ICA

Lecture 6: Linear Subspace Projections: Independent Component Analysis

Machine Learning, Summer Term 2019

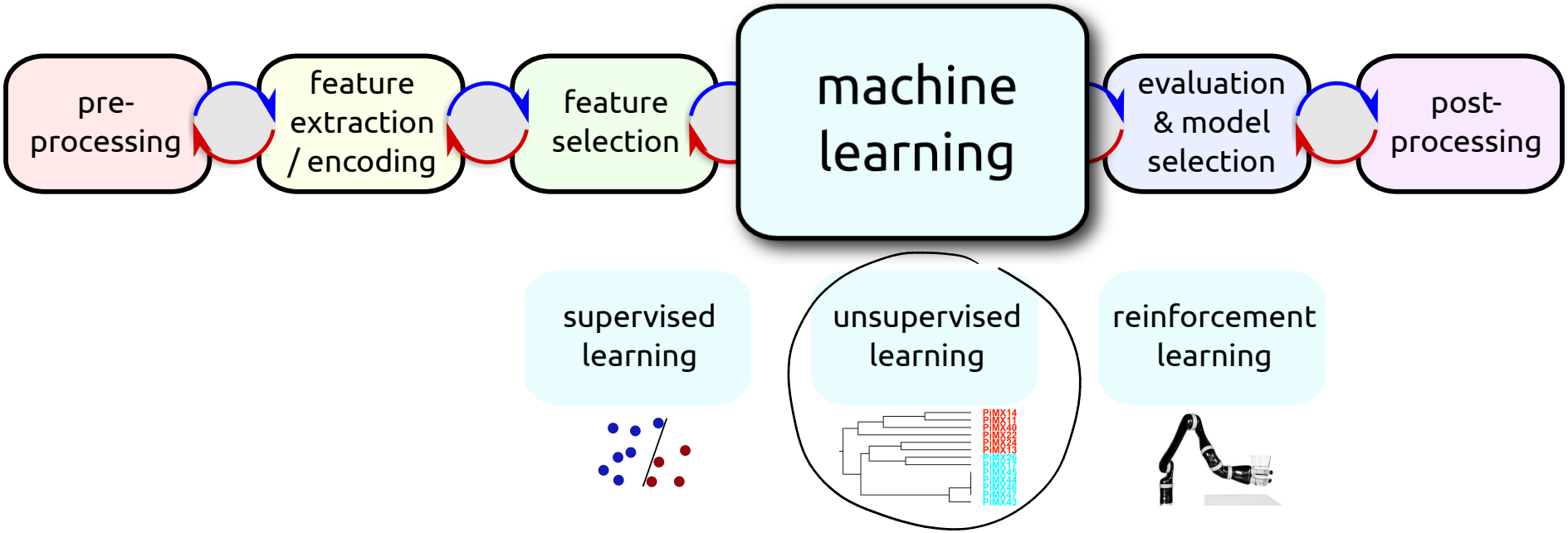
Michael Tangermann Frank Hutter Marius Lindauer

University of Freiburg

Lecture Overview

Lecture Overview

ML Design Cycle



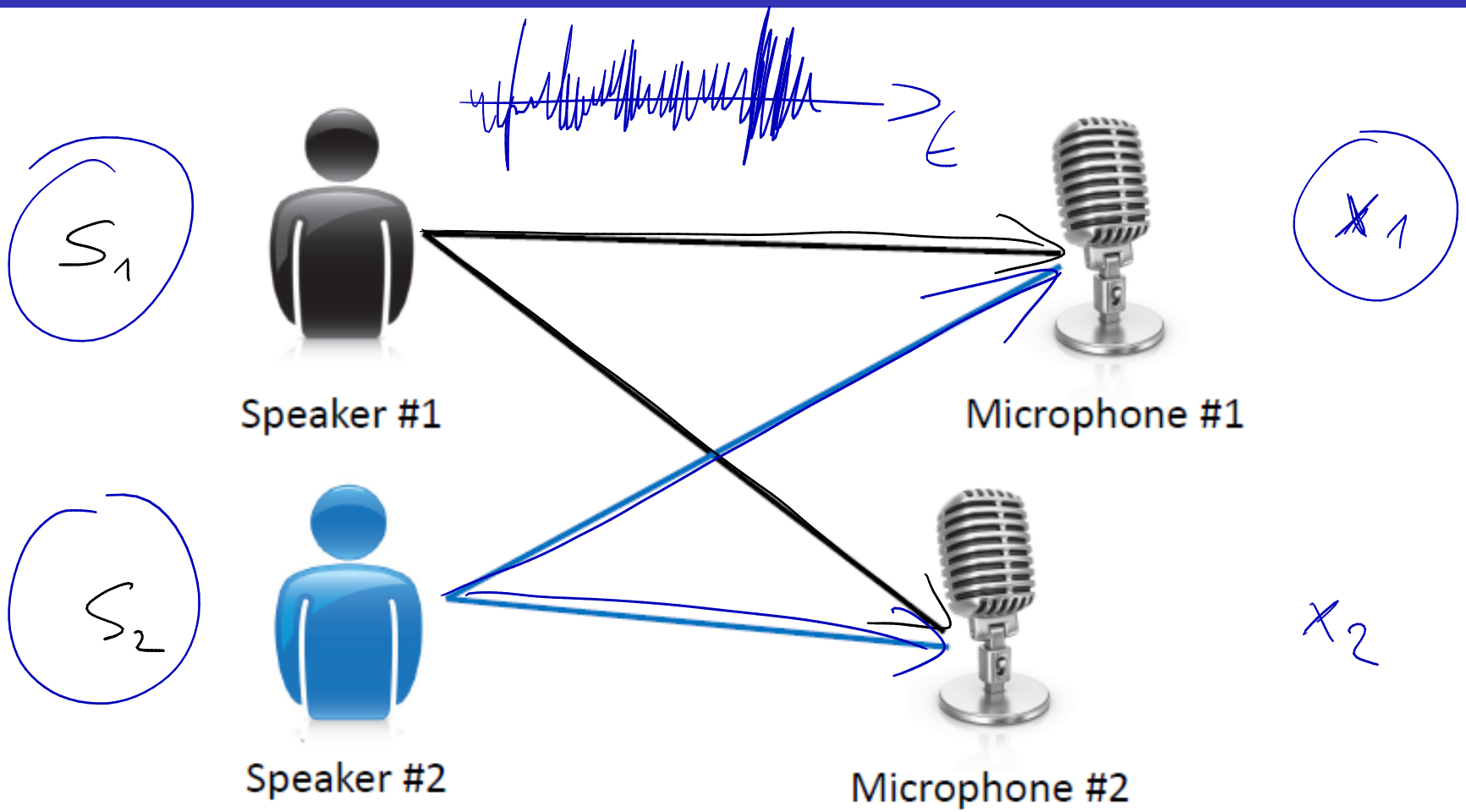
Today's topic is how to separate mixed data sources into *subspaces* or *components* making use of the unsupervised independent component analysis (ICA):

- Labels are not required

Motivation: Cocktail Party Problem

Using a number of microphones -
can we separate the single speakers (sources)?
For a computer: very hard problem due to reverberation!

Motivation: Cocktail Party Problem

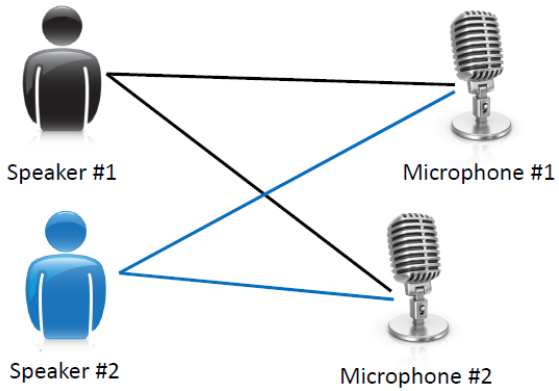


Assumptions: no reverberation,
number of speakers and microphones is matched.

Assumptions for ICA

Independent Component Analysis

Blind Source Separation

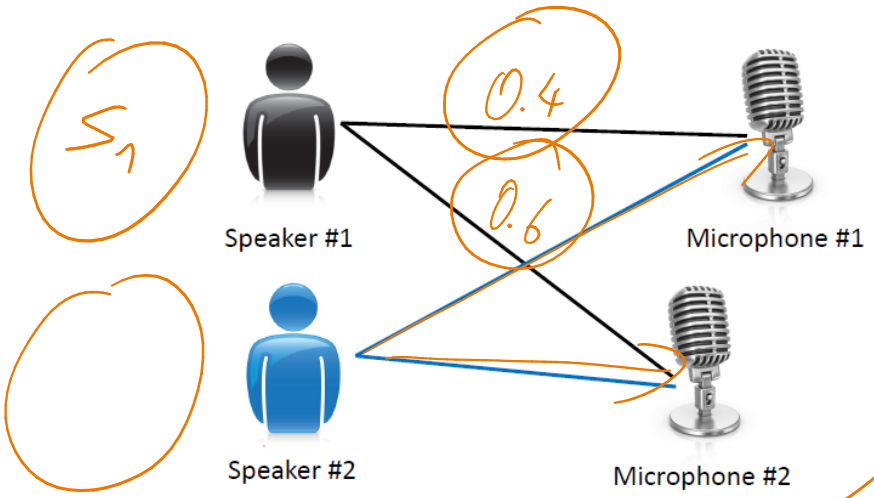


Demo: http://cnl.salk.edu/~tewon/Blind/blind_audio.html
Demo: http://d-kitamura.net/en/demo_rank1_en.htm
Demo: http://paris.cs.illinois.edu/demos/index.html
Demo: https://cnl.salk.edu/~tewon/Blind/blind_audio.html

General problem setting:

- N **independent** sound sources s_j, which may be active simultaneously.
- N sensors (microphones) are spread in the room. They capture different mixtures x_i of the sources.
- Sources $s_j(t)$ and observed mixtures $x_i(t)$ are time series signals. (with $t = \{1, \ldots, T\}$, but this is usually omitted)

Assumptions for ICA



Helpful prerequisites:

- The observed signals are mean-free.
- The observed signals have been whitened.

Strict assumptions for ICA:

- The sources mix linearly into the observations.
- At least n-1 of the sources have a non-Gaussian distribution.
- The sources are statistically independent at each time point t.

Key Observations for ICA

From the assumptions:

- The sources mix linearly into the observations.
- At least n-1 of the sources have a non-Gaussian distribution.
- The sources are statistically independent at each time point t.

... and from the central limit theorem we can expect,

- that any mixture of sources is more Gaussian than (n-1-many of the) the original sources.

How can we make use of this?

Key Observations for ICA

From the assumptions:

- The sources mix linearly into the observations.
- At least n-1 of the sources have a non-Gaussian distribution.
- The sources are statistically independent at each time point t.

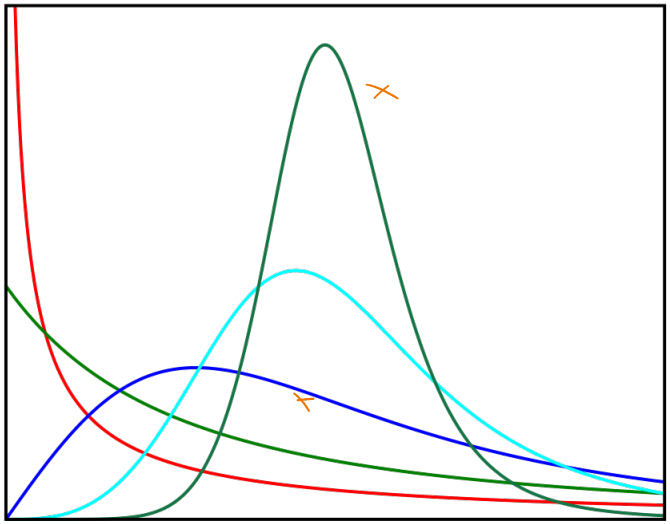
... and from the central limit theorem we can expect,

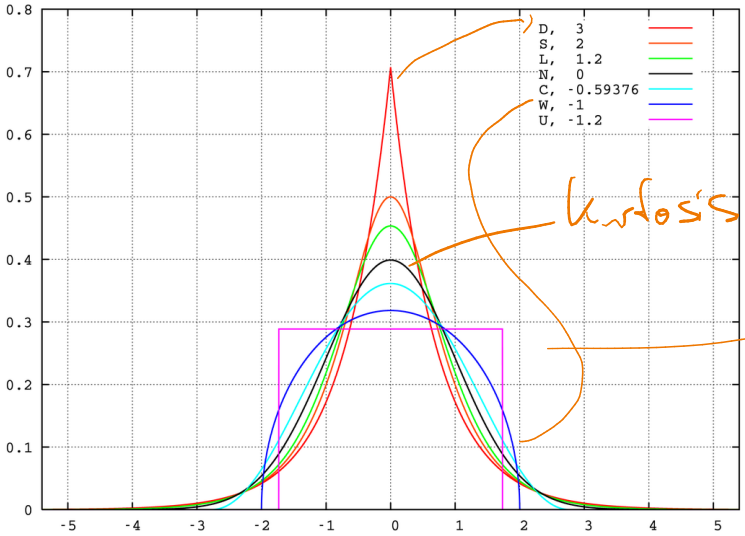
- that any mixture of sources is more Gaussian than (n-1-many of the) the original sources.

How can we make use of this? 

- Conversely: the independent sources will be less Gaussian than their mixture.
- For undoing the mixing, we should search for components, which are as non-Gaussian as possible!

Reminder: Non-Gaussian Distributions





- Distributions can be described by their moments (mean, variance, skewness, kurtosis)
- Gaussian distributions have kurtosis of zero, while non-Gaussian distributions have non-zero kurtosis!
- Kurtosis of y is defined by $kurt(y) = E\{y^4\} - 3(E\{y^2\})^2$
- Assumption of non-Gaussian distributions is often met by real data

[https://en.wikipedia.org/wiki/Kurtosis],

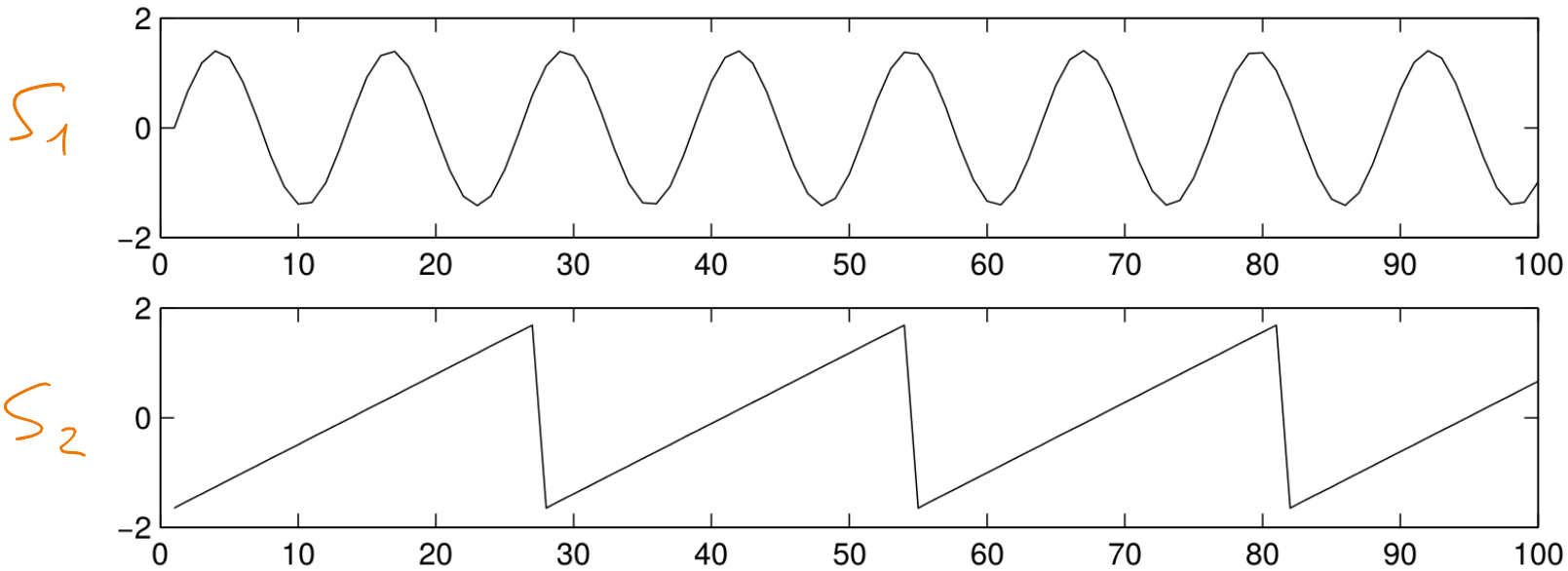
[https://en.wikipedia.org/wiki/Moment_(mathematics)]

The Simplest Cocktail Party Problem

Task:

- Try to recover (unmix) the unknown original sources $s_j(t)$ based on the recorded /observed mixtures $x_i(t)$.

Let's assume optimal conditions (no reverberation), just two sources and two microphones. These are the time series of the original sources:

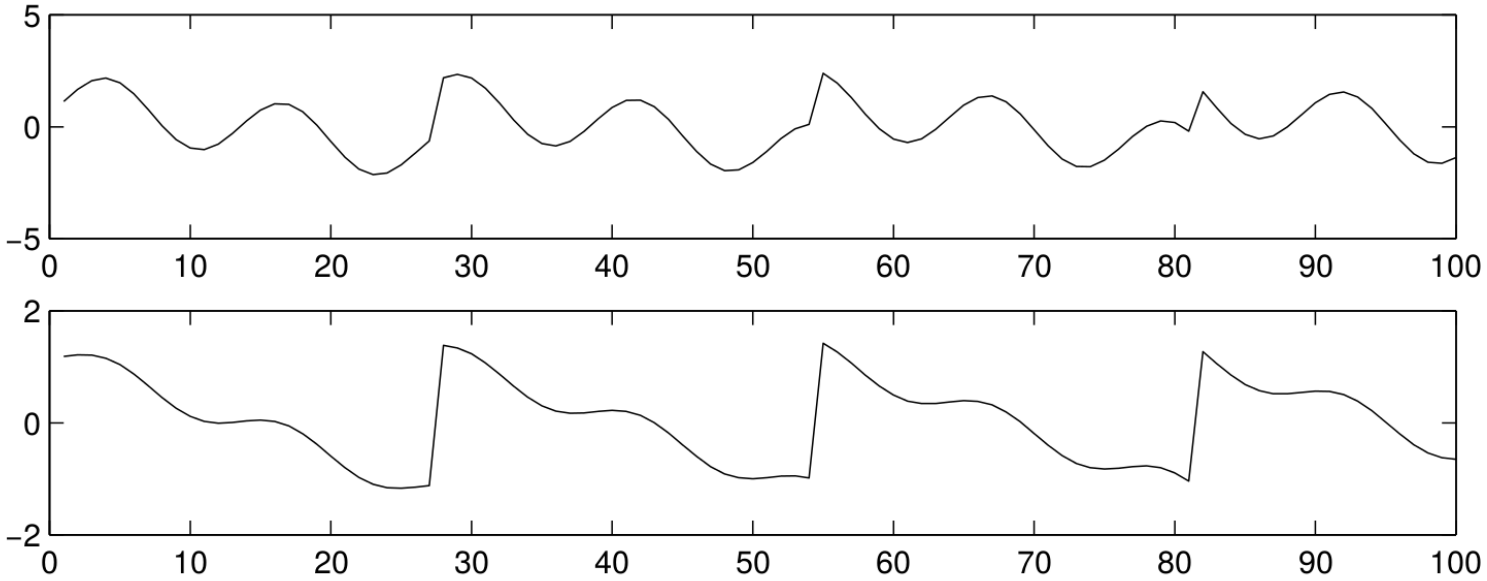


The Simplest Cocktail Party Problem

Task:

- Try to recover (unmix) the unknown original sources $s_j(t)$ based on the recorded /observed mixtures $x_i(t)$.

Let's assume optimal conditions (no reverberation), just two sources and two microphones. These are the mixed signals:



x_1

x_2

The Simplest Cocktail Party Problem

Task:

- Try to recover (unmix) the unknown original sources $s_j(t)$ based on the recorded /observed mixtures $x_i(t)$.

Let's assume optimal conditions (no reverberation, neither sources nor sensor move), just two sources and two microphones.

If we knew know the mixture coefficients, we could express the relationship between sources and observed microphone signals:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned}$$

The Simplest Cocktail Party Problem

Task:

- Try to recover (unmix) the unknown original sources $s_j(t)$ based on the recorded /observed mixtures $x_i(t)$.

Let's assume optimal conditions (no reverberation, neither sources nor sensor move), just two sources and two microphones.

If we knew know the mixture coefficients, we could express the relationship between sources and observed microphone signals:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned}$$

- If x_i and a_{ij} were provided - how could we recover the sources s_j?

Lecture Overview

How we get to an ICA Model

As we only know the mixed sensor signals x_i but not the mixing weights a_{ij}, solving the equations for s_j is impossible.
$\rightarrow$ We need additional information!

How we get to an ICA Model

As we only know the mixed sensor signals x_i but not the mixing weights a_{ij}, solving the equations for s_j is impossible.
$\rightarrow$ We need additional information!

Use the assumptions! We expected the sources s_i to have (mostly) non-Gaussian distributions and that they are statistically independent at all time points t.

How we get to an ICA Model

As we only know the mixed sensor signals x_i but not the mixing weights a_{ij}, solving the equations for s_j is impossible.
$\rightarrow$ We need additional information!

Use the assumptions! We expected the sources s_i to have (mostly) non-Gaussian distributions and that they are statistically independent at all time points t.

Thus we could try to extremize the kurtosis of estimated sources (while ensuring their independence) to solve the ICA problem.

Compare: what is minimized or maximized to solve PCA?

The ICA Model

For N sources and N sensors, the mixing or forward model can conveniently be written as (omitting time t for convenience):

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

$\mathbf{A} \in \mathbb{R}^{N \times N}$ is the so called mixing matrix and is assumed to be unknown (as are the sources $\mathbf{s}_1, \ldots, \mathbf{s}_N$).

The model is *generative*, as it describes, how the observed data $\mathbf{x}$ is generated by a mixing process of the underlying sources $\mathbf{s}$.

If we would be able to estimate $\mathbf{A}$, then its inverse $\mathbf{W}$ would tell us, how to obtain the hidden components based on our observed signals:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$$

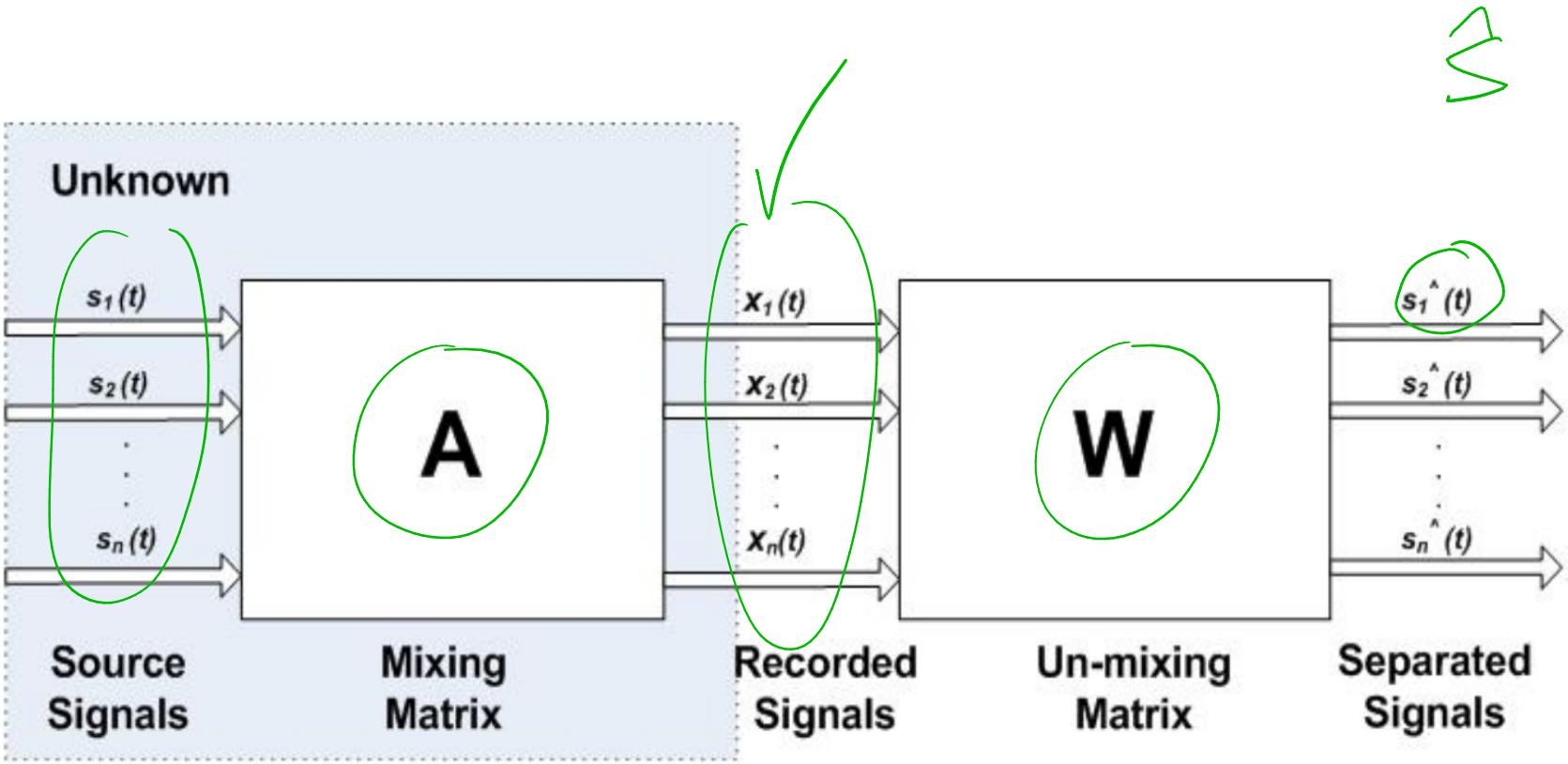
The ICA Model

If we would be able to estimate $\mathbf{A}$, then its inverse $\mathbf{W}$ would tell us, how to obtain the hidden components based on our observed signals:

$$\mathbf{s} = \mathbf{W}\mathbf{x}$$

This is sometimes called the backward model, with $\mathbf{W}$ being the unmixing matrix.

Example of Forward and Backward Model



How can we estimate $\mathbf{A}$ and $\mathbf{W}$?

Lecture Overview

Estimation of Model Parameters

Simplest approach: optimize the vector $\mathbf{w}$ (one of the columns of $\mathbf{W}$), which is used to project to column vector $\mathbf{x}$ onto an estimated source $\hat{s}$:

$$\hat{s} = \mathbf{w}^T \mathbf{x}$$

To optimize $\mathbf{w}$, loop until convergence:

- Initialize weight vector $\mathbf{w}$
- Determine direction, in which kurtosis of $\hat{s}$
  - grows most strongly (for positive kurtosis) or
  - decreases most strongly (for negative kurtosis)
- Run a step with a gradient descent method to get improved vector $\mathbf{w}$

Estimation of Model Parameters

Remarks:

- Iterative approach can be expanded to multivariate case
- Once $\mathbf{W}$ is estimated, $\mathbf{A}$ is available, too.

Estimation of Model Parameters

Remarks:

- Iterative approach can be expanded to multivariate case
- Once $\mathbf{W}$ is estimated, $\mathbf{A}$ is available, too.
- Unfortunately kurtosis is hard to estimate in a robust way, thus other measures of non-Gaussianity are preferred in practical implementations:
  - negentropy (to be maximized)
  - mutual information (to be minimized)
  - likelihood (to be maximized)
  - ...

Estimation of Model Parameters

Remarks:

- Iterative approach can be expanded to multivariate case
- Once $\mathbf{W}$ is estimated, $\mathbf{A}$ is available, too.
- Unfortunately kurtosis is hard to estimate in a robust way, thus other measures of non-Gaussianity are preferred in practical implementations:
  - negentropy (to be maximized)
  - mutual information (to be minimized)
  - likelihood (to be maximized)
  - ...
- Popular implementation of ICA (available in most toolboxes): FastICA

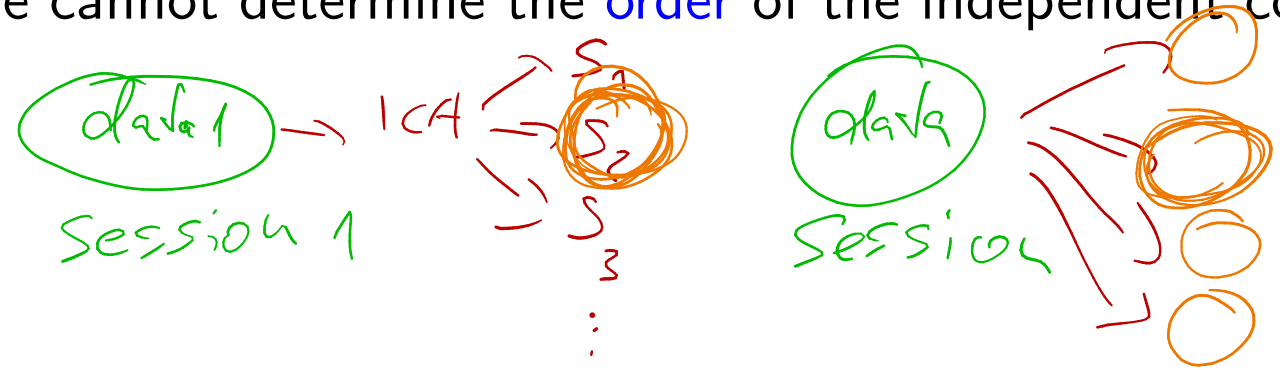
Lecture Overview

Ambiguity of ICA Solution

The ICA problem is **under-determined**, as we only know sensor time series $\mathbf{X}$.

Though solutions can be found by making strong assumptions, they are ambiguous:

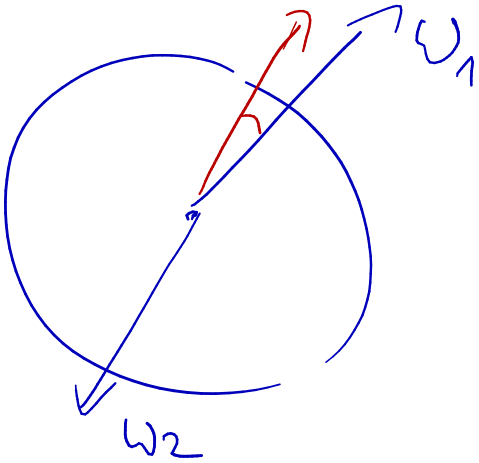
- We cannot determine the variances (energies) of the independent components.
- We cannot determine the signs of the independent components.
- We cannot determine the order of the independent components.



Typical Pitfalls for ICA on Real Data

Your data is noisy, and repeated runs of ICA on data of successive experimental sessions will deliver slightly different components. However, you would like to compare sessions...

How can you deal with this problem?



Typical Pitfalls for ICA on Real Data

Your data is noisy, and repeated runs of ICA on data of successive experimental sessions will deliver slightly different components. However, you would like to compare sessions...

How can you deal with this problem? 

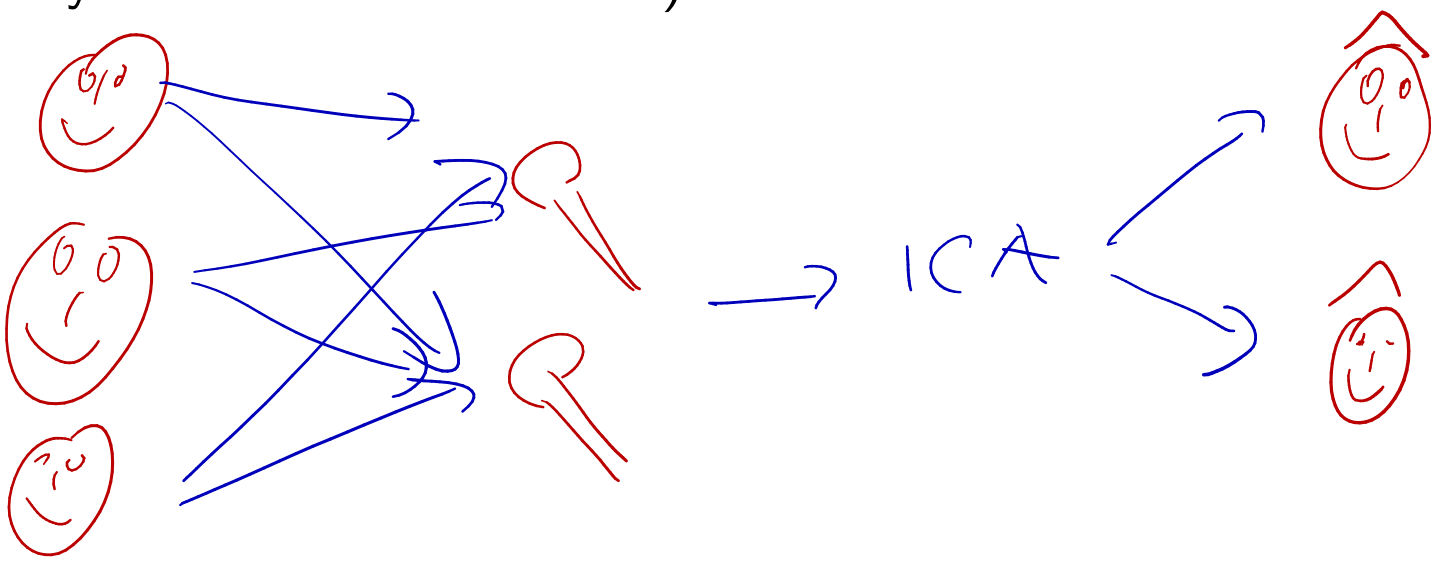
Helpful strategies:

- Try to match the components obtained from repeated runs of the ICA. (difficult! Non-Euclidian space...)
- Train the unmixing matrix based on data of one session and apply it to data of another session.

Typical Pitfalls for ICA on Real Data

The number of independent sources and the number of your sensors may not match.

- Case 1: You have more sources than sensors – cp. natural cocktail party problem with a human listener (two sensors!) and more than 2 speakers (very bad for standard ICA!). What is the effect?



Typical Pitfalls for ICA on Real Data

The number of independent sources and the number of your sensors may not match.

- Case 1: You have more sources than sensors – cp. natural cocktail party problem with a human listener (two sensors!) and more than 2 speakers (very bad for standard ICA!). What is the effect?
- Case 2: you have less sources than sensors

Please discuss with your neighbours: what may be the result in case 2?

Typical Pitfalls for ICA on Real Data

The number of independent sources and the number of your sensors may not match.

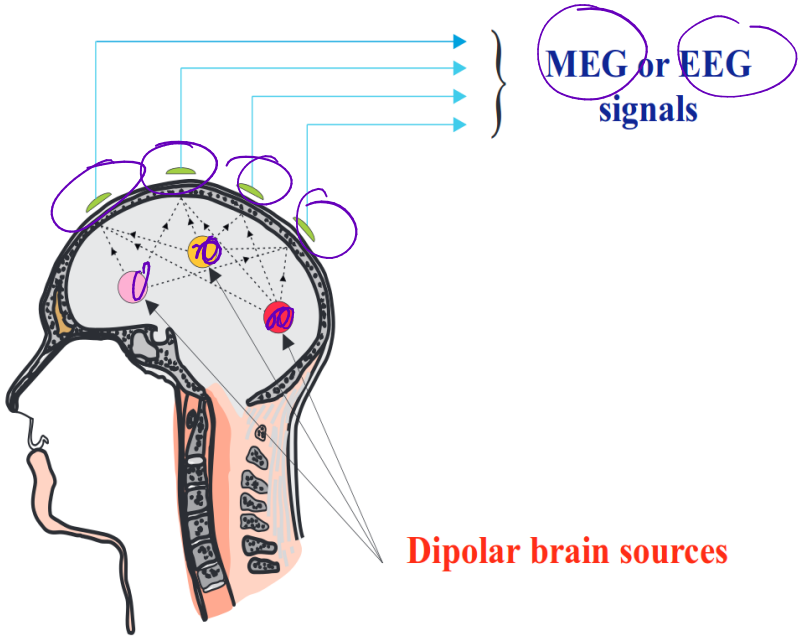
- Case 1: You have more sources than sensors – cp. natural cocktail party problem with a human listener (two sensors!) and more than 2 speakers (very bad for standard ICA!). What is the effect?
- Case 2: you have less sources than sensors

Please discuss with your neighbours: what may be the result in case 2?

As data is usually noisy, ICA will start to create arbitrary splits of meaningful components. Potential solutions are:

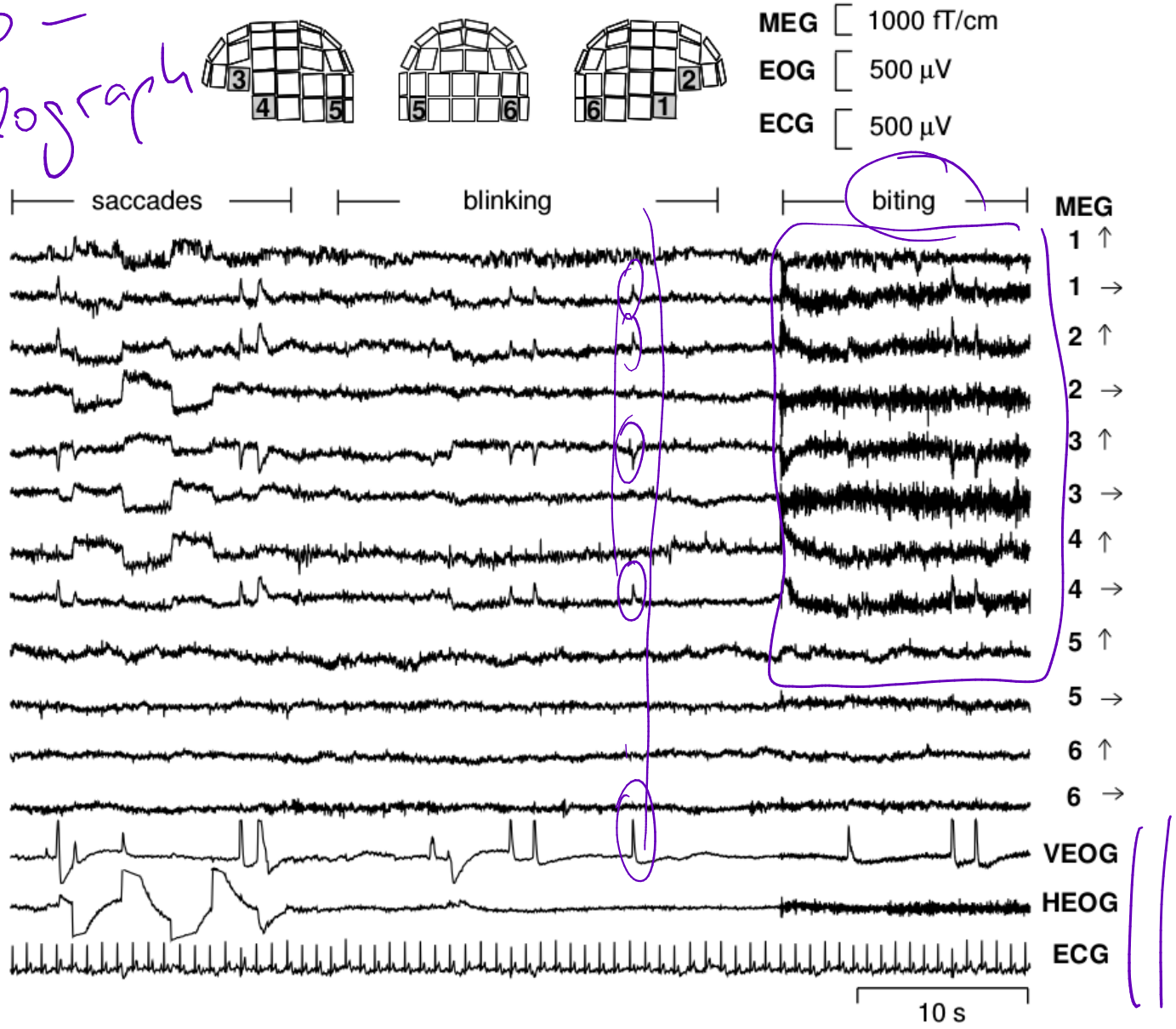
- Try to merge corresponding subspaces post-hoc
- Reduce the dimensionality of your data prior to applying ICA. PCA may do a good job...

Application Fields for ICA



- Neuroscience: interpretation of brain signals and their sources, separation of neural signals from artifacts
- Hearing aid research
- Prediction of stock market prices
- Telecommunications
- Geology
- Radioastronomy
- Image denoising

Typical ICA Workflow on (Neuro-)Physiological Data

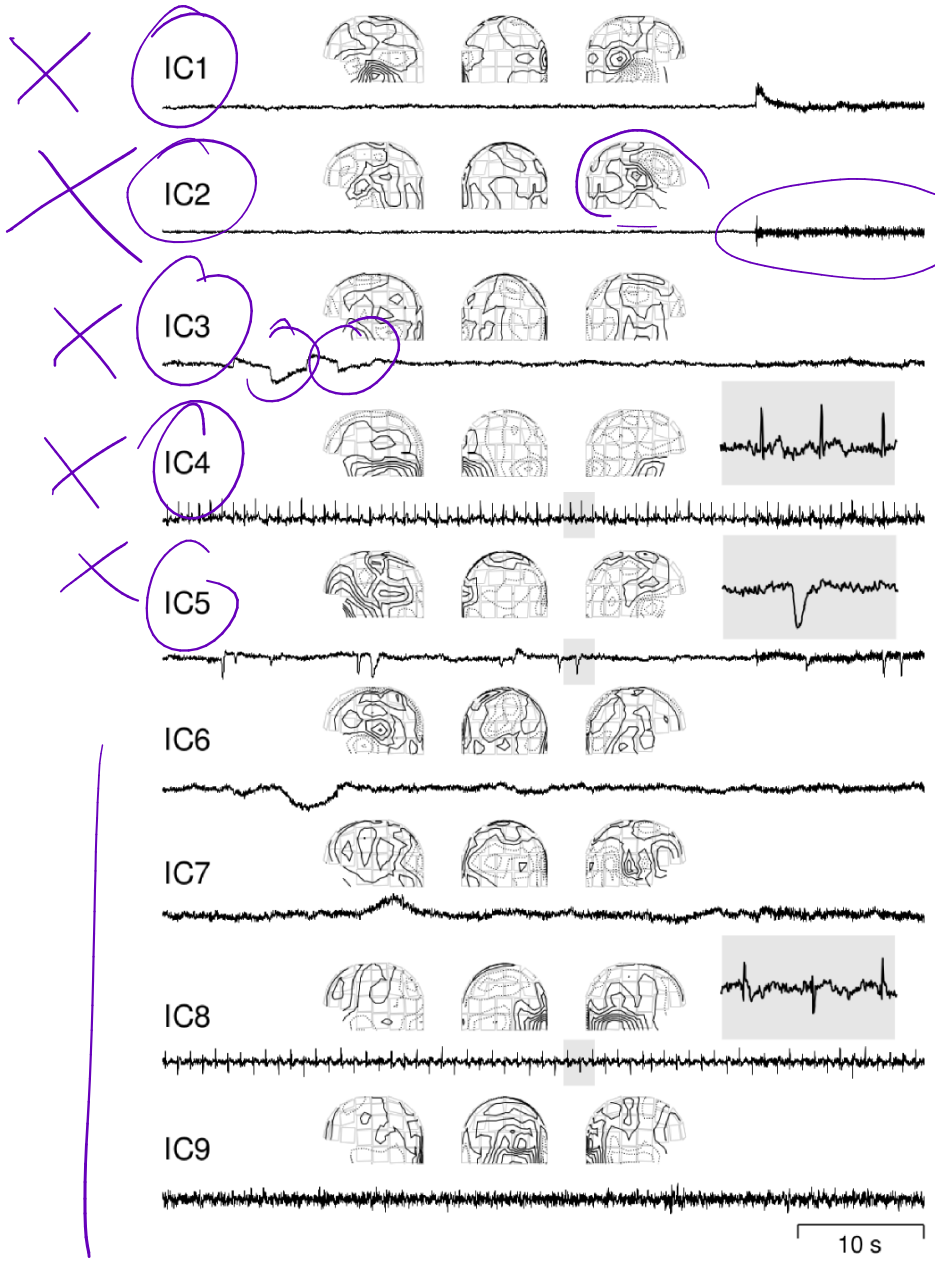


Typical Workflow for Independent Component Analysis

Perform the following steps:

- Unmix the recorded data to obtain the original sources
- Inspect the components visually into desired and undesired sources (remark: automated approaches exist to classify neural- from non-neural sources, e.g. [MARA toolbox by Winkler et al.,J. Neural Eng. 2014])

Typical Workflow of Independent Component Analysis



Typical Workflow for Independent Component Analysis

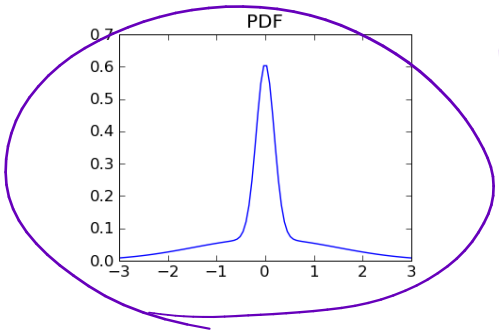
Perform the following steps:

- Unmix the recorded data to obtain the original sources
- Inspect the components visually into desired and undesired sources (remark: automated approaches exist to classifiy neural- from non-neural sources, e.g. [MARA toolbox by Winkler et al.,J. Neural Eng. 2014])
- Option 1: Use desired sources only, continue to work in the lower-dimensional ICA-space
- Option 2: Reconstruct the sensor data in the original space using only the desired sources

Attention: Option 2 leads to reconstructed data,
which may not have full rank any more!

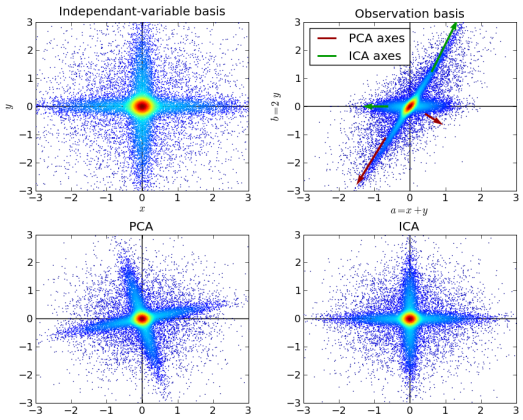
Comparison with PCA

Let's choose two variables, which are independent and are sampled from this non-Gaussian probability density function:



kurtosis > 0

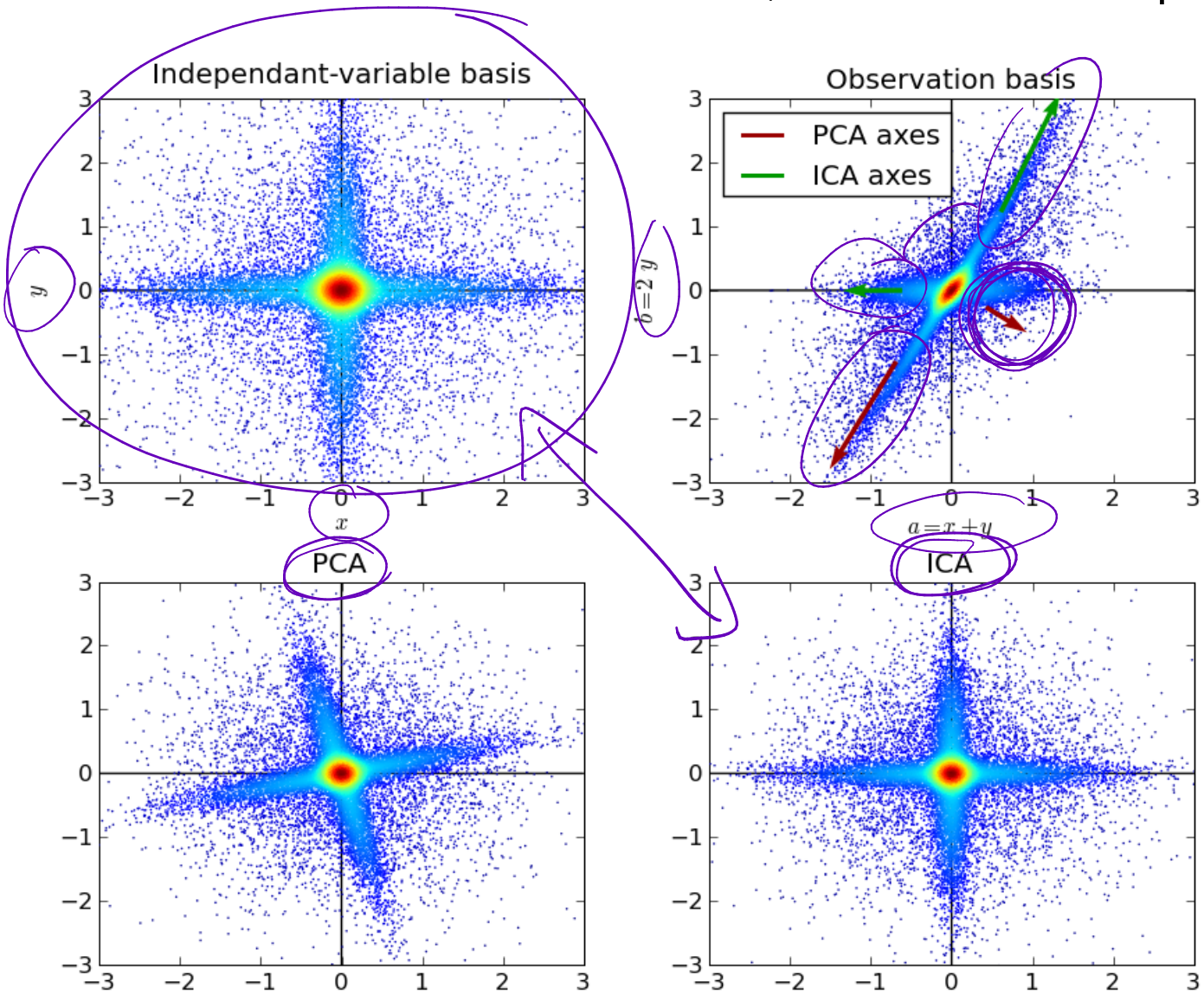
In some situations, PCA can not recover the original sources, while ICA is able to separate them:



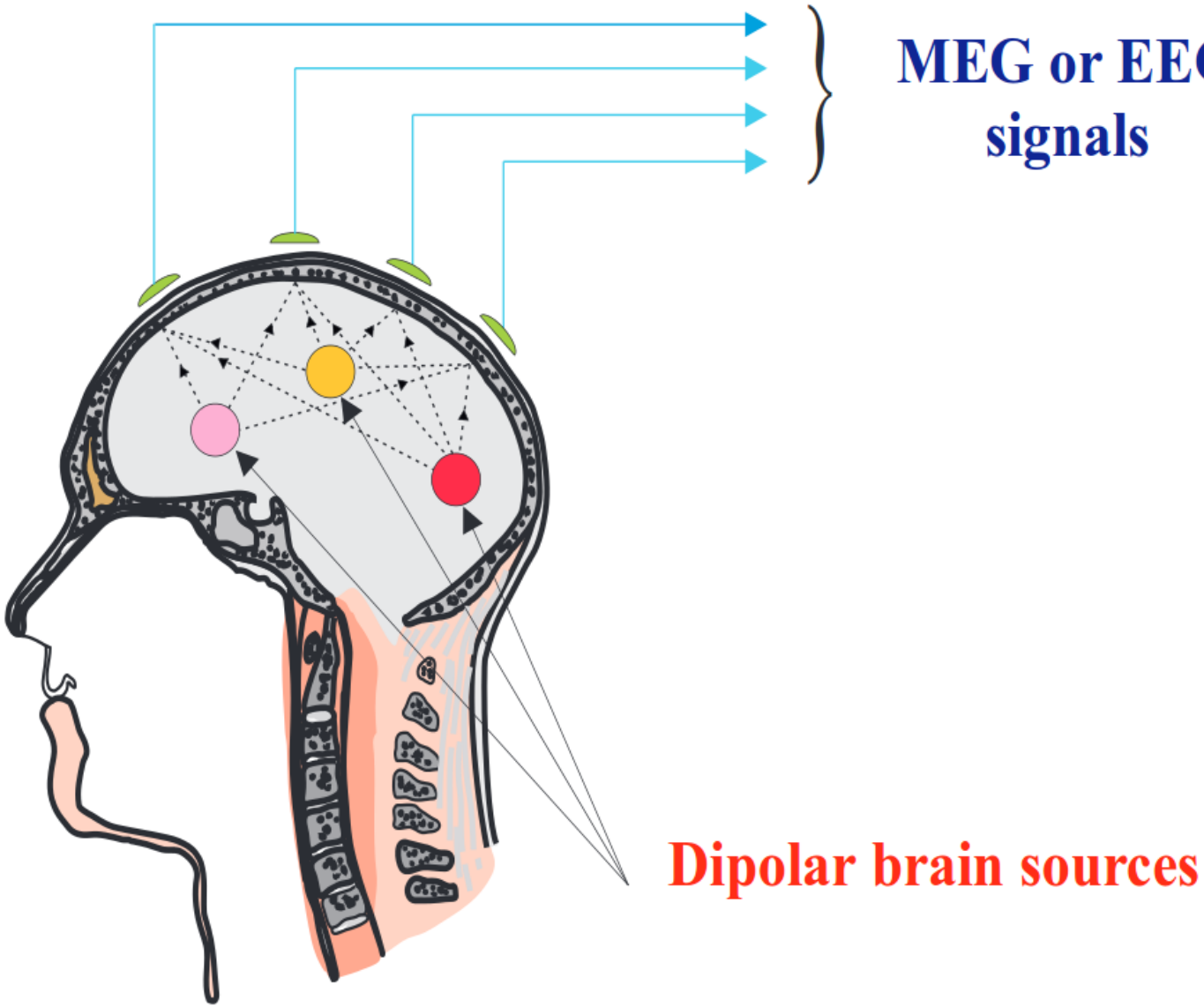
(zoom in...)

Comparison with PCA

While PCA can not recover the two sources, ICA is able to separate them:



Application Fields for ICA



Lecture Overview

Pros and Cons of ICA

- Con: Components have arbitrary sign, arbitrary order and amplitude.
  $\rightarrow$ Finding the matching components e.g. for two experimental sessions is not trivial.

Pros and Cons of ICA

- Con: Components have arbitrary sign, arbitrary order and amplitude.
  $\rightarrow$ Finding the matching components e.g. for two experimental sessions is not trivial.
- Pro: ICA is a linear method - once trained,
  it can be applied extremely fast for online systems

Pros and Cons of ICA

- Con: Components have arbitrary sign, arbitrary order and amplitude.
  $\rightarrow$ Finding the matching components e.g. for two experimental sessions is not trivial.
- Pro: ICA is a linear method - once trained,
  it can be applied extremely fast for online systems
- Pro: Being a linear method, independent components can be visualized. Experts can judge the quality of the unmixing.

Pros and Cons of ICA

- Con: Components have arbitrary sign, arbitrary order and amplitude. $\rightarrow$ Finding the matching components e.g. for two experimental sessions is not trivial.
- Pro: ICA is a linear method - once trained, it can be applied extremely fast for online systems
- Pro: Being a linear method, independent components can be visualized. Experts can judge the quality of the unmixing.
- Pro: Many variants of ICA exist. They use different assumptions on what "statistical independence" means. Thus for most unmixing problems (specific forms of noise, exploit temporal correlations within sources, perform ICA across data of several subjects etc.), you will probably find a variant, which can deal with your data.

Related Subspace Methods

- Whitening / sphereing: transform data to zero mean and unit covariance (preprocessing step for ICA)
- Factor analysis FA (incorporate domain-specific assumptions)
- Canonical correlation analysis CCA (relate two data sources to a common subspace which maximizes cross-covariance)
- Kernel-ICA (non-linear extension of ICA)
- Blind Source Separation (BSS) - actually ICA is a special case of BSS.

Ressources on ICA

- Great collection of demos, easy and detailed papers on ICA on the webpage of Aapo Hyvärinen's group in Finland: [http://research.ics.aalto.fi/ica/].
  (The lecture was mostly based on his great tutorial paper!)
- Cool applications of ICA and advanced algorithms for blind source separation on website of Paris Smaragdis: [http://paris.cs.illinois.edu/]
- Wikipedia.org on ICA for a great top-down overview!
- Tutorial provided for the EEGlab matlab toolbox by the group of Scott Makeig

Summary by learning goals

Having heard this lecture, you can now . . .

- Formulate the type of problems, that can be solved by ICA (and which can not)
- Formulate the assumptions made by ICA
- Formulate and solve the optimization problem for ICA
- Explain the difference between ICA and PCA