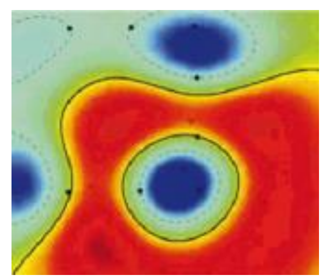
Independent Component Analysis (ICA)



Lecture by Klaus-Robert Müller, TUB 2013

Recap: PCA

PCA finds a linear transformation $\mathbf{V}$ in the data space such that the components $y_i(t)$ of

$$\mathbf{y}(t) = \mathbf{V}\mathbf{x}(t) \tag{1}$$

are uncorrelated. In other words, PCA diagonalizes the Covariance matrix:

$$\mathbf{C}_{\mathrm{y}} = E\{\mathbf{y}\mathbf{y}^\top\} = \mathbf{V}E\{\mathbf{x}\mathbf{x}^\top\}\mathbf{V}^\top = \mathbf{V}\mathbf{C}_{\mathrm{x}}\mathbf{V}^\top = diag. \tag{2}$$

This is a symmetric eigenvalue problem, so columns of the matrix $\mathbf{V}^\top$ are the eigenvectors of $\mathbf{C}_{\mathrm{x}}$. The matrix $\mathbf{V}$ is orthogonal:

$$\mathbf{V}\mathbf{V}^\top = \mathbf{I} \tag{3}$$

```
» [U,D] = eig(C);
» V = U';
» y = V*x;
```

Modelling EEG as superposition of independent components

Assumptions

- The EEG signal is composed of a number of independent neuronal sources.
- Each source i produces a scalar source signal $s_i(t)$.
- Each source can be measured at different places on the scalp with different intensity, i.e. each source has a distinct scalp map, represented by the vector $\mathbf{a}_i$.
- The source signals are mapped linearly and instantaneously to the scalp by multiplication of the source signal with its scalp map:

$$\mathbf{x}_i(t) = \mathbf{a}_i s_i(t) \tag{4}$$

- The voltage measured at each electrode is the sum of the contributions of all sources:

$$\mathbf{x}(t) = \sum_i \mathbf{x}_i(t) = \sum_i \mathbf{a}_i s_i(t) \tag{5}$$

The Independent Component Analysis (ICA) model

The ICA model can also be written in matrix notation:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{6}$$

where the field patterns $\mathbf{a}_i$ are the columns of the matrix $\mathbf{A}$ and the components of the vector $\mathbf{s}(t)$ are the source signals $s_i(t)$.

Or, even more matrix-like:

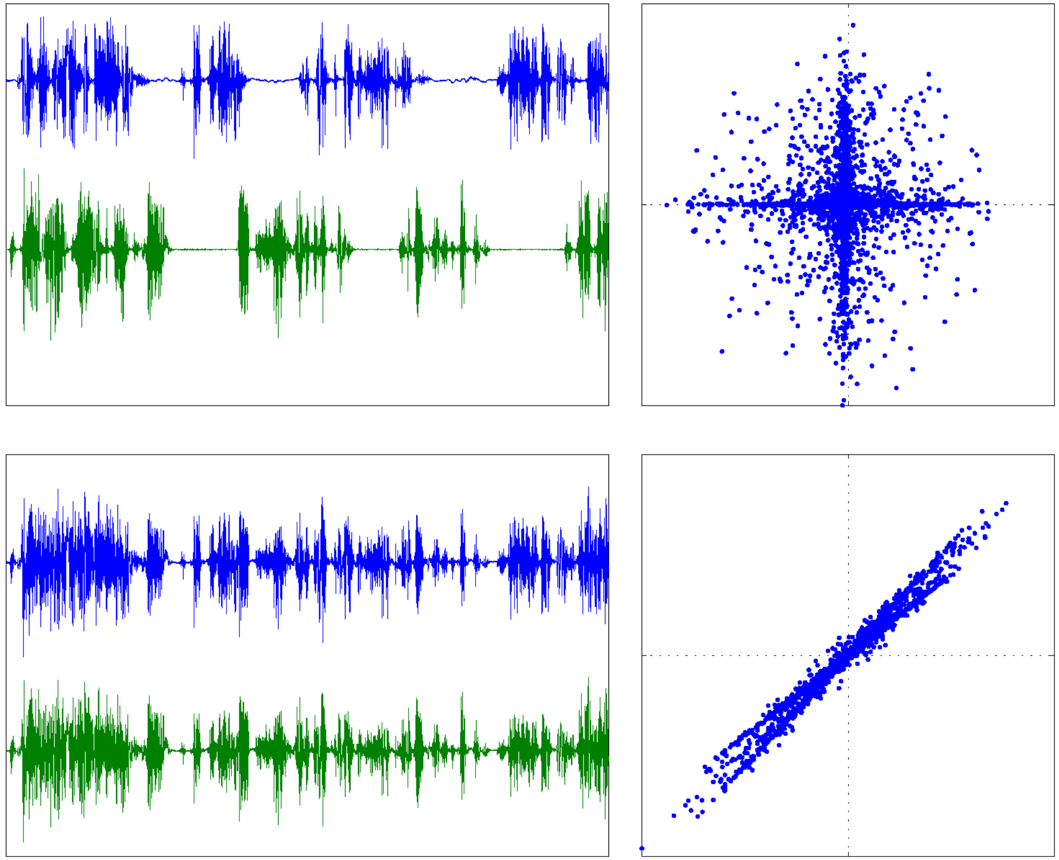
$$\mathbf{X} = \mathbf{A}\mathbf{S} \tag{7}$$

where $\mathbf{X}$ and $\mathbf{S}$ are $N \times T$-matrices (each row is a time series from one electrode).

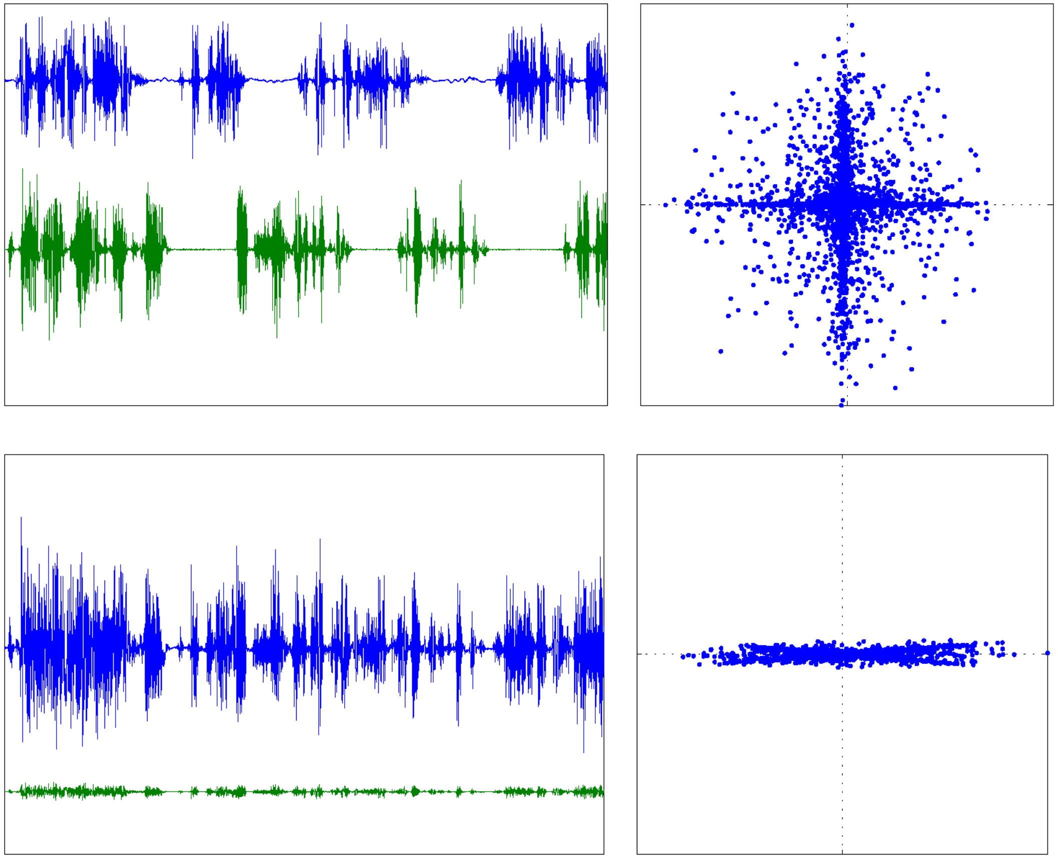
Task:

Estimate both the field pattern matrix/mixing matrix $\mathbf{A}$ and the source signals $\mathbf{s}(t)$ given only the observed EEG signal $\mathbf{x}(t)$ while assuming independence of the sources.

Why not just using PCA?



Why not just using PCA?



Recap: Cleaning fMRI with PCA

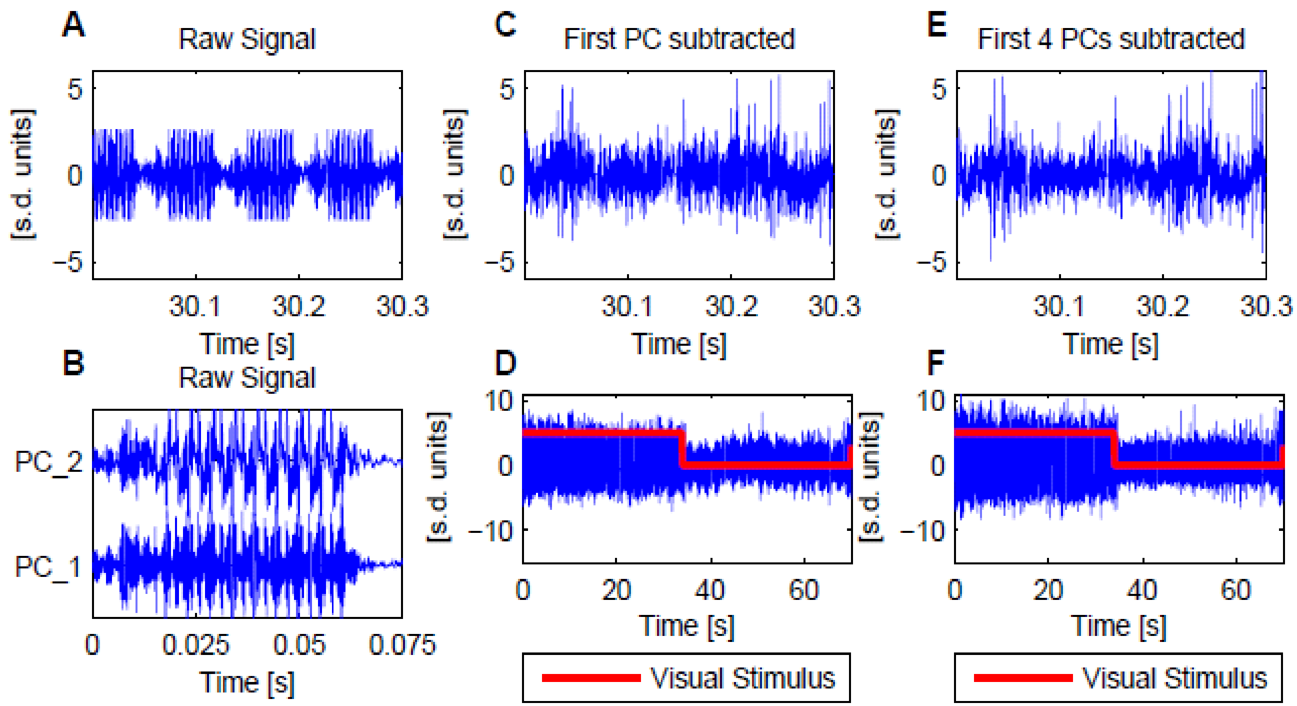
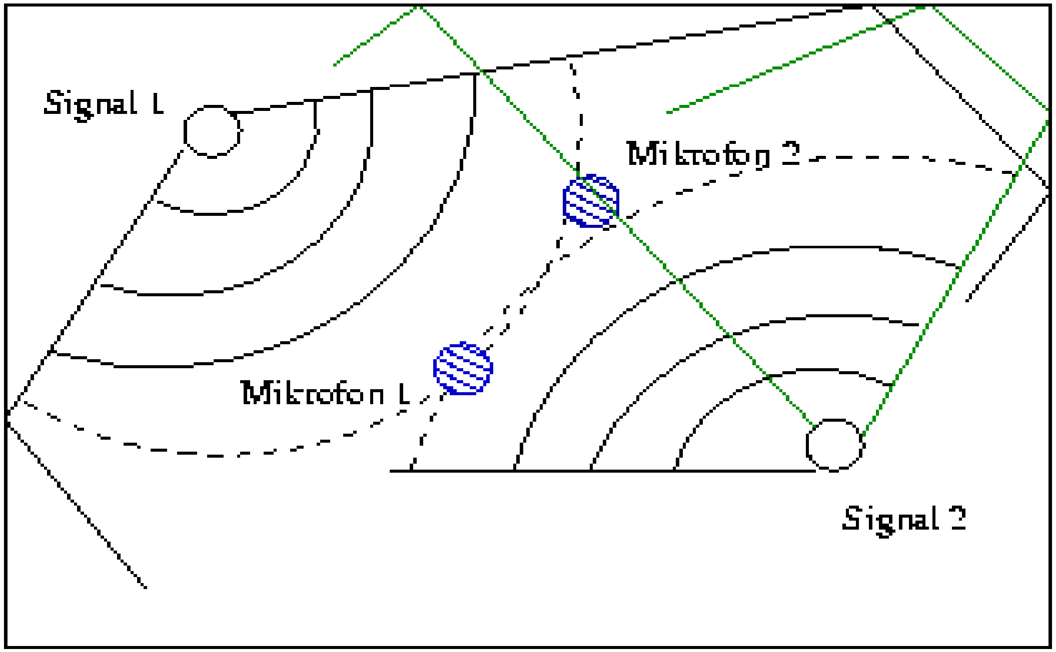


Figure 2: Example of PCA based artefact removal from neurophysiological recordings recorded during fMRI data acquisition; **A**: example of raw contaminated time series; **B**: first two principal components of an artefact induced by switching magnetic field gradients; **C**: same data as in **A** without the first principal component; **D**: same as **C** but for the length of an entire stimulation period; **E**, **F**: same as **C**, **D** but after subtraction of the first 4 principal components;

Blind Source Separation



applications:
cocktailparty problem, biomedical measurements (EEG, MEG), etc.

question:
decomposition & analysis of **superimposed** signals, robust denoising.

Acoustic Demo: "Cocktail party"

- 3 mixed signals (music, speech, street noise)

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

- problem: **demixing**!

“Blind” Source Separation I

• microphones $\mathbf{x}(t)$ measure **unknown** mixtures of **unknown** (sound) sources

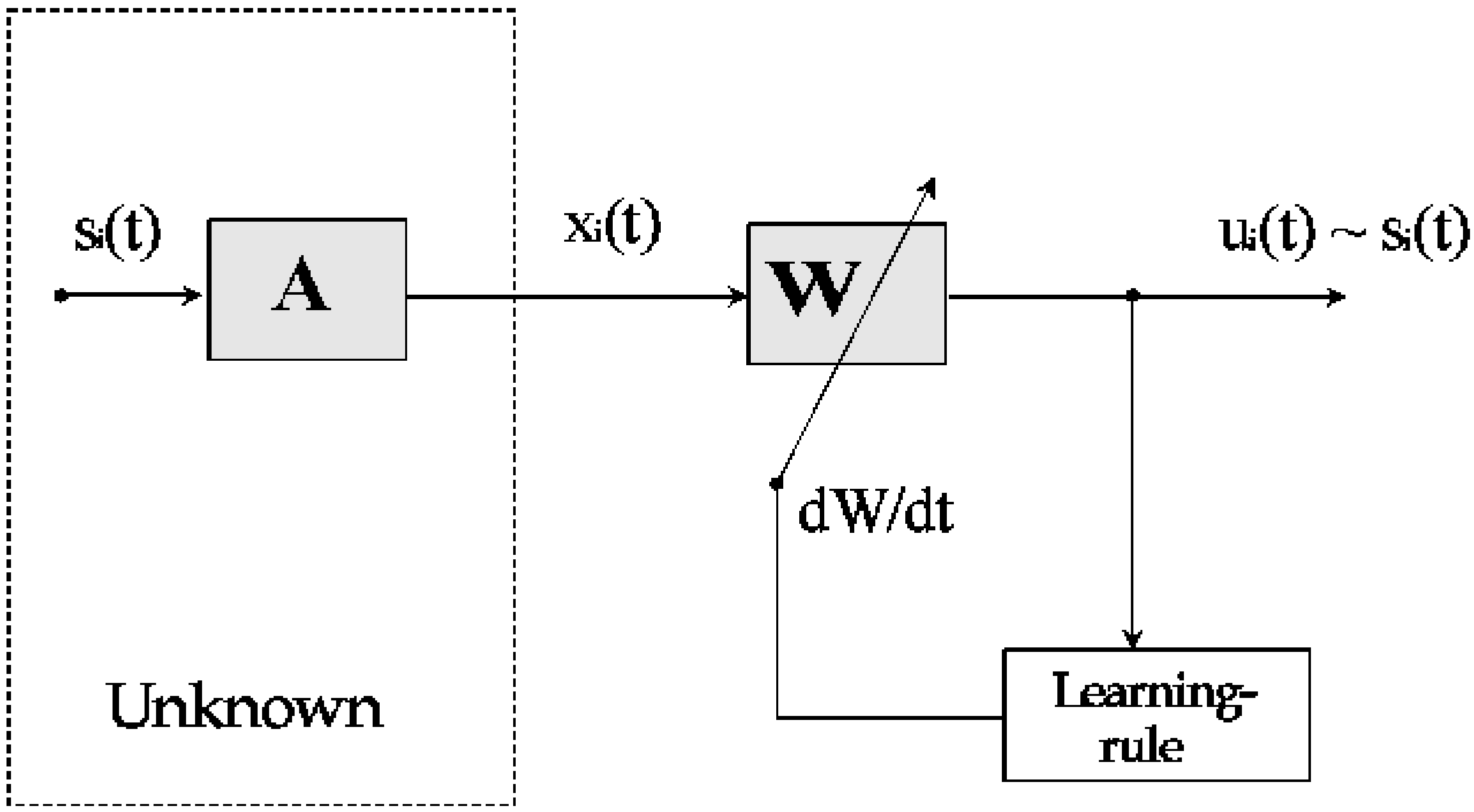
$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$$

• **assumption**: statistical independence of the source signals (**ICA**)

• **Ansatz**: invert mixing process **A** by **learning** of **W** and enforce statistical independence of unmixed signals $\mathbf{u}(t)$!

$$\mathbf{u}(t) = \mathbf{W}\,\mathbf{x}(t)$$

"Blind" Source Separation II



"Blind" Source Separation III

- **assumption**: statistical independence of sources

$$p(\boldsymbol{u}) \quad = \quad \prod_{i=1}^{n} p_i(u_i)$$

- higher cross-moments should vanish

- minimize distance between distributions

$$D(\mathbf{W}) = \int p(\boldsymbol{u}) \log \left(\frac{p(\boldsymbol{u})}{\prod_{i=1}^{n} p_i(u_i)} \right) d\boldsymbol{u}$$

“Blind” Source Separation IV

- Gram Chalier expansion

$$p_i(u_i) \sim \frac{1}{\mathcal{N}} e^{-(u_i)^2/2} \left(1 + \frac{m_i^{(3)}}{3!} H_3(u_i) + \frac{[m_i^{(4)} - 3]}{4!} H_4(u_i) + \ldots \right)$$

- Edgeworth expansion:

$$\begin{aligned} p_i(u_i) \quad &\sim \quad \frac{1}{\mathcal{N}} e^{-(u_i)^2/2} (1 + \frac{m_i^{(3)}}{3} H_3(u_i) + \frac{m_i^{(4)}}{4} H_4(u_i) + \\ &+ \quad \frac{10}{6} (m_i^{(3)})^2 H_6(u_i) + \frac{1}{5} m_i^{(5)} H_5(u_i) + \frac{35}{8} m_i^{(3)} m_i^{(4)} H_7(u_i) \\ &+ \quad \ldots) \end{aligned}$$

where $m_i^{(k)}$ is kth order moment of u_i and $H_k(u_i)$ are Chebyshev-Hermite polynomials (order k).

“Blind” Source Separation V

- after tedious but straight forward calculation, we get

$$
\begin{aligned}
D(\mathbf{W}) \quad \sim \quad & -\int p(\boldsymbol{x}) \log(p(\boldsymbol{x})) - \log \|\det(\mathbf{W})\| + \frac{n}{2}\log(2\pi e) + \dots \\
& -\sum_{i=1}^{n}\Big[\frac{(m_i^{(3)})^2}{2\cdot 3!} + \frac{[m_i^{(4)}-3]^2}{2\cdot 4!} - \frac{5}{8}(m_i^{(3)})^2[m_i^{(4)}-3] + \dots \\
& -\frac{1}{16}[m_i^{(4)}-3]^3\Big]
\end{aligned}
$$

$$
\frac{d\mathbf{W}}{dt} = \eta(t)\{\mathbf{I} - \mathbf{f}(\boldsymbol{u})\boldsymbol{u}^T\}\mathbf{W} \qquad \text{(e.g. Amari et al. 96)}
$$

$$
f(u) = 3/4u^{11} + 25/4u^9 - 47/4u^5 + 29/4u^3.
$$

"Blind" Source Separation with Temporal Information

- **model:**

  x(*t*) = **A** **s**(*t*), **u**(*t*) = **W** **x**(*t*)

- **define** covariance matrices over time:

$$\mathbf{V} = \langle \boldsymbol{x}_t \boldsymbol{x}_t^T \rangle \qquad \mathbf{V}_\tau = \langle \boldsymbol{x}_t \boldsymbol{x}_{t-\tau}^T \rangle \qquad \forall\, i \neq j,$$

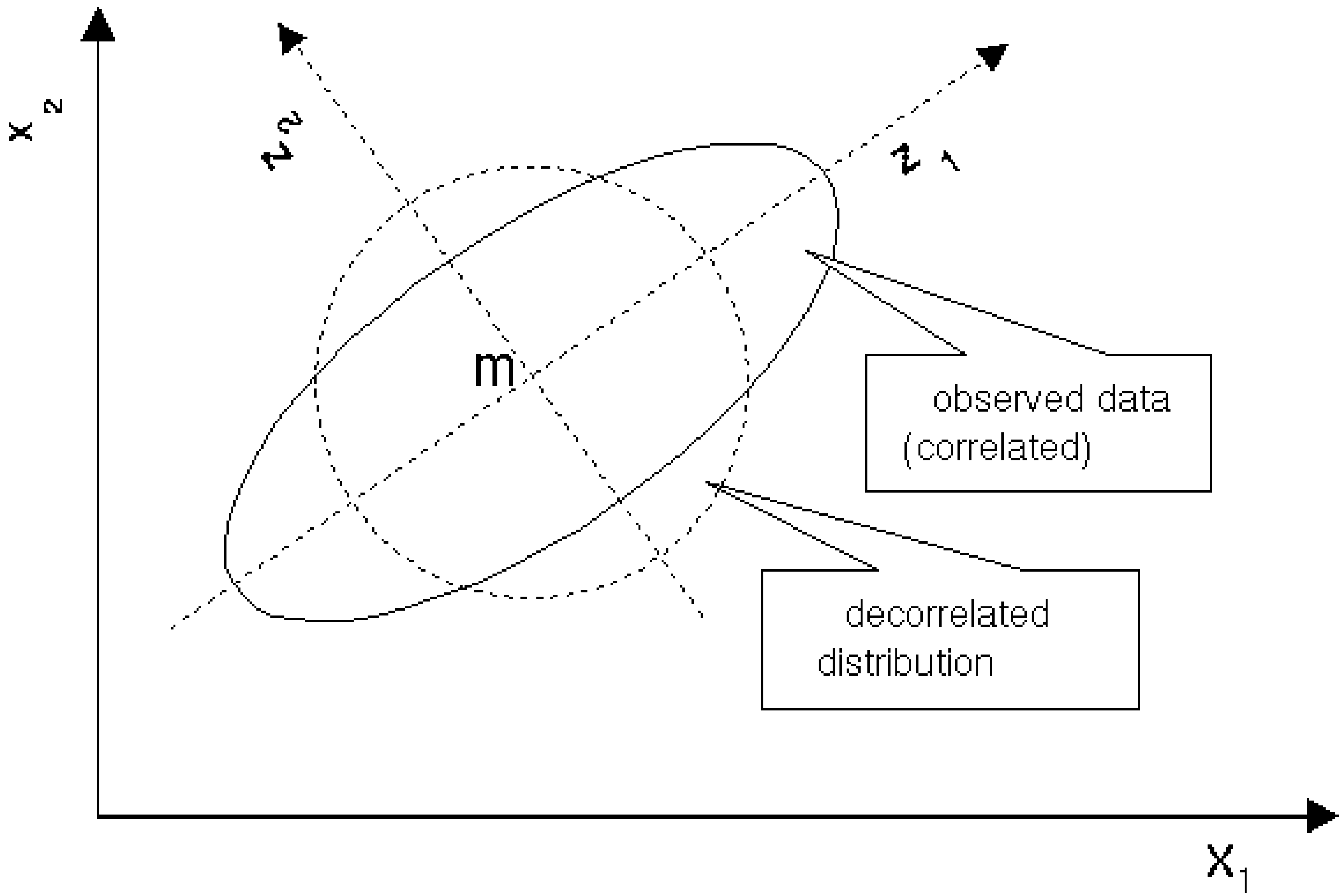
- **assumption**: s has significant autocorrelation

- **algorithm**: **TDSEP** minimizes error

$$L\{\mathbf{W}\} = \sum_{i \neq j} \langle u_i(t) u_j(t) \rangle^2 + \sum_{\{\tau\}} \langle u_i(t) u_j(t-\tau) \rangle^2$$

- **solution**: linear algebra vs. gradient descent
- simultaneous diagonalisation of $\{\mathbf{V},\ \mathbf{V}_\tau, \ldots\}$

Whitening and Jacobi Rotations I



Whitening and Jacobi Rotations II

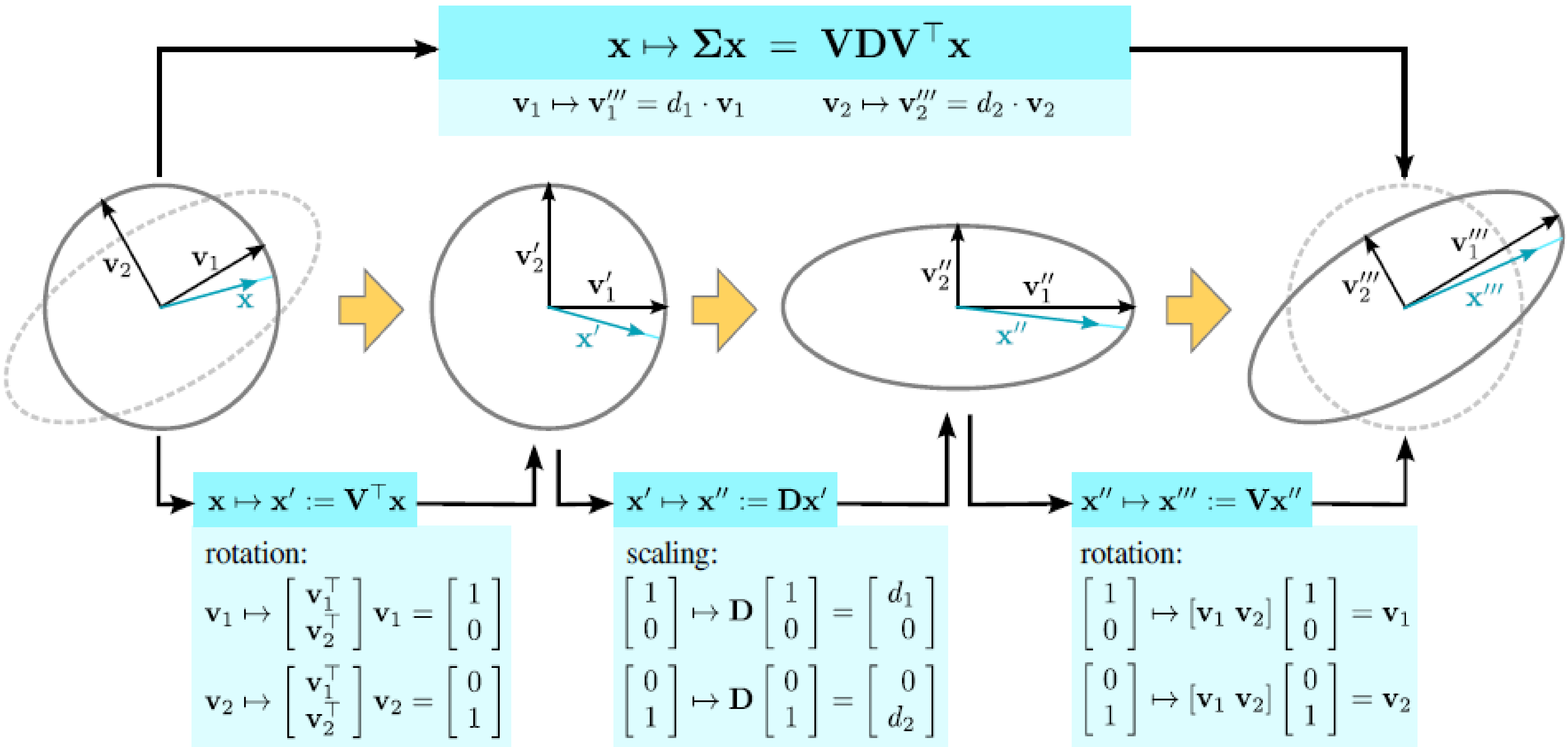
- whitening transformation **K** is e.g. determined as inverse square root of the covariance matrix

$$\mathbf{K} = \langle \boldsymbol{x}\boldsymbol{x}^T \rangle^{-\frac{1}{2}} = (\boldsymbol{v}\,\Lambda\,\boldsymbol{v}^T)^{-\frac{1}{2}} = \boldsymbol{v}\,\Lambda^{-\frac{1}{2}}\,\boldsymbol{v}^T.$$

- then approximative simultaneous diagonalisation of transformed time-delayed covariance matrix

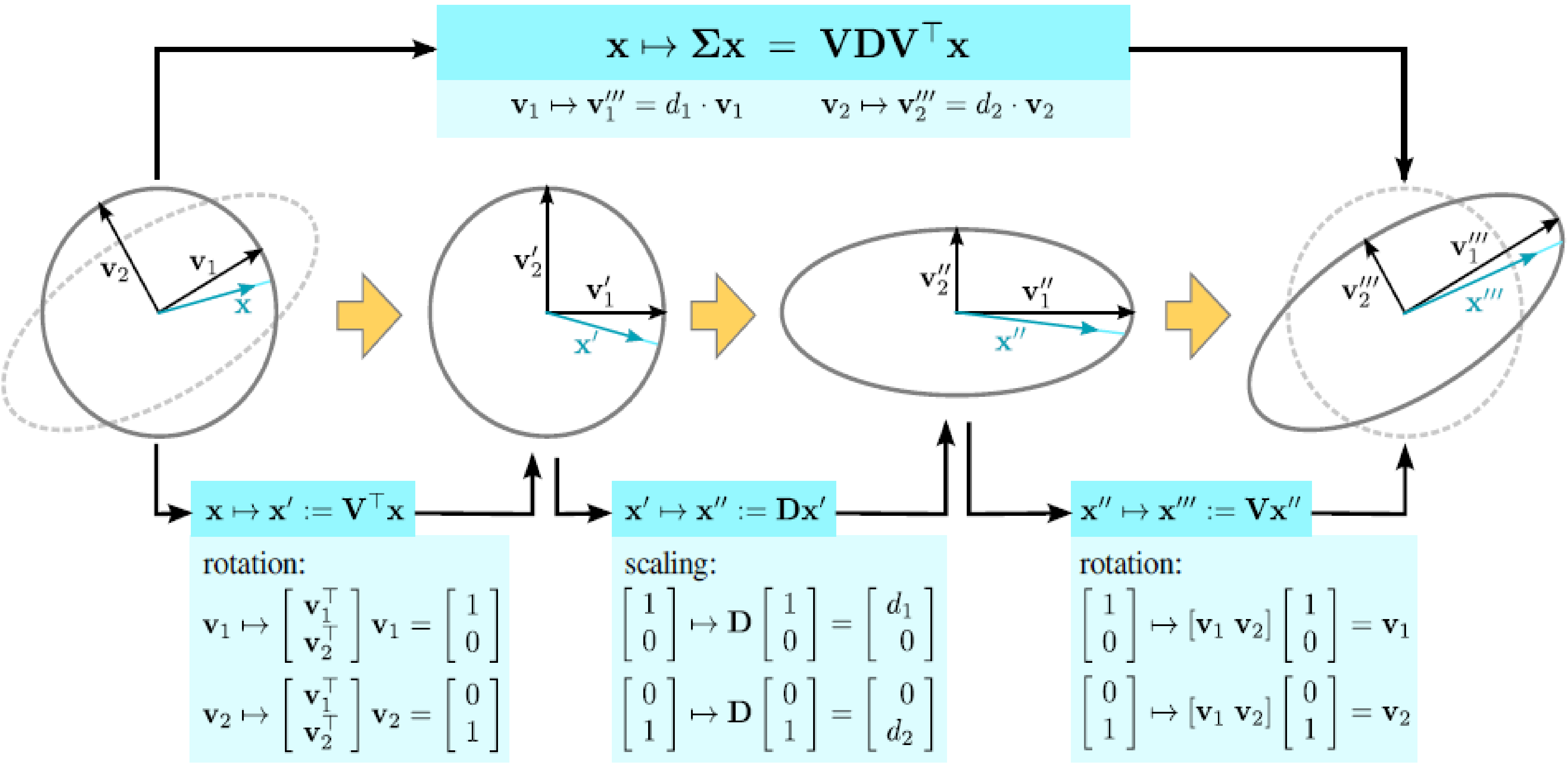
$$\mathbf{V}_{\tau(z)} \;=\; \langle z_t z_{t-\tau}^T \rangle = \boldsymbol{Q}^T\,\mathbf{V}_{\tau(s)}\,\boldsymbol{Q} = \boldsymbol{Q}^T \Lambda_\tau\,\boldsymbol{Q}.$$

- **solution**: $\mathbf{A} = \mathbf{K}^{-1}\boldsymbol{Q}$

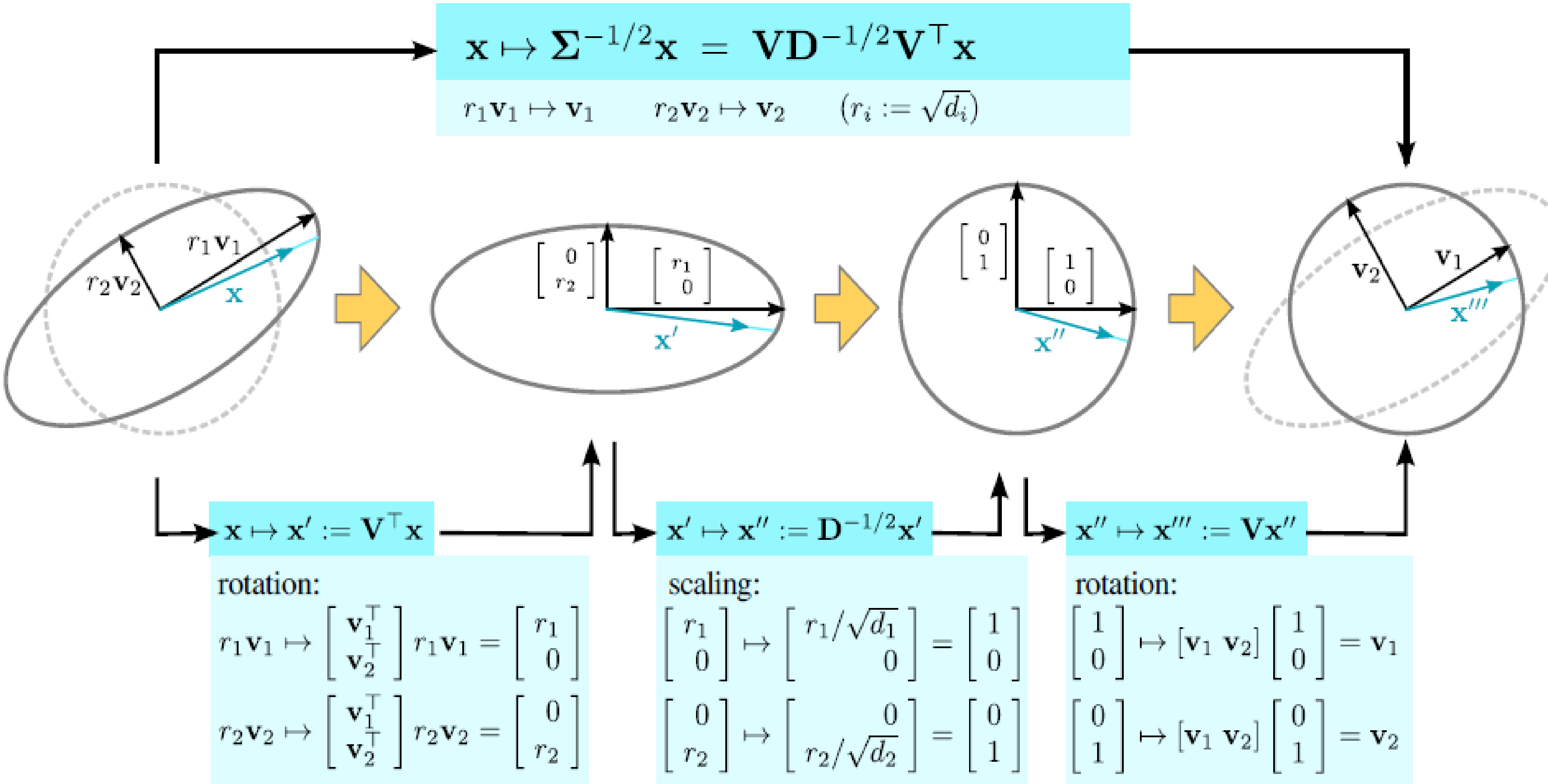


This Figure shows that the multiplication is transformation of the space which maps the unit sphere to an ellipsoid, which is defined by the covariance matrix. (But note that the radii here are defined by the Eigenvalues, not by the square root of the Eigenvalues as on the last slide. In other words, it is a scaling along the principle axes of the ellipsoid defined by Σ.

BBCI berlin brain computer interface



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



The whitening transform maps the space such that a Gaussian distribution with the given covariance matrix becomes a stanard 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.

How to enforce statistical independence?

- **model:** $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$ $\mathbf{u}(t) = \mathbf{W}\,\mathbf{x}(t)$
- higher order statistics (expansions)

$$\frac{d\mathbf{W}}{dt} = \eta(t)\{\mathbf{I} - \mathbf{f}(\boldsymbol{u})\boldsymbol{u}^T\}\mathbf{W} \qquad \text{(e.g. Amari et al. 96)}$$

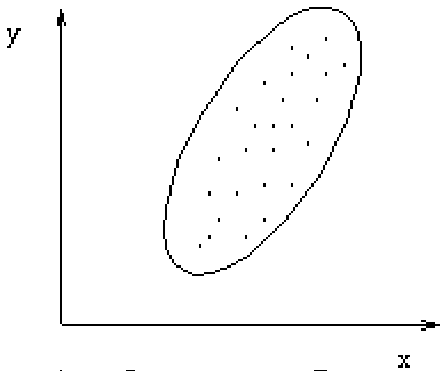
$$f(u) = 3/4u^{11} + 25/4u^9 - 47/4u^5 + 29/4u^3.$$

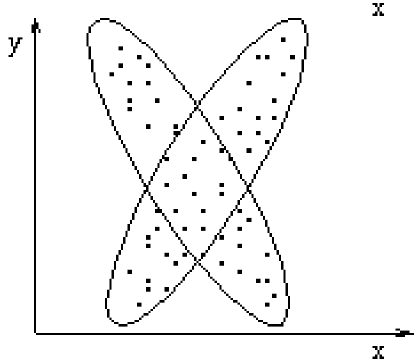
- second order statistics & **temporal information**

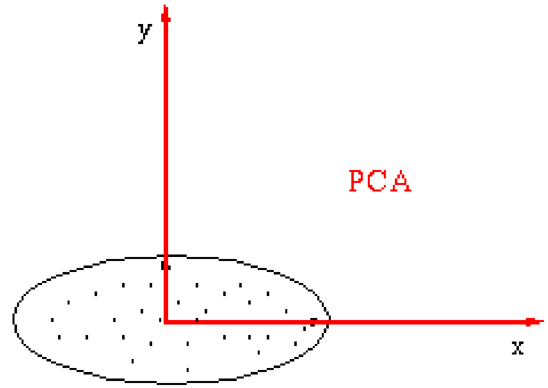
$$L\{\mathbf{W}\} = \sum_{i \neq j} \langle u_i(t) u_j(t) \rangle^2 + \sum_{\{\tau\}} \langle u_i(t) u_j(t-\tau) \rangle^2$$

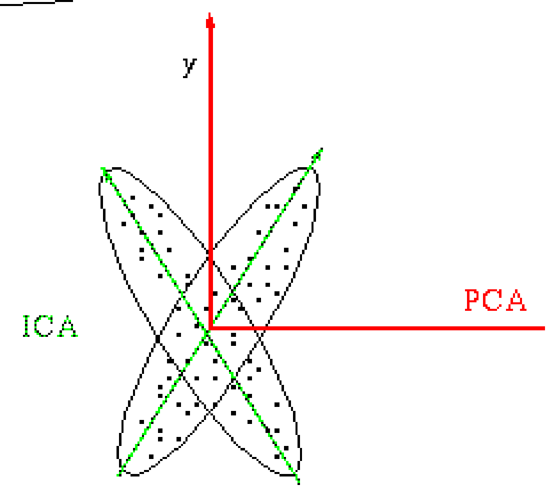
- **(simultaneous diagonalisation of matrices, TDSEP)**

PCA vs ICA









Acoustic Demo II

- 3 mixed signals (music, speech, street noise)

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

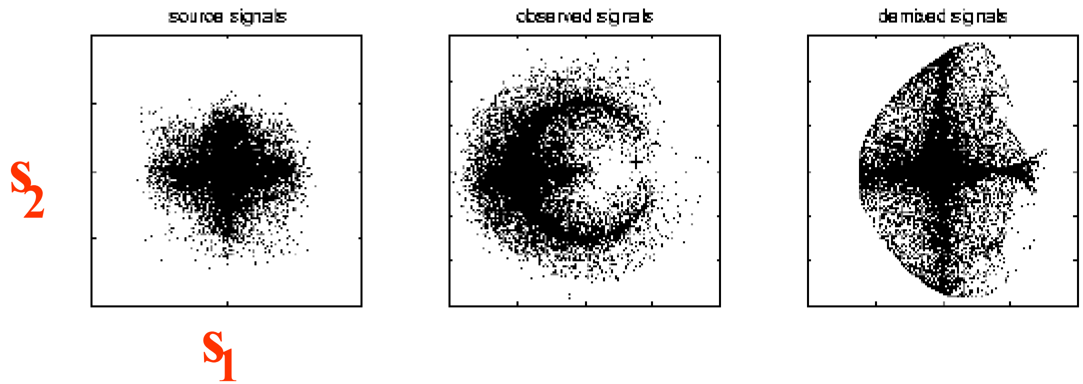
- problem: music signal has **very small** amplitude, i.e. hidden signal

- question: which music instrument?

- *mixed* *unmixed signal*

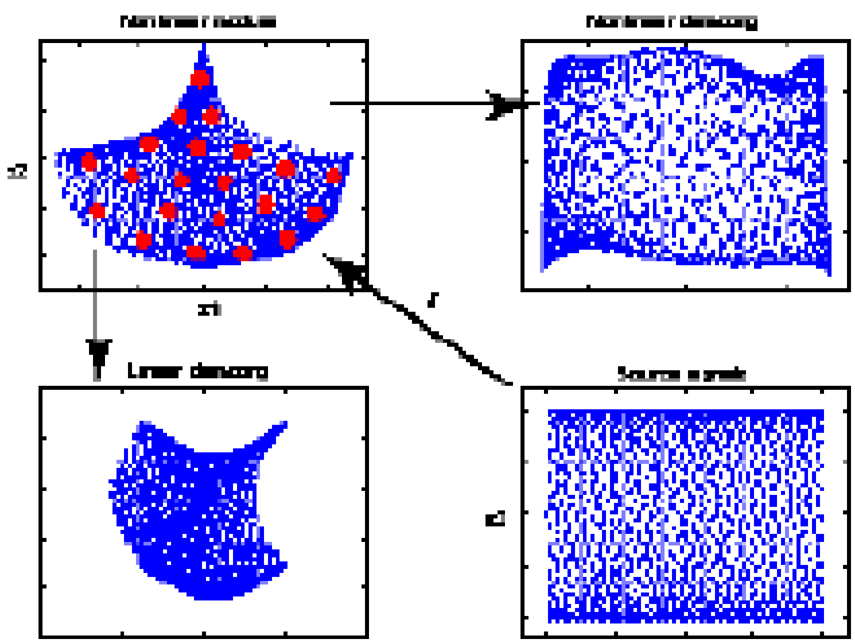
Cf. cerebral cocktail party problem

Nonlinear source separation



$$\mathbf{x}[t] = f(\mathbf{s}[t])$$

BSS of nonlinearly distorted mixtures with kernel based learning methods [Harmeling et al. 2001, 2002, Ziehe et al. 2001]



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

Reliability assessment

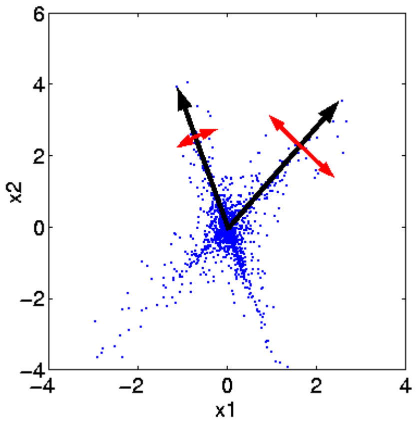
Unsupervised learning techniques like ICA always return an answer/estimate that is found within their model class.

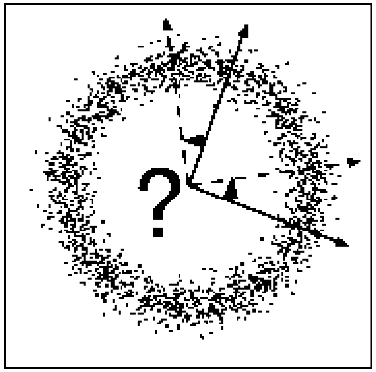
However:

- Is the used model appropriate?
- Can we assess the quality of our separation?
- Can we specify errorbars of our estimates?

→ **Is the result reliable?**

What does Reliability mean?





- How sure is the algorithm about its result?
  - One source could be more reliable than others
- Is the result reproduceable?
- Are there higher dimensional independent components?

Resampling approach

Produce surrogate data sets that can be written as mixtures of independent sources with the same mixing matrix *A.*

observed data

$$\{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)\} \quad \longrightarrow \quad \hat{A}$$

$$\Downarrow$$

surrogate data

$$\{\boldsymbol{x}^{*1}(1), \ldots, \boldsymbol{x}^{*1}(T)\} \quad \longrightarrow \quad \hat{A}^{*1}$$

$$\vdots$$

$$\{\boldsymbol{x}^{*k}(1), \ldots, \boldsymbol{x}^{*k}(T)\} \quad \longrightarrow \quad \hat{A}^{*k}$$

Reliability assessment

1. Do blind source separation with some ICA algorithm.
   $Y = \hat{A}^{-1} X$
2. Produce surrogate data from Y, whiten these data sets.
3. For each surrogate data set: Do BSS.
   This produces a set of rotation matrices.
4. Decompose rotation into rotation angles via matrix logarithm $\alpha = ln(R)$
5. Standard deviations of the rotation angles define a separability matrix

The separability matrix

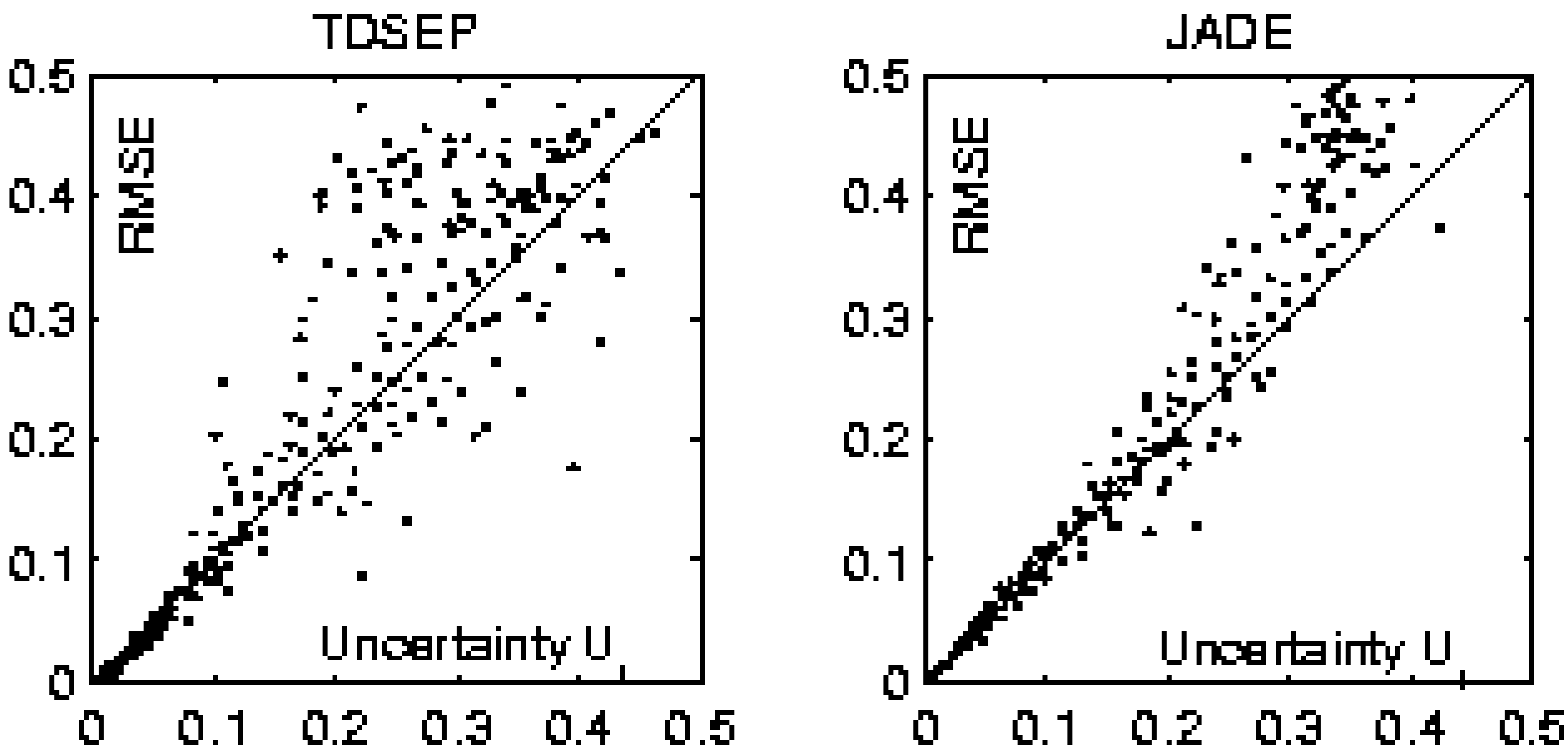
$$S_{ij} = \sqrt{\langle \alpha_{ij}^2 \rangle}$$

measures, how unstable the estimated mixing matrix is w.r.t. a rotation in the plane spanned by the estimated components i and j

$$U_i = \max_j S_{ij}$$

uncertainty of the estimated projection direction i, approximates RMSE

RMSE vs. Uncertainty



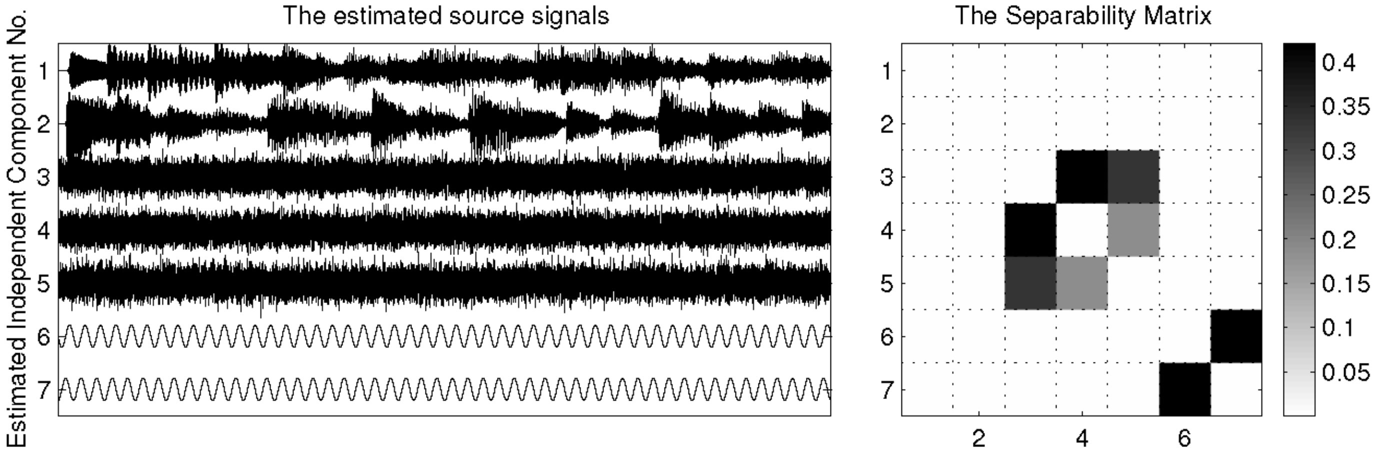
Experimental Results: The (real) RMSE is nicely correlated to the (estimated) uncertainty.

A toy example

- 7 channel mixture of
  - two harmonic oscillations (sin and cos)
  - two speech signals
  - two white Gaussian noise processes
  - one uniformly distributed white noise

- Source separation based on
  - temporal decorrelation (TDSEP)
  - higher-order statistics (JADE)

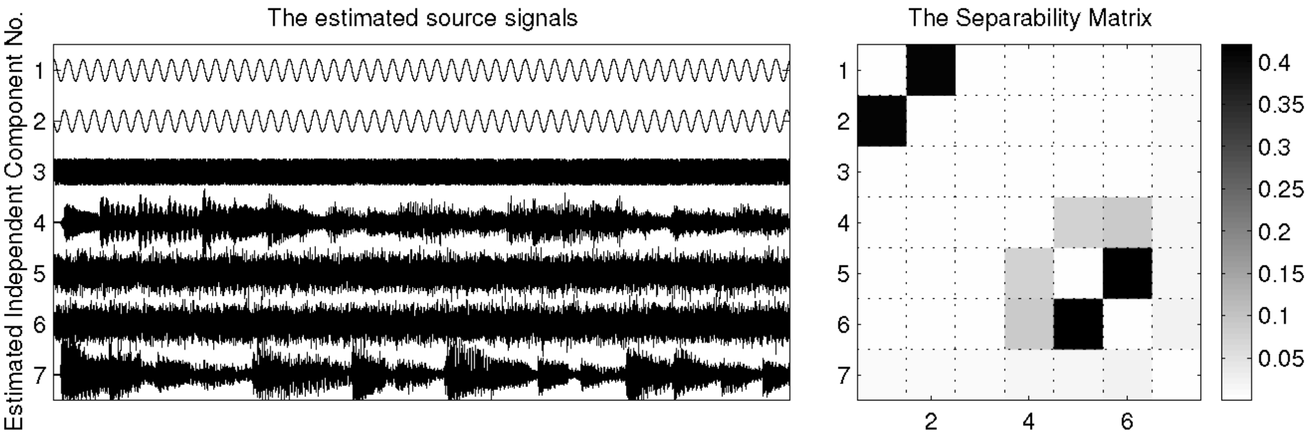
Separability results TDSEP



Separability matrix indicates stable subspaces for

- speech signals (one dimensional)
- sinusodial signals (two dimensional)

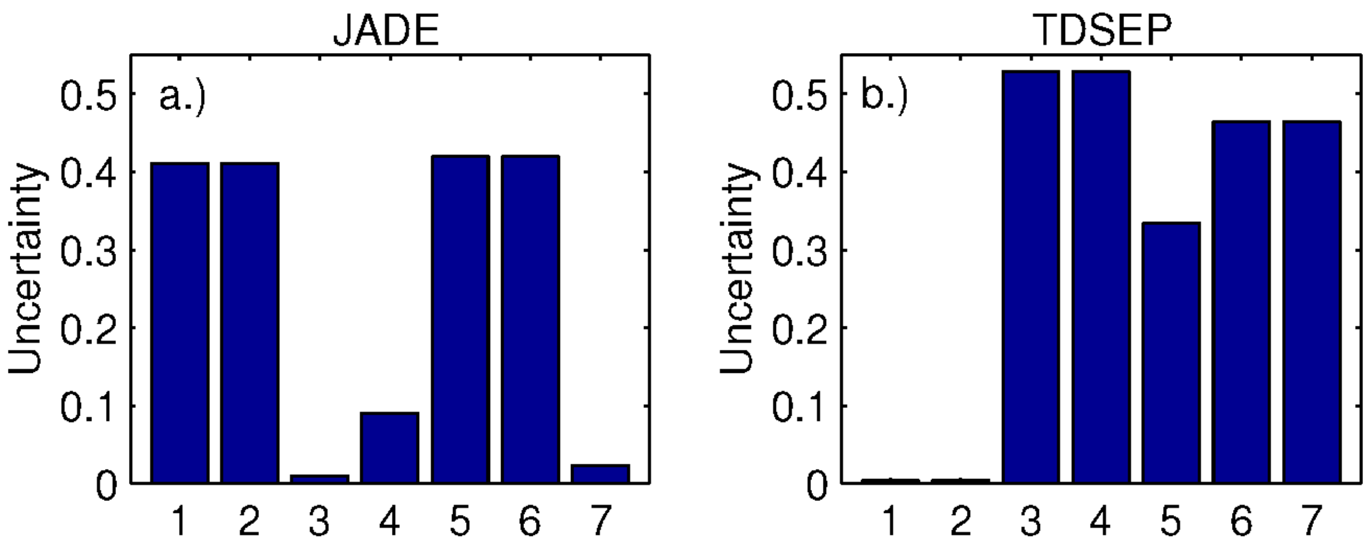
Separability results JADE



Separability matrix indicates stable subspaces for

- sinusodial signals (two dimensional)
- uniform white noise (one dimensional)
- speech signals (one dimensional)

Uncertainties for toy data



- JADE yields reliable estimates for audio sources (4; 7) and non-gaussian random source (3)
- TDSEP only for audio sources (1; 2)

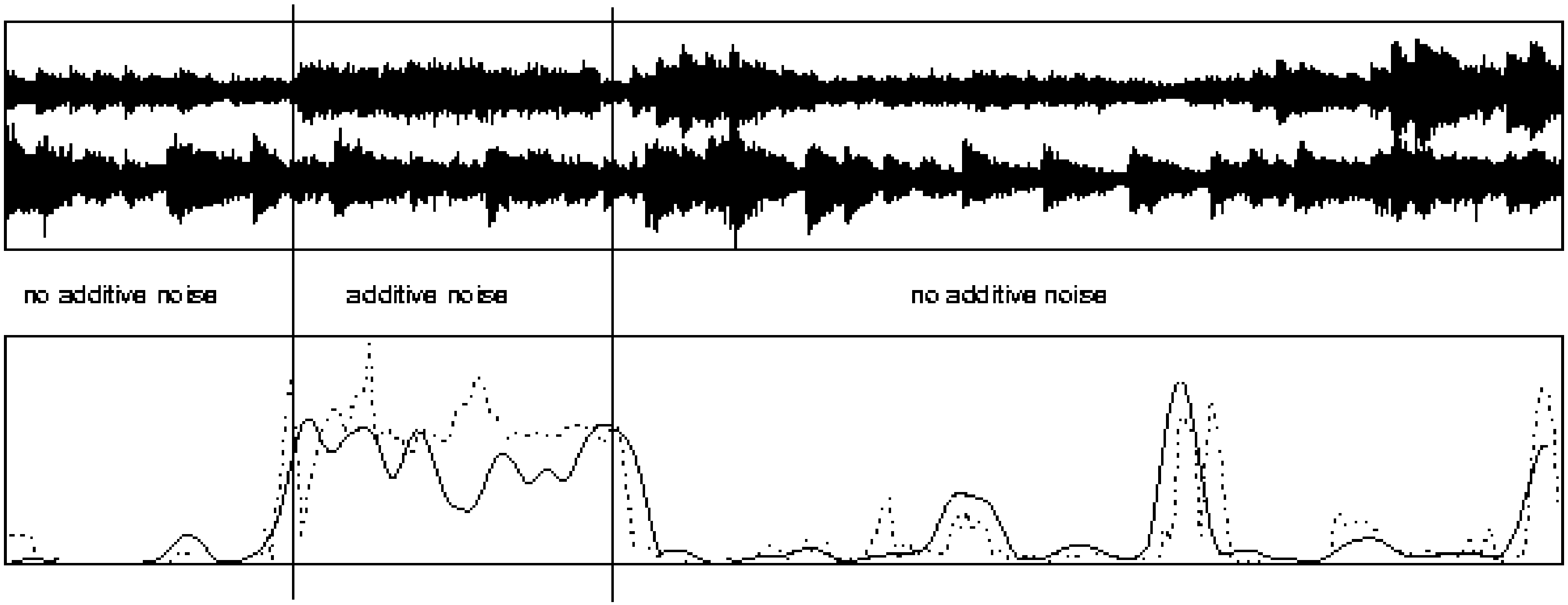
Improving the separation performance

1. Using different models on different sources
   - TDSEP: audio sources and (maybe) the sin/cos – subspace
   - JADE on orthogonal subspace: non-gaussian random source.

2. Using only “good” parts of a time series
   - Moving-window reliability analysis
   - Discard unreliable parts
   - Useful e.g. if noiselevel changes with time

- Tests on artificially generated data show remarkable separation improvement

Moving Window Reliability Analysis

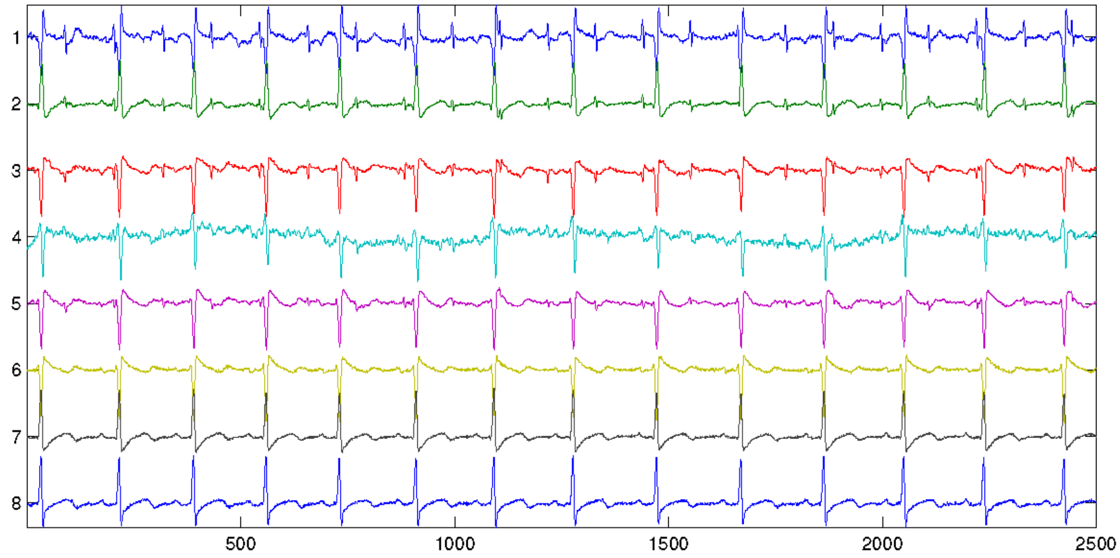


Solid line: Uncertainty, dotted line: Separation error

Testbed: Fetal ECG extraction

- The experimental setup
  - Electrodes located on abdomen and thorax of a pregnant woman
  - ECG is measured at sampling rate of 500Hz
- The data set:
  - 8 channels, 2500 data points
- Data analysis:
  - Source separation with JADE, Reliability analysis

Application: Fetal Electrocardiogram (ECG)

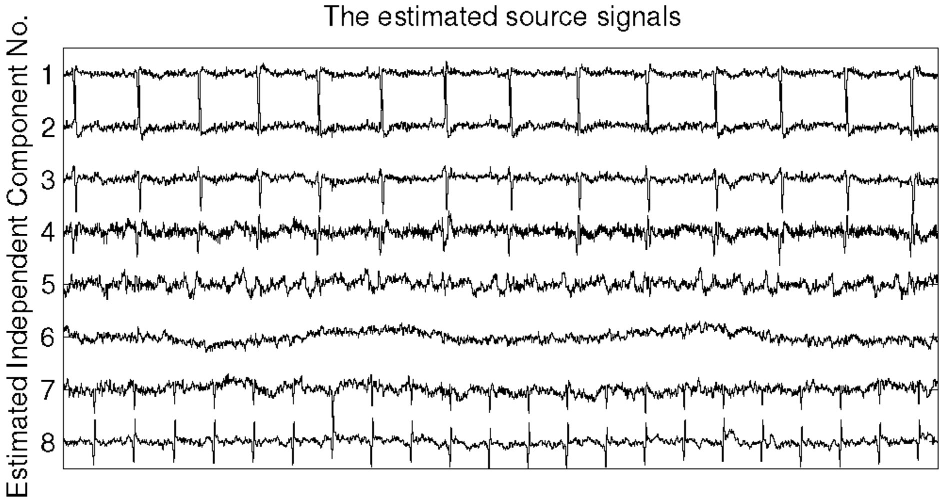


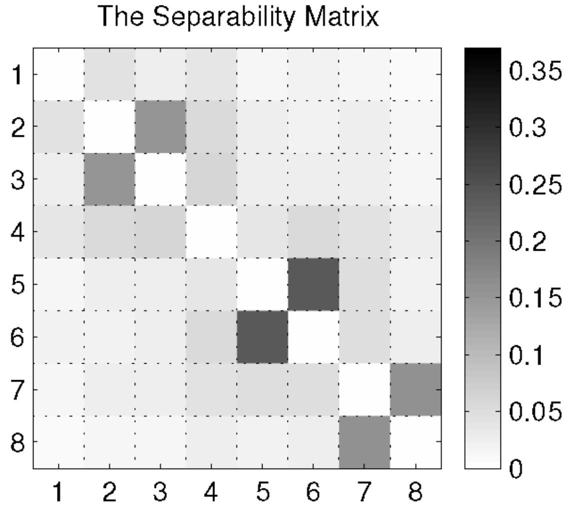
Cutaneous potential recordings of a pregnant woman

(8 channels; 1-5: abdominal; 6,7,8: thoracic)

`ftp.esat.kuleuven.ac.be/pub/SISTA/data/biomedical/foetal_ecg`

Results





- ICA (Jade) decomposition separates cardiac signals of mother and fetus
- Block structure of the separability matrix is of physiological relevance: indicates independent multi-dimensional subspaces

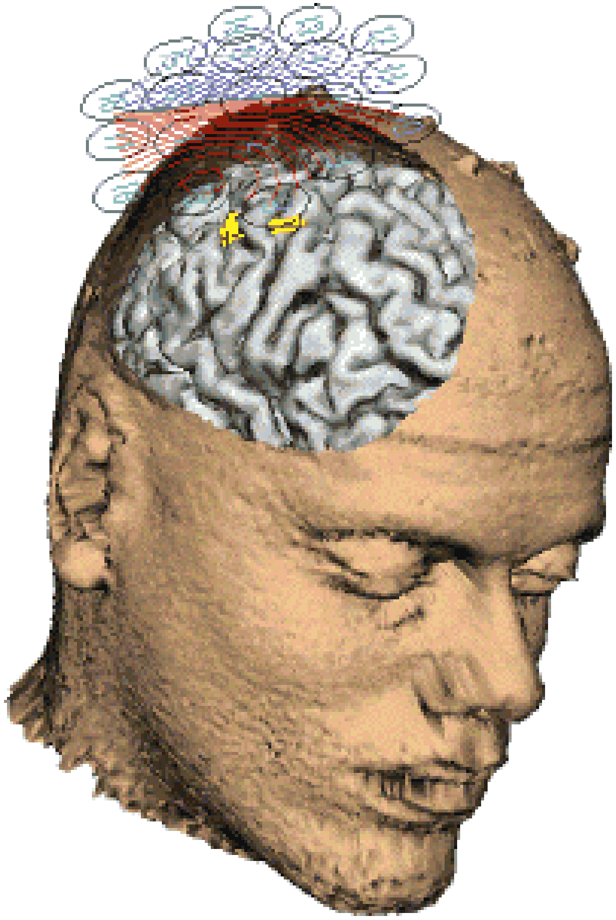
ICA Analysis of Non-invasively Recorded DC-fields in Humans

Klaus-Robert Müller, Andreas Ziehe, Gerd Wübbeler,
Bruno-Marcel Mackert, Lutz Trahms, Gabriel Curio

TUB, AG Neurophysics, Charite Berlin
and PTB Berlin

Cortical Signals

- brain works distributed and parallel
- idea: discriminate "speakers in brain"
- signal processing problem analog to
- Cocktailparty problem



Cortical Signals II

•***GOAL***: identification and extraction of small brain signals despite of noise (external or physiological “noise”, i.e. background activity)

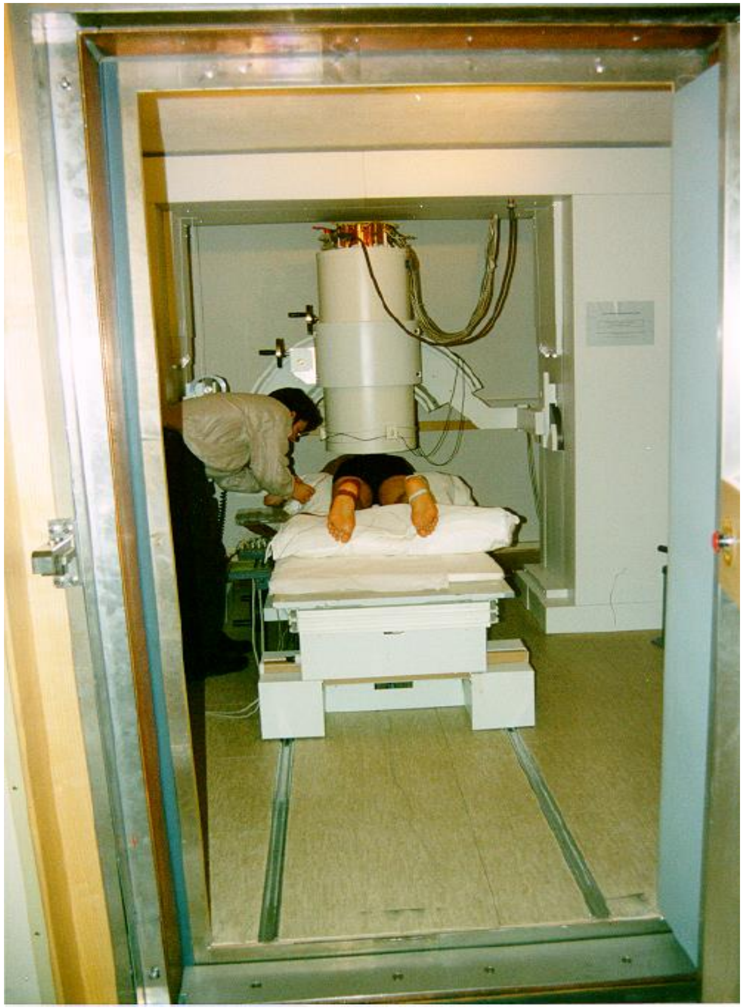
•denoised signals as basis for neurophysiological modeling

•challenge for signal processing, time series prediction and machine learning

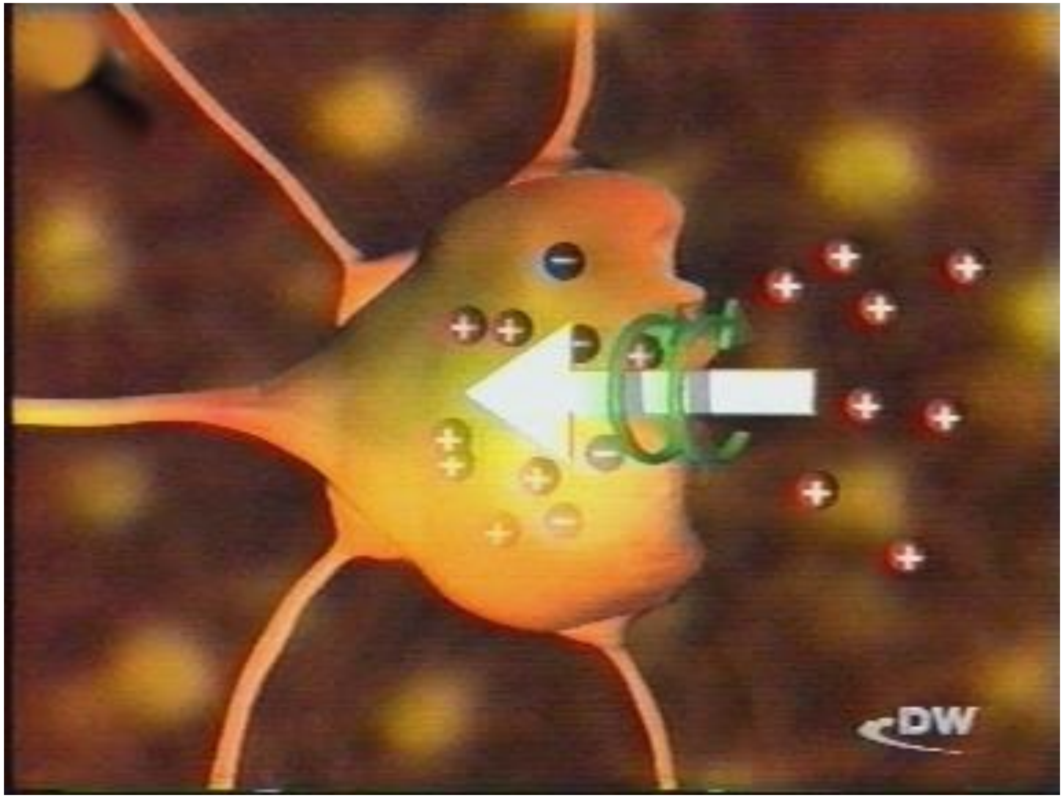
•reliability of the analysis

•relevant signals are often extremely weak compared to the noise, i.e. a factor of 10000!

Setup: shielded MEG chamber



Why do we measure magnetic fields?



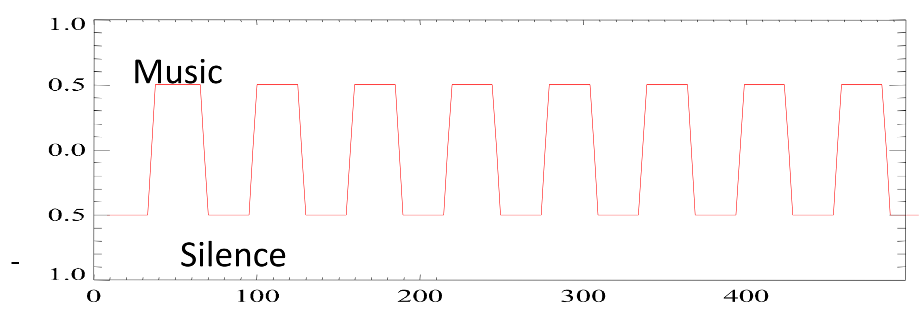
Magnetic fields show brain activity:
single neurons depolarize → synchronous active neuron populations alow a non-invasive monitoring of macroscopic currents.

Setup II

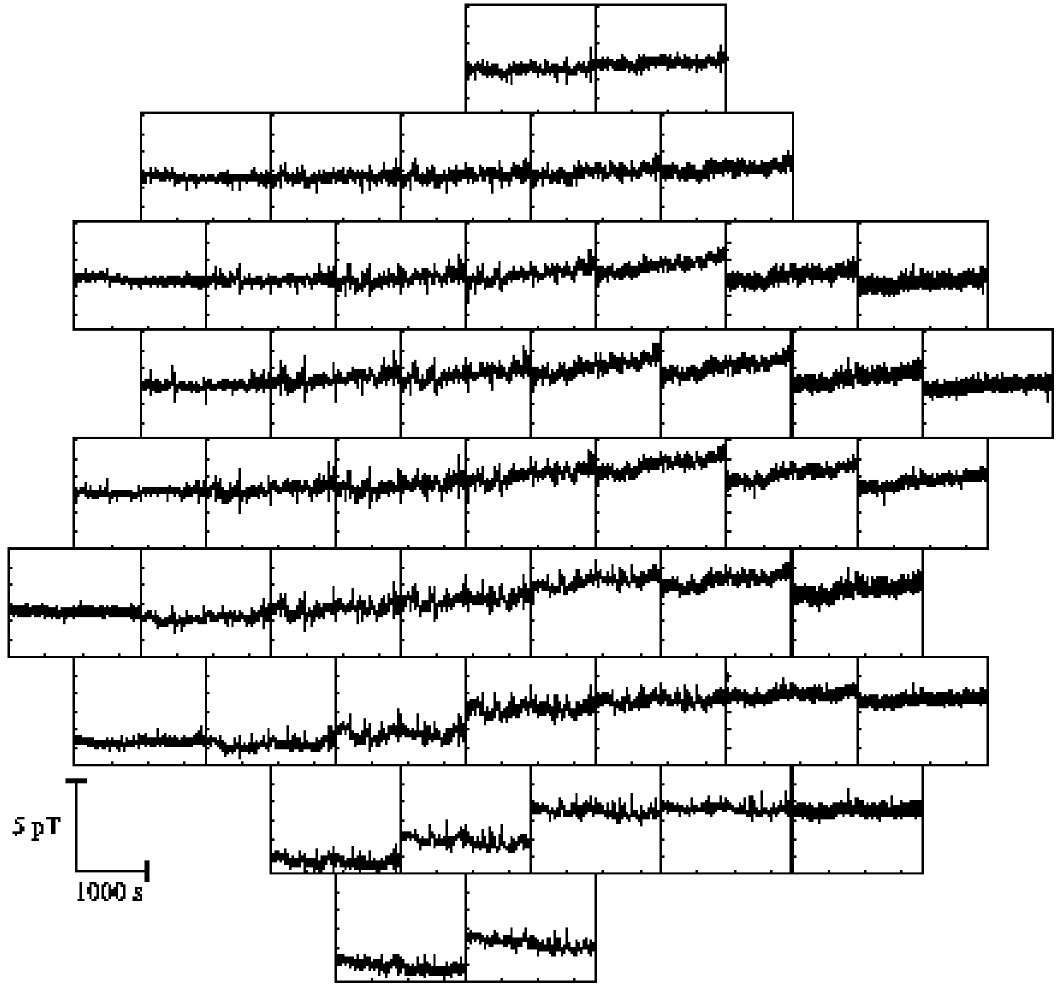
MEG positioning near the auditory cortex

Analysis of DC MEG

- **Paradigm**: acoustic stimulation by presentation of alternating periods of music and silence, each for 30 s;
- Non-invasive measurements of magnetic fields over the left auditory cortex for 30 min with 49 channel SQUID gradiometer
- mechanical horizontal modulation of the body position with a frequency of 0.4 Hz,
- transposes DC magnetic field into higher frequency to improve the signal-to-noise ratio
- **data**: reconstructed DC magnetic fields, sampling with modulation frequency 0.4 Hz → 720 points/channel for 30 minutes

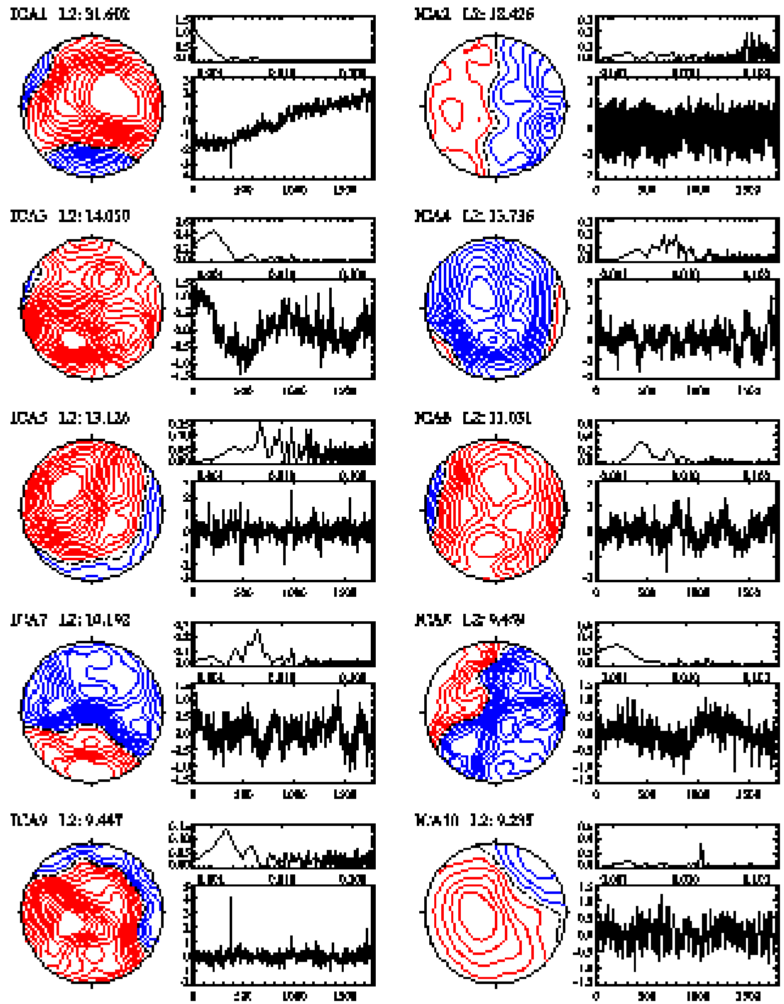


Data

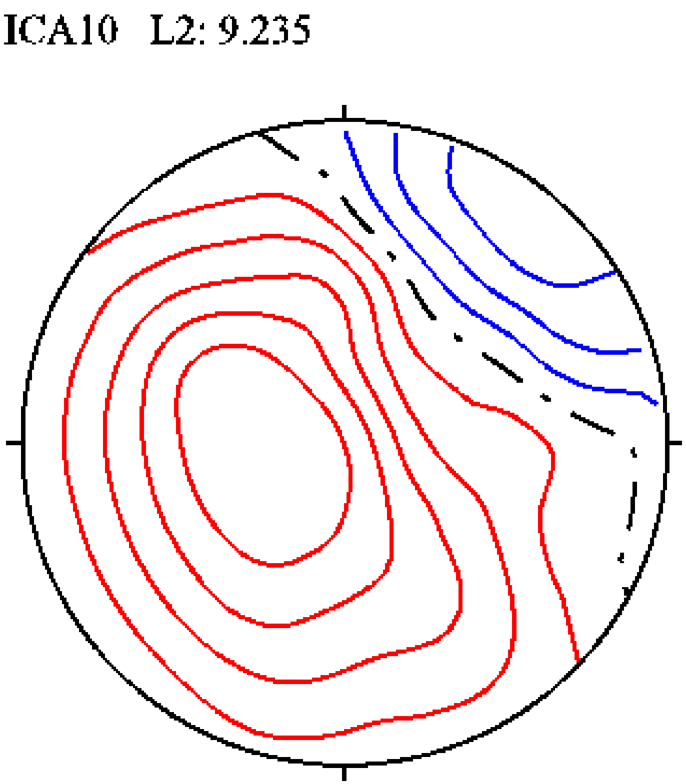


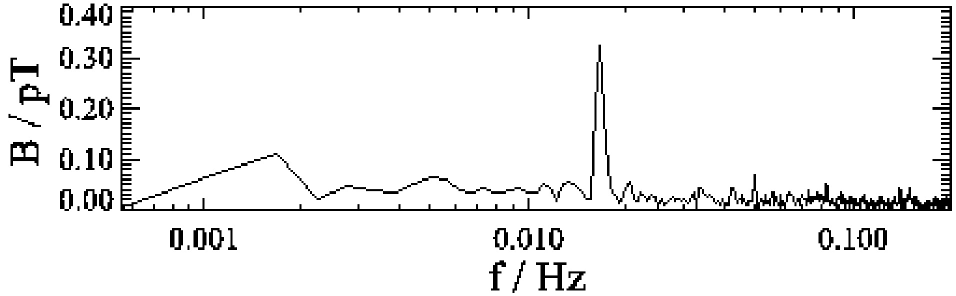
measured data ordered according to sensor position.

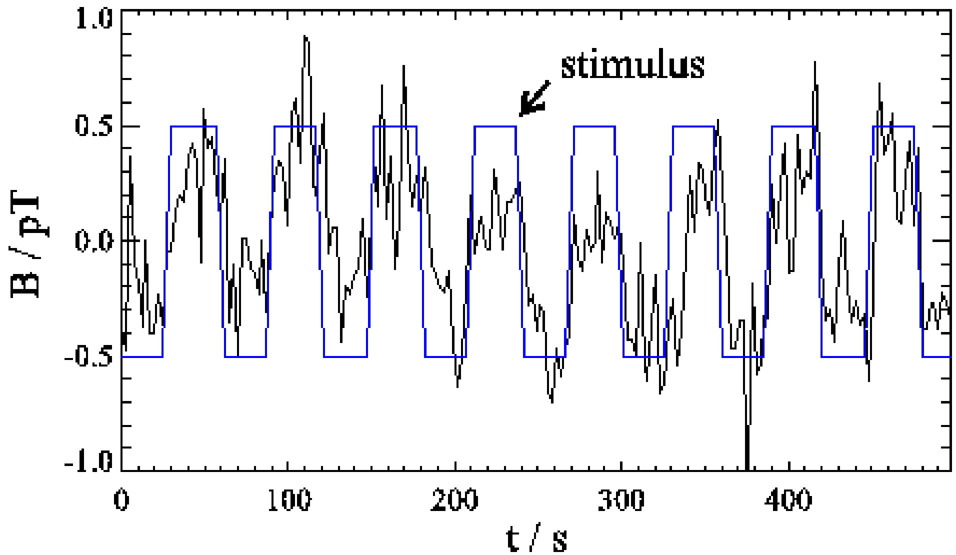
Several ICA Components



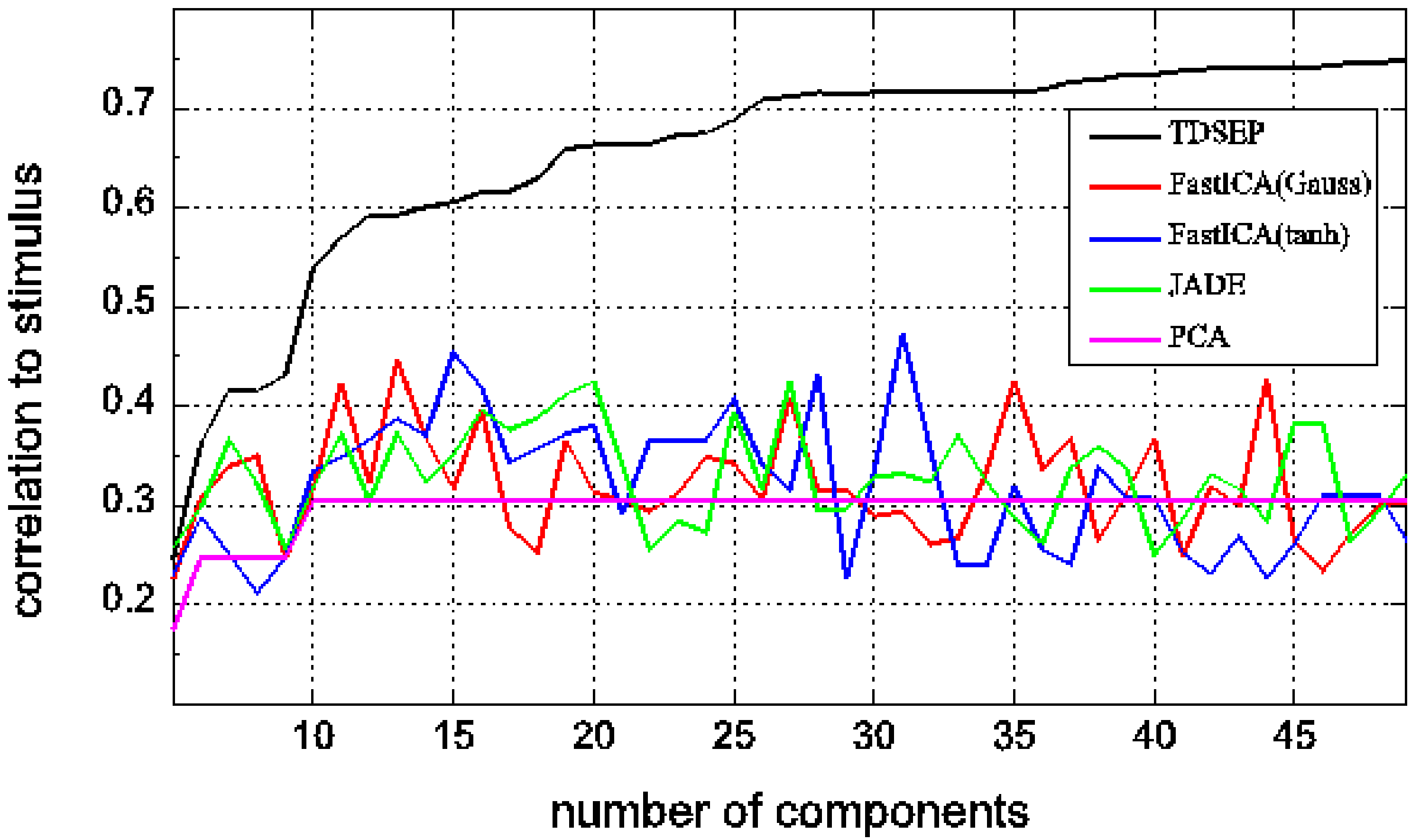
Component 10





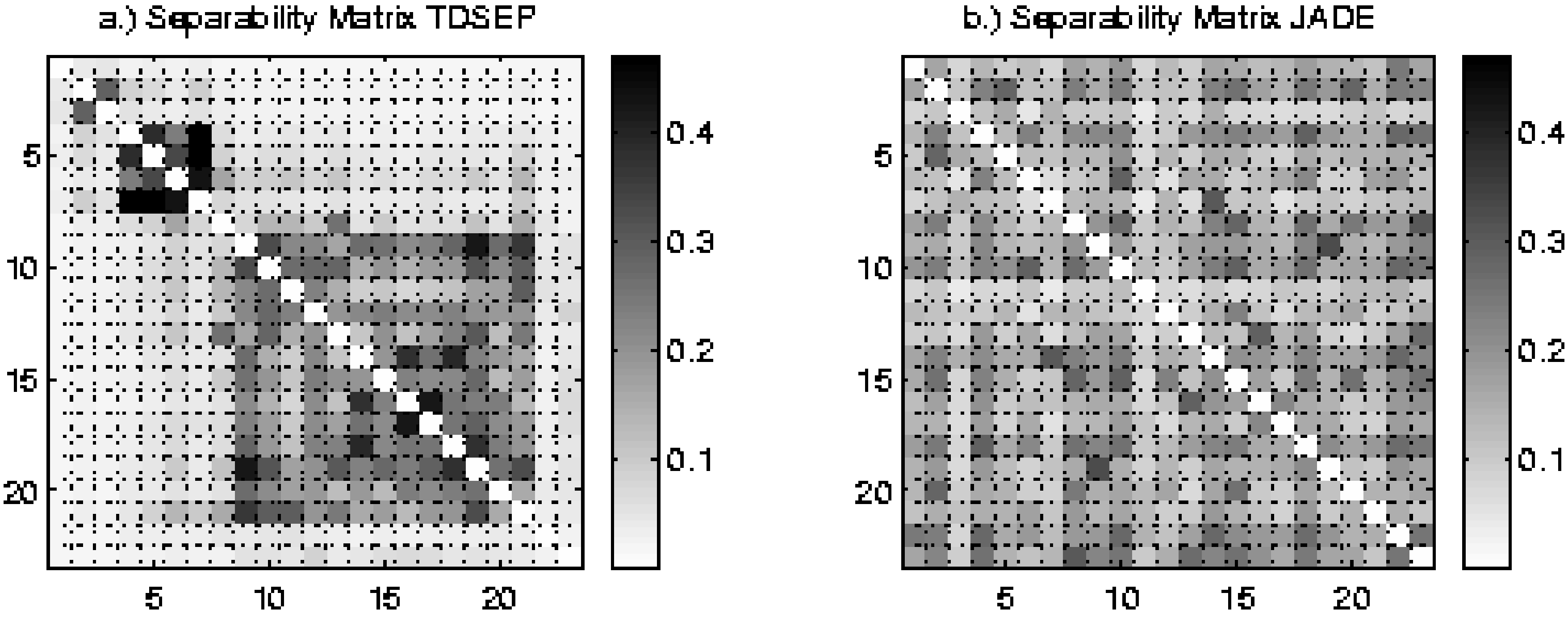


Comparing three Algorithms



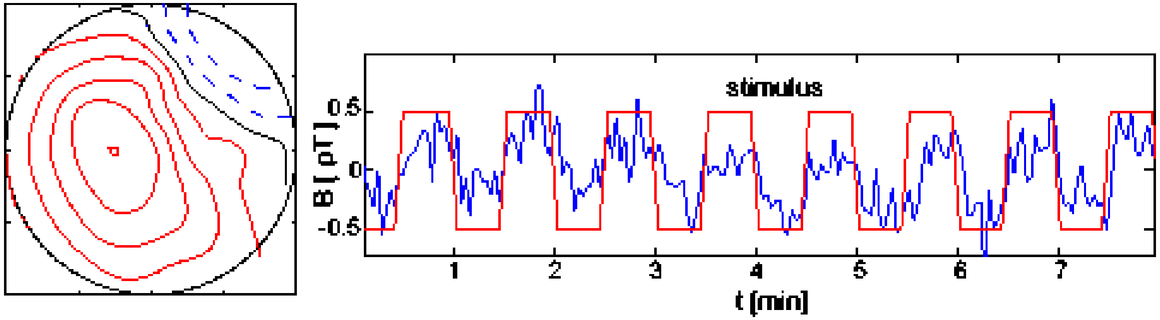
MEG-DC Experiment

- Separability matrices for TDSEP and JADE

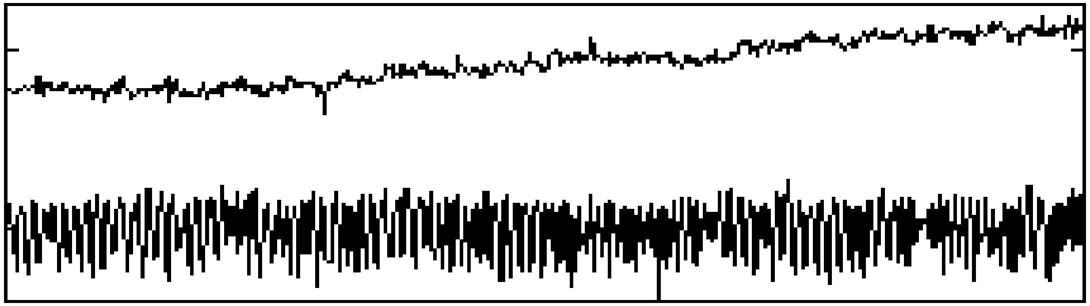


MEG-DC Experiment

- TDSEP 22: field pattern and time course



- TDSEP 1: Drift, TDSEP 23: measurement artifact



Summary

- ICA demixes the data into „minimally statistically dependent“ sources
- Different measures of statistical dependence

  → different algorithms
- Resampling methods can be used to assess the quality of ICA projections

  → identication of the appropriate ICA-Model possible
- Application of ICA often reveals stable subspaces which are physiologically plausible

Shannon-Entropy: an alternative path to ICA

The Shannon-Entropy of a discrete random variable X is defined as

$$H(X) = -\sum P(X = a_i) \log P(X = a_i), \qquad (11)$$

where the a_i denote the possible values of X. The Shannon-Entropy is

- non-negative: $H(X) \geq 0$
- small, if all probabilities $P(X = a_i)$ are close to 0 or 1 and maximal for a uniform distribution $P(X = a_1) = P(X = a_2) = \ldots$ (measure of 'randomness' or information content).

The differential entropy of a continuous random variable x with density $p(x)$ is defined as

$$H(x) = -\int p(\xi) \log p(\xi) d\xi. \qquad (12)$$

(note, that the differential entropy can be negative)

Mutual Information

An ICA-algorithm seeks for a linear invertible transformation $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$ that minimizes the mutual dependence between the components $y_i(t)$. A suitable measure of the dependence is the *mutual information*:

$$I(y_1, y_2, \ldots, y_N) = \sum_i H(y_i) - H(\mathbf{y}) \tag{13}$$

The mutual information is

- non-negative: $I(y_1, y_2, \ldots, y_N) \geq 0$
- zero if and only if the variables are statistically independent.

The mutual information measures the amount of information shared between different random variables.

Entropy of a linear transformation

How does the entropy $H(\mathbf{y})$ of $\mathbf{y} = \mathbf{B}\mathbf{x}$ depend on the transformation $\mathbf{B}$?
We know that the corresponding densities transform by

$$p_y(\mathbf{y}) = p_x(\mathbf{x}) |\det \mathbf{B}|^{-1} \tag{14}$$

Therefore, the entropy of the transformed quantity is given by

$$\begin{aligned} H(\mathbf{y}) &= -E\{\log p_y(\mathbf{y})\} \\ &= -E\{\log p_x(\mathbf{x})\} + E\{\log |\det \mathbf{B}|\} \\ &= H(\mathbf{x}) + \log |\det \mathbf{B}| \end{aligned} \tag{15}$$

Using this, the mutual information between the y_i is given by

$$I(y_1, y_2, \ldots, y_N) = \sum_i H(y_i) - H(\mathbf{x}) - \log |\det \mathbf{B}| \tag{16}$$

Independence and non-Gaussianity

We will now only allow those transformations $\mathbf{B}$ that yield uncorrelated signalsof variance 1. (This is possible since uncorrelatedness is a necessary condition for independence and we are free to choose a scaling.)

We obtain

$$\begin{aligned} & \mathbf{I} = E\{\mathbf{y}\mathbf{y}^\top\} = \mathbf{B}E\{\mathbf{x}\mathbf{x}^\top\}\mathbf{B}^\top \\ \Rightarrow \quad & 1 = (\det \mathbf{B})(\det E\{\mathbf{x}\mathbf{x}^\top\})(\det \mathbf{B}^\top) \end{aligned} \tag{17}$$

which implies that $\det \mathbf{B}$ does not depend on the choice of $\mathbf{B}$ but only on the data $\mathbf{x}$.

Using this, the mutual information between the y_i is given by

$$I(y_1, y_2, \ldots, y_N) = \sum_i H(y_i) + const. \tag{18}$$

This means we have to minimize the entropy in each channel to minimize the mutual information. For a fixed mean and variance the gaussian distribution has the highest entropy, so ICA tries to find **non-Gaussian projections**.

Algorithms: InfoMax

Observations:

- Probabilityofobservationsis $\mathrm{p}(\mathbf{x}) = |\det(\mathbf{B})|\ \mathrm{p}(\mathbf{y})$
- Forindependentsources: $\mathrm{p}(\mathbf{y}) = \prod_{\mathrm{i}} \mathrm{p}_{\mathrm{i}}(\mathrm{y}_{\mathrm{i}})$

Likelihood: $L(\mathbf{B}, \mathbf{y}) = \log|\det(\mathbf{B})| + \sum_i \log \mathrm{p}_{\mathrm{i}}(\mathrm{y}_{\mathrm{i}})$

Maximizing*L* w.r.t.**B**usingnaturalgradientdescentleadsto update rule

$$\mathbf{B} \leftarrow \mathbf{B} + \lambda(\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^{\mathrm{T}})\mathbf{B}\ ,$$

where $\varphi(\mathbf{y}) = \left(-\frac{\partial \log p_1(y_1)}{\partial y_1}, \ldots, -\frac{\partial \log p_n(y_n)}{\partial y_n}\right).$

[Bell & Sejnowski, 1995; Amari et al., 1996]

Algorithms: InfoMax

Difficulty: $\varphi(\mathbf{y})$depend on unknownprobabilitydistributionsp$(\mathbf{y})$

Idea:
replace$\varphi(\mathbf{y})$ bypredefinednonlinearfunctionscorrespondingtoreasonabledis tributions

Convenientchoice:
$\varphi(\mathrm{y}) = \mathrm{y} + \tanh(\mathrm{y})$for super-Gaussiansources, and
$\varphi(\mathrm{y}) = \mathrm{y} - \tanh(\mathrm{y})$for sub-Gaussiansources

Update formulatransformsto

$$\mathbf{B} \leftarrow \mathbf{B} + \lambda(\mathbf{I} - \mathbf{K}\tanh(\mathbf{y})\mathbf{y}^{\mathbf{T}})\mathbf{B}\,,$$

where**K** is a diagonal matrixencodingthe sign.

Algorithms: FastICA

Idea: maximizethenegentropyofthesources

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y})$$

(negentropymeasuresthedeviationfromtheGaussiandistribution)

Approximation: $\mathrm{J}(\mathrm{y_i}) \approx \mathrm{c}[\mathrm{E}\{\mathrm{g}(\mathrm{y_i})\} - \mathrm{E}\{\mathrm{g}(\mathrm{v})\}]^2$,

whereg is a non-quadratic function, c is an irrelevant constant, and ν is a Gaussian variable of zero mean and unit variance.

[Hyvärinen, 1999]

Algorithms: FastICA

Optimize Jforgivenfunction g(y) w.r.t. *w* toobtainone IC*:*

$$\max_{\mathbf{w}} \left[\mathrm{E}\{ \mathrm{g}(\mathbf{w}^{\mathrm{T}}\mathbf{x}) \} - \mathrm{E}\{ \mathrm{g}(\mathrm{v}) \} \right]^2$$

s.t.

$$E\left\{ \left(\mathbf{w}_k{}^{\mathrm{T}} \mathbf{x} \right)^2 \right\} = 1\,.$$

Fixed-point algorithm:

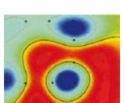
- Initialize **w**randomly
- Perform update step

$$\mathbf{w}^+ \leftarrow E\left\{ \mathbf{x} g(\mathbf{w}^T\mathbf{x}) \right\} - E\left\{ g'(\mathbf{w}^T\mathbf{x}) \right\} \mathbf{w}$$

$$\mathbf{w} \leftarrow \mathbf{w}^+ / \|\mathbf{w}^+\|$$

- Performnormalizationtomeetconstraints

Note: multiple componentsbydeflationofstraighforwardextensionofthecostfunction.



Choice of nonlinearities for fastICA

$$g(y) = \log\cosh(y)$$

„good general-purposecontrastfunction“ (Hyvärinen)

$$g(y) = -\exp(\frac{-y^2}{2})$$

„maybebetterwhen the independent components are highly super-Gaussian, or when robustness is very important”

$$g(y) = \frac{1}{4}y^4$$

„onlyjustifiedfor estimating sub-Gaussian independent components when there are no outliers”

Using temporal decorrelation for ICA

We have seen, that decorrelation (PCA) is not enough to solve the ICA problem. However, if two sources i and j are independent, their source signals $s_i(t)$ and $s_j(t)$ are uncorrelated even if one signal ist time-shifted:

$$E\{s_i(t)s_j(t-\tau)\} = 0 \quad \text{if} \quad i \neq j \tag{21}$$

or, in matrix notation

$$E\{\mathbf{s}(t)\mathbf{s}^\top(t-\tau)\} = diag. \tag{22}$$

On the diagonals of this matrices are the autocovariance functions of the source signals.

We define the symmetrized *time-lagged covariance matrix* of the vector-valued time series $\mathbf{s}(t)$ as

$$\mathbf{C}_\mathbf{s}(\tau) \equiv \frac{1}{2}\left(E\{\mathbf{s}(t)\mathbf{s}^\top(t-\tau)\} + E\{\mathbf{s}(t-\tau)\mathbf{s}^\top(t)\}\right) \tag{23}$$

A simplistic variant of TDSEP

- Use only two different time lags τ_1 and τ_2.
- Direct joint diagonalization of the two matrices as generalized eigenvalue problem:

```
» [V,D] = eig(C1,C2);
» B = V';
» y = B*x;
```

- This simpler approach works, but depends strongly on the choice of the τ-parameters (they have to capture the temporal structure). The full TDSEP is much more stable and is quite powerful for data with temporal structure.
- For the solution of the general problem, a techniques for approximate joint diagonalization has to be used.

Summary

- ICA demixes the data into „minimally statistically dependent“ sources
- Different measures of statistical dependence

  → different algorithms
- Resampling methods can be used to assess the quality of ICA projections

  → identication of the appropriate ICA-Model possible
- Application of ICA often reveals stable subspaces which are physiologically plausible