Brain-Computer Interfacing

WS 2018/2019 – Vorlesung #07

Benjamin Blankertz

Lehrstuhl für Neurotechnologie, TU Berlin

benjamin.blankertz@tu-berlin.de

28 · Nov · 2018

The Blessing and Curse of Machine Learning

Machine learning provides **multivariate** techniques for the analysis of EEG data involving optimization of user-specific models.

This results in a considerably **increased sensitivity** in the discovering of neural correlates.

The down side is that the **interpretation** of *where* the discriminative information originates from is not always straight forward...

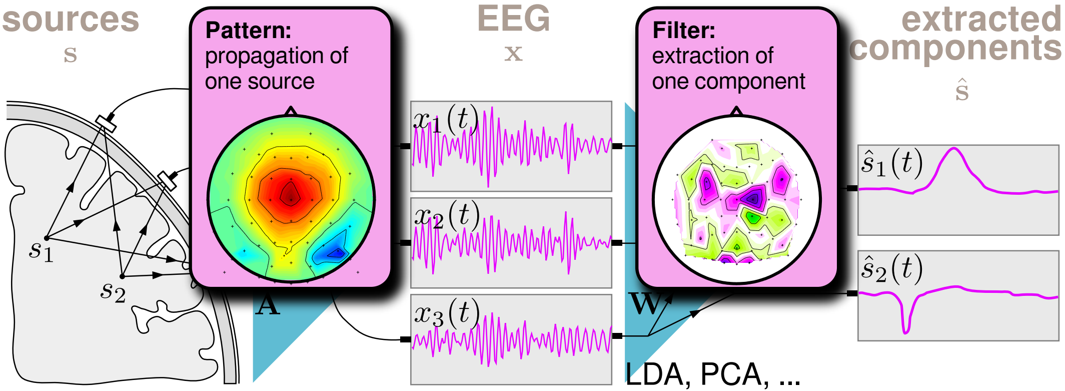
... nevertheless very important. Do not use ML techniques as black box.

In the this lectures we will learn one aspect of how to *open the black box* in ERP classification.

Today’s Topics

- Clearer definition of spatial patterns and spatial filters
- Better understanding of spatial patterns and spatial filters
- Interpretation of spatial filters (?!)
- Reuptake: Spatial patterns corresponding to given filters

Recap: Patterns and Filters in the Linear Model of EEG



sources

$\mathbf{s}$

Pattern:
propagation of
one source

EEG

$\mathbf{x}$

Filter:
extraction of
one component

extracted components

$\hat{\mathbf{s}}$

LDA, PCA, ...

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

forward model

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

backward model

Definition of a Spatial Filter

Given data with C channels $\mathbf{x}(t) \in \mathbb{R}^C$, in principle any vector of the corresponding dimensionality $\mathbf{w} \in \mathbb{R}^C$ can be a **spatial filter**. We call it so, if it is applied to the signal in the sensor space

$$y(t) = \mathbf{w}^\top \mathbf{x}(t)$$

If several filters are applied simultaneously, we have a matrix of filters $\mathbf{W} \in \mathbb{R}^{C \times P}$ and we obtain a vector of several components $\mathbf{y}$:

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

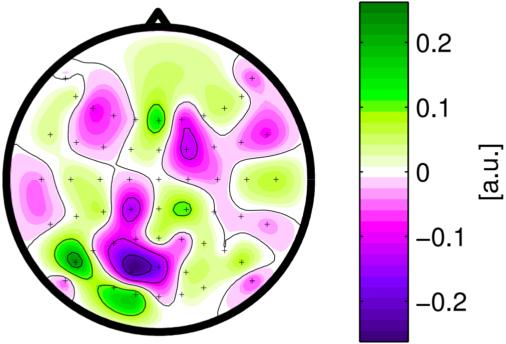
When data is given for a fixed number of time points, we also use the matrix notation $\mathbf{X} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_T)] \in \mathbb{R}^{C \times T}$ and

$$\mathbf{Y} = \mathbf{W}^\top \mathbf{X}$$

In particular, any linear classifier trained on spatial features is a spatial filter. In a broader sense we use the term also in the context of spatio-temporal features.

Problem Setting: Interpretation of Classifier Weights?

The weights of a linear classifier can be visualized in the domain of the input features. For spatial features, they can be depicted as scalp topography.

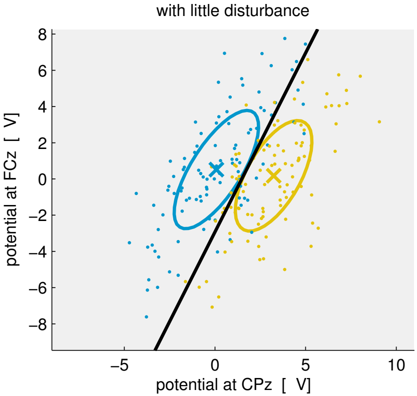


LDA trained on spatial features extracted from the time interval 380–410 ms. It is tempting to interpret the prominent weights of this map wrt neurophysiology.

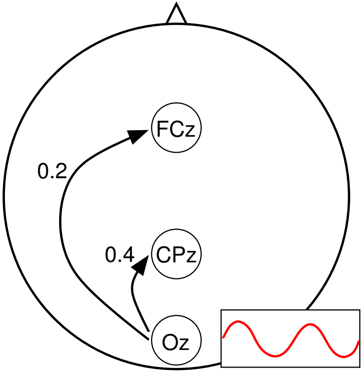
(Temporal and spatio-temporal features suggest interpretations an analogue way.) The following two lectures will show that it is not that easy, but there is a solution to it.

Understanding Spatial Filters

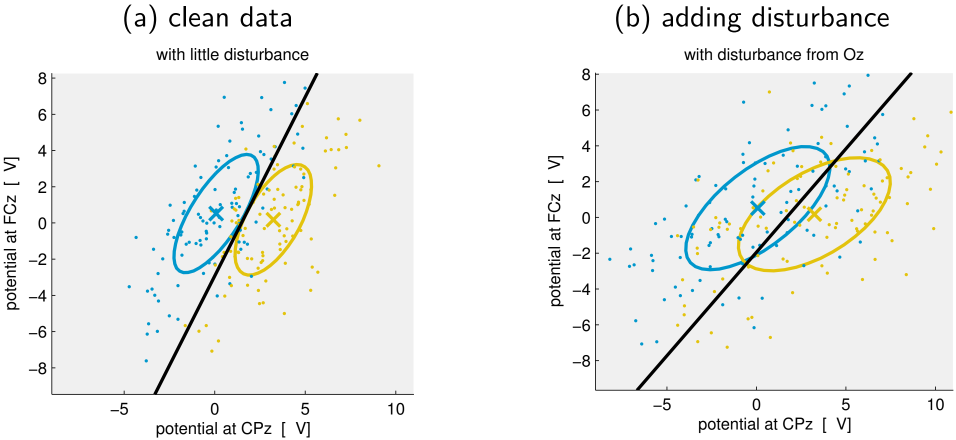
(a) clean data



(b) adding disturbance



Understanding Spatial Filters


(a) clean data

(b) adding disturbance

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.

Definition of a Spatial Pattern

We use the term **spatial pattern** for a propagation vector $\mathbf{a} \in \mathbb{R}^C$ that describes how the activity of a source is propagated to the C sensors. Spatial patterns are applied to signals in the source space:

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

Spatial Patterns and Filters in the Linear Source Model

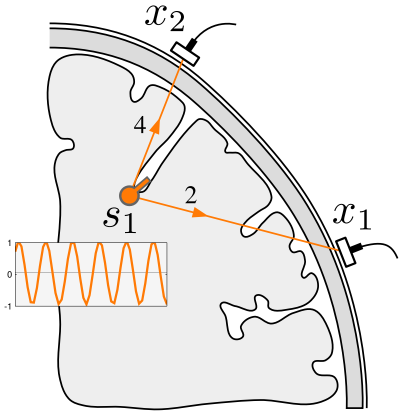
In the previous lecture, we obtained a **correspondence** between spatial filters (row vectors of the backward model $\mathbf{W}^\top$) and spatial paterns (column vectors of the forward model $\mathbf{A}$).

This can be used to check whether a given spatial filter corresponds to a plausible spatial pattern. If a spatial filter is obtained, e.g., from a classifier, it is of high interest to obtain the corrsponding pattern for interpretation.

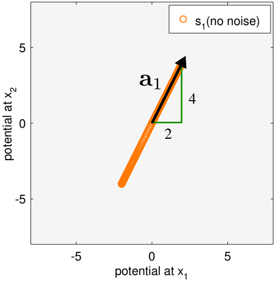
Illustration of Spatial Patterns and Filters

The following part will illustrate spatial patterns and spatial filters in a simple linear model with **two sources** and **two sensors**.

Explaining Spatial Patterns and Spatial Filters

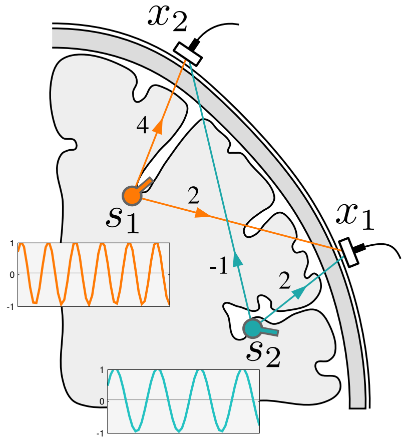


$$\mathbf{x}(t) = \mathbf{a}_1 s_1(t)$$



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

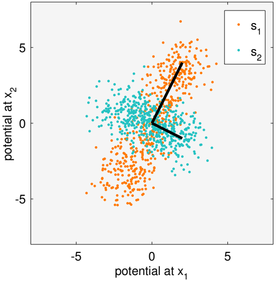
Explaining Spatial Patterns and Spatial Filters



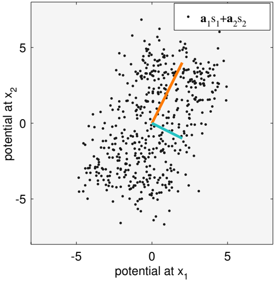
$$\mathbf{a}_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad \mathbf{a}_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

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

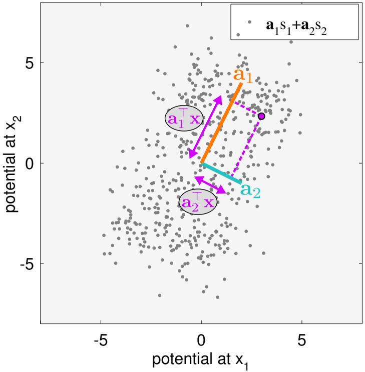
$$\mathbf{x}(t) = \mathbf{a}_2 s_2(t) + \mathbf{n}(t)$$



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

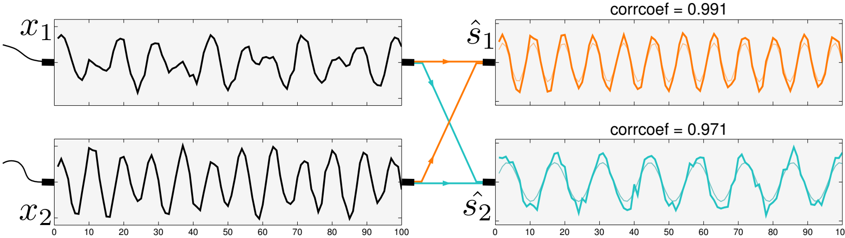


Explaining Spatial Patterns and Spatial Filters

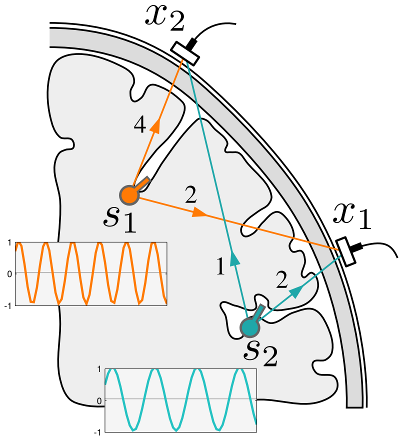


$$
\begin{aligned}
\hat{s}_1 &= \mathbf{a}_1^\top \mathbf{x} \\
&= \mathbf{a}_1^\top \mathbf{a}_1 s_1 + \underbrace{\mathbf{a}_1^\top \mathbf{a}_2}_{=0} s_2 + \mathbf{a}_1^\top \mathbf{n} \\
&\sim s_1
\end{aligned}
$$

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

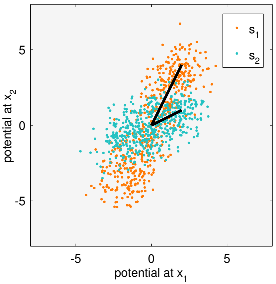


Explaining Spatial Patterns and Spatial Filters

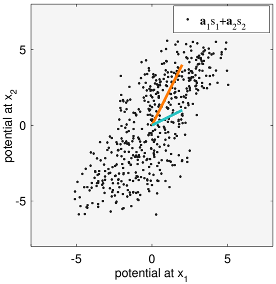


$$\mathbf{a}_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad \tilde{\mathbf{a}}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

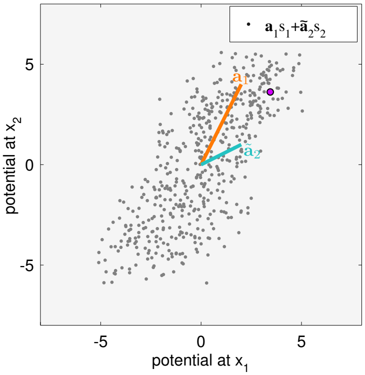
$$\mathbf{x}(t) = \mathbf{a}_1 s_1(t) + \mathbf{n}(t)$$
$$\mathbf{x}(t) = \tilde{\mathbf{a}}_2 s_2(t) + \mathbf{n}(t)$$



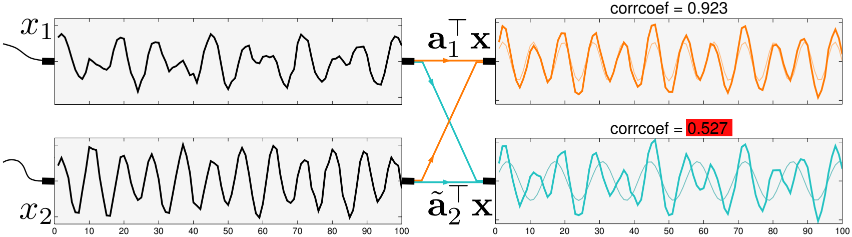
$$\mathbf{x}(t) = \mathbf{a}_1 s_1(t) + \tilde{\mathbf{a}}_2 s_2(t) + \mathbf{n}(t)$$



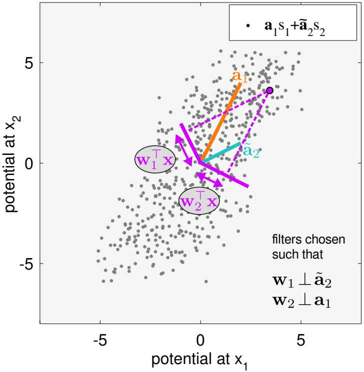
Explaining Spatial Patterns and Spatial Filters



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



Explaining Spatial Patterns and Spatial Filters

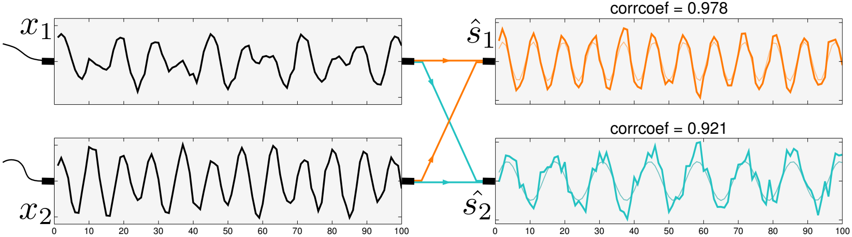


$$\hat{s}_1 = \mathbf{w}_1^\top \mathbf{x}$$

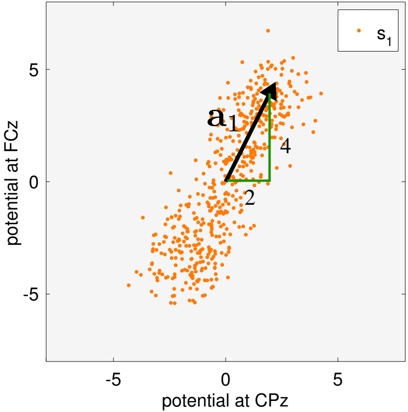
$$= \mathbf{w}_1^\top \mathbf{a}_1 s_1 + \underbrace{\mathbf{w}_1^\top \mathbf{a}_2}_{=0} s_2 + \mathbf{w}_1^\top \mathbf{n}$$

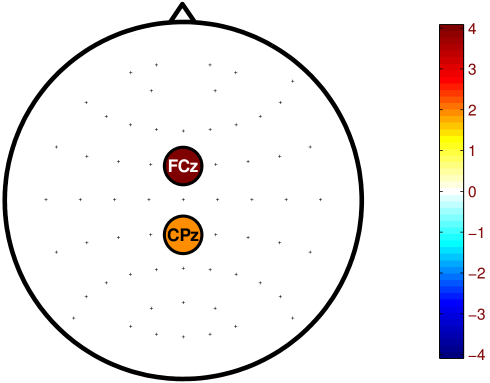
$$\sim s_1$$

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

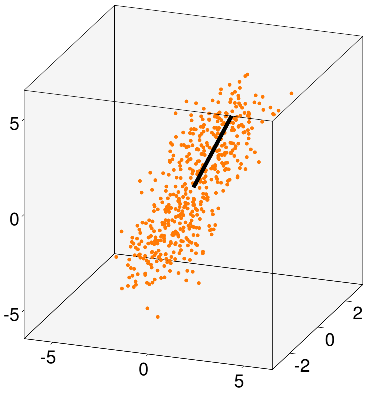


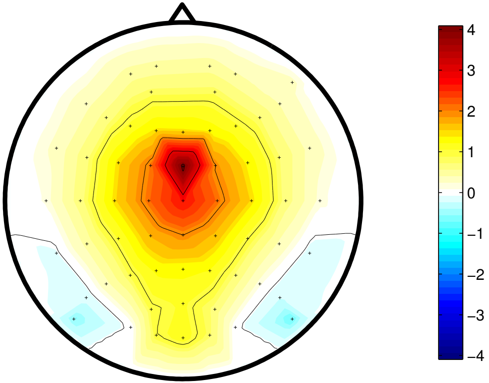
Correspondence of Vectors in Feature Space and Patterns





Correspondence of Vectors in Feature Space and Patterns





Left: High-dimensional features spaces are hard to visualize.

Right: A vector in the feature space (such as weight vectors) can be represented as a scalp topography (for spatial features).

Discussion of Spatial Filters and Patterns

The next slide repeat the point of the previous slides using text without illustration. They are not presented in the lecture and intended for offline reading.

[Blankertz et al, 2011]

Interpretation of Spatial Filters (Same Story as Before)

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

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

and the task is to find a spatial filter $\mathbf{w}$ to recover s_1. Applying the (yet to be determined) filter $\mathbf{w}$ to $\mathbf{x}(t)$ yields

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

To recover s_1 (i.e., to eliminate the contribution of s_2), the filter $\mathbf{w}$ needs to be chosen such that $\mathbf{w}^\top \mathbf{a}_2 = 0$: the filter $\mathbf{w}$ is orthogonal to $\mathbf{a}_2$.

In the untypical case of orthogonal propagation vectors $(\mathbf{a}_1^\top \mathbf{a}_2 = 0)$ $\mathbf{w} = \mathbf{a}_1$ does the job: The best filter corresponds to the propagation direction of the source, i.e., a pattern.

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 $\mathbf{a}_2$.

Interpretation of Spatial Filters (Concrete Example)

In order to discuss what the previous result means in view of interpretability of spatial filters, let's take an 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.

Finding Patterns for given Filters of a Discriminative Model

Having learnt about the problem of taking the map of a spatial filter for neurophysiological interpretation, we are faced with the following task:

> *Given a spatial filter, determine a corresponding pattern ('corresponding' in the sense of the linear model)!*

This was derived in the previous lecture:
Let data $\mathbf{X}$ and a filter matrix $\mathbf{W}$ (backward model) be given and define $\mathbf{S} = \mathbf{W}^\top\mathbf{X}$.
Then we obtain a matrix of corresponding patterns $\hat{\mathbf{A}}$ (forward model) by

$$\hat{\mathbf{A}} = \mathbf{\Sigma}_\mathbf{x}\mathbf{W}\mathbf{\Sigma}_\mathbf{s}^{-1}$$

[Haufe et al, 2014]

Patterns Corresponding to Given Filters

The result

$$\hat{\mathbf{A}} = \mathbf{\Sigma}_{\mathbf{x}} \mathbf{W} \mathbf{\Sigma}_{\mathbf{s}}^{-1}$$

has the following consequences for special cases:

- If $K = 1$ (in particular for LDA), we obtain

  $$\hat{\mathbf{a}} \simeq \mathbf{\Sigma}_{\mathbf{x}} \mathbf{w}.$$

- If the components are uncorrelated (e.g., PCA, ICA), we get

  $$\hat{\mathbf{A}} \simeq \mathbf{\Sigma}_{\mathbf{x}} \mathbf{W}.$$

- The patterns and filters coincide if additionally the observations $\mathbf{X}$ are uncorrelated

  $$\hat{\mathbf{A}} \simeq \mathbf{W}.$$

  However, this assumption is rather unrealistic for EEG due to volume conduction, see lecture #01.

Consequence for LDA

After having trained an LDA, you (should) want to know which areas in the brain are tapped for the descrimination. For this question, you need to know the pattern corresponding to the LDA filter.

According to the previous results, we obtain the pattern as

$$\begin{aligned} \hat{\mathbf{a}} &\simeq \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{w}_{\mathsf{LDA}} \\ &= \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ &= \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \end{aligned}$$

The solution is simple: the pattern is just the difference of the means.

Patterns and Filters Coincide

We have seen in the last lecture: Patterns and filters of the linear model coincide in PCA where the sources are (assumed to be) **uncorrelated**, i.e., propagation vectors are orthogonal.

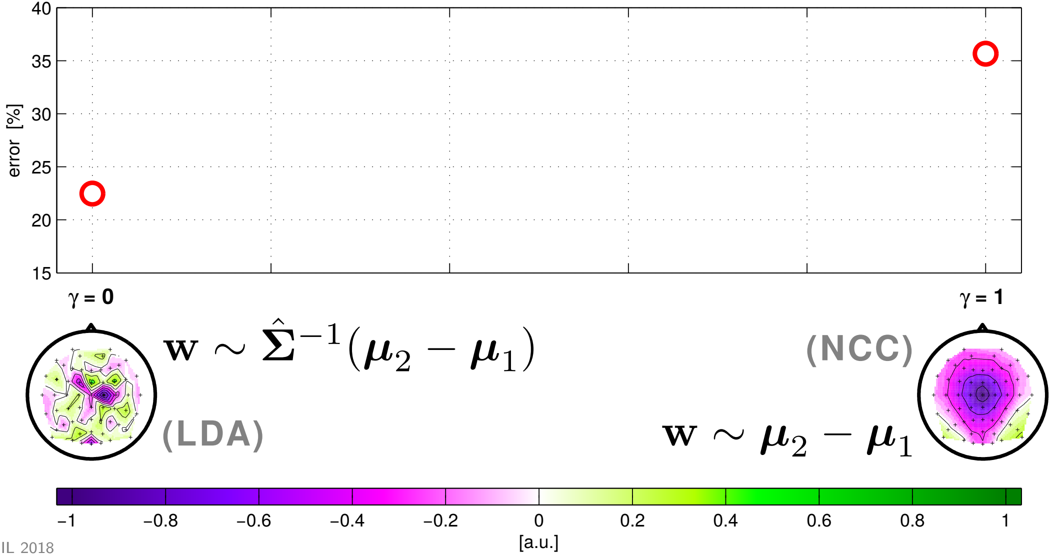
Moreover, when in NCC the structure of the noise is assumed to be spherical (e.g., because it cannot be estimated reliably) patterns and filters coincide as well.

Note:

For varying γ from 0 to 1 in Shrinkag-LDA, the weight vector that can be regarded as filter converges towards a pattern (NCC case, where pattern and filter coincide), see illustration on the next slide.

Shrinkage Interpolates between Filter and Pattern

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

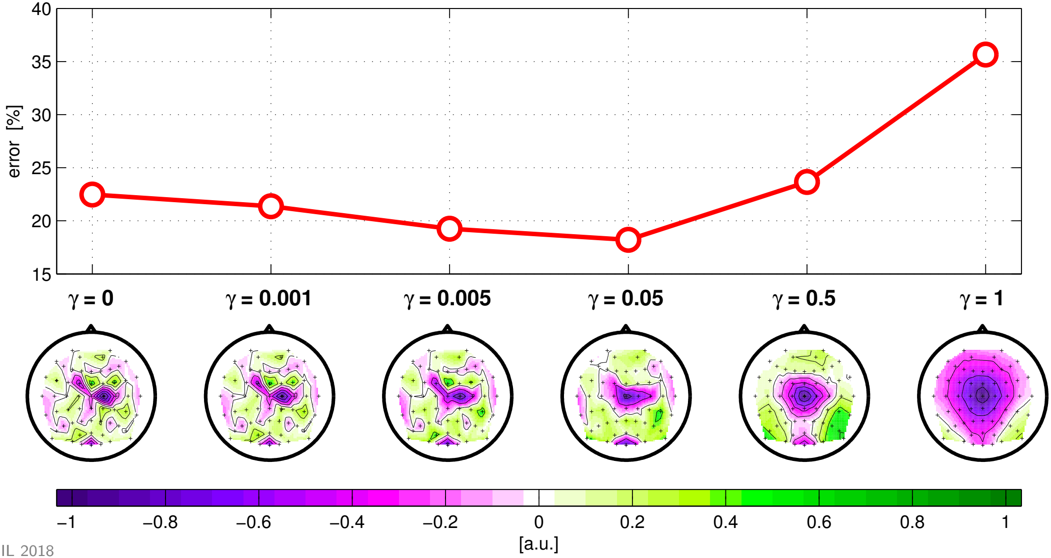


$$\mathbf{w} \sim \hat{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

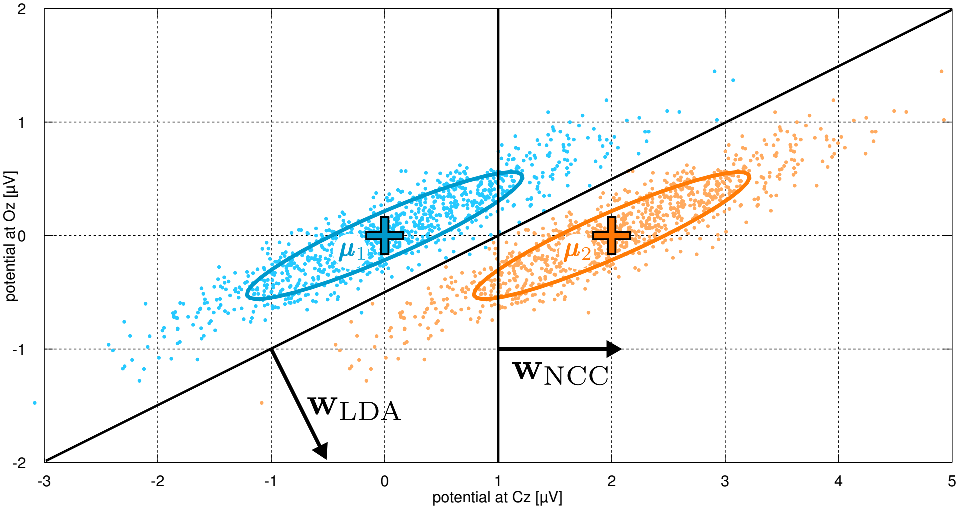
$$\mathbf{w} \sim \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$$

Shrinkage Interpolates between Filter and Pattern

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



Remember: NCC and LDA



$\mathbf{w}_{\text{NCC}}$ is a pattern and therefore suboptimal for classification.

$\mathbf{w}_{\text{LDA}}$ has stronger weight on the vertical dimension, although it only contains noise wrt the classification task (filter discussion!).

Lessons Learnt

After this lecture you should

- have a clear idea about spatial filters and spatial patterns
- know about the difficulties in interpreting spatial filters
- be capable of determining a spatial pattern which corresponds to a given spatial filter

Warming up for the Exam – Quick Questions

There will be regular tasks and so-called *quick questions*.

The quick questions should be answered very briefly and to the point, one sentence or some bullet points.

- Assume cross-validation applied to features of an ERP speller data set yields about 20% misclassifications for *targets* vs. *non-targets*. Why does this result not conclusively indicate better-than-chance performance of the classifer?
- Given an EEG data set with ERPs of two classes, how can a temporal profile of discriminability be obtained?

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.
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014).
  On the interpretation of weight vectors of linear models in multivariate neuroimaging.
  NeuroImage, 87:96–110. Neuroimage single best paper of 2014 Award.