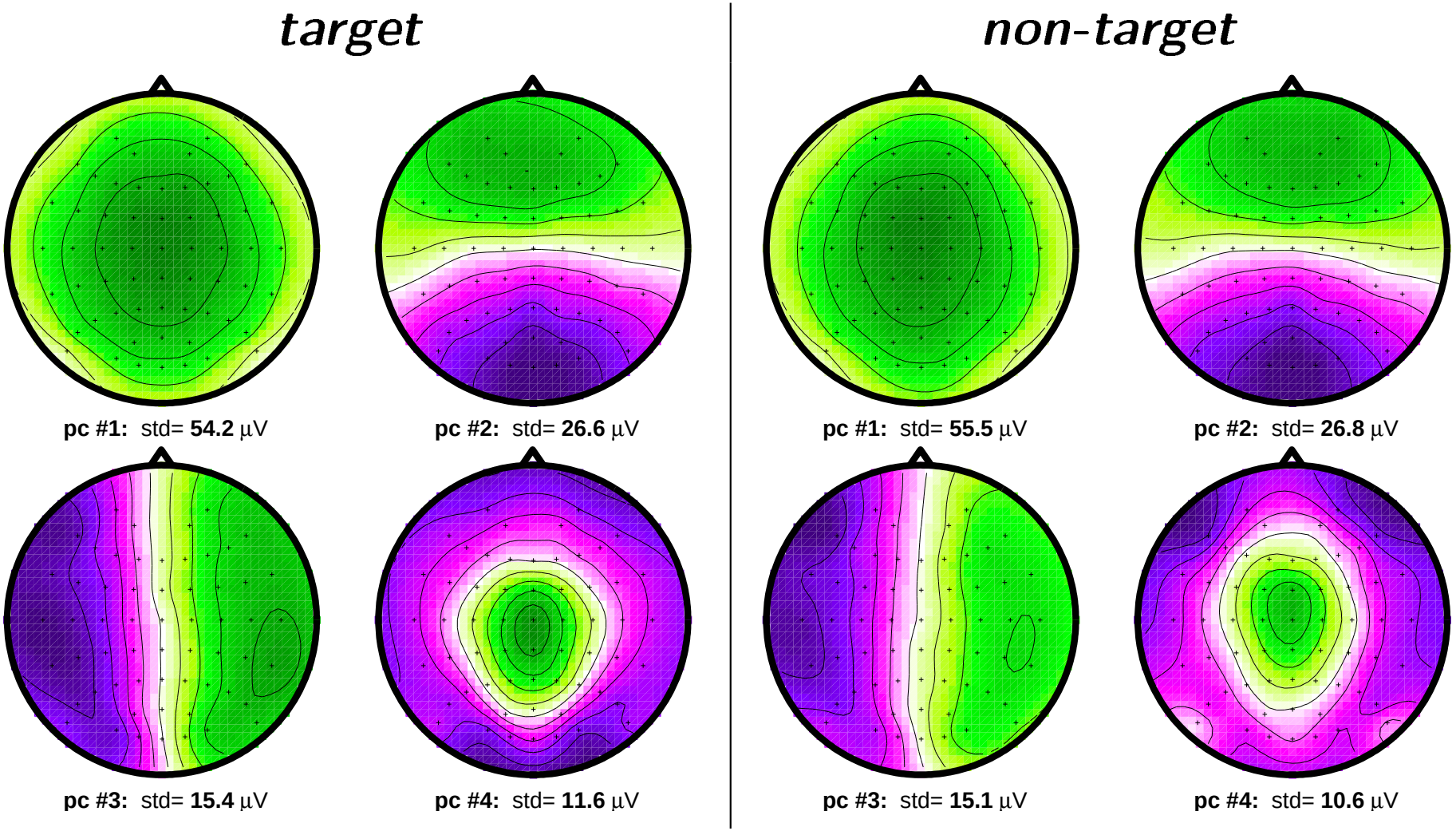


Distribution of the Noise

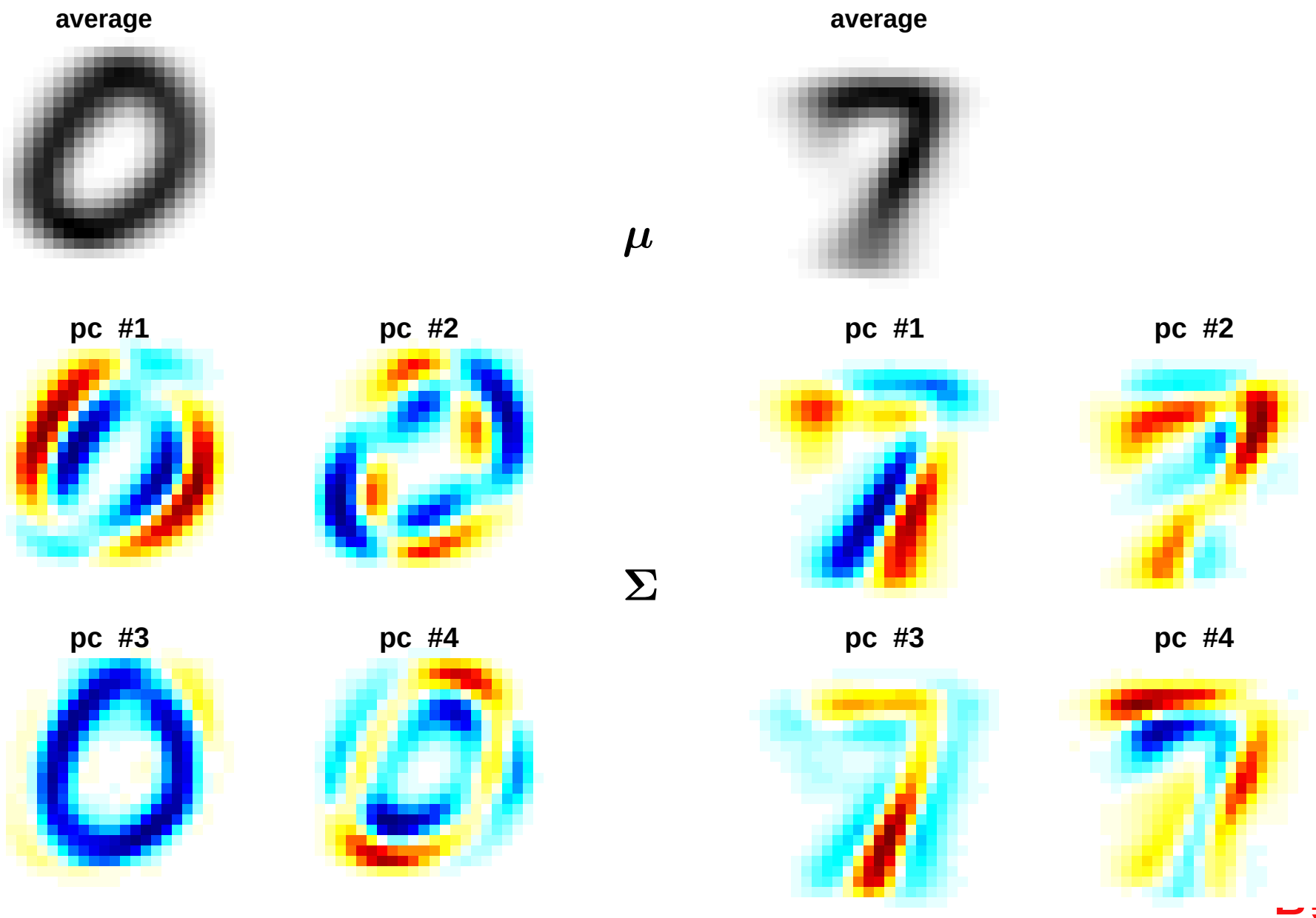
Scatter plots of projections on PCs:



The Structure of the Noise



For Comparison: Covariances in Handwritten Digits



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.

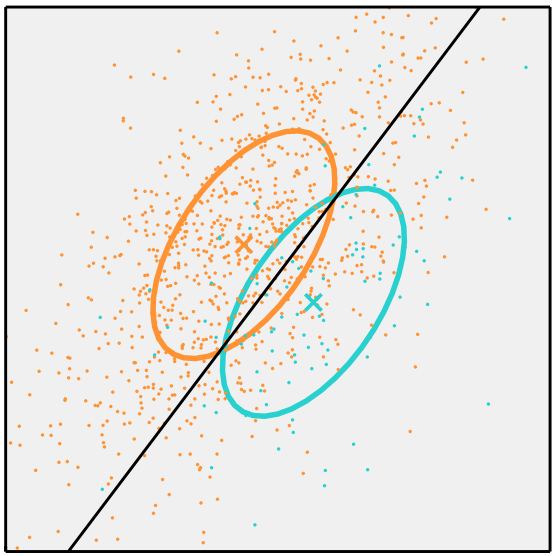
These sets of samples have to be disjoint and **INDEPENDENT**.

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

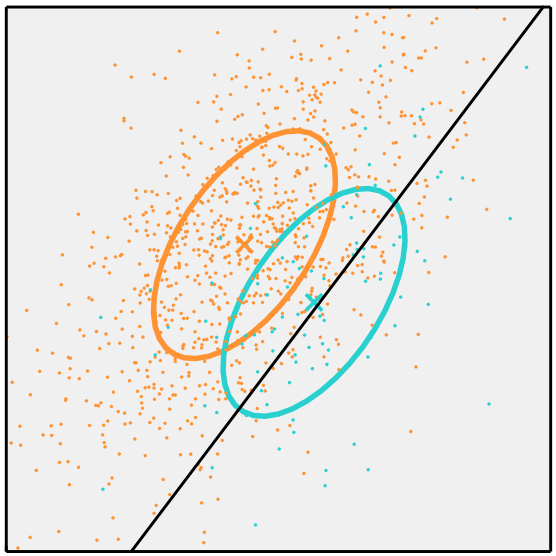
Loss Function for Unbalanced Classes

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

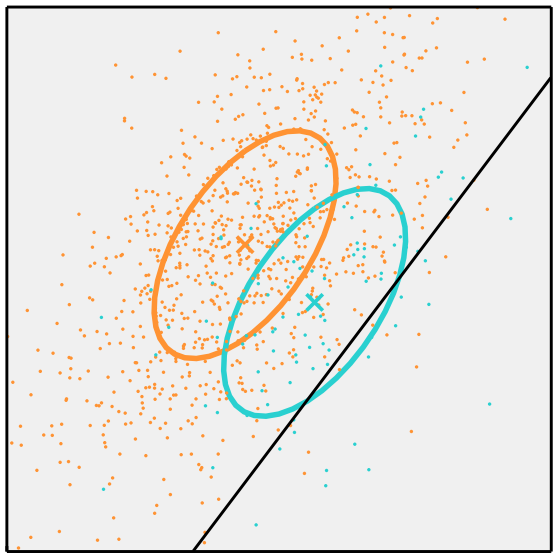
Examples of weighted and unweighted error – bias of classifier is varied:



Error
Unweighted: 26.6%
Weighted: 22.4%
ROC-based: 16.7%



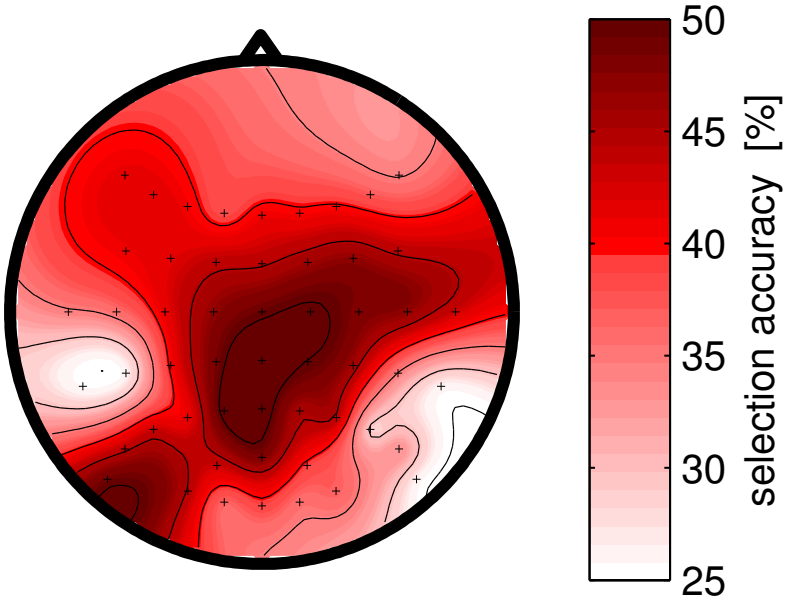
Error
Unweighted: 12.8%
Weighted: 29.8%
ROC-based: 16.7%



Error
Unweighted: 9.8%
Weighted: 41.9%
ROC-based: 16.7%

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* is to investigate the spatial distribution of discriminative information:

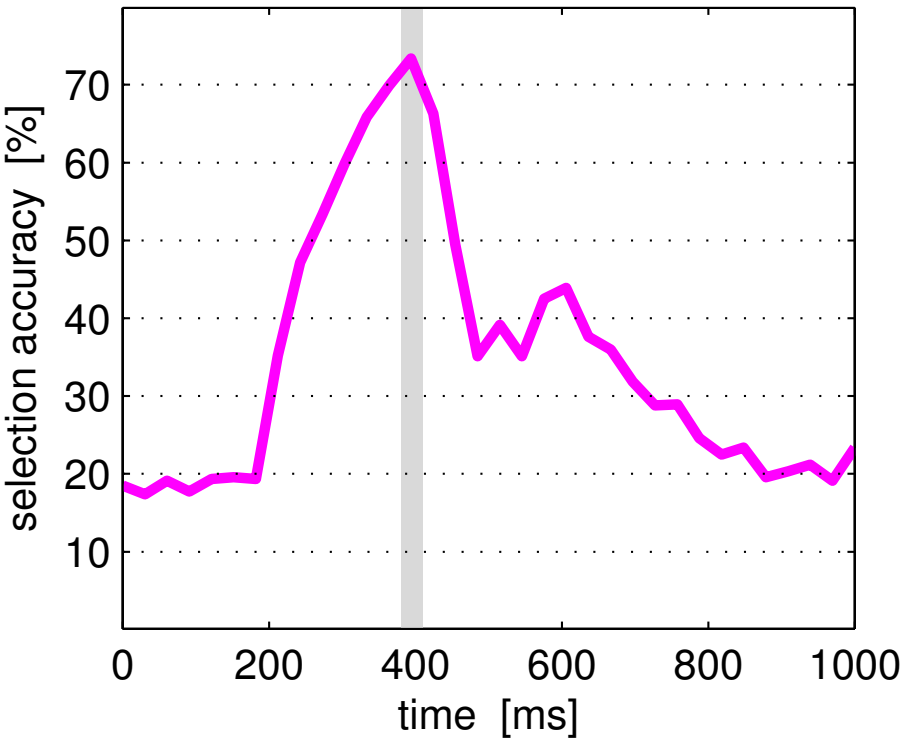


For each single channel the classification performance is determined for temporal features with LDA by cross validation. The resulting error values can be visualized as scalp topography.

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:

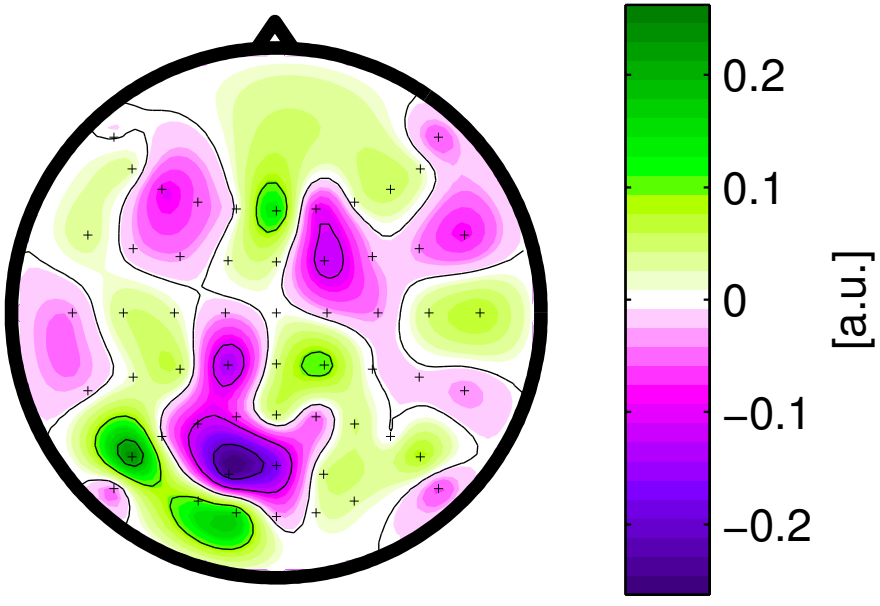


The classification error of spatial features was determined for each time interval of 30 ms duration, shifted from 0 to 1000 ms.

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:

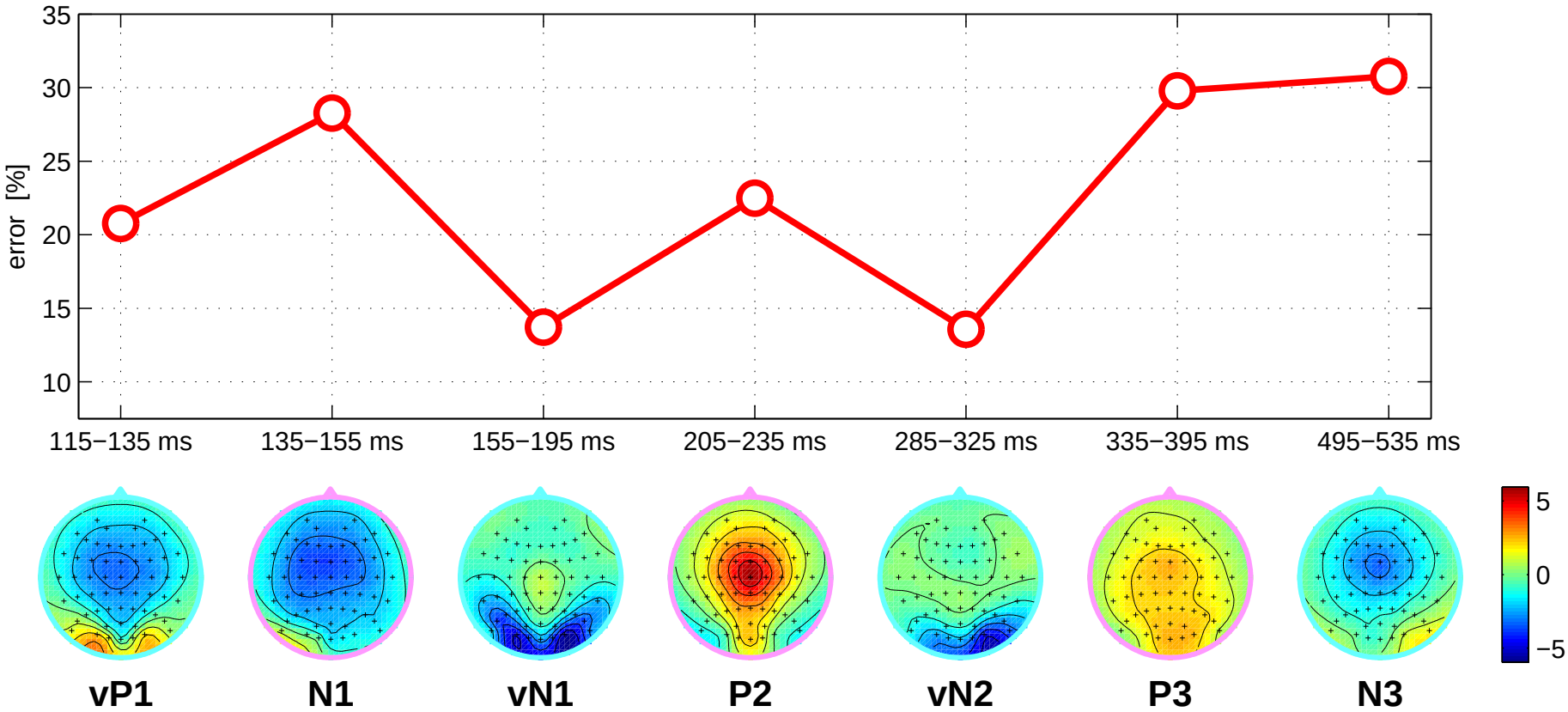


LDA trained on spatial features extracted from the time interval 380–410 ms. The resulting weight vector can be visualized as a topography and can be regarded as a spatial filter.

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

Results of Classifying Spatial Features

Classifying on spatial features corresponding to the prominent discriminative components to error rates between 14% and 31%:



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.

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.

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.

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:

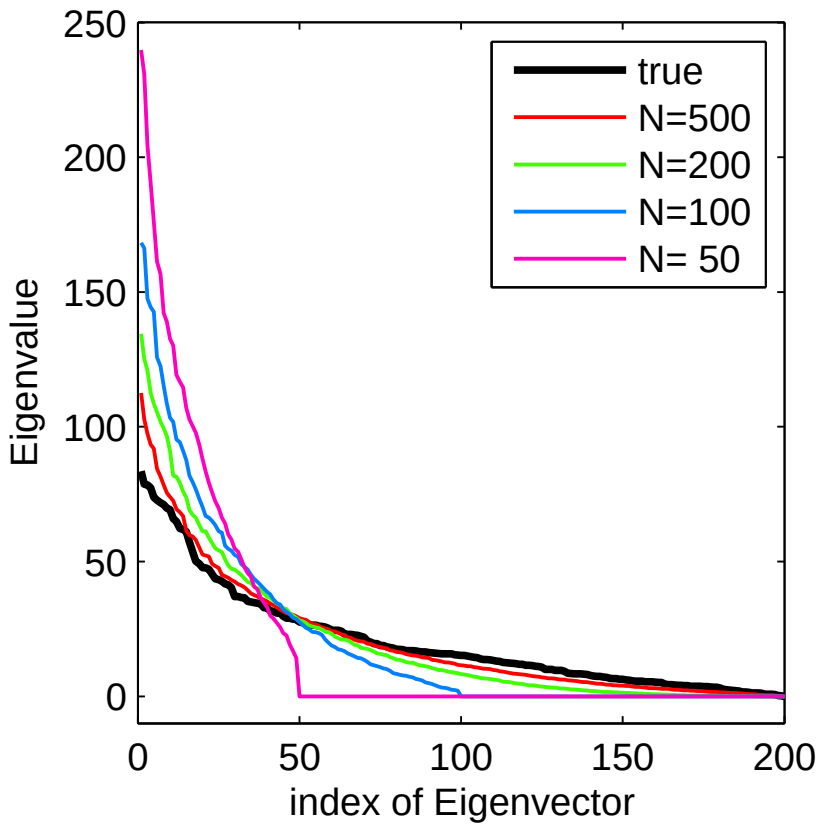
- Large Eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are too large
- Small Eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are too small

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

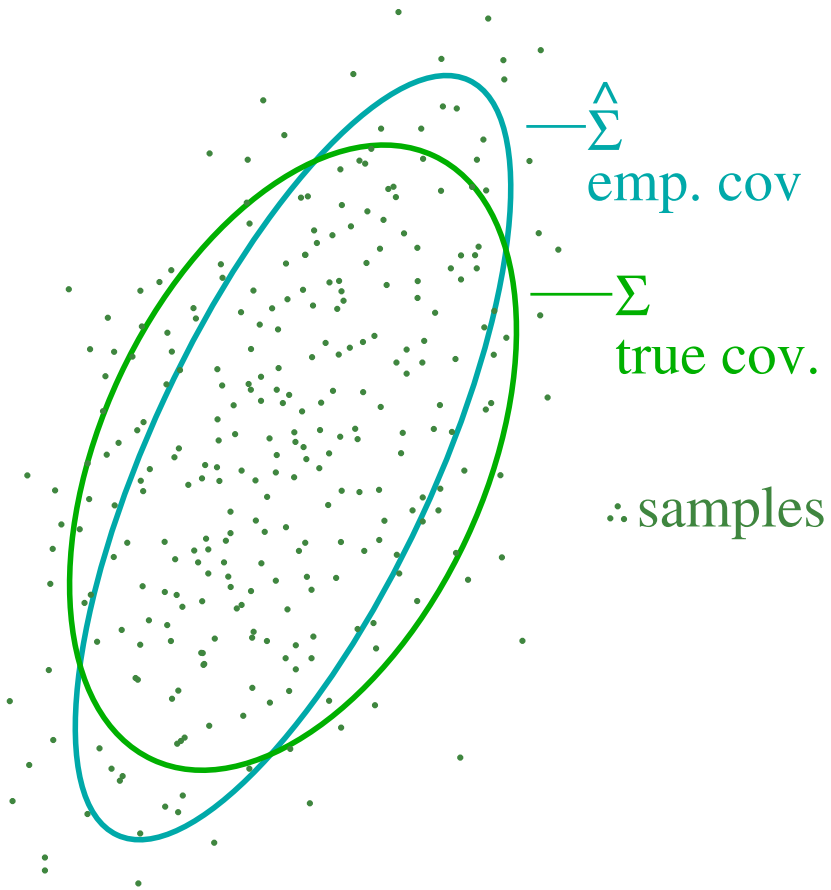
BBCI

Bias in Estimating Covariances (2)

Simulation for $d = 200$:



Cartoon in 2D:

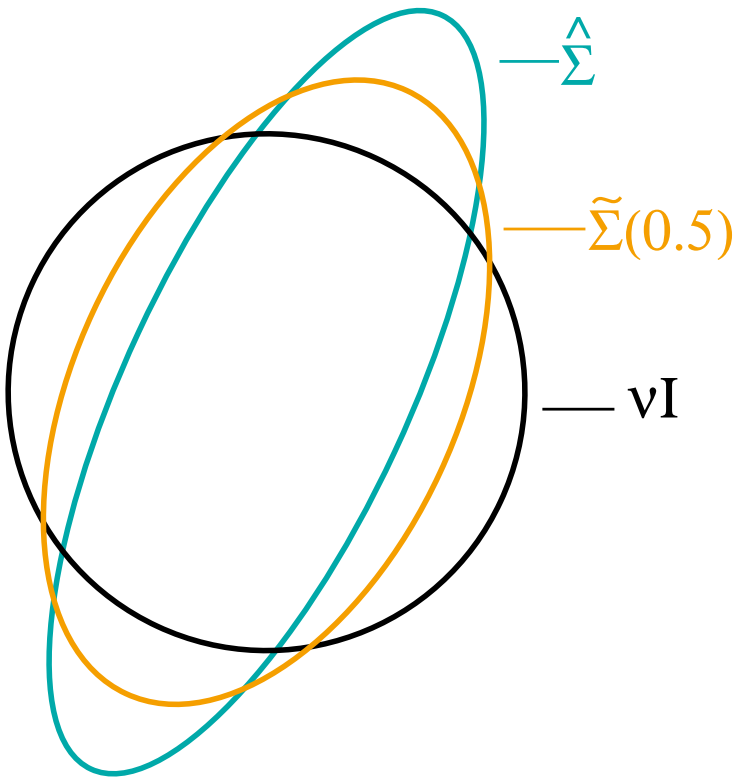


A Remedy for the Estimation Bias

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:

Properties of the Shrunk Covariance Matrix

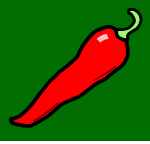
From the Eigenvalue decomposition of the empirical covariance matrix $\hat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ with orthonormal $\mathbf{V}$ and diagonal $\mathbf{D}$, we get an Eigenvalue decomposition of $\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$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

BBCI



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 $\tilde{\mathbf{\Sigma}}(\gamma)$:

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

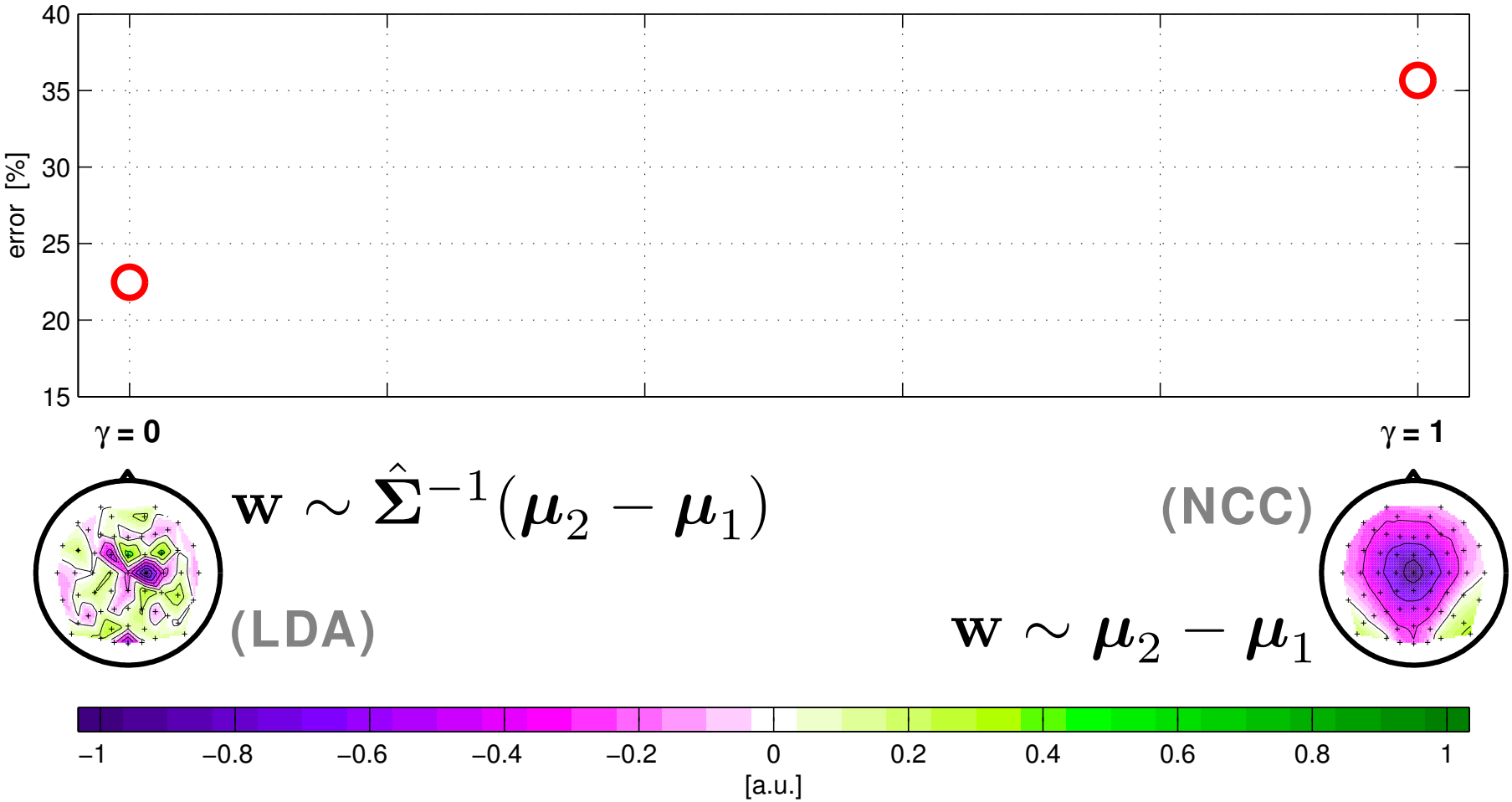
Here, γ is a hyper parameter that has to be selected between 0 and 1.

- $\gamma = 0$ 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.

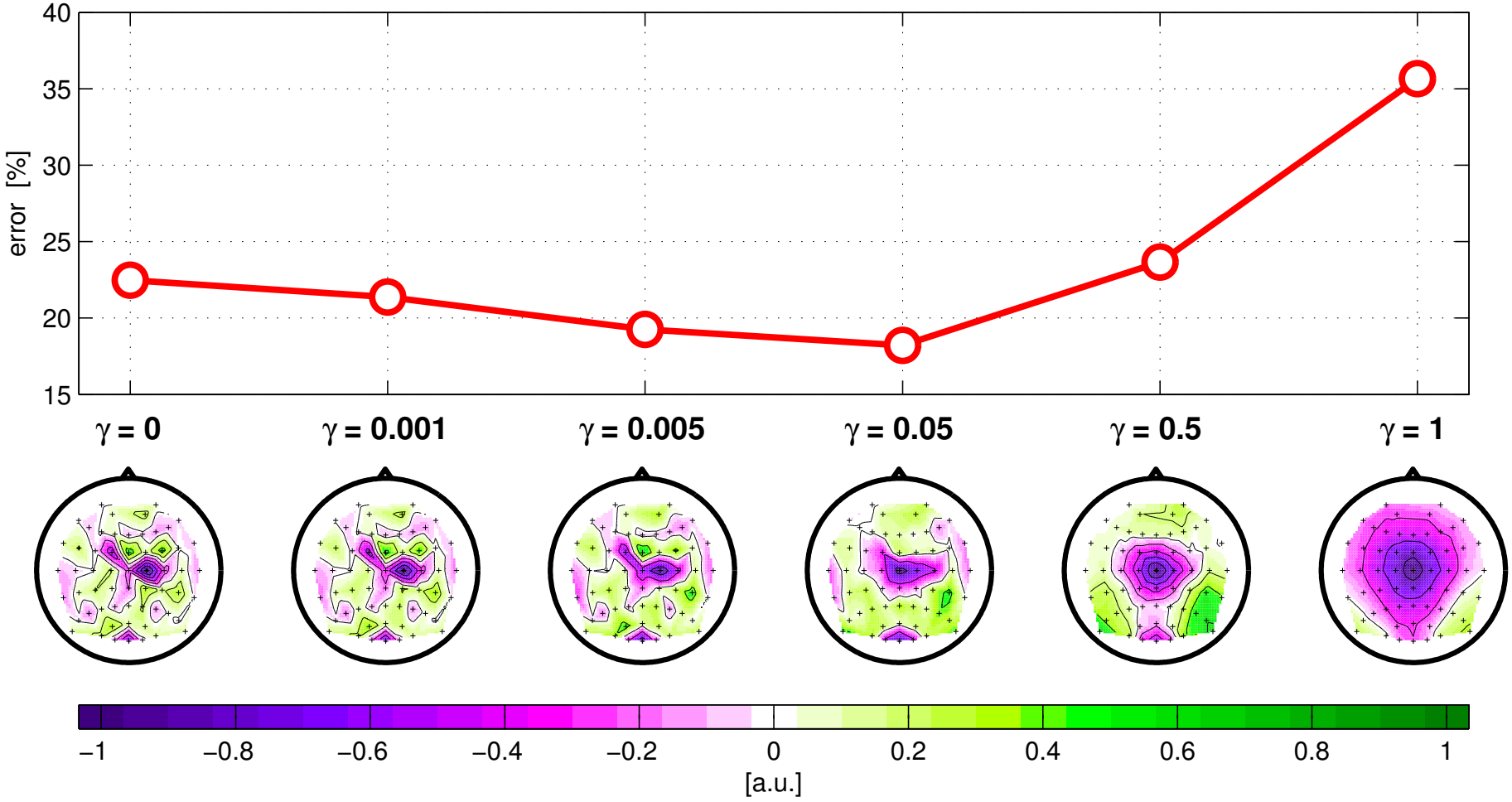
Impact of Shrinkage as Trade-off

LDA with shrinkage: $\mathbf{w} = \tilde{\Sigma}(\gamma)^{-1}(\mu_2 - \mu_1); \quad \tilde{\Sigma}(\gamma) = (1-\gamma)\hat{\Sigma} + \gamma\nu\mathbf{I}$



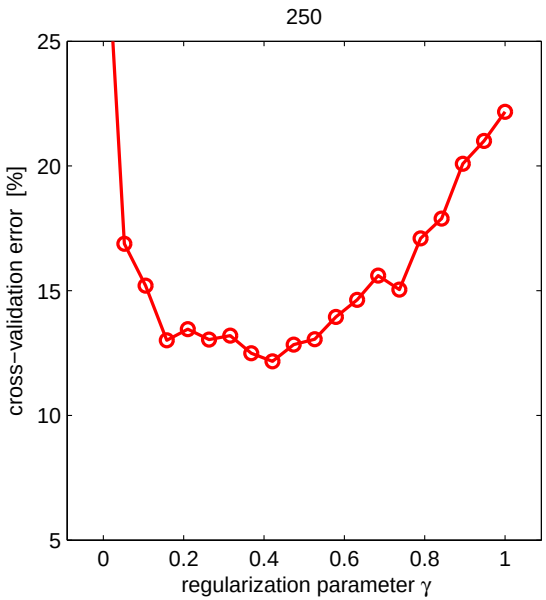
Impact of Shrinkage as Trade-off

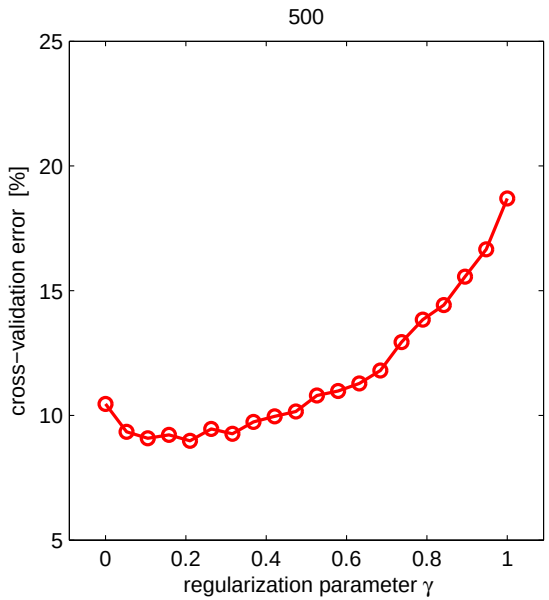
With increasing shrinkage, the spatial filters (classifier) look smoother, but classification may degrade with too much shrinkage.

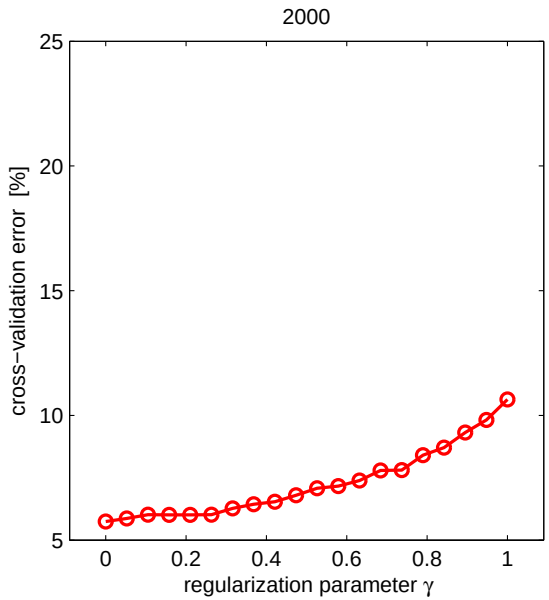


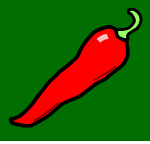
Regularized LDA at Work

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









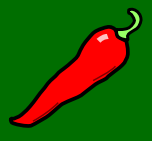
Optimal Selection of Shrinkage Parameter

LDA with shrinkage of the covariance matrix has one free parameter (γ), also called hyperparameter, that needs to be selected. There is no general way to do it.

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 correlation coefficient of feature i and j in trial k:

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

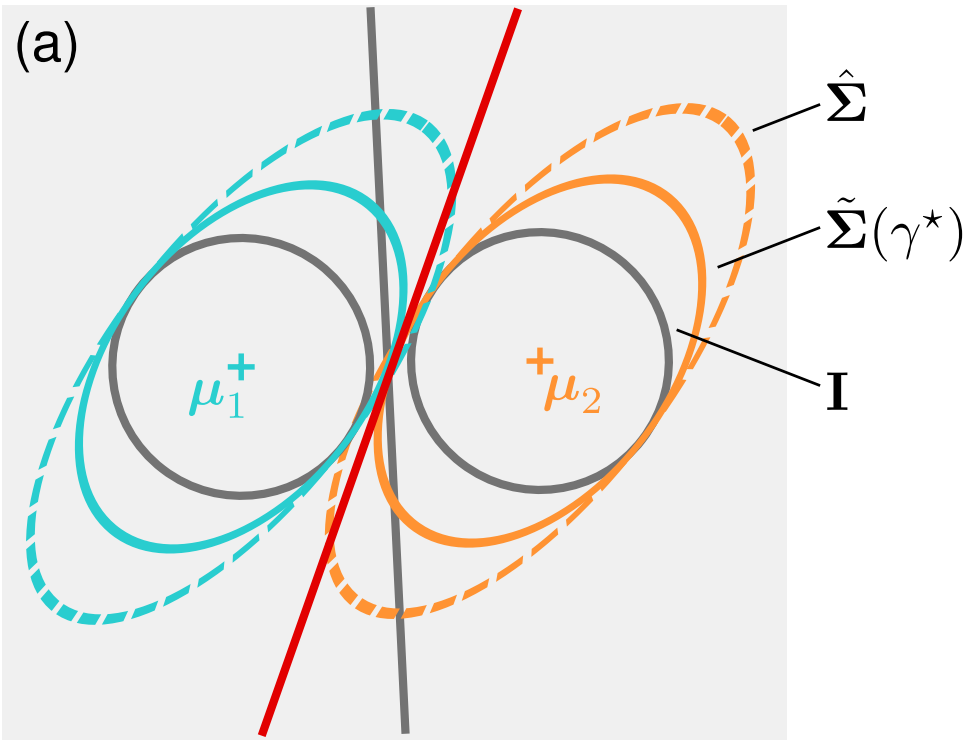
Denoting by s_{ij} the element in the i-th row and j-th column of the matrix $\hat{\mathbf{\Sigma}} - \nu\mathbf{I}$, the optimal shrinkage parameter $\gamma^\star = \operatorname{argmin}_\gamma \|\tilde{\mathbf{\Sigma}}(\gamma) - \mathbf{\Sigma}\|_F^2$ can be analytically calculated as [Schäfer & Strimmer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

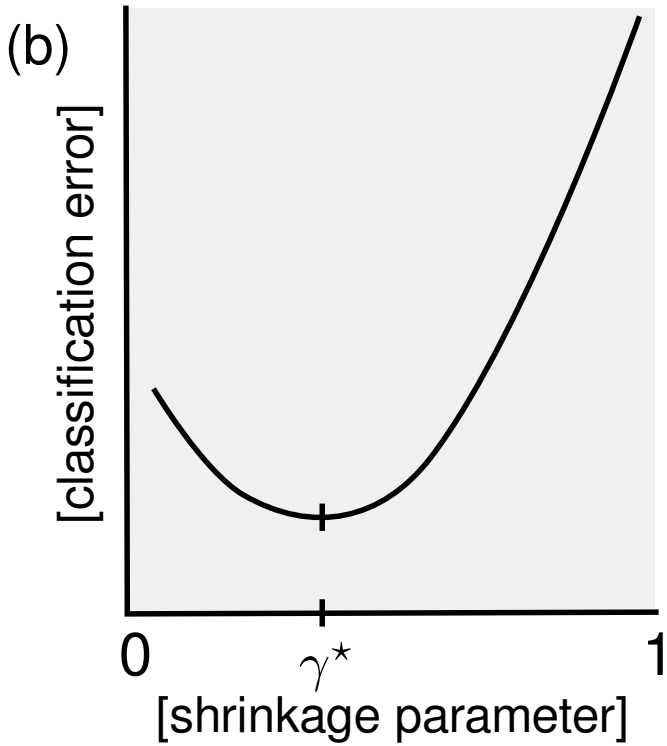
$$\gamma^\star = \frac{n}{(n-1)^2}\,\frac{\sum_{i,j=1}^{d} \operatorname{var}_k(z_{ij}(k))}{\sum_{i,j=1}^{d} s_{ij}^2}.$$

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

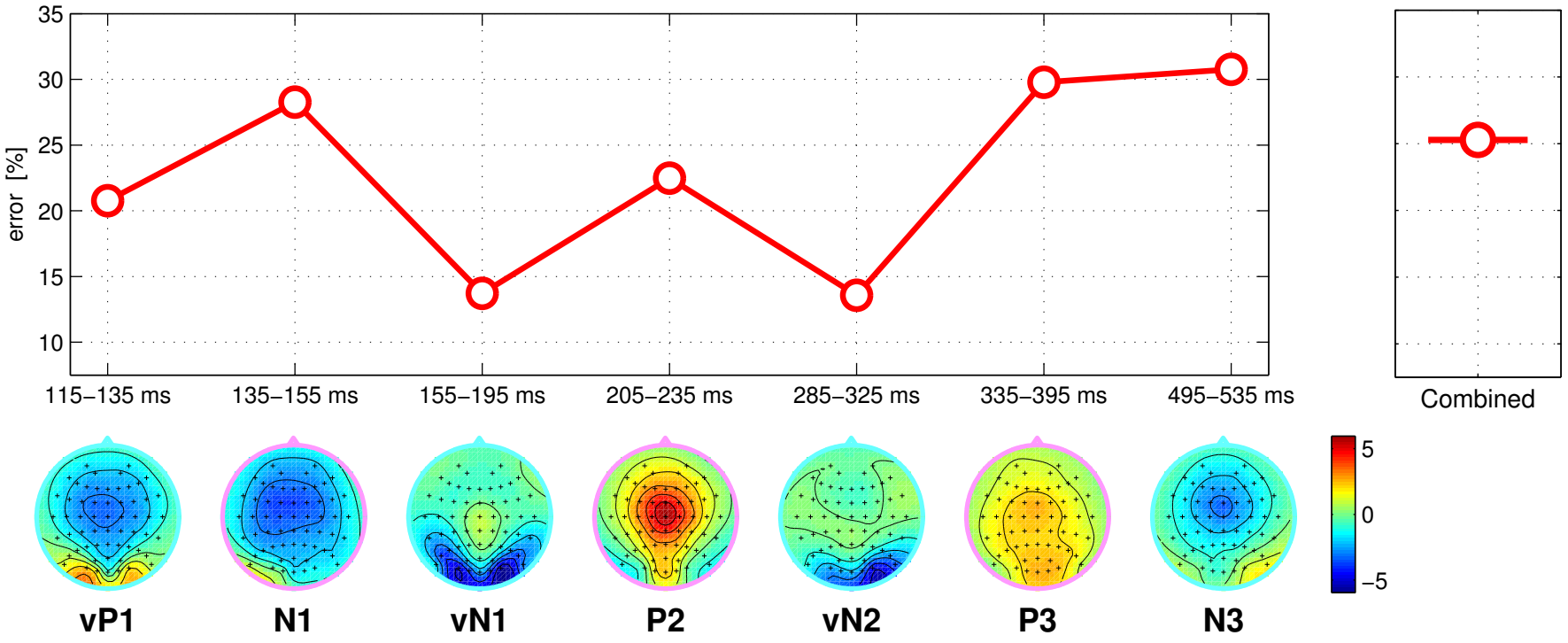
BBCI

Classification with Shrinkage-LDA





Classification on Single Components and Combined

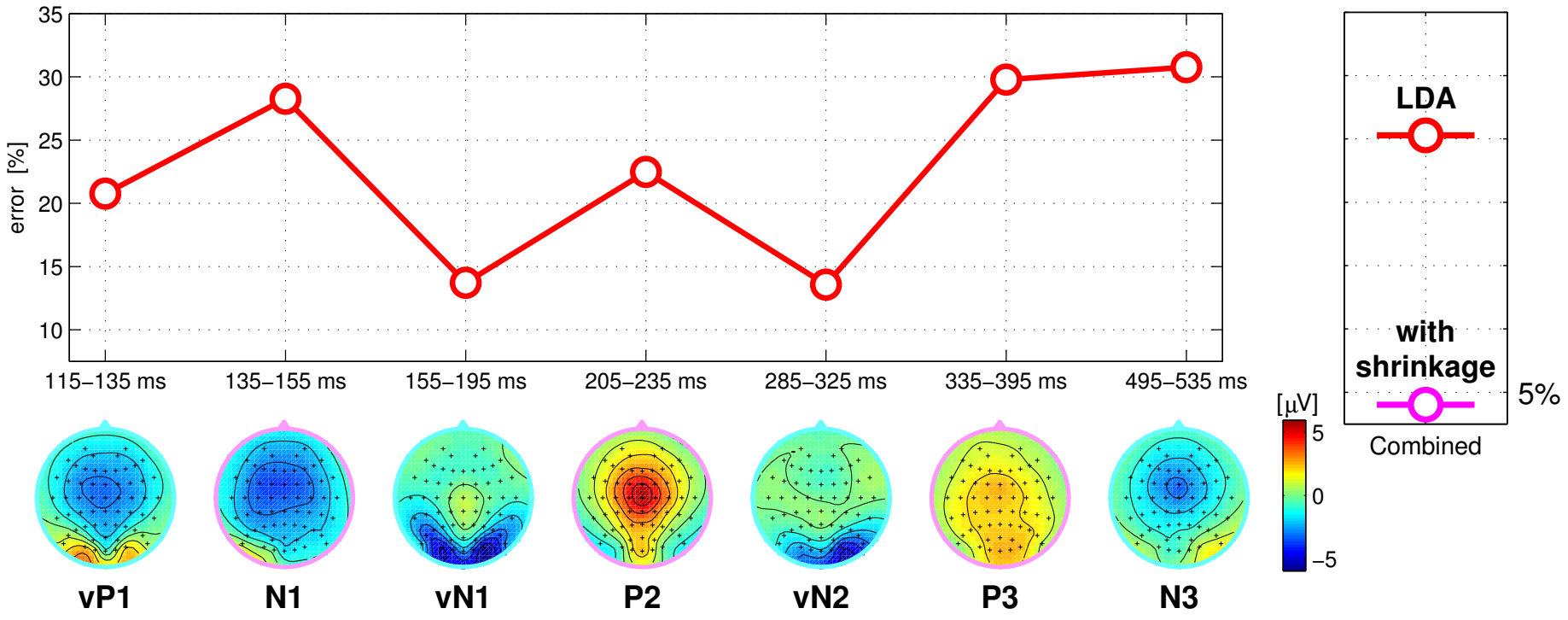


Classification (with $N = 750$ training samples) on seven different single components ($d = 55$) yields errors between **14%** and 31%.

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

[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*.

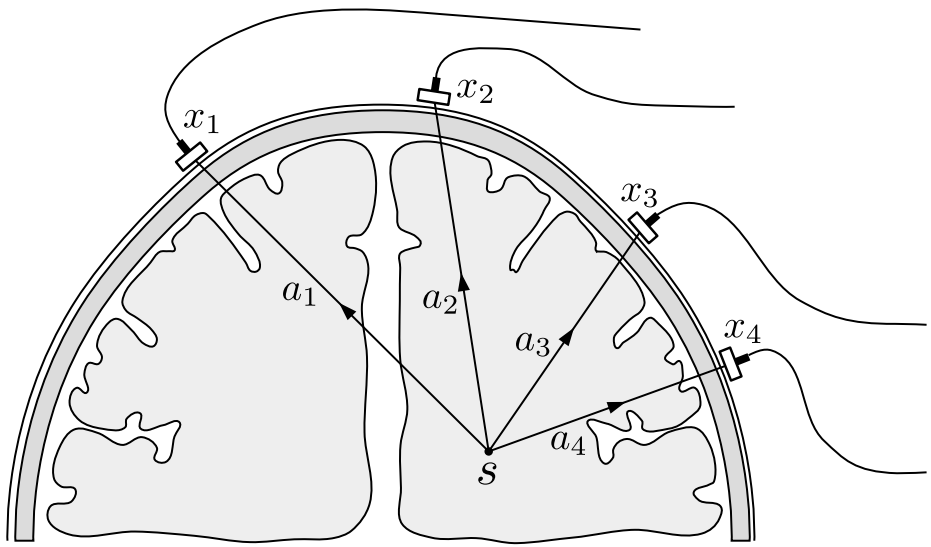
Shrinkage-LDA: only **4%** error. [Blankertz et al, NeuroImage 2011]

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 $\mathbf{a}_1, \ldots, \mathbf{a}_k$.
- Potentials are additive. Defining the matrix $\mathbf{A}$ as being composed of the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_k$ (i.e., $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_k]$), the **Forward Model** is

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

- 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

More generally, recovering of sources is the **backward model**:

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

Linear Model of EEG: Backward Model

More generally, recovering of sources is the **backward model**:

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

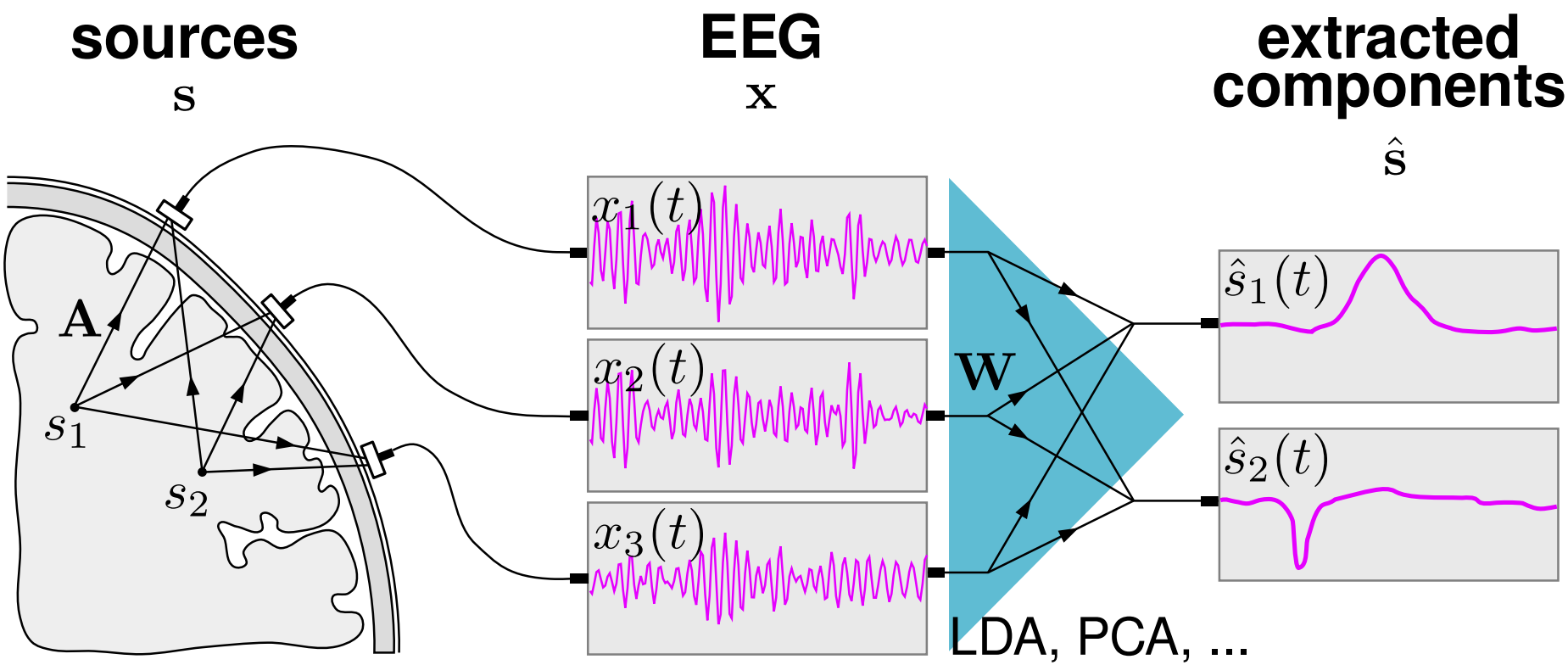
Given a forward model $\mathbf{A}$, taking $\mathbf{W}^\top$ as $\mathbf{A}^\# = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top$, the pseudo inverse of $\mathbf{A}$, is the least mean squares estimator:

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

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

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

Linear Model of EEG



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

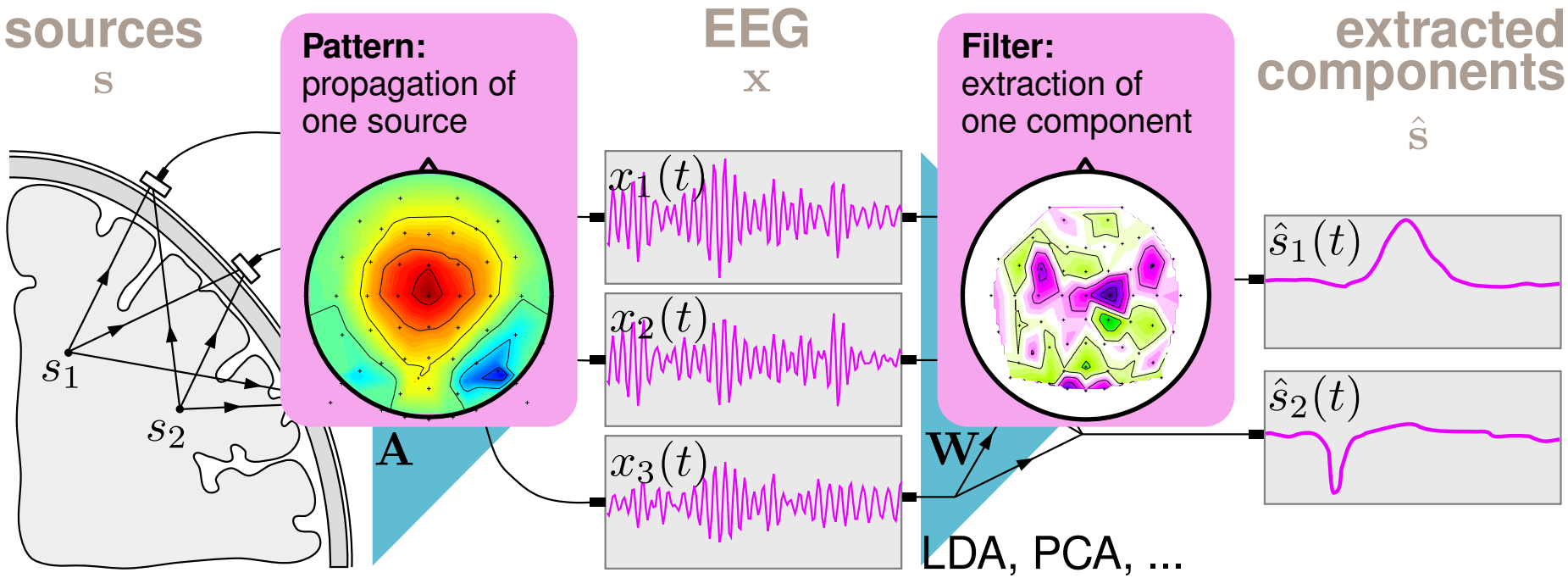
forward model

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

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



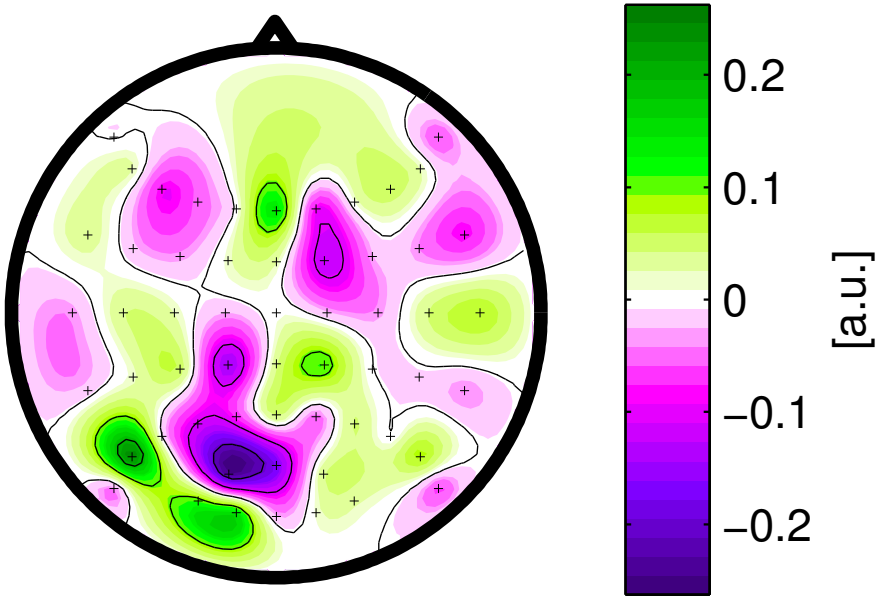
$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t)$$

forward model

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

backward model

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

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 (2)

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

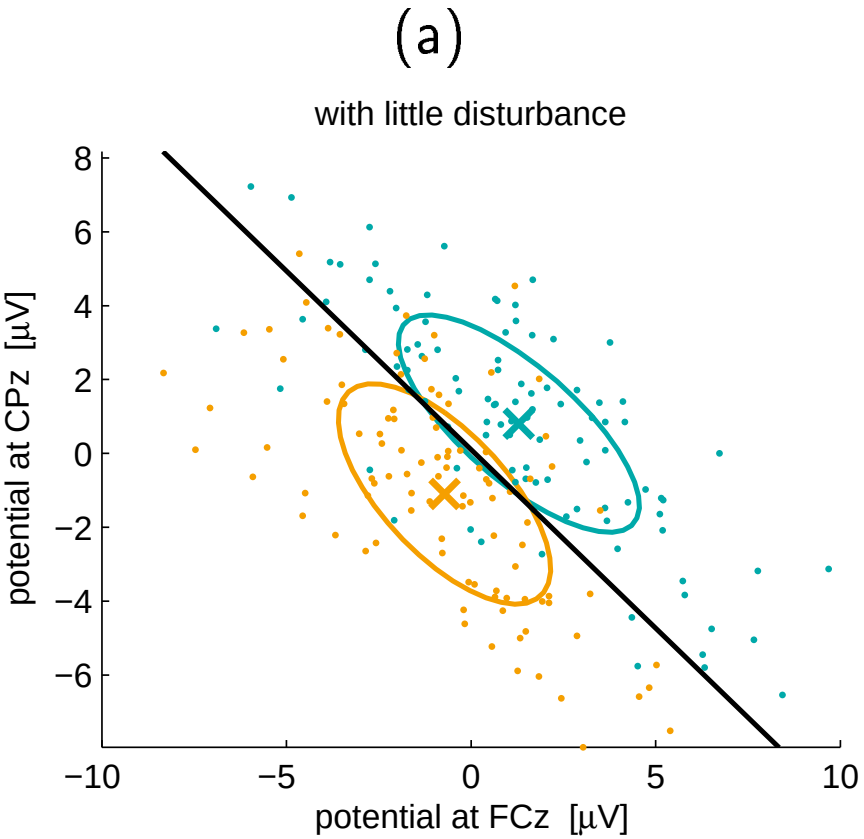
- s_1, the cognitive P300 component

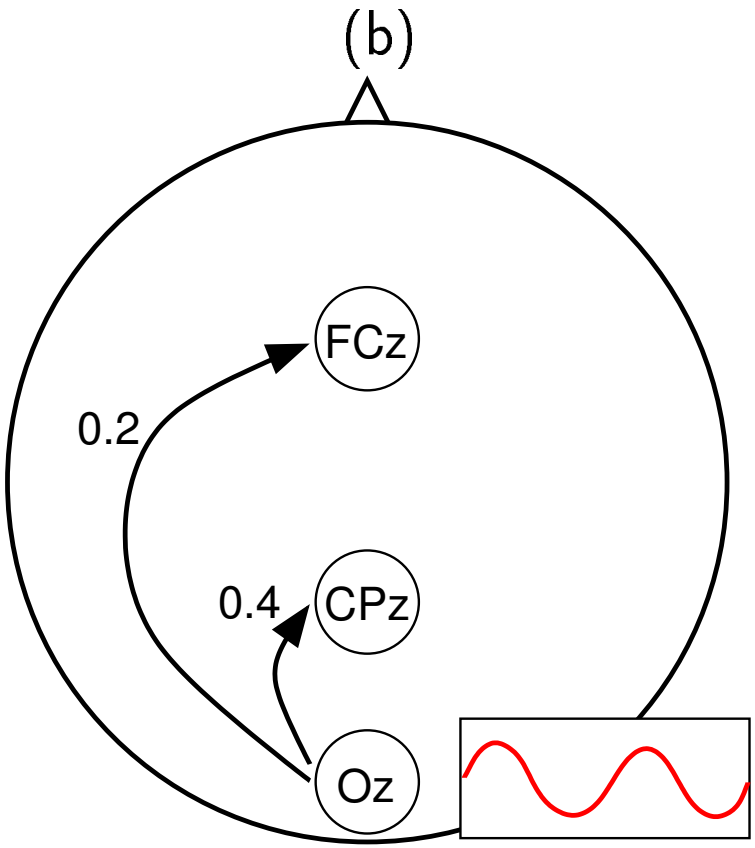
but there is interference from

- s_2, the visual area.

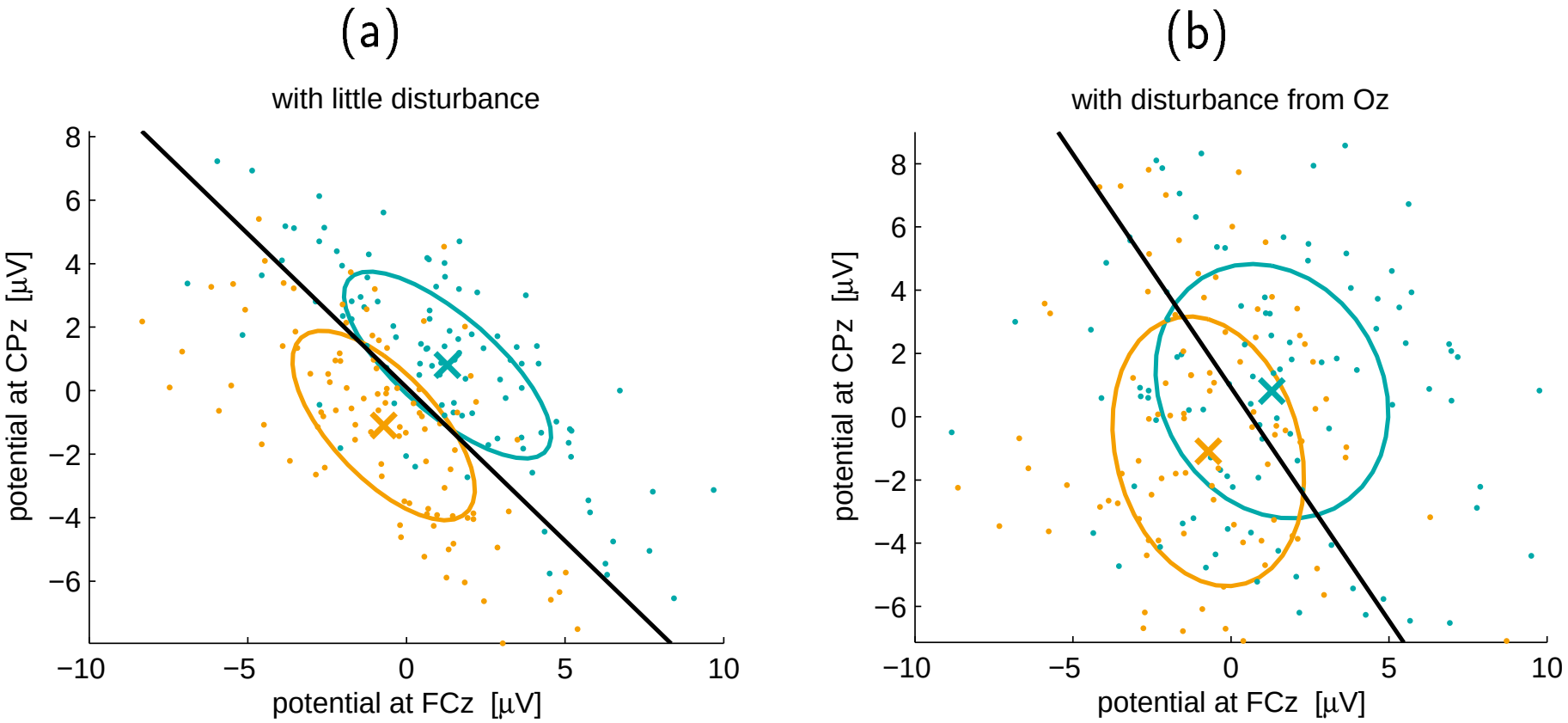
The best filter to recover the P300 component (s_1) depends also on the interfering source of the visual area (s_2). In particular, the spatial map of the filter probably shows strong weights over occipital area, although the P300 component originates from the central region.

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.

References

P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch, "EEG coherency I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales", *Electroencephalogr Clin Neurophysiol*, 103(5): 499–515, 1997.

S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging", *Neuroimage*, 56: 387–399, 2011, URL `http://dx.doi.org/10.1016/j.neuroimage.2010.11.004`.

J. Schäfer and K. Strimmer, "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics", *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components – a tutorial", *Neuroimage*, 56: 814–825, 2011, URL `http://dx.doi.org/10.1016/j.neuroimage.2010.06.048`.

BBCI