Brain-Computer Interfacing

WS 2018/2019 – Lecture #03

Benjamin Blankertz

Lehrstuhl für Neurotechnologie, TU Berlin

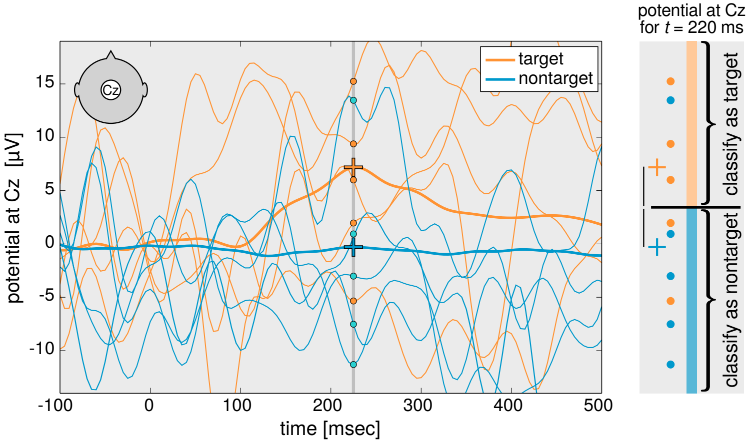
benjamin.blankertz@tu-berlin.de

31 · Okt · 2018

Today's Topics

- Multivariate distributions of EEG features
- Eigenvalue decomposition (EVD)
- Characterization of Gaussian distributions
- ERP features: temporal, spatial, and spatio-temporal

Reminder of Lecture #02: Univariate Feature



- The potential measured 220ms post-stimulus at **Cz** is a one-dimensional observation variable: a *univariate* feature.
- This feature can be regarded as a simple *classifier* to discriminate between target and nontarget trials.

Classification

A binary **classifier** is a function f that maps elements of a *feature space* (here $\mathbb{R}^n$) to one out of two classes (resp. class labels):

$$f_{\mathbf{w}} : \mathbb{R}^n \to \{-1, +1\}; \quad \mathbf{x} \mapsto f_{\mathbf{w}}(\mathbf{x})$$

Since the function f depends on parameters (vector $\mathbf{w}$), we write $f_{\mathbf{w}}$.

Using machine learning techniques, these parameters $\mathbf{w}$ are determined from training data (also called calibration data). This procedure will be discussed in the next lecture.

Method of Previous Lecture as Classifier

In the previous lecture, we used a univariate feature, i.e., $n = 1$.

The classifier is just defined by a threshold $b \in \mathbb{R}$:

$$f_b : \mathbb{R} \to \{-1, +1\}; \quad x \mapsto \mathrm{sign}(x - b)$$

The threshold b is the 'middle' of the class means:

$$b := (\mu_{-1} + \mu_{+1})/2$$

where μ_i is the mean of the 'features' of class i: $\mu_1 = \mathrm{mean}\,\langle x_i \rangle_{y_i=-1}$ and $\mu_2 = \mathrm{mean}\,\langle x_i \rangle_{y_i=+1}$.

In order to determine those values (and therefore b), we need calibration data where the class labels is known.

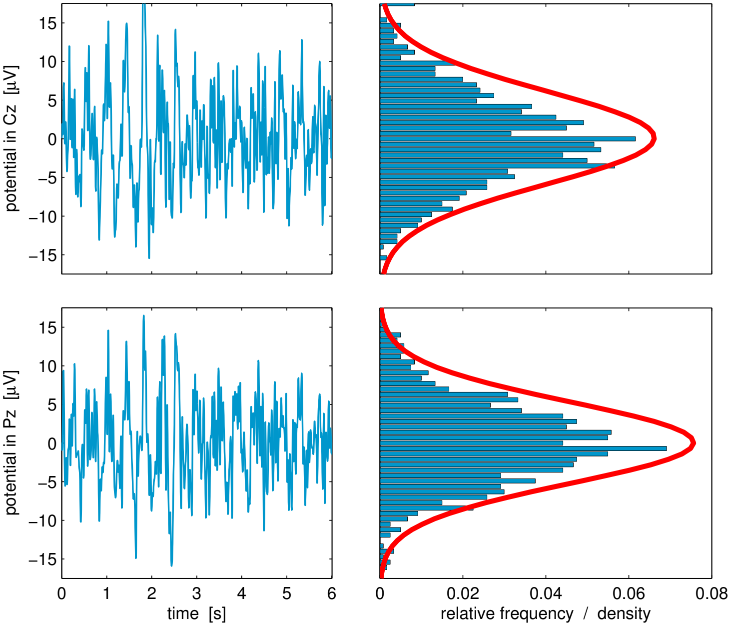
From Univariate to Multivariate Features

As mentioned in the previous lecture, we need to employ **multivariate features** in order to obtain better classification performance than with the simple threshold criterium:

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

To that end, we discuss multivariate distributions.

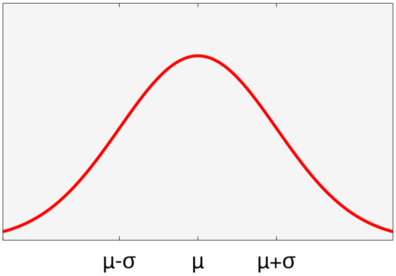
Univariate Distributions of Single-Channel EEG



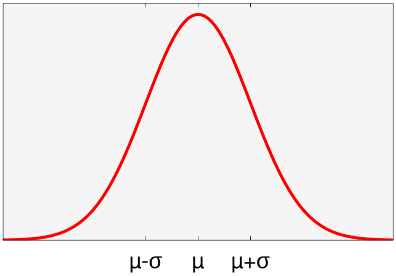
Two Univariate Gaussian Distributions

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

Component #1

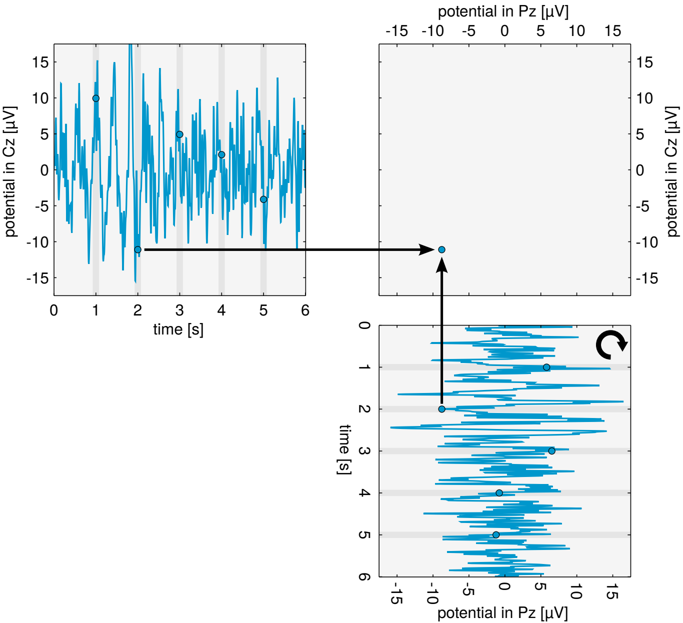


Component #2

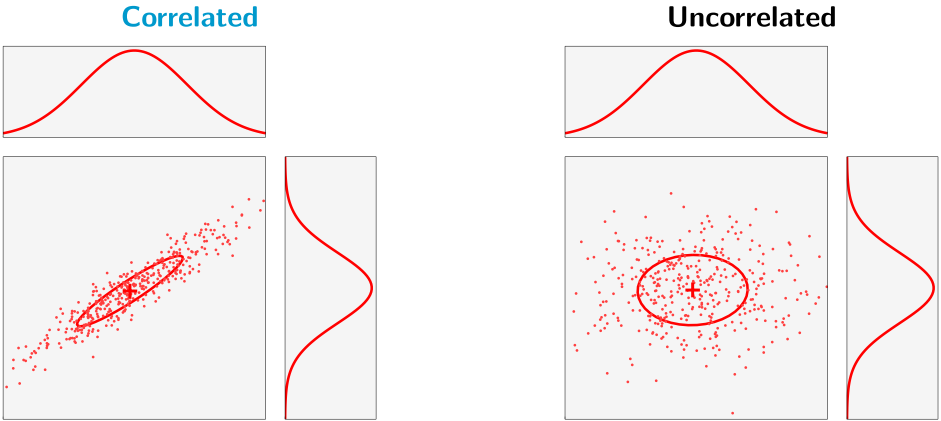


But what might their joint multivariate distribution look like?

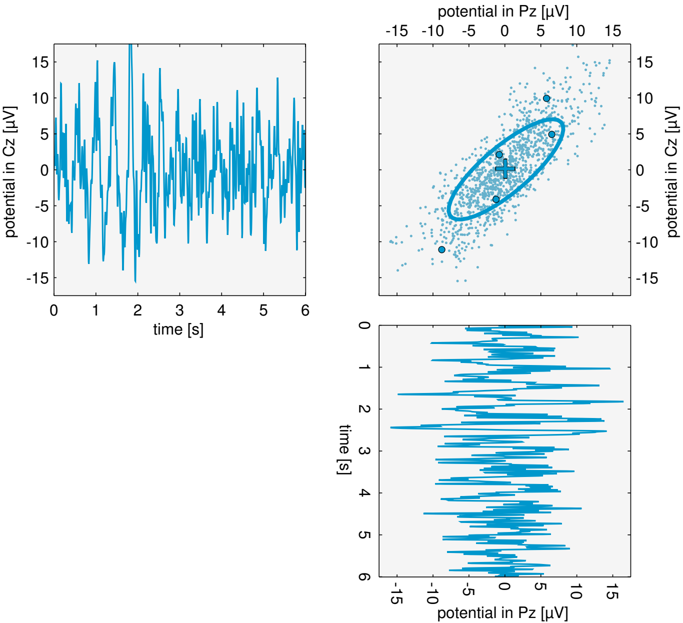
Two-Dimensional Distribution



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).
- **Remember spatial smearing!**


potential in Cz [µV]
time [s]
potential in Pz [µV]
potential in Cz [µV]
time [s]
potential in Pz [µV]

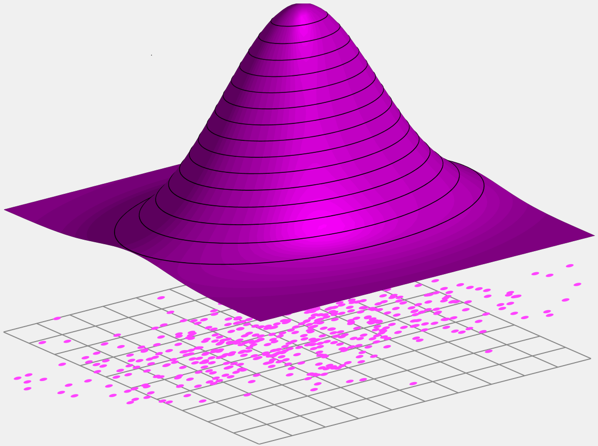
Multivariate Gaussian Distributions

Such multivariate Gaussian distributions, but typically with much more dimensions than two, as used for visualization, are the basis for our classification methods.

Therefore, we will look with more detail on the characterization of Gaussian distributions.

(a)

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



Some New Notions and a Bit of Math

For handling a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_T$ (each $\mathbf{x}_i$ is called *feature vector* or *sample*) we often concatenate them into a matrix

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T].$$

Same notation is used for (multivariate) time series $\mathbf{x}(t) \;\; (t = 1, \ldots, T)$.
Each column of $\mathbf{X}$ is an observation, each row of $\mathbf{X}$ is called a factor.

The mean is defined as $\boldsymbol{\mu}_\mathbf{X} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$.

To estimate the covariance, typically the unbias estimator is used

$$\boldsymbol{\Sigma}_\mathbf{X} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{x}_t - \boldsymbol{\mu}_\mathbf{X})(\mathbf{x}_t - \boldsymbol{\mu}_\mathbf{X})^\top$$

To ease calculations, one often assumes that the rows of $\mathbf{X}$ have zero mean. (Proofs in this course using the zero mean assumption are true also in general.) This simplfies the formula to

$$\boldsymbol{\Sigma}_\mathbf{X} = \frac{1}{T-1}\mathbf{X}\mathbf{X}^\top$$

Some New Notions and a Bit of Math (2)

We define the Kronecker vector $\boldsymbol{\delta}_i \in \mathbb{R}^p$ to be the vector with all entries being 0, except for the i-th entry being 1.

$$\delta_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i\text{-th row}$$

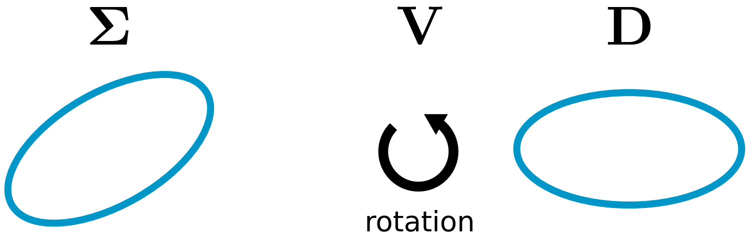
Eigenvalue Decomposition (EVD)

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

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

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

The decomposition can be thought of as rotation of an uncorrelated distribution:



The role of $\mathbf{V}$ and $\mathbf{D}$ are explained on a subsequent slide.

Variance in Direction of Eigenvectors

Let samples $\mathbf{X} \in \mathbb{R}^{p \times T}$ and a vector $\mathbf{w} \in \mathbb{R}^{P}$ of length 1 be given.

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

$$\begin{aligned} \operatorname{var}(\mathbf{w}^\top \mathbf{X}) &= \frac{1}{T-1} \mathbf{w}^\top \mathbf{X} (\mathbf{w}^\top \mathbf{X})^\top = \frac{1}{T-1} \mathbf{w}^\top \mathbf{X} \mathbf{X}^\top \mathbf{w} \\ &= \mathbf{w}^\top \boldsymbol{\Sigma}_\mathbf{X} \mathbf{w} \end{aligned} \tag{1}$$

Now, assume an EVD of the covariance matrix $\boldsymbol{\Sigma}_\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$.

- Let $\mathbf{v}_i$ be the i-th Eigenvector of $\boldsymbol{\Sigma}_\mathbf{X}$. **What is the variance of $\mathbf{X}$ in the direction of $\mathbf{v}_i$?**

$$\operatorname{var}(\mathbf{v}_i^\top \mathbf{X}) \overset{(1)}{=} \mathbf{v}_i^\top \boldsymbol{\Sigma}_\mathbf{X} \mathbf{v}_i = \mathbf{v}_i^\top \mathbf{V}\mathbf{D}\mathbf{V}^\top \mathbf{v}_i = \boldsymbol{\delta}_i^\top \mathbf{D} \boldsymbol{\delta}_i = d_i \tag{2}$$

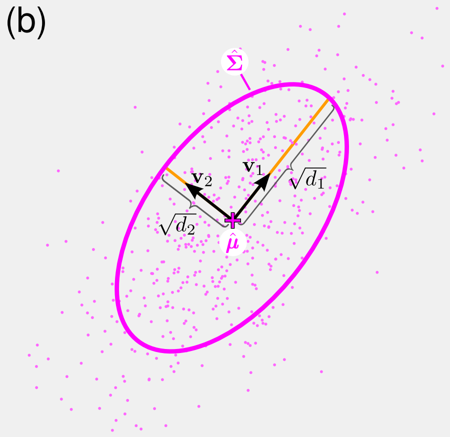
where the orthonormality of $\mathbf{V}$ is exploited in $\mathbf{v}_i^\top \mathbf{V} = \boldsymbol{\delta}_i$.

This material is thought to be just repetition from other courses.

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

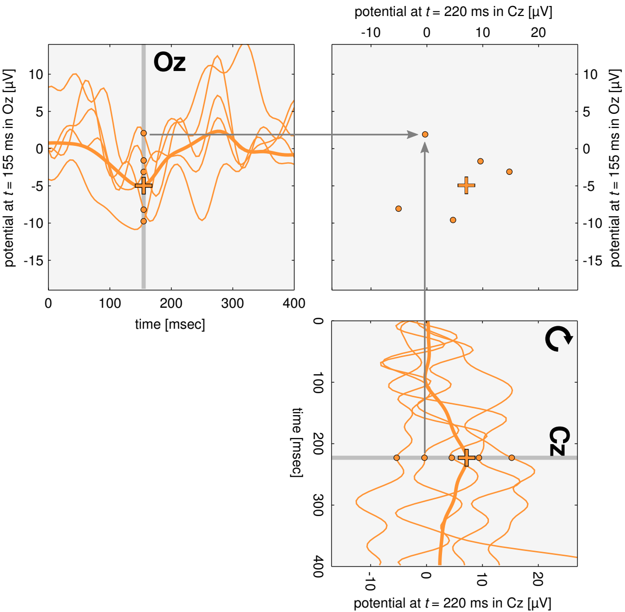
Eigenvalue decomposition of the covariance $\hat{\boldsymbol{\Sigma}}$ of $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$:

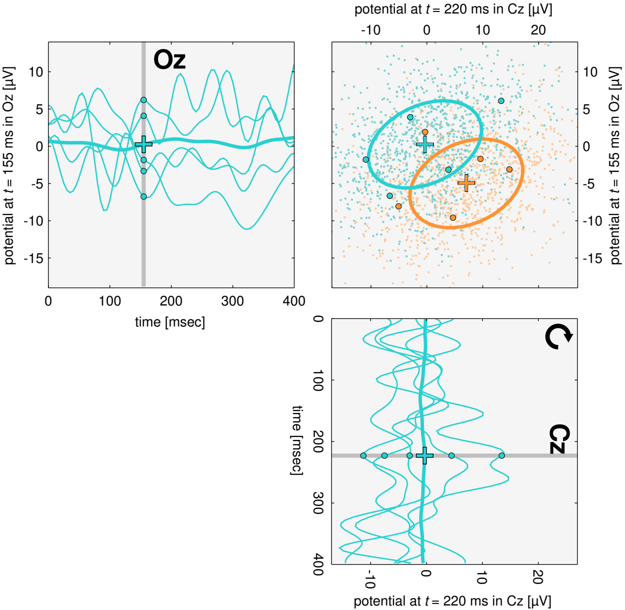
$$\hat{\boldsymbol{\Sigma}} = \mathbf{V}\mathbf{D}\mathbf{V}^\top, \quad \text{with orthonormal } \mathbf{V} \text{ and diagonal } \mathbf{D}.$$

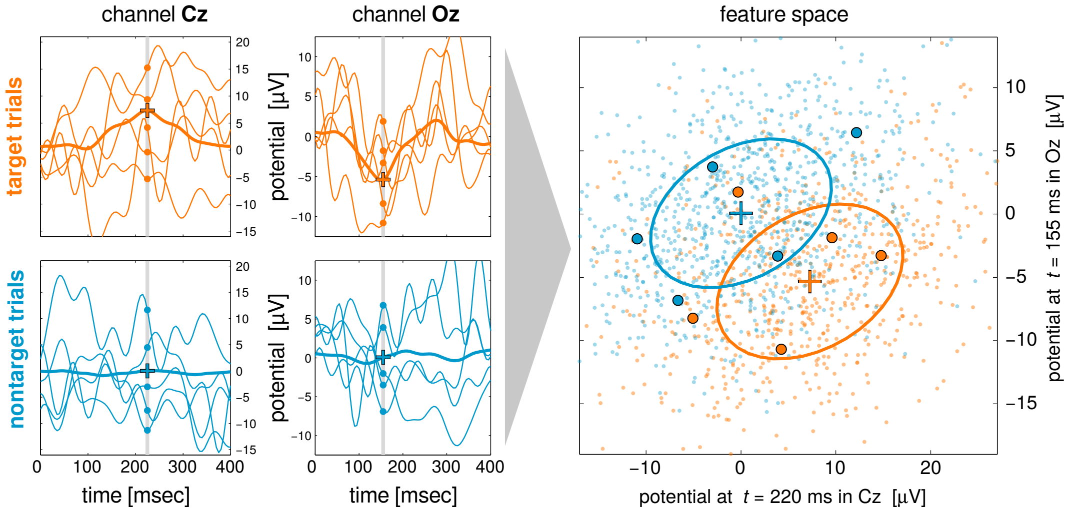


- Eigenvectors are columns of $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$.
- Eigenvalues are diagonal elements d_i of $\mathbf{D}$.
- $\sqrt{d_i} = \mathrm{std}(\mathbf{v}_i^\top \mathbf{X})$
- In $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ typically $\boldsymbol{\mu}$ is considered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

Representation of Multivariate Distributions: Scatter Plot




Oz
potential at t = 155 ms in Oz [µV]
time [msec]
0
100
200
300
400
10
5
0
-5
-10
-15
potential at t = 220 ms in Cz [µV]
-10
0
10
20
Cz


channel **Cz**
channel **Oz**
feature space
target trials
nontarget trials
potential [µV]
time [msec]
potential at t = 220 ms in Cz [µV]
potential at t = 155 ms in Oz [µV]

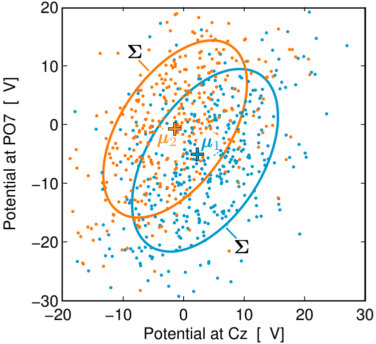
Distribution of ERP Features

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

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

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

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



For features of ERP data:

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

[Blankertz et al, NeuroImage 2011]

Interlude: About Matrices

Orthonormal matrices $\mathbf{V} \in \mathcal{O}(p)$ satisfy

$$\mathbf{V}^\top \mathbf{V} = \mathbf{I} = \mathbf{V}\mathbf{V}^\top.$$

For a diagonal matrix $\mathbf{D} \in \mathrm{Diag}(p)$ we have

$$\mathbf{D}^{1/2} = \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_p}).$$

The square root of a sym. and pos. definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ is uniquely defined. Employing the EVD $\mathbf{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ we can confirm the identity

$$\mathbf{\Sigma}^{1/2} = \mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^\top$$

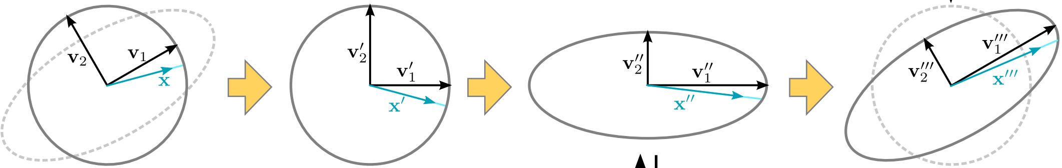
by checking the property:

$$\begin{aligned}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2} &= \mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^\top\mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^\top = \mathbf{V}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{V}^\top \\ &= \mathbf{V}\mathbf{D}\mathbf{V}^\top = \mathbf{\Sigma}\end{aligned}$$

Illustration of Multiplication by sqrt-Covariance Matrix

$$\mathbf{x} \mapsto \mathbf{\Sigma}^{1/2}\mathbf{x} \;=\; \mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^\top \mathbf{x}$$

$\mathbf{v}_1 \mapsto \mathbf{v}_1''' = \sqrt{d_1} \cdot \mathbf{v}_1 \qquad \mathbf{v}_2 \mapsto \mathbf{v}_2''' = \sqrt{d_2} \cdot \mathbf{v}_2$



$\mathbf{x} \mapsto \mathbf{x}' := \mathbf{V}^\top \mathbf{x}$

rotation:

$$\mathbf{v}_1 \mapsto \begin{bmatrix} \mathbf{v}_1^\top \\ \mathbf{v}_2^\top \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{v}_2 \mapsto \begin{bmatrix} \mathbf{v}_1^\top \\ \mathbf{v}_2^\top \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$\mathbf{x}' \mapsto \mathbf{x}'' := \mathbf{D}^{1/2}\mathbf{x}'$

scaling:

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \mapsto \mathbf{D}^{1/2} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \sqrt{d_1} \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} \mapsto \mathbf{D}^{1/2} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{d_2} \end{bmatrix}$$

$\mathbf{x}'' \mapsto \mathbf{x}''' := \mathbf{V}\mathbf{x}''$

rotation (inverse to first):

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \mapsto [\mathbf{v}_1 \ \mathbf{v}_2] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \mathbf{v}_1$$

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} \mapsto [\mathbf{v}_1 \ \mathbf{v}_2] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \mathbf{v}_2$$

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 Eigevectors $\mathbf{v}_i$ to the coordinate axes. *2. Step.* The multiplication of a vector with the diagonal matrix $\mathbf{D}^{1/2}$ 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.

Some Notions for Defining Feature Extraction

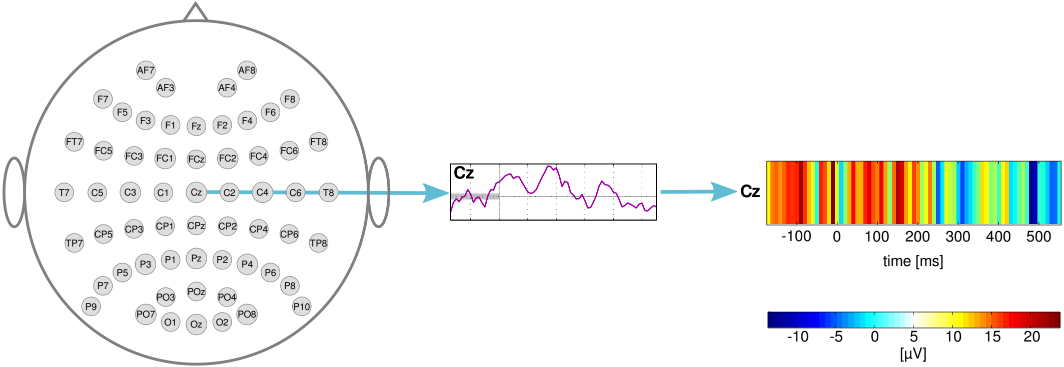
On the following slides, different kinds of features are introduced. On one hand, this is done in a very formal way, for which some notions are introduced below. On the other hand, the features (and the implementation of their calculation) can be understood without that formalism.

- $x_c(t)$ scalp potential at channel c and time point t within a trial
- $\mathcal{C} = \{c_1, \ldots, c_M\}$ subset of channels
- $\mathbf{x}_\mathcal{C}(t) = [x_{c_1}(t), \ldots, x_{c_M}(t)]^\top$ data vector for the subset of channels
- $\mathcal{T} = \langle \mathcal{T}_i \rangle_{i=1,\ldots,I}$ sequence of time intervals, with each $\mathcal{T}_i$ being a subset of time indices

Extraction of Temporal Features

Given a set of time intervals $\mathcal{T}$ and one channel c_1, we define the **temporal feature**

$$\mathbf{X}(\{c_1\}, \mathcal{T}) = [\text{mean}\,\langle x_{c_1}(t)\rangle_{t\in\mathcal{T}_1}; \ldots; \text{mean}\,\langle x_{c_1}(t)\rangle_{t\in\mathcal{T}_I}].$$

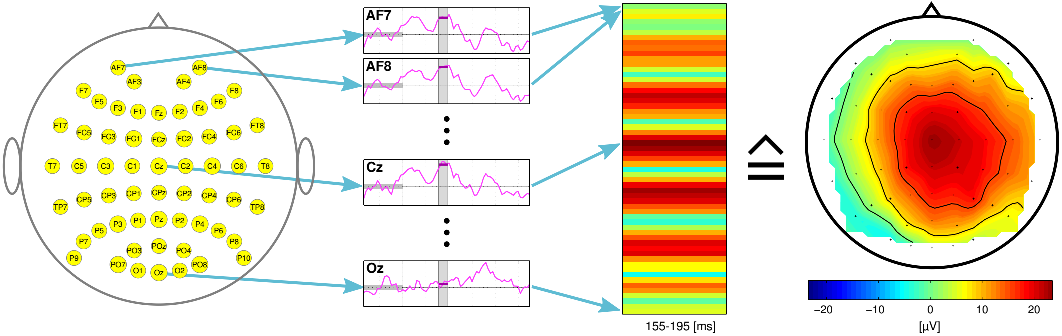


The dimensionality of the feature vector coincides with the number of sample points within one epoch (when using one-sample intervals for $\mathcal{T}_i$):

Given a set of channels $\mathcal{C}$ and one time interval $\mathcal{T}_1$, we define the **spatial feature**

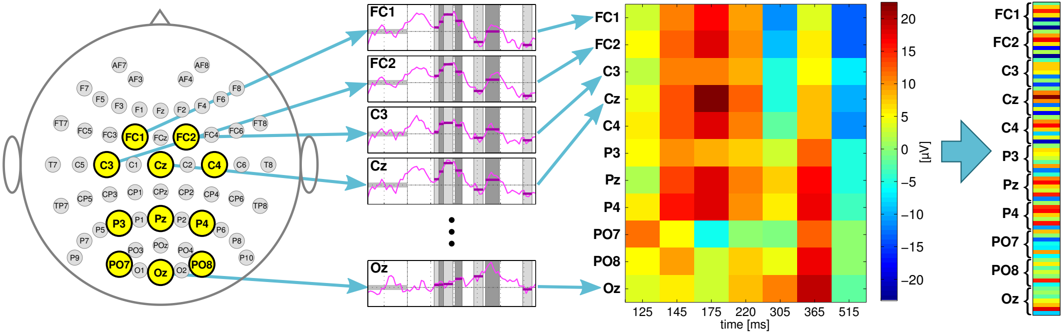
$$\mathbf{X}(\mathcal{C}, \langle \mathcal{T}_1 \rangle) = [\mathrm{mean}\, \langle \mathbf{x}_{\mathcal{C}}(t) \rangle_{t \in \mathcal{T}_1}].$$

The dimensionality of the feature vector coincides with the number of channels (when using all channels as $\mathcal{C}$.). Each value of the feature vector is the (average) amplitude in one time interval of the respective channel:

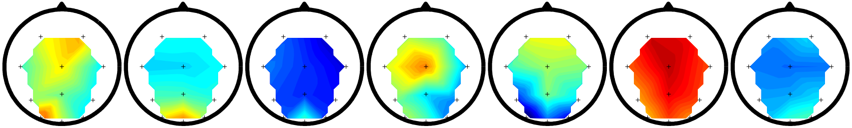


Generally, we define the *spatio-temporal* feature as

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



This feature can be visualized as a sequence of scalp maps:



Lessons Learnt

After this lecture you should

- be familiar with the different kinds of ERP features
- have an idea about multivariate Gaussian distribution and
  how they look like for ERP features
- be acquainted with the square root of a covariance matrix
- know Eigenvalue Decomposision (EVD) formally and
  the meaning of Eigenvectors and Eigenvalues
- be able to sketch a 2D Gaussian distribution with help of the EVD
- have notice about the formal definition of a binary classifier

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.