

# Brain-Computer Interfacing

## WS 2018/2019 – Lecture #04



Benjamin Blankertz

Lehrstuhl für Neurotechnologie, TU Berlin

[benjamin.blankertz@tu-berlin.de](mailto:benjamin.blankertz@tu-berlin.de)

07 · Nov · 2018



- ▶ Classification of **spatio-temporal** ERP features:
- ▶ Nearest Centroid Classifier (NCC)
- ▶ Linear Discriminant Analysis (LDA)
- ▶ Validation, loss functions
- ▶ Investigation of discriminability with purely spatial/temporal features



# Distributions of Spatial, Temporal and Spatio-Temporal Features

**Single-trials** of spatial, temporal, and even spatio-temporal features can be visualized very well.

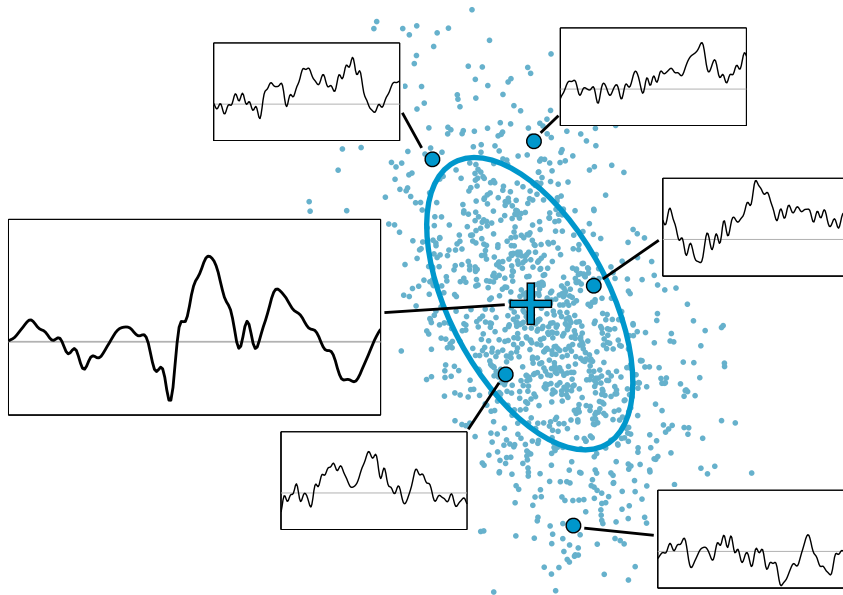
(These are scalp topographies, time courses, and a series of scalp maps, as seen before.)

The **mean** of those features can be visualized in the same way.

But what about whole **distributions** (scatter plots)?

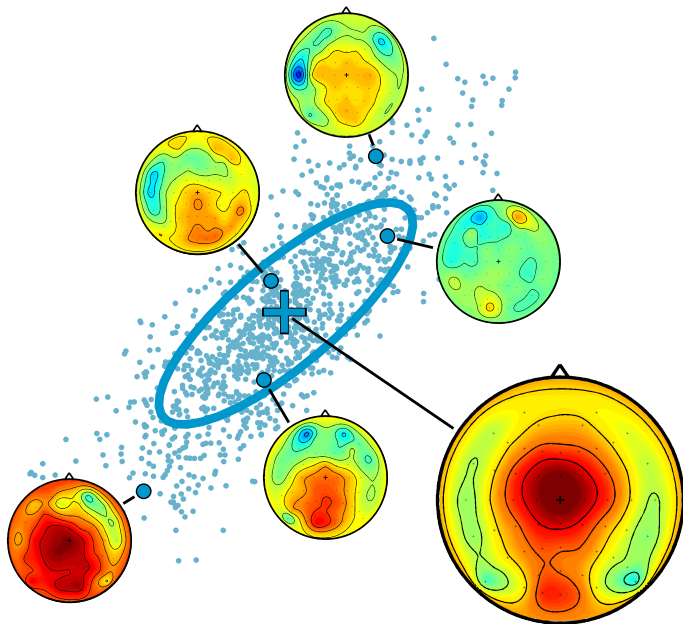
... and **covariance matrices**?

# Scatter Plot of Temporal Features



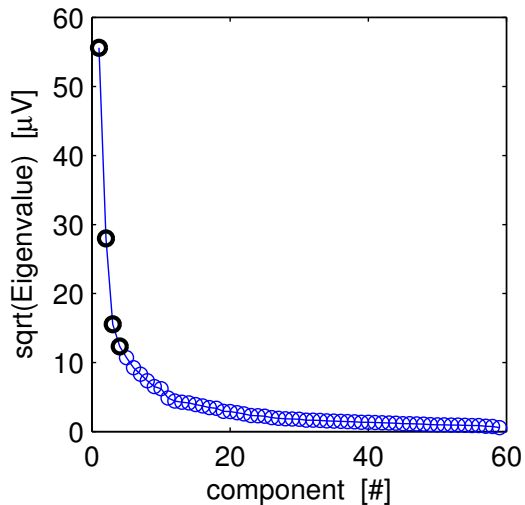
But actually, the features live in a very high dimensional space.

# Scatter Plot of Spatial Features

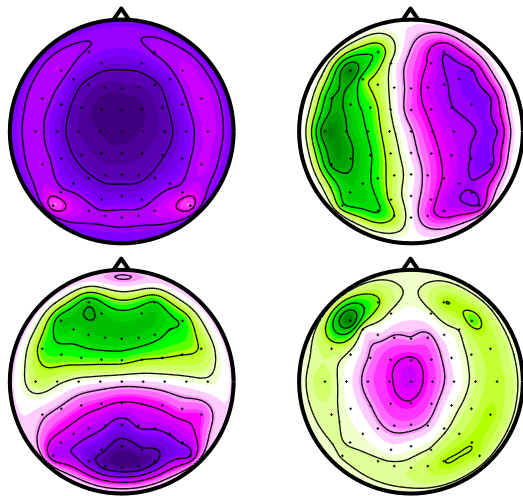


# A Glance at the Covariance Matrix of Spatial Features

**Eigenvalue Spectrum**



**Eigenvectors 1-4 as maps**



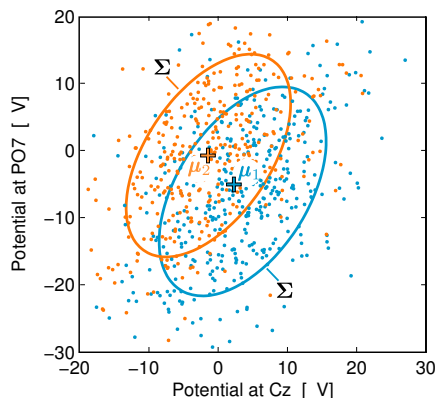
## Reminder: Distribution of ERP Features

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

$$\mathbf{x}_k(t) = \mathbf{p}_1(t) + \mathbf{r}_k(t) \quad \text{for trials } k \text{ of condition 1}$$

$$\mathbf{x}_k(t) = \mathbf{p}_2(t) + \mathbf{r}_k(t) \quad \text{for trials } k \text{ of condition 2}$$

with Gaussian noise:  $\mathbf{r}_k(t) \sim \mathcal{N}(0, \Sigma)$ .



For features of ERP data:

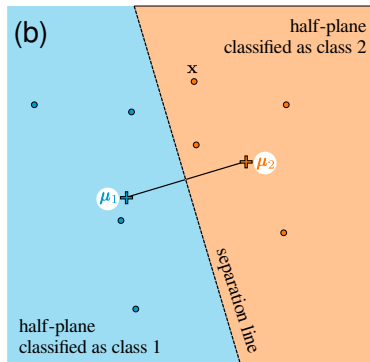
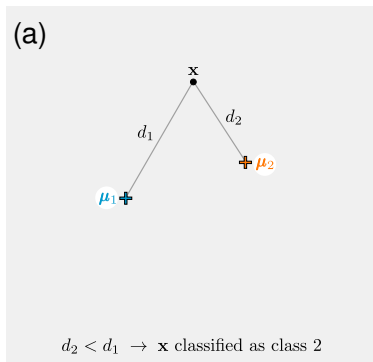
- ▶  $\mu_1$ : ERP of condition 1
- ▶  $\mu_2$ : ERP of condition 2
- ▶  $\Sigma$ : noise: non-phase-locked activity (independent of condition!)

[Blankertz et al, NeuroImage 2011]



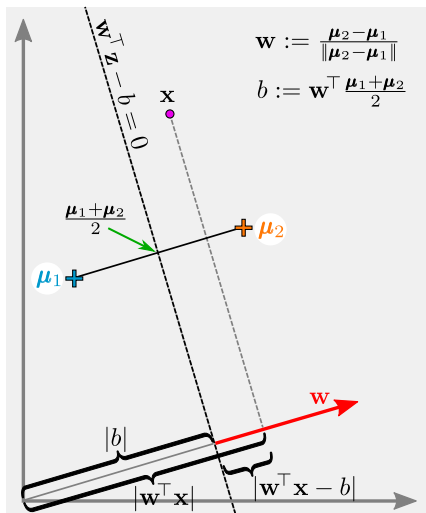
# Nearest Centroid Classifier (NCC)

**(a)** Let us assume a simple setting of a classification problem with little information: Only the means (or centroids)  $\mu_1$  and  $\mu_2$  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.

# Formalization of Separating Hyperplanes

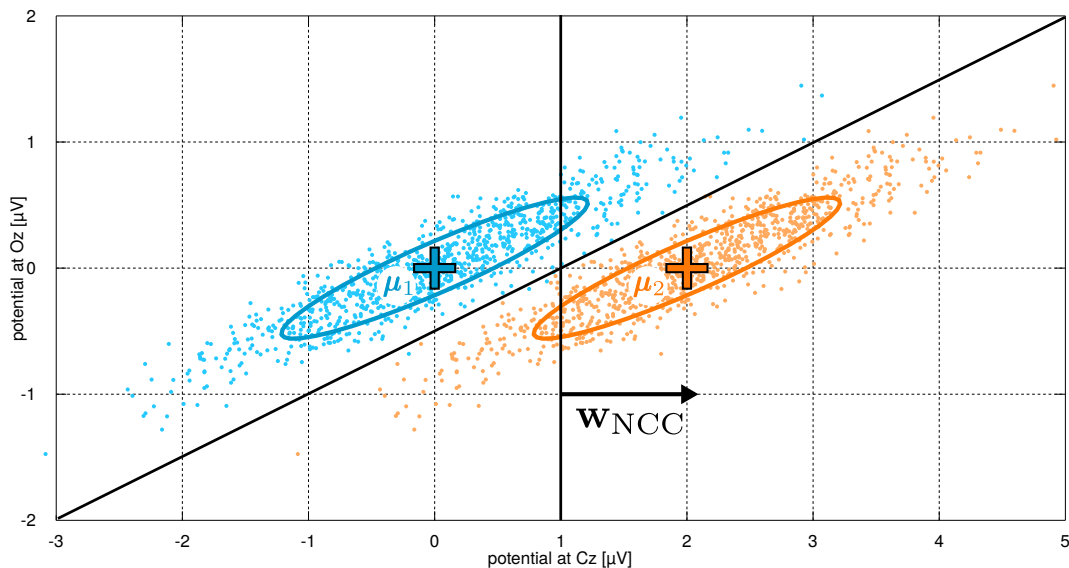


Note: The normalization in the definition of  $\mathbf{w}$  is just done to have easier expressions in the illustration.

For the sake of classification it is sufficient to define:  $\mathbf{w}_{\text{NCC}} := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$

$$\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} - b \mapsto \begin{cases} \text{class 1} & \text{if } \mathbf{w}^\top \mathbf{x} - b < 0 \\ \text{class 2} & \text{if } \mathbf{w}^\top \mathbf{x} - b \geq 0 \end{cases}$$

## 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.

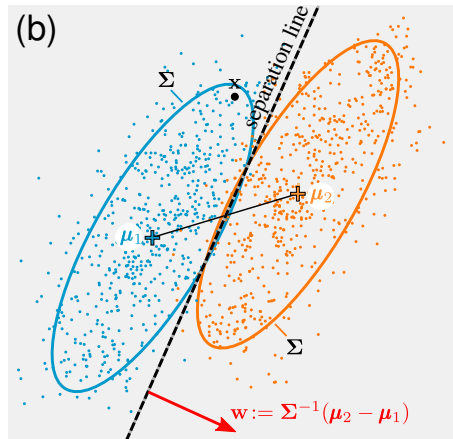
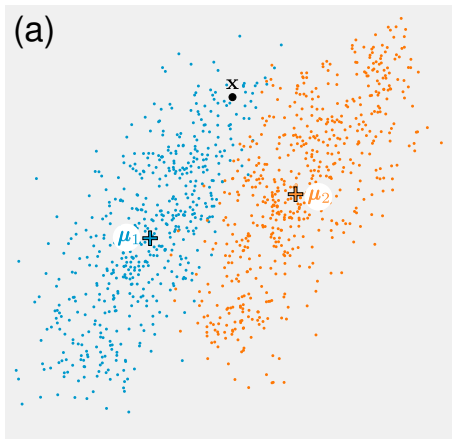
This optimal classifier is called **Linear Discriminant Analysis** (LDA): Given two Gaussian distributions  $\mathcal{N}(\mu_1, \Sigma)$  and  $\mathcal{N}(\mu_2, \Sigma)$ , LDA is defined by the normal vector

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

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

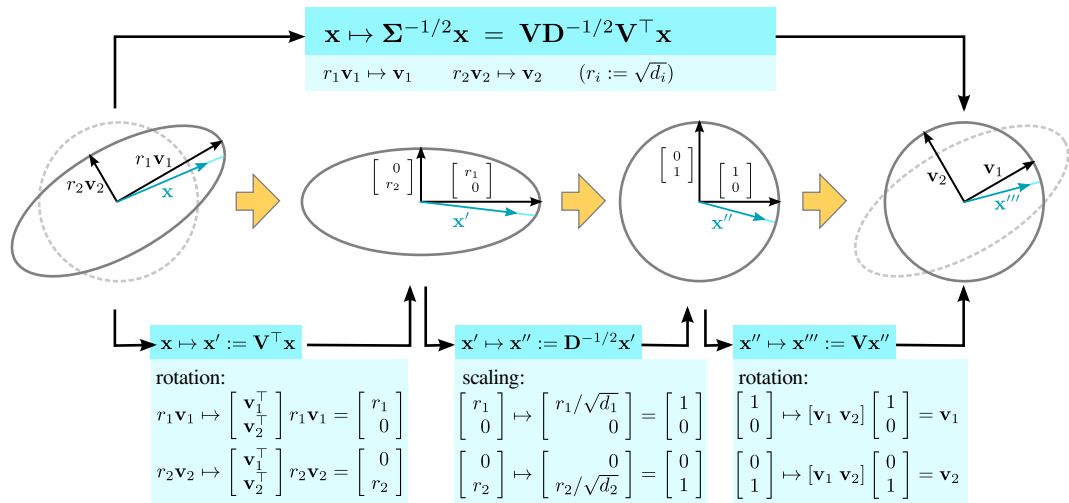
# Linear Discriminant Analysis

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



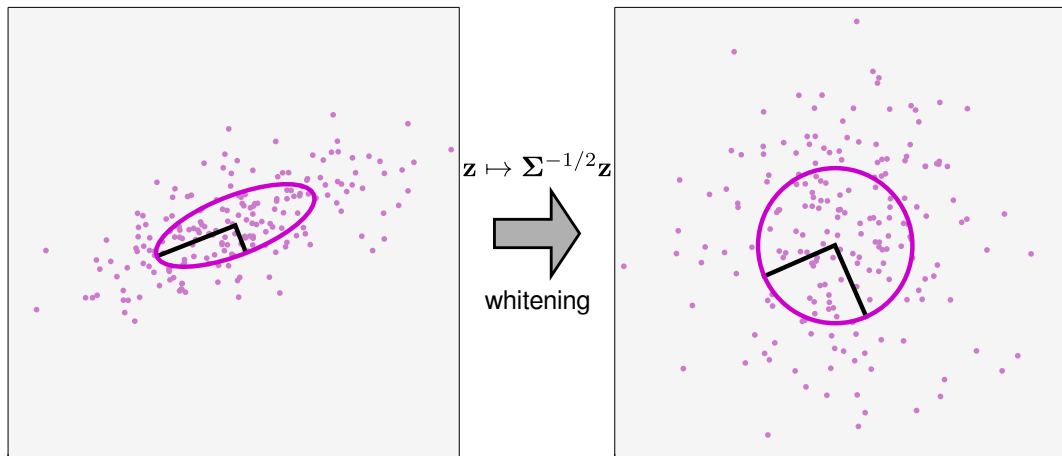
**(b)** In Linear Discriminant Analysis, a common covariance matrix for both classes is estimated, which describes the (class-independent) noise. Note, that  $x$  is classified here differently with LDA than with NCC.

# Interlude: Illustration of Whitening Transform



The whitening transform maps the space such that a Gaussian distribution with the given covariance matrix becomes a standard normal distribution, i.e., the variance in all directions is 1. It maps the ellipsoid given by the standard isodensity line of the Gaussian distribution to the unit sphere.

# Illustration of Whitening



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

## Whitening - The Math

We define “ $\mathbf{P} = \Sigma^{-1/2}$ ” formally via EVD of  $\Sigma = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ :

$$\mathbf{P} := \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top$$

Note, that  $\mathbf{P} = \mathbf{P}^\top$  and  $\mathbf{D}^{-\frac{1}{2}} = \text{diag}(\frac{1}{\sqrt{d_1}}, \dots, \frac{1}{\sqrt{d_k}})$ .

To see that this is indeed a definition of a square root of matrix  $\Sigma^{-1}$  calculate:

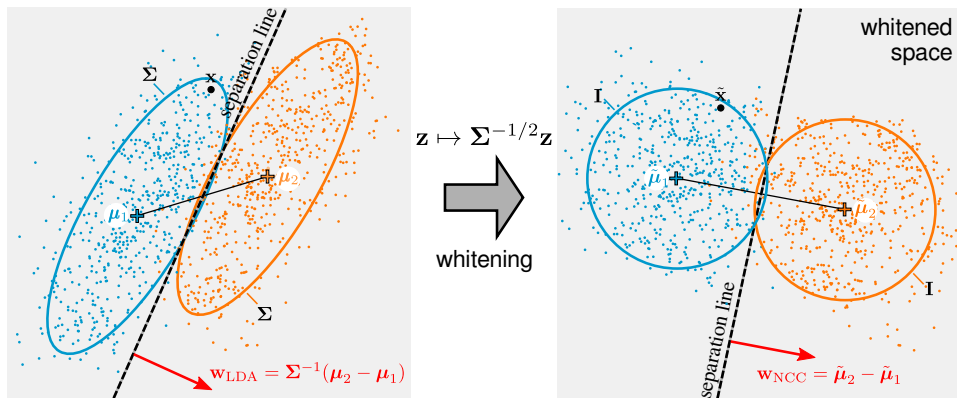
$$\mathbf{P}\mathbf{P} = \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top\mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top = \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^\top$$

We obtain as the covariance of the projected data

$$\begin{aligned}\Sigma_{\mathbf{P}^\top\mathbf{X}} &= \mathbf{P}^\top\Sigma_{\mathbf{X}}\mathbf{P} \\ &= \mathbf{P}^\top(\mathbf{V}\mathbf{D}\mathbf{V}^\top)\mathbf{P} \\ &= \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top\mathbf{V}\mathbf{D}\mathbf{V}^\top\mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top \\ &= \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{D}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}\end{aligned}\tag{2}$$



# Correspondence between NCC and LDA



Classification with LDA in the original space ( $\mathbf{x}$ ,  $\mu_i$ ,  $\mathbf{w}_{\text{LDA}}$ ) is equivalent to classification with NCC in the whitened space ( $\tilde{\mathbf{x}}$ ,  $\tilde{\mu}_i$ ,  $\mathbf{w}_{\text{NCC}}$ ). Note, that  $\Sigma^{-1/2} = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^\top$  is symmetric.

$$\begin{aligned}\mathbf{w}_{\text{NCC}}^\top \tilde{\mathbf{x}} &= (\tilde{\mu}_2 - \tilde{\mu}_1)^\top \tilde{\mathbf{x}} = (\Sigma^{-1/2} \mu_2 - \Sigma^{-1/2} \mu_1)^\top \Sigma^{-1/2} \mathbf{x} \\ &= (\mu_2 - \mu_1)^\top \Sigma^{-1/2} \Sigma^{-1/2} \mathbf{x} = \mathbf{w}_{\text{LDA}}^\top \mathbf{x}\end{aligned}$$

# Linear Discriminant Analysis – Assumptions for Optimality

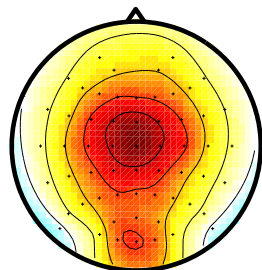
Now, we come back to the assumptions which are required to warrant optimality of the LDA classifier:

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

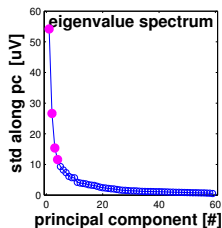
We will verify the first two assumptions empirically by investigating one exemplary dataset. The last assumption will be discussed in the next lecture.

# Mean and Eigenvalue Spectrum for a P300 Data Set

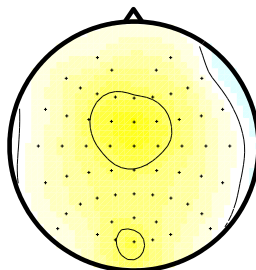
*target*



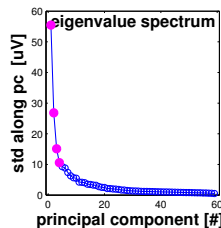
**average target**



*non-target*



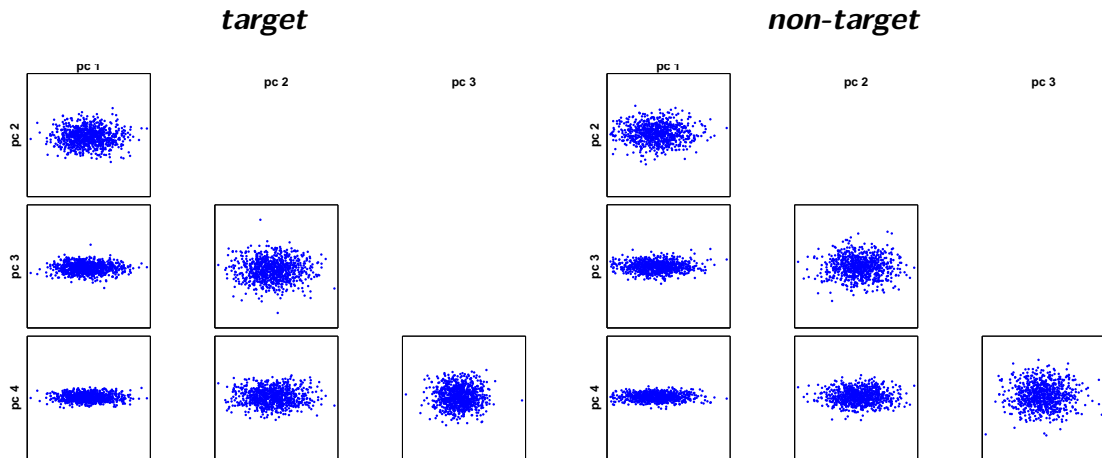
**average nontarget**



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

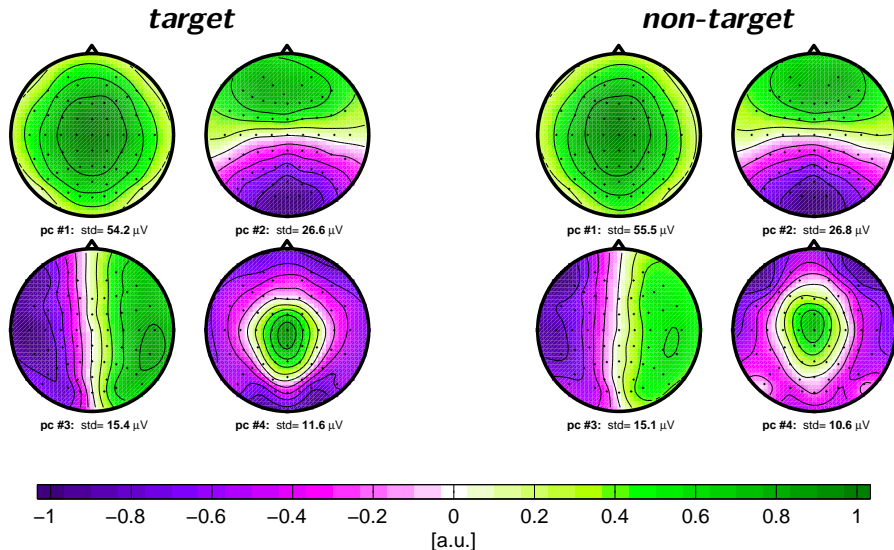
# Distribution of the Noise

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



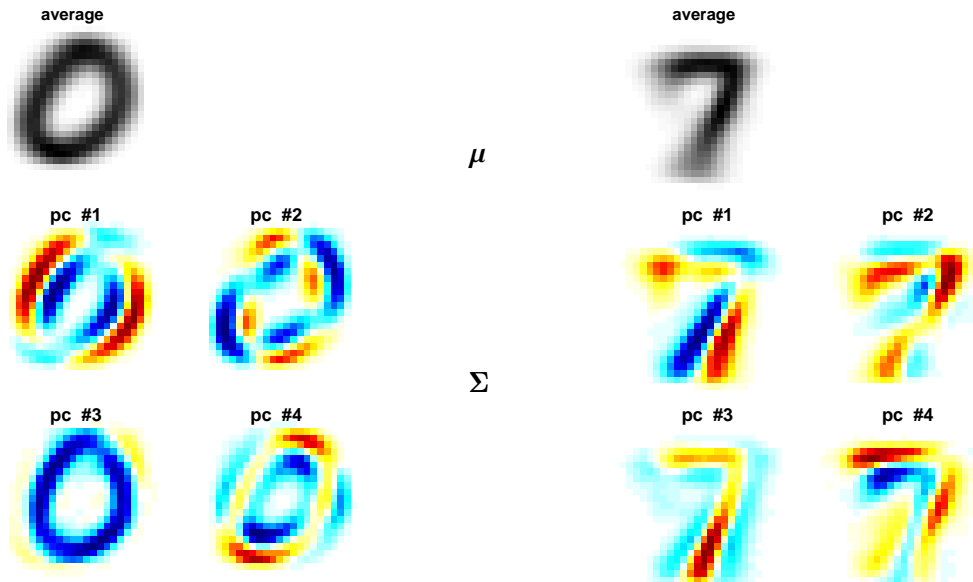
➤ These projections look very much Gaussian.

# 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, and why cross-validation does not always warrant independence of training and validation set.

One particular issue requiring attention for validation (which also affects the exercises) is discussed on the next slide.

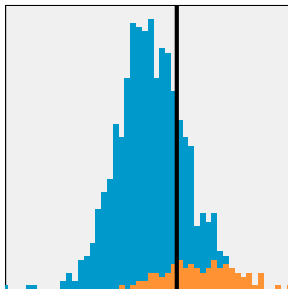
# Loss Functions for Unbalanced Classes

Blue class:  $N_1 = 900$  samples, orange class:  $N_2 = 100$  samples. 

**Weighted error:**  $\text{err}_{\text{weighted}} = \frac{1}{2} (\text{err}_{\text{class 1}} + \text{err}_{\text{class 2}})$

**AUC-based:** AUC value of classifier outputs

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

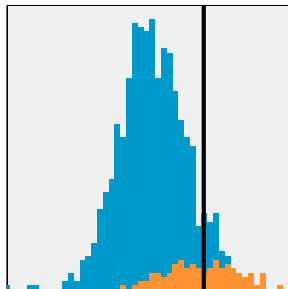


Error rate

Unweighted: 23.6%

Weighted: 25.1%

AUC-based: 16.6%

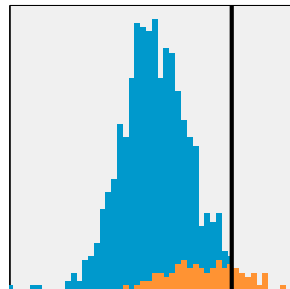


Error rate

Unweighted: 12.8%

Weighted: 30.0%

AUC-based: 16.6%



Error rate

Unweighted: 9.5%

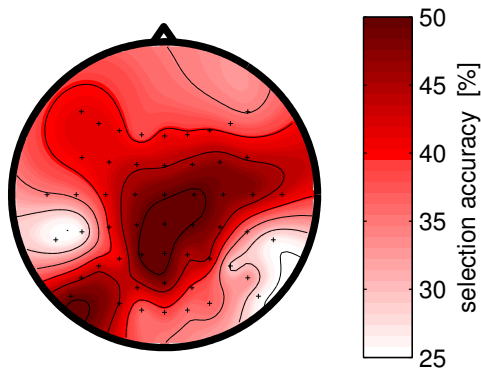
Weighted: 39.5%

AUC-based: 16.6%



## Application of (Purely) Temporal Features

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

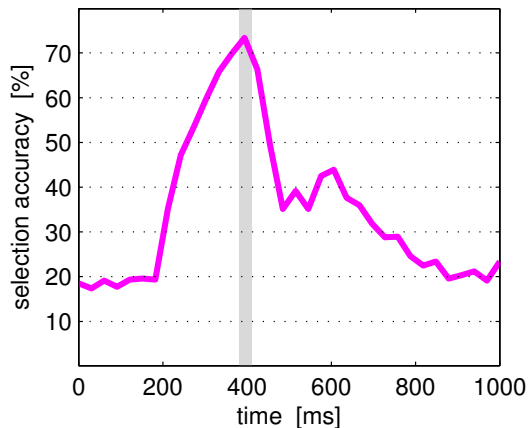


For each single channel the classification performance is determined for temporal features with LDA by cross validation. The resulting error values can be visualized as scalp topography.

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.

## Practical Remark: Estimating $\Sigma$ for LDA

To implement LDA, we need to estimate  $\Sigma$  from the data. We have samples of two classes given (assuming equal covariance matrices):

- 1 Estimate the covariance matrices of both classes  $\Sigma_1$  and  $\Sigma_2$  and average them

$$\Sigma := \frac{1}{2}(\Sigma_1 + \Sigma_2)$$

- 2 Calculate the weighted average of  $\Sigma_1$  and  $\Sigma_2$  (with  $N_i$  being the number of samples in class  $i$ )

$$\Sigma := \frac{N_1-1}{N_1+N_2-1}\Sigma_1 + \frac{N_2-1}{N_1+N_2-1}\Sigma_2$$

- 3 Subtract from each sample the mean of the respective class (centering) and calculated the covariance matrix across all those centered samples.

Methods 2 and 3 are equivalent. They might provide better results than 1, if the classes are imbalanced.

After this lecture you should

- ▶ be familiar with the whitening transform
- ▶ know well the linear classifiers NCC and LDA,  
be aware of their differences  
and appreciate the connection (via whitening)
- ▶ have notice about the assumptions of LDA (warranting optimality)
- ▶ be acquainted with the inspection of high dimensional distributions using EVD
- ▶ be aware of the essentials of validating classifiers
- ▶ have an idea of investigating discriminability in the spatial and in the temporal domain

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