Lecture 16
Independent Component Analysis

Jochen Braun

Otto-von-Guericke-Universität Magdeburg,
Cognitive Biology Group

Theoretical Neuroscience II
SS 2020

Credits: https://research.ics.aalto.fi/ica/icademo/
Hyvarinen & Oja, 2000

16. Independent component analysis (ICA)

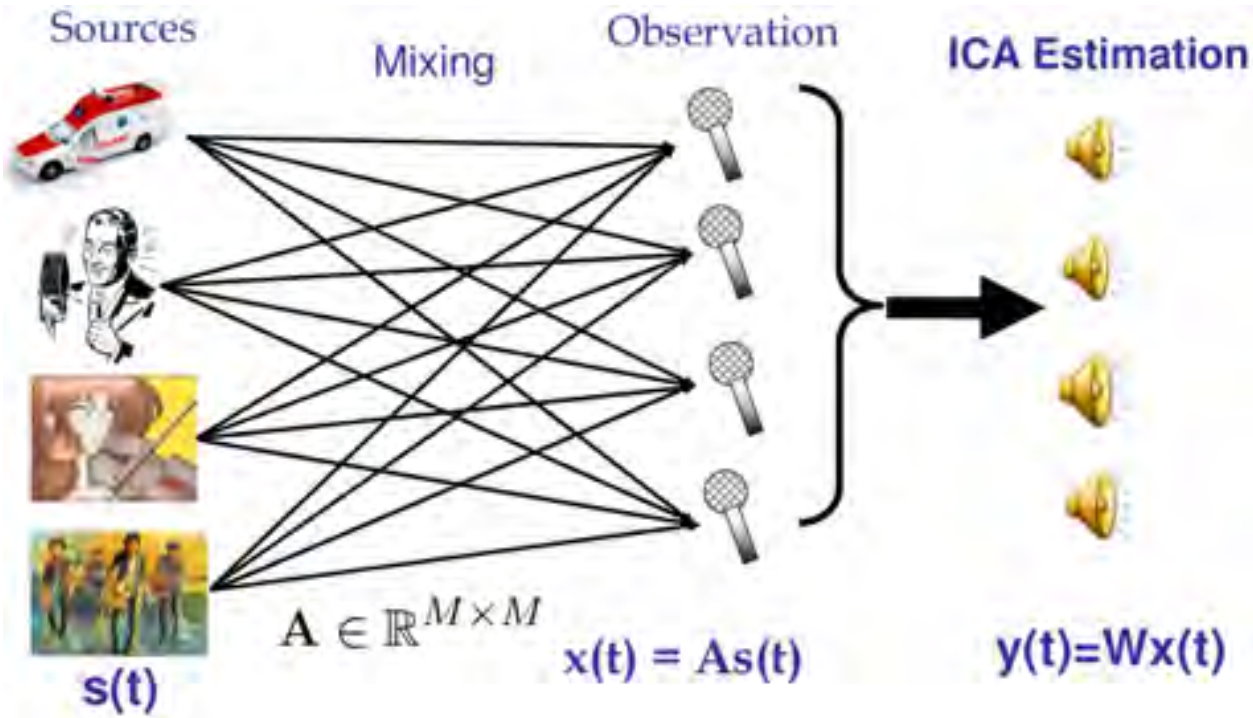
Typical natural causes have rich statistics (more complex than Gaussian). **Blind source separation** *decomposes observed signal mixtures into independent and maximally non-Gaussian sources (causes). Statistical dependencies may remain even after signals have been whitened (scaled to zero mean and unit variance) and decorrelated (by Principal Component Analysis).* **ICA** *decomposes signals such as to eliminate these remaining dependencies (in higher moments), maximizing both independence and non-Gaussianity. This heuristic approach can be justified by information theory.* **Fast ICA** *is a convenient computational implementation. ICA has many successful applications, for example, in functional brain imaging.*

Overview

1. **Blind source separation**
2. **ICA demo**
3. **Theoretical justification**
4. **Processing steps of ICA**
5. **Imaging applications**

1. Blind source separation

Consider a 'cocktail party' situation, with a cacophony of sounds, musical performances, and voices ...



Assume that multiple microphones record various mixtures (linear combinations) $x(t)$ of these independent sources $s(t)$.

Other examples

The blind source separation problem occurs in many contexts:

- Electrical recordings of brain activity (EEG). Each scalp electrode measures a mixture of potential sources (dipoles) inside the brain.
- To better understand neural processes, it would be useful to identify independent potential sources.
- Natural signals and images may be a mixture of prototypical components (sound or image features).
- To efficiently represent and analyze natural signals and images, it would be useful to have a dictionary of such components.

Mixing matrix

Observed time-courses $\boldsymbol{x}(t) = \boldsymbol{x}_i$ are a linear mixture of latent sources $\boldsymbol{s}(t) = \boldsymbol{s}_i$, where $i = 1 \ldots n$ indexes time-samples:

$$\boldsymbol{x}_i = \boldsymbol{A} \cdot \boldsymbol{s} = \begin{pmatrix} \boldsymbol{a}_1 \cdot \boldsymbol{s}_i \\ \boldsymbol{a}_2 \cdot \boldsymbol{s}_i \\ \vdots \\ \boldsymbol{a}_n \cdot \boldsymbol{s}_i \end{pmatrix} = \begin{pmatrix} a_{11}s_{1i} + \ldots + a_{1m}s_{mi} \\ a_{21}s_{1i} + \ldots + a_{2m}s_{mi} \\ \vdots \\ a_{n1}s_{1i} + \ldots + a_{nm}s_{mi} \end{pmatrix}$$

where n is the number of observed time-courses, m is the number of latent sources, and k is the number of time-samples.

$$\boldsymbol{A} \in \mathcal{R}^{n \times m}, \qquad \boldsymbol{x} \in \mathcal{R}^{n \times k}, \qquad \boldsymbol{s} \in \mathcal{R}^{m \times k}$$

Typically, we assume that the number of observations n is at least as large as the number of sources m.

Several ambiguities

As neither mixing matrix $\boldsymbol{A}$ nor sources $\boldsymbol{s}$ are known, the situation is intrinsically ambiguous:

- We cannot know the absolute variance (power) of any source $\boldsymbol{s}_j$, because we observe only $a_{1j}\boldsymbol{s}_{jk}$, $a_{2j}\boldsymbol{s}_{jk}$, $\ldots a_{nj}\boldsymbol{s}_{jk}$.
- Accordingly, we assume that all sources have unit variance (power).
- We cannot order or rank the sources in any way, because any permutation yields the same observations.
- Accordingly, we accept any ordering of sources as equivalent.

Unmixing matrix

We seek an unmixing matrix $\boldsymbol{W}$ which renders the recovered sources $\boldsymbol{y}(t) = \boldsymbol{y}_i$ as independent as possible:

$$\boldsymbol{y}_i = \boldsymbol{W} \cdot \boldsymbol{x} = \begin{pmatrix} \boldsymbol{w}_1 \cdot \boldsymbol{x}_i \\ \boldsymbol{w}_2 \cdot \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{w}_m \cdot \boldsymbol{x}_i \end{pmatrix} = \begin{pmatrix} w_{11}x_{1i} + \ldots + w_{1n}x_{ni} \\ w_{21}x_{1i} + \ldots + w_{2n}x_{ni} \\ \vdots \\ w_{m1}x_{1i} + \ldots + w_{mn}x_{ni} \end{pmatrix}$$

$$\boldsymbol{W} \in \mathcal{R}^{m \times n}, \qquad \boldsymbol{x} \in \mathcal{R}^{n \times k}, \qquad \boldsymbol{y} \in \mathcal{R}^{m \times k}$$

The number of recovered sources m cannot be larger than the number of observations n!

Independence

Two variables $y_{1,2}$ are independent if the joint probability density is separable

$$p(y_1, y_1) \quad = \quad p(y_1)\, p(y_2)$$

or, equivalently, if the mutual information is zero

$$I_m = \int \int p(y_1, y_2) \ln \frac{p(y_1, y_2)}{p(y_1)\, p(y_2)}\, dy_1\, dy_2 = 0$$

or, equivalently, if expectations of all transformations are separable

$$E\left\{h_1(y_1)h_2(y_2)\right\} = E\left\{h_1(y_1)\right\} E\left\{h_2(y_2)\right\}, \qquad \forall h_1(y), h_2(y)$$

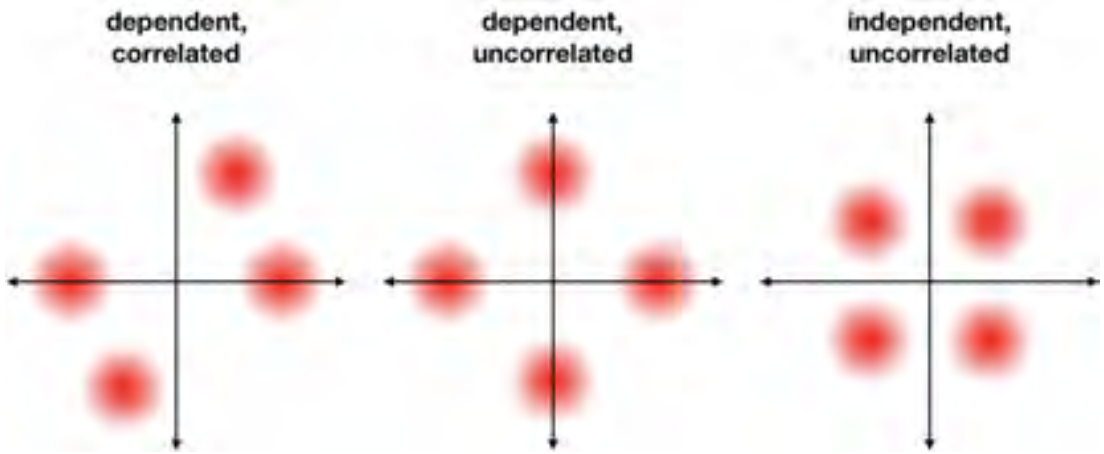
Uncorrelatedness

Uncorrelatedness is a far weaker constraint than independence.
Two variables $y_{1,2}$ are uncorrelated if the covariance is zero:

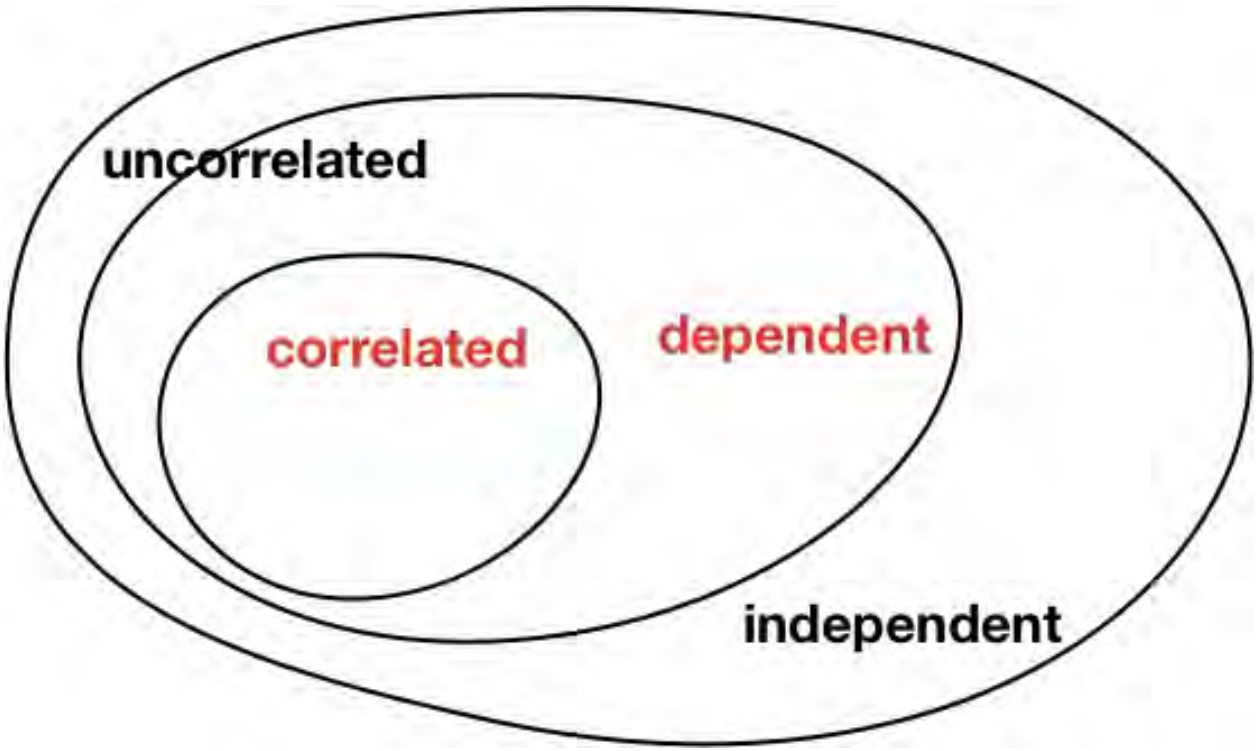
$$E\{y_1\, y_2\} - E\{y_1\}\, E\{y_2\} = 0 \qquad \Leftrightarrow \qquad E\{y_1\, y_2\} = E\{y_1\}\, E\{y_2\}$$

This does not imply separability of higher moments

$$E\left\{y_1^k\, y_2^k\right\} \neq E\left\{y_1^k\right\}\, E\left\{y_2^k\right\} \qquad k > 1$$

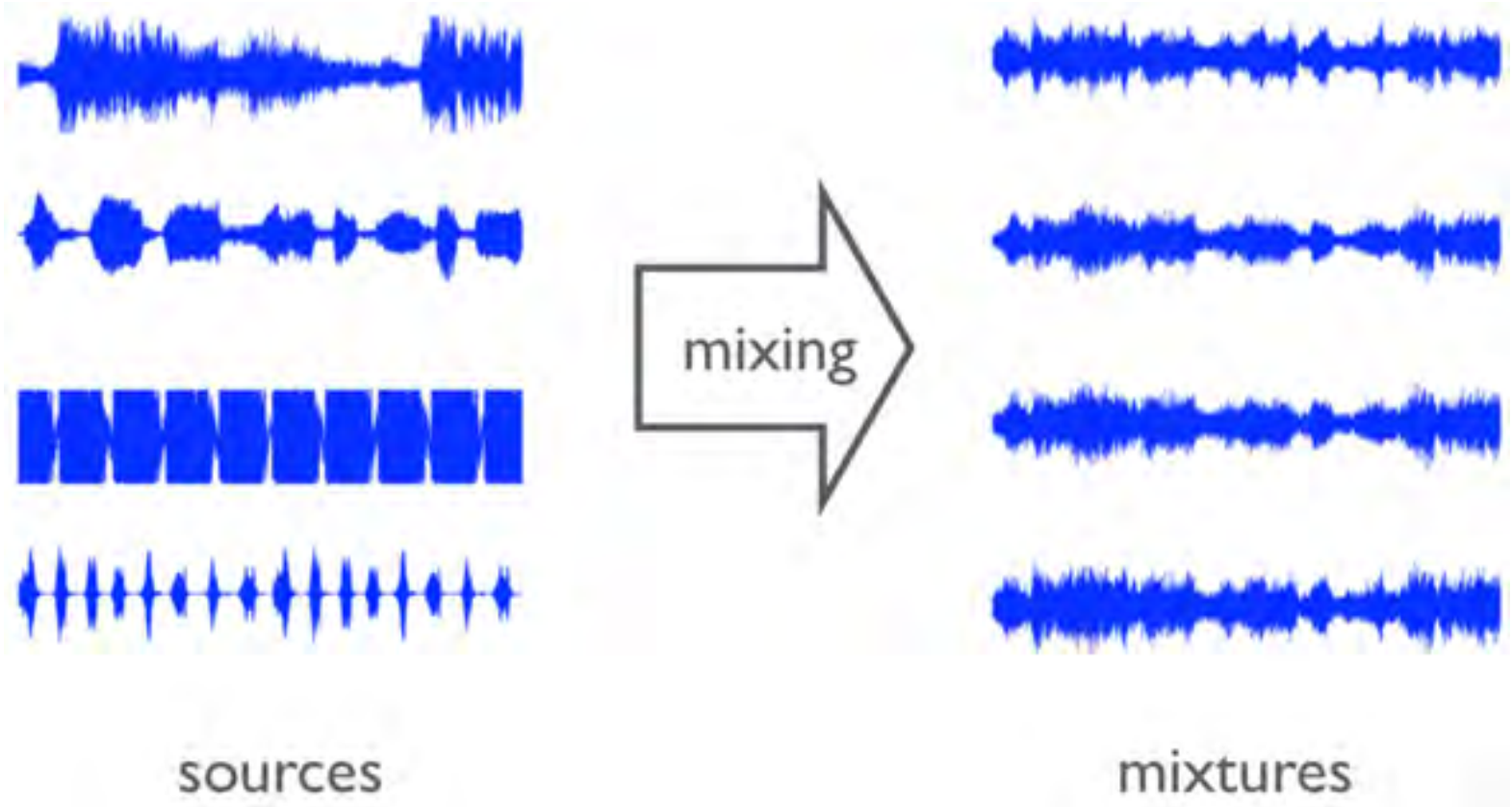


Correlation and dependence

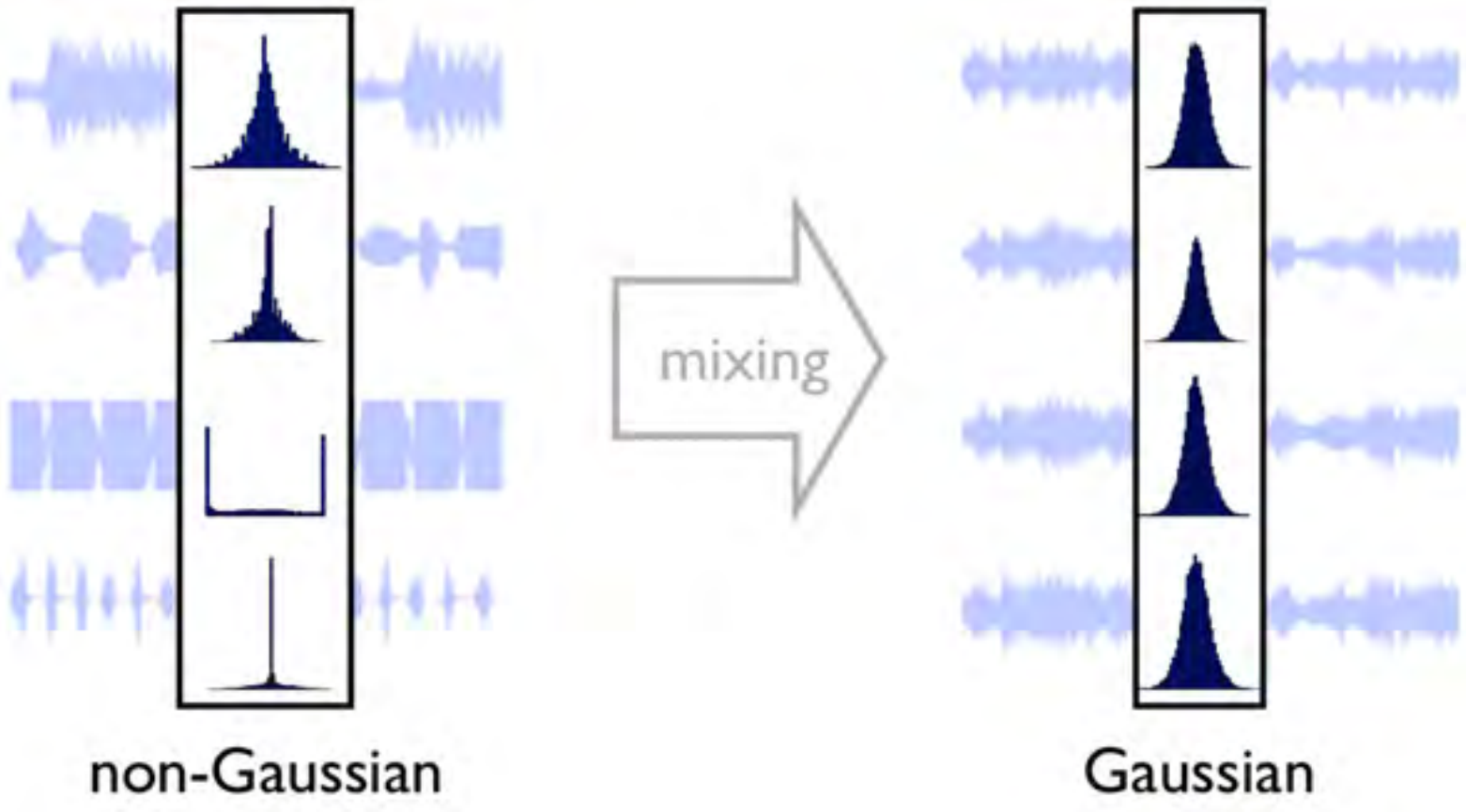


Central limit theorem

Random variables may be ‘Gaussian’ or ‘non-Gaussian’ or anything in between.



Linear combinations of *independent* 'non-Gaussian' are *more* 'Gaussian' than the original variables.



Independent = maximally non-Gaussian

Given unknown mixing matrix $\boldsymbol{x} = \boldsymbol{A} \cdot \boldsymbol{s}$ and linear combination (attempted partial unmixing): $y = \boldsymbol{w}^T \cdot \boldsymbol{x}$.

Compare our attempt at unmixing $\boldsymbol{w}^T$ with the ground truth $\boldsymbol{A}^T$ (true unmixing):

$$\boldsymbol{z} = \boldsymbol{A}^T \cdot \boldsymbol{w} \quad \Rightarrow \quad y = \boldsymbol{w}^T \cdot \boldsymbol{x} = \boldsymbol{w}^T \cdot \boldsymbol{A} \cdot \boldsymbol{s} = \left(\boldsymbol{A}^T \cdot \boldsymbol{w}\right)^T \cdot \boldsymbol{s} = \boldsymbol{z}^T \cdot \boldsymbol{s}$$

As a linear combination of s_i, y is more Gaussian than any s_i. To be maximally non-Gaussian, y must be equal to one s_i.

Thus, by choosing $\boldsymbol{w}$ such as to maximize non-Gaussianity of

$$y = \boldsymbol{w}^T \cdot \boldsymbol{x} = \boldsymbol{z}^T \cdot \boldsymbol{s}$$

we can identify one independent component!

Procedure of blind source separation

- Assume that sources are *independent* and have *non-Gaussian* statistics (dependent and/or Gaussian sources would not be distinguishable).
- Assume that mixing matrix is square and orthogonal (number of observations equals number of independent sources).
- Find first unmixing vector $\boldsymbol{w}$ such as to maximize non-Gaussianity of $\boldsymbol{w}^T \cdot \boldsymbol{x}$.
- This identifies first independent component s_1 ($-s_1$ is also a solution).
- Find second unmixing vector $\boldsymbol{v}$, orthogonal to $\boldsymbol{w}$, such as to maximize non-Gaussianity of $\boldsymbol{v}^T \cdot \boldsymbol{x}$.
- This identifies second independent component s_2 ($-s_2$ is also a solution).
- Repeat until the number of 'independent components' equals the number of observations.

Points to note

- Given a linear mixture of unknown signals, 'blind source separation' seeks an 'unmixing' matrix.
- Some ambiguities cannot be resolved: signal power, dependent signals, Gaussian signals.
- Important distinction between dependence (all moments) and correlatedness (only second moment).
- Generally, *mixing* signals makes statistics *more* Gaussian.
- Thus, a matrix that makes statistics *less* Gaussian, is presumably an *unmixing* matrix.
- 'Less Gaussian' implies non-zero higher moments.
- Abstract procedure of 'blind source separation'

2. ICA Demo

cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/

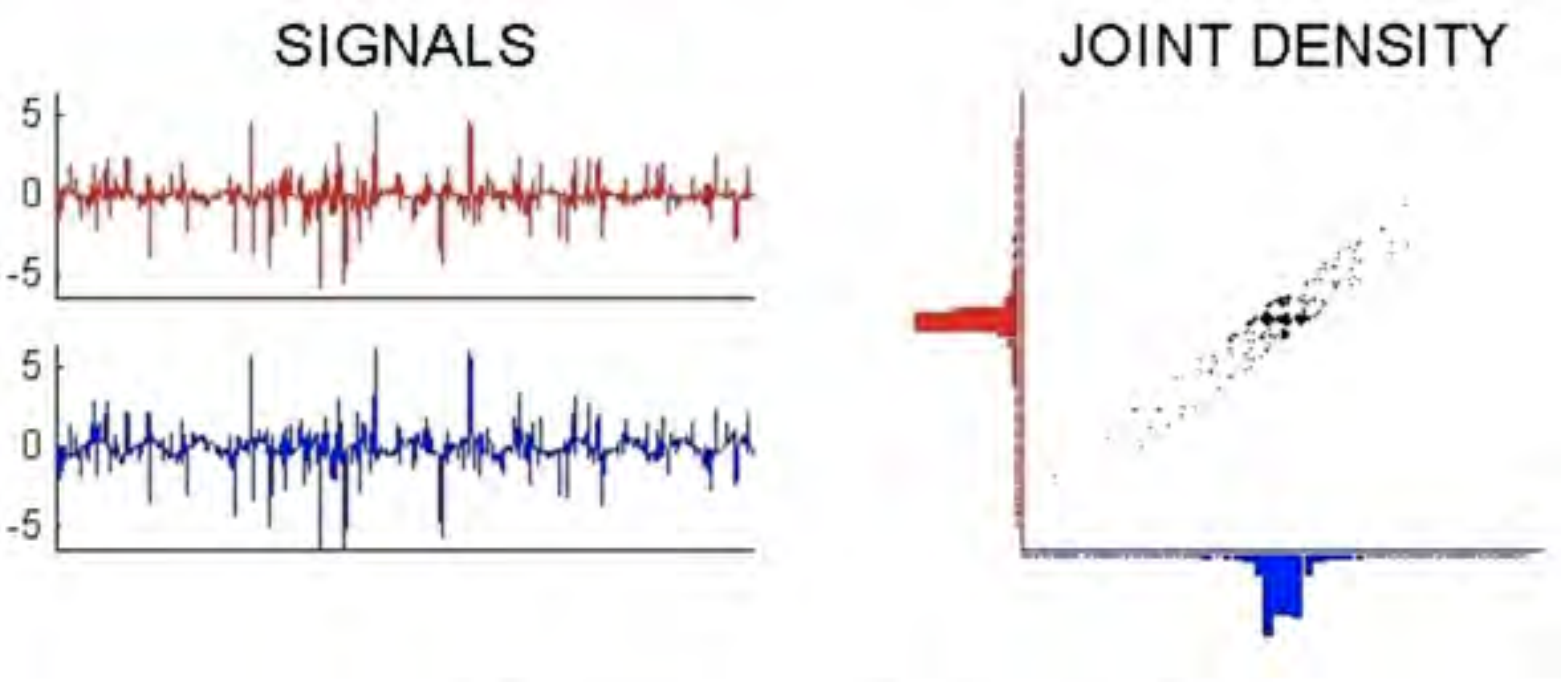
Here, we demonstrate ICA for solving the blind source separation problem. We are given two linear mixtures of two source signals which we know to be independent of each other, i.e. observing the value of one signal does not give any information about the value of the other. We seek to determine the source signals given only the mixtures.

Putting this into mathematical notation, we model the problem by

$$\boldsymbol{x} = \boldsymbol{A} \cdot \boldsymbol{s}$$

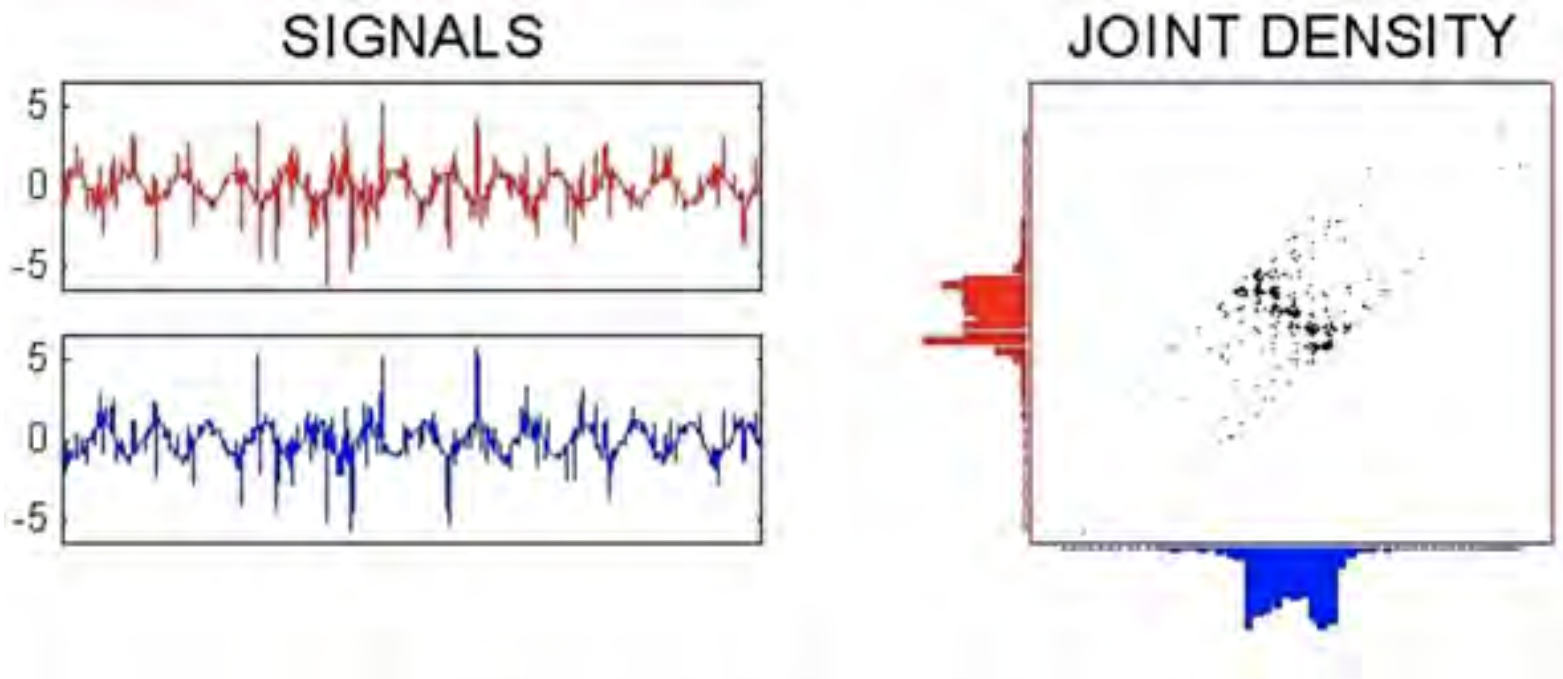
where $\boldsymbol{s}$ is a two-dimensional random vector containing the independent source signals, $\boldsymbol{A}$ is the two-by-two mixing matrix, and $\boldsymbol{x}$ contains the observed (mixed) signals.

This first plot (below) shows the signal mixtures on the left and the corresponding joint density plot on the right. That is, at a given time instant, the value of the top signal is the first component of x, and the value of the bottom signal is the corresponding second component. The plot on the right is then simply constructed by plotting each such point x. The marginal densities are also shown at the edge of the plot.



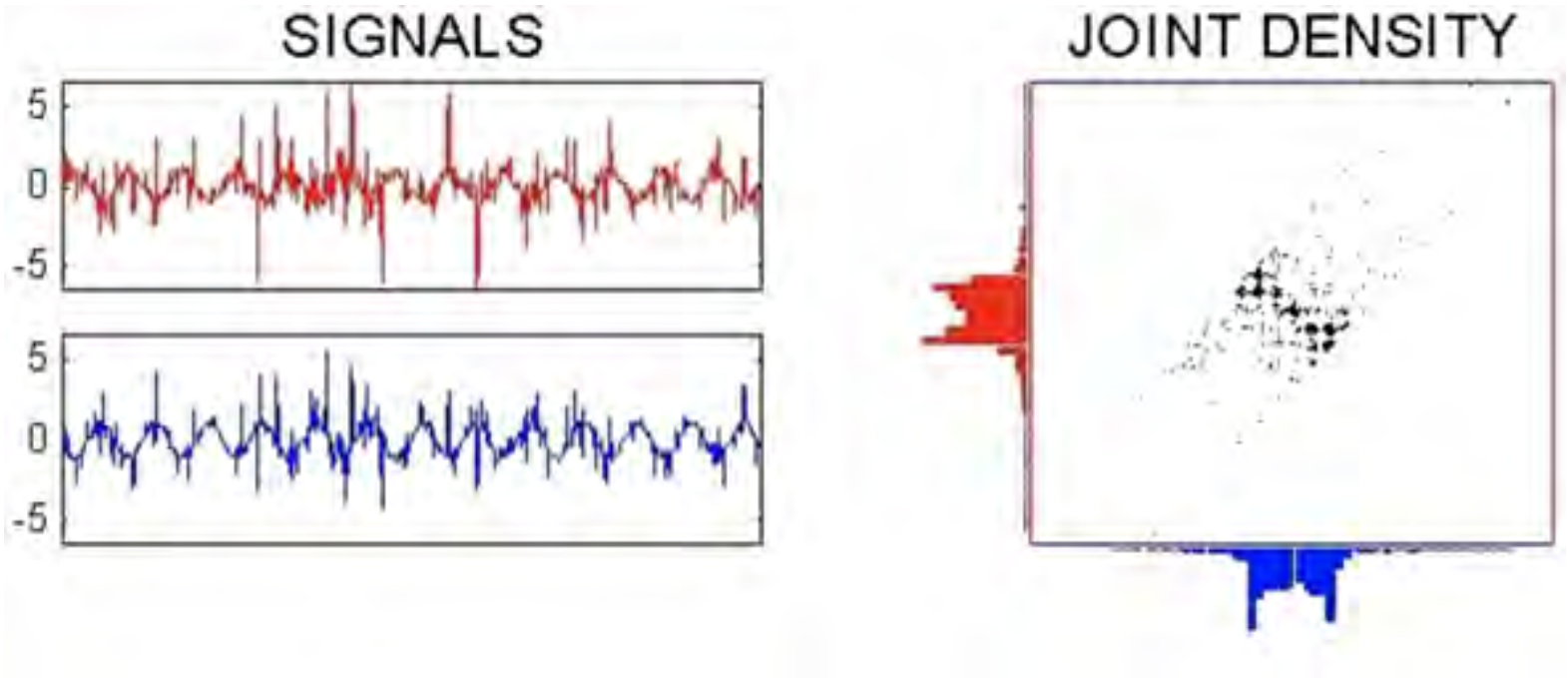
Input signals and density

The first step to remove any correlations in the data ('whitening' or PCA). We seek a linear transformation $\boldsymbol{V}$ such that when $\boldsymbol{y} = \boldsymbol{V} \cdot \boldsymbol{x}$ we obtain $E\left(\boldsymbol{y} \cdot \boldsymbol{y}^T\right) = \boldsymbol{I}$. To this end, we choose $\boldsymbol{V} = \boldsymbol{C}_x^{-1/2}$ as the 'scaled eigenvectors' of the covariance matrix $\boldsymbol{C}_x = \left(\boldsymbol{x} \cdot \boldsymbol{x}^T\right)$, ensuring $E\left(\boldsymbol{y} \cdot \boldsymbol{y}^T\right) = E\left(\boldsymbol{V} \cdot \boldsymbol{x}\boldsymbol{x}^T \cdot \boldsymbol{V}^T\right) = \boldsymbol{C}_x^{-1/2} \cdot \boldsymbol{C}_x^{-1/2} = \boldsymbol{I}$. The figure shows the whitened signal $\boldsymbol{y}$ and joint density $p(\boldsymbol{y})$



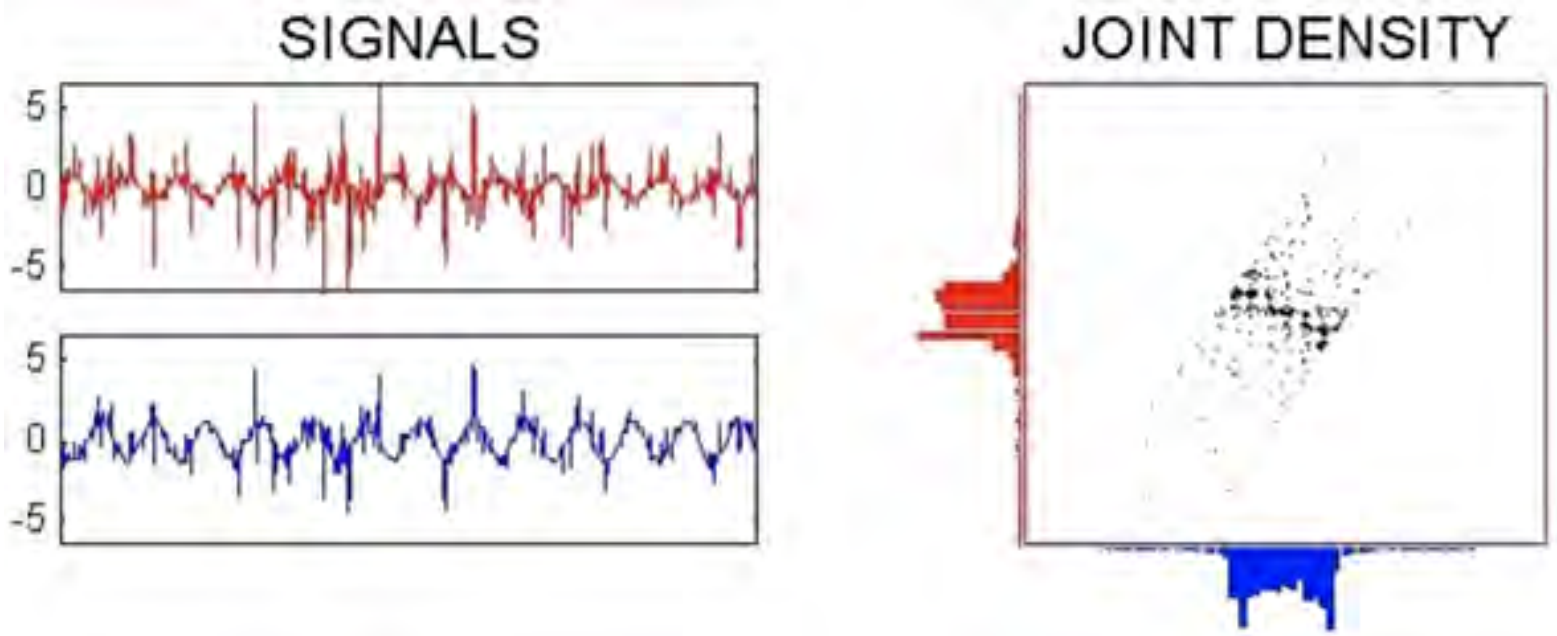
Whitened signals and density

After whitening, we seek to increase the degree of independence by performing a rotation (orthogonal transformation). The appropriate rotation maximizes the non-Gaussianity of the marginal densities.



Separated signals after 1 step of FastICA

There are many algorithms for performing ICA, but a particularly efficient one is FastICA, which was developed by Hyvarinen and Oja (1999).

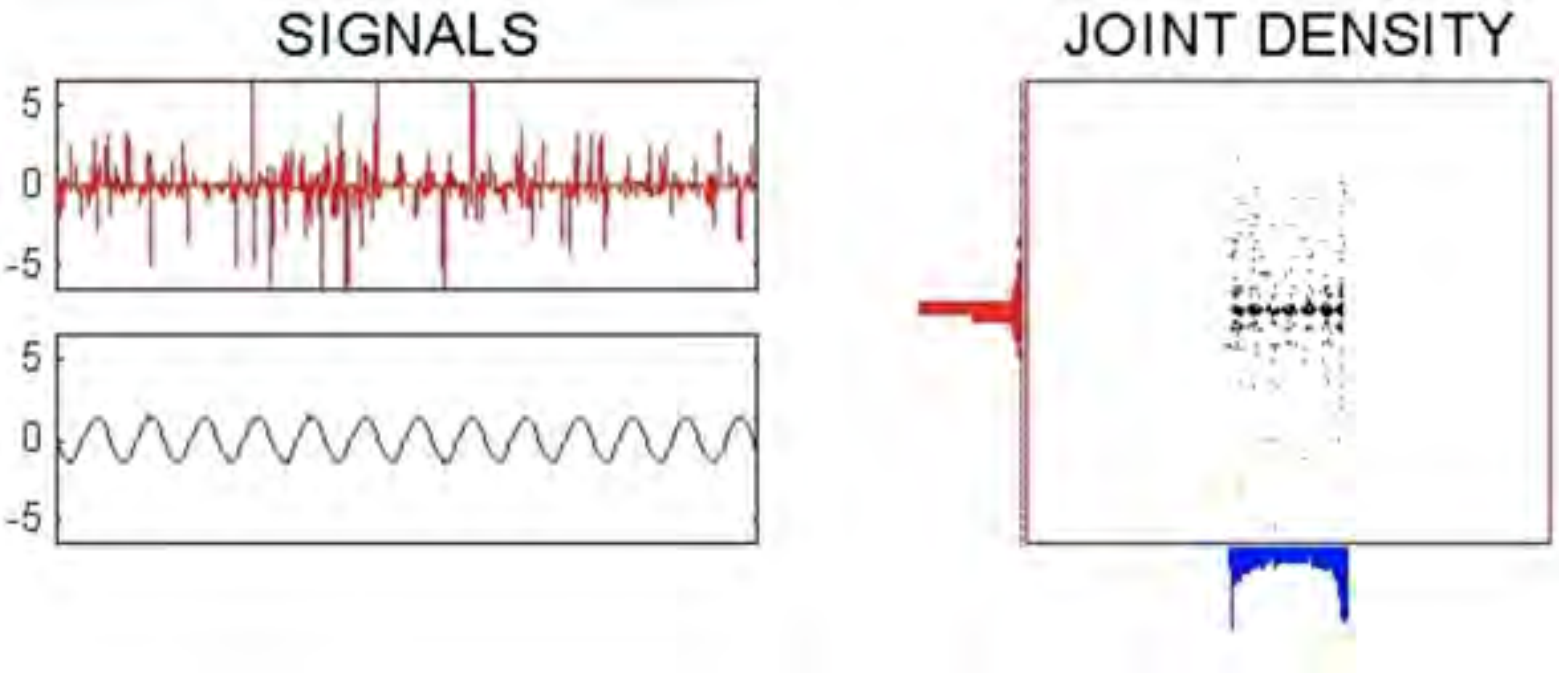


Separated signals after 2 steps of FastICA

.. the rotation continues ...

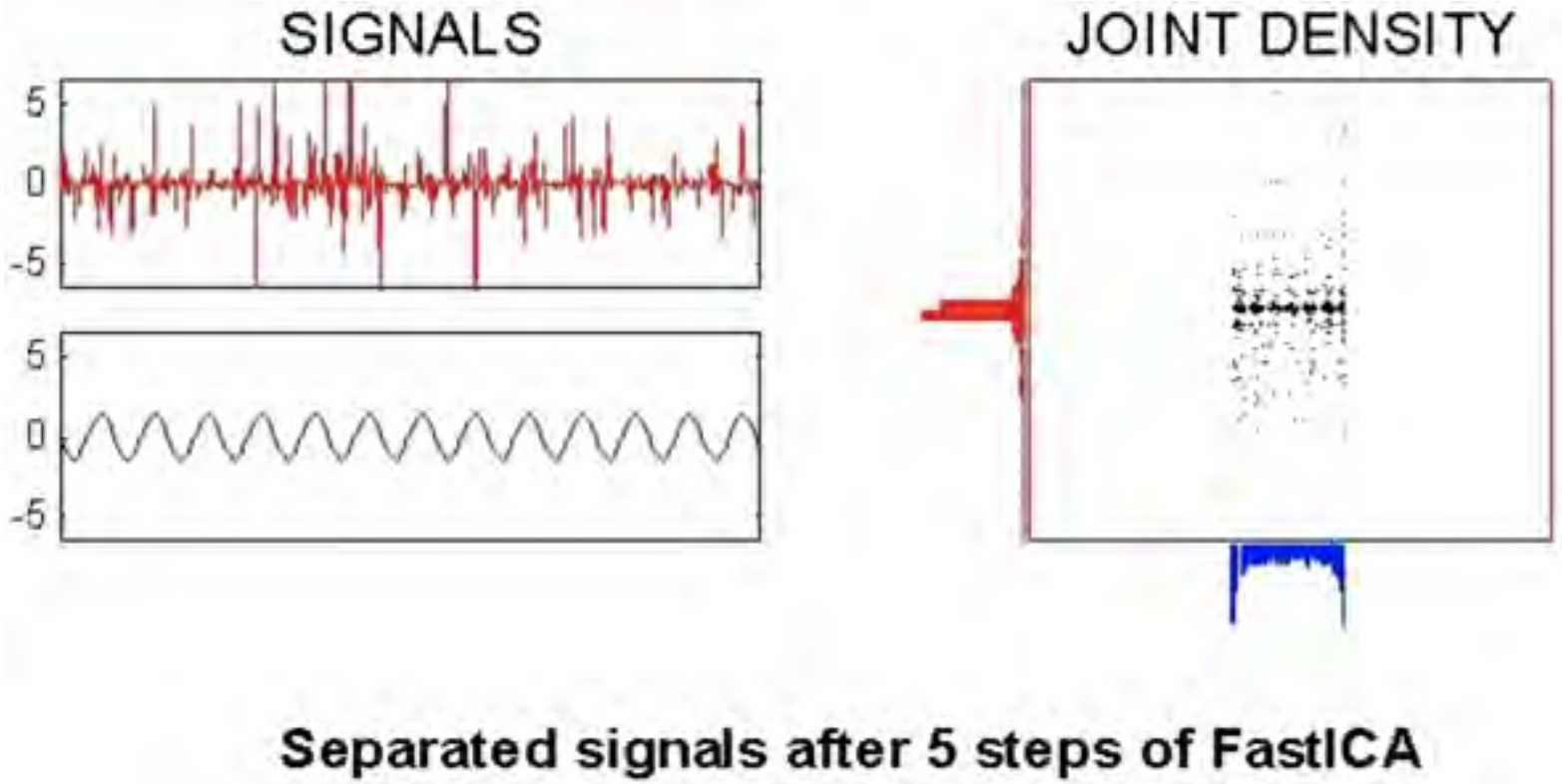
Separated signals after 3 steps of FastICA

... until it begins to converge ...



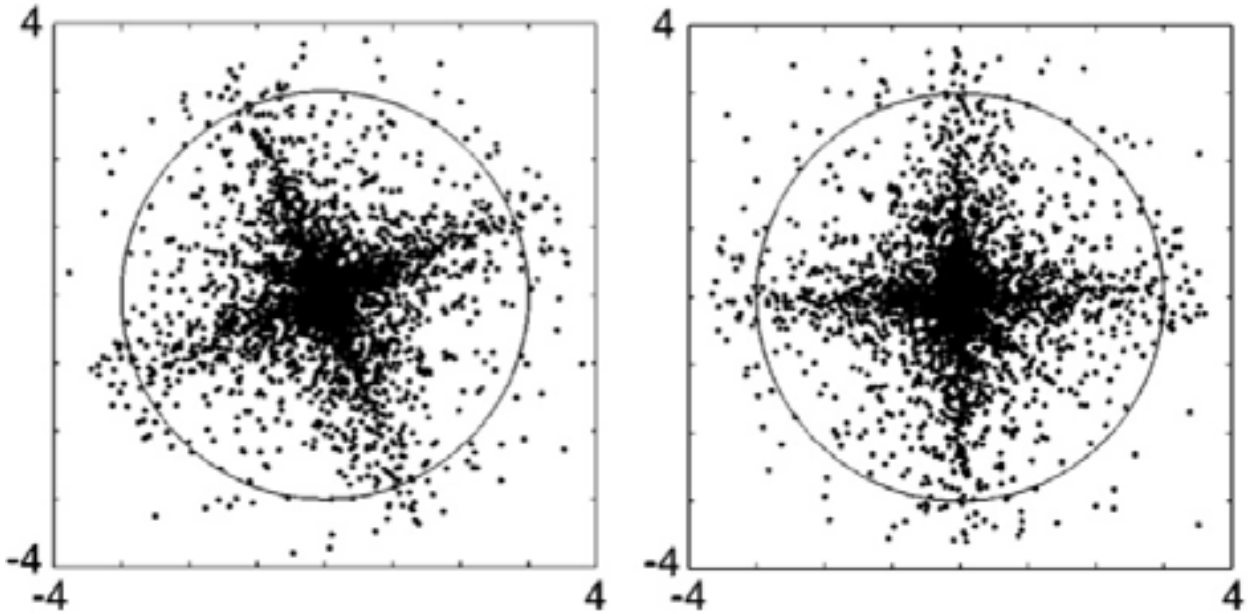
Separated signals after 4 steps of FastICA

Convergence! In this example, the two source signals were a sinusoid and impulse noise. The joint density can be seen to be the product of marginal densities, $p(\boldsymbol{y}) = p(y_1)p(y_2)$. Thus we have recovered two independent source signals.



Separated signals after 5 steps of FastICA

ICA rotates coordinates such as to maximize higher-order moments and independence!



Points to note

- Fast ICA is a computational scheme for solving the blind source separation problem.
- Hyvarinnen and Oja (2000) illustrated its operation graphically.
- First step is to whiten and decorrelate signals (perform principal component analysis, PCA).
- In general this, leaves residual dependencies in higher-order moments.
- Several steps of Fast ICA rotate coordinate system.
- After convergence, both higher-order moments and independence are maximal.

3. Theoretical justification

Several quantitative measures of nongaussianity are possible:

- Excess kurtosis
- Negentropy
- Approxmate negentropy
- Mutual information
- Maximum likelihood
- Infomax principle

We assume a random variable y with zero-mean and unit-variance:

$$E\{y\} = 0, \qquad E\{y^2\} = 1$$

Excess kurtosis

$$kurt(y) = E\left\{y^4\right\} - 3\left(E\left\{y^2\right\}\right)^2 = E\left\{y^4\right\} - 3$$

$$kurt(y_1 + y_2) = kurt(y_1) + kurt(y_2), \qquad kurt(\alpha\, y_1) = \alpha^4\, kurt(y_1)$$

In terms of the transformation $\boldsymbol{z} = \boldsymbol{A}^T \cdot \boldsymbol{w}$, we seek the values $z_{1,2}$ maximizing the kurtosis of $y = z_1 s_1 + z_2 s_2$:

$$|kurt(y)| = |kurt(z_1 s_1)| + |kurt(z_2 s_2)| = z_1^4\, |kurt(s_1)| + z_2^4\, |kurt(s_2)|$$

with

$$E\left\{y^2\right\} = z_1^2 + z_2^2 = 1$$

It can be shown that the maxima are where one element of $\boldsymbol{z}$ is 0 and the other is ± 1.

Negentropy

Negentropy measures how much less entropy a variable has than a comparable Gaussian variable. It is based on differential entropy

$$H(\boldsymbol{y}) = -\int p(\boldsymbol{y}) \log p(\boldsymbol{y})\, d\boldsymbol{y}$$

A Gaussian variable has the largest differential entropy of all variables with constant variance. Negentropy is defined as the entropy difference to a Gaussian variable with identical covariance:

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

Negentropy is well justified by statistical theory. However, negentropy is computationally difficult and requires the (high-dimensional) density $p(\boldsymbol{y})$.

Approximate negentropy

Improved approximations of negentropy were developed by Hyvarinen (1998):

$$J(y) \approx \sum_{i=1}^{p} k_i \left[E\left\{G_i(y)\right\} - E\left\{G_i(\nu)\right\}\right]^2$$

where k_i are positive constants, ν is a Gaussian variable of zero-mean and unit variance, and functions G_i are suitable non-quadratic functions.
Note that $G(y) = y^4$ retrieves a kurtosis-based approximation

$$J(y) \propto \left[E\left\{G(y)\right\} - E\left\{G(\nu)\right\}\right]^2$$

Suitable choices of G yield robust estimates of negentropy, e.g.:

$$G_1(u) = \frac{1}{a_1} \log\cosh(a_1 u), \qquad G_2(u) = -\exp\left(-u^2/2\right)$$

Mutual information

The mutual information between m random variables y_i is

$$I(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(\boldsymbol{y})$$

It is equivalent to the Kullback-Leibler divergence between the joint density $p(\boldsymbol{y})$ and the product of marginal densities $\prod_{i=1}^{m} p(y_i)$. For an invertible transform $\boldsymbol{y} = \boldsymbol{W} \cdot \boldsymbol{x}$, it can be shown that

$$I(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(\boldsymbol{x}) - \log |\det \boldsymbol{W}|$$

Mutual information and negentropy differ only by a constant and the sign:

$$I(y_1, y_2, \ldots, y_m) = C - \sum_{i} J(y_i)$$

ICA of a random vector $\boldsymbol{x}$ is an invertible transformation $\boldsymbol{y} = \boldsymbol{W} \cdot \boldsymbol{x}$ which minimizes the mutual information of the transformed (unmixed) components y_i.

Finding this invertible transformation is roughly equivalent to finding the 1-D subspaces with maximum negentropy.

Rigorously, ICA estimation by minimization of mutual information is equivalent to maximizing the sum of nongaussianities of estimates, where the estimates are constrained to be uncorrelated.

Points to note

- ICA is based on a heuristic approach: maximizing non-gaussianity (higher order moments).
- Several related theoretical arguments justify this approach.
- We outlined excess kurtosis, negentropy, and mutual information.

4. Processing steps of ICA

The processing steps are as follows:

- Centering
- Whitening
- Filtering
- Fast ICA (repeatedly)
- Decorrelation (repeatedly)

Centering

Observations $\boldsymbol{x}$ must have zero mean. If necessary, this can be ensured by subtracting the mean:

$$\boldsymbol{x} = \boldsymbol{x}_{ori} - \boldsymbol{m}, \qquad \boldsymbol{m} = E\left\{\boldsymbol{x}_{ori}\right\}$$

After estimating the mixing matrix $\boldsymbol{A}$ and zero-mean sources $\boldsymbol{s}$, we can add the mean back to obtain non-zero-mean sources

$$\boldsymbol{s}_{ori} = \boldsymbol{s} + \boldsymbol{A}^T \cdot \boldsymbol{m}$$

Whitening

Observations $\boldsymbol{x}$ must be 'whitened' by rotation to the principle component axes. Recall eigenvalue decomposition of covariance matrix

$$E\left\{\boldsymbol{x}\cdot\boldsymbol{x}^T\right\} = \boldsymbol{E}\cdot\boldsymbol{D}\cdot\boldsymbol{E}^T, \qquad \tilde{\boldsymbol{x}} = \boldsymbol{E}\cdot\boldsymbol{D}^{-1/2}\cdot\boldsymbol{E}^T\cdot\boldsymbol{x}$$

$$\Rightarrow \qquad E\left\{\tilde{\boldsymbol{x}}\cdot\tilde{\boldsymbol{x}}^T\right\} = \boldsymbol{I}$$

Whitening transforms the mixing matrix $\boldsymbol{A}$ to a new one, $\tilde{\boldsymbol{A}}$, which is orthogonal:

$$\tilde{\boldsymbol{x}} = \boldsymbol{E}\cdot\boldsymbol{D}^{-1/2}\cdot\boldsymbol{E}^T\cdot\boldsymbol{x} = \boldsymbol{E}\cdot\boldsymbol{D}^{-1/2}\cdot\boldsymbol{E}^T\cdot\boldsymbol{A}\cdot\boldsymbol{s} = \tilde{\boldsymbol{A}}\cdot\boldsymbol{s}$$

$$\Rightarrow \qquad E\left\{\tilde{\boldsymbol{x}}\cdot\tilde{\boldsymbol{x}}^T\right\} = \tilde{\boldsymbol{A}}\cdot E\left\{\boldsymbol{s}\cdot\boldsymbol{s}^T\right\}\cdot\tilde{\boldsymbol{A}}^T = \tilde{\boldsymbol{A}}\cdot\tilde{\boldsymbol{A}}^T = \boldsymbol{I}$$

Fast ICA (repeatedly)

Seek unit vector $\boldsymbol{w}$ that maximizes nongaussianity of projected observations $\boldsymbol{w}^T \cdot \boldsymbol{x}$, as measured by negentropy $J(\boldsymbol{w}^T \cdot \boldsymbol{x})$.
Denote by $g_{1,2}$ the derivatives of non-quadratic functions $G_{1,2}$, for example

$$g_1(u) = \tanh(u), \qquad g_2(u) = u \exp(-u^2/2)$$

1. Choose an initial (random) vector $\boldsymbol{w}$.
2. Rotate $\boldsymbol{w}^+ = E\left\{\boldsymbol{x}\, g\left(\boldsymbol{w}^T \cdot \boldsymbol{x}\right)\right\} - E\left\{g'\left(\boldsymbol{w}^T \cdot \boldsymbol{x}\right)\right\}$
3. Renormalize $\boldsymbol{w} = \boldsymbol{w}^+ / \|\boldsymbol{w}^+\|$.
4. Repeat until converged.

Convergence means that old and new $\boldsymbol{w}$ are almost identical (dot product almost unity).
For derivation see Hyvarinen & Oja (2000).

Decorrelation (repeatedly)

To estimate several independent components, we need to iterate with several orthogonal projection vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n$.

To prevent different vectors from converging to the same maxima, we must decorrelate the outputs $\boldsymbol{w}_1 \cdot \boldsymbol{x}, \boldsymbol{w}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{w}_n \cdot \boldsymbol{x}$ after every iteration, such as to keep them orthogonal!

There are several methods for achieving this.

Advantages of Fast ICA

According to its inventor, the advantages of Fast ICA include:

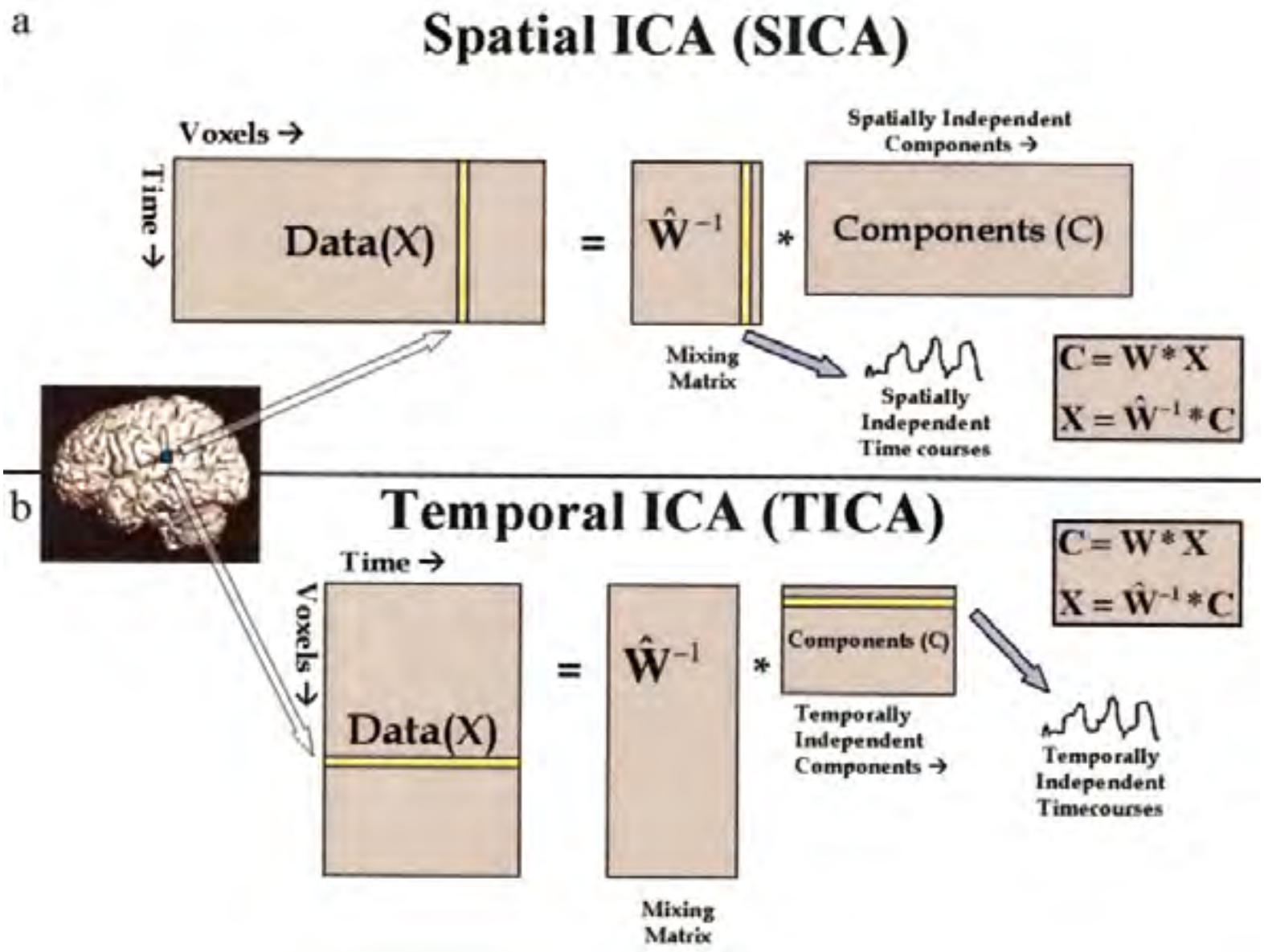
- Convergence is faster (cubic or quadratic) than for gradient descent (linear).
- No step-size parameter.
- Finds directly independent components of practically any non-Gaussian distribution, using any non-linearity g.
- Method can be optimized by choosing non-linearity g.
- Independent components may be estimated one by one, reducing computational load in an exploratory analysis.
- Matlab implementation is available.

5. Example applications

ICA is useful in many brain recording and brain imaging applications, for example

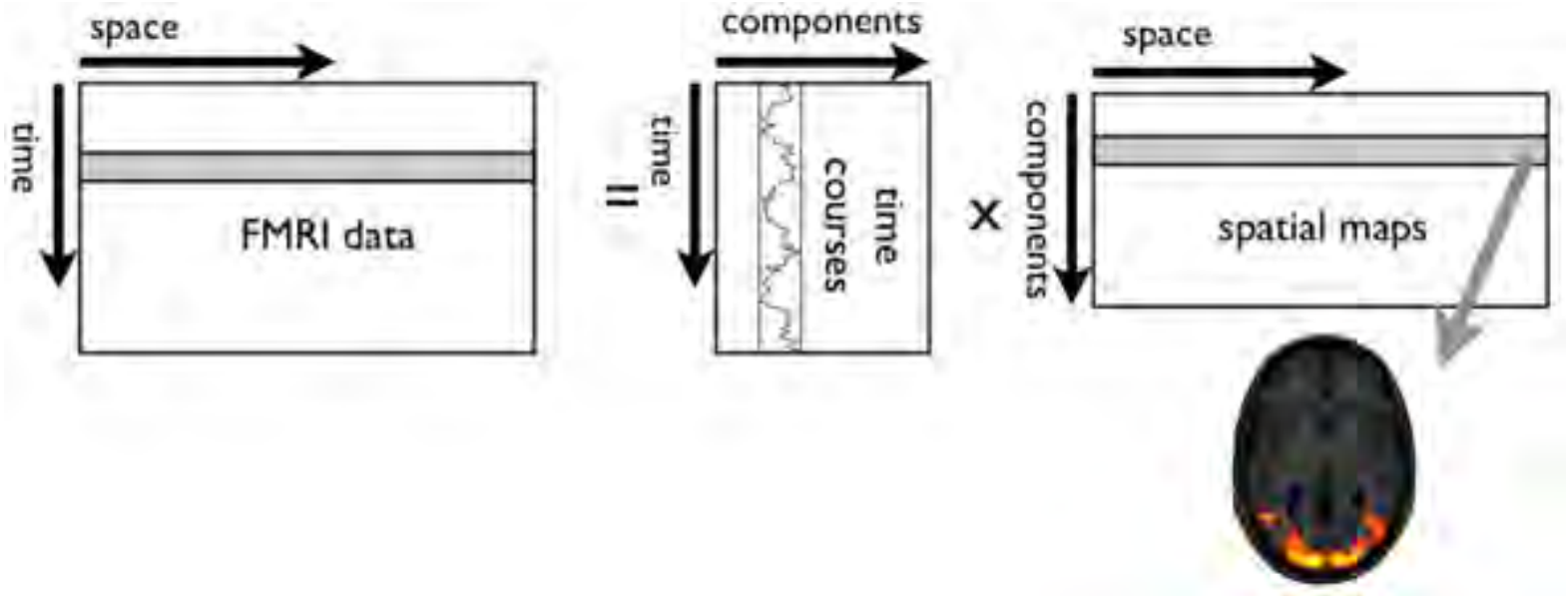
- Spatial ICA
- Temporal ICA
- Artifact removal

Spatial and temporal ICA

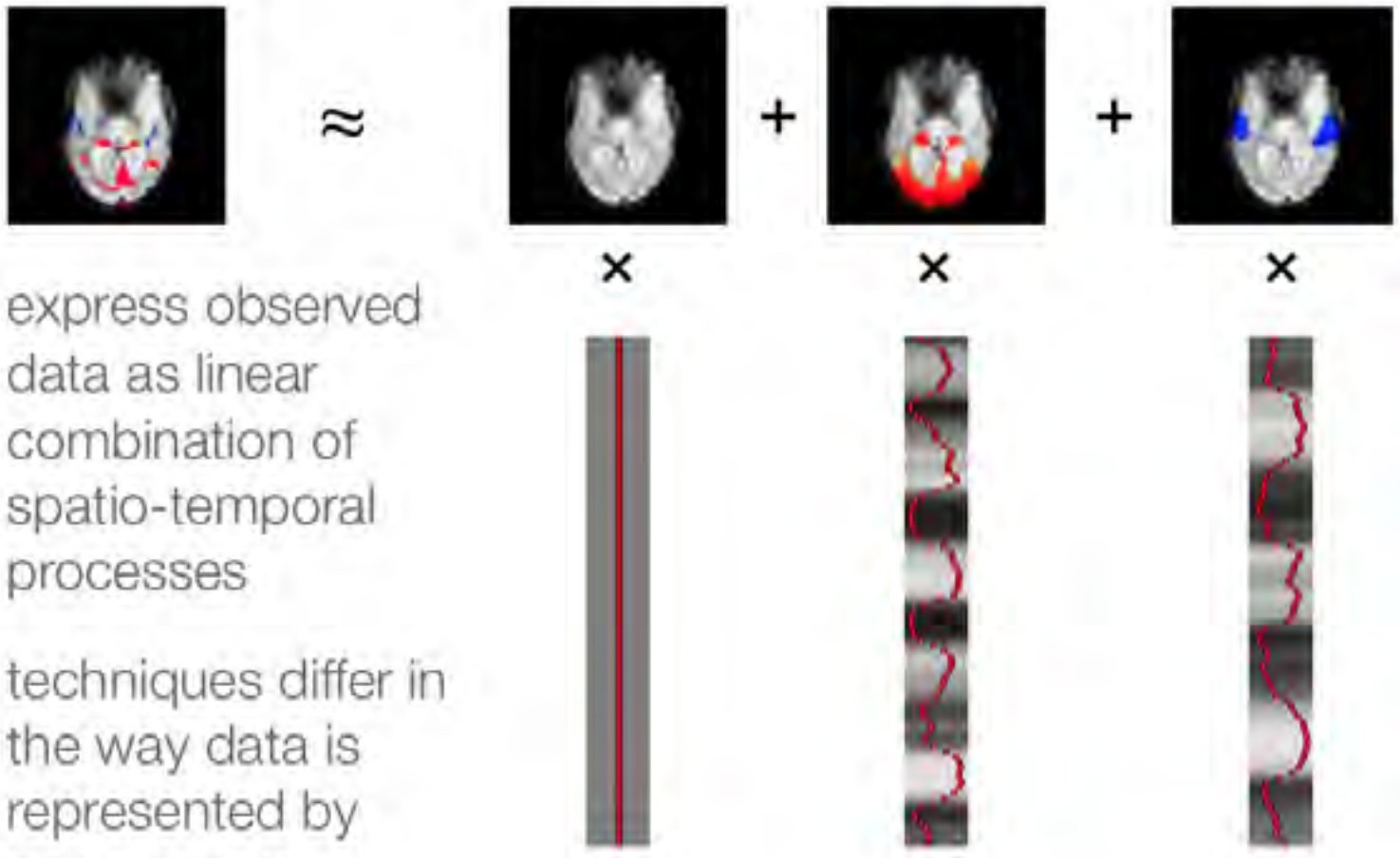


Calhoun VD, Cadali T, et al, (2001) Human Brain Mapping.

Independent spatial components, modulated in time

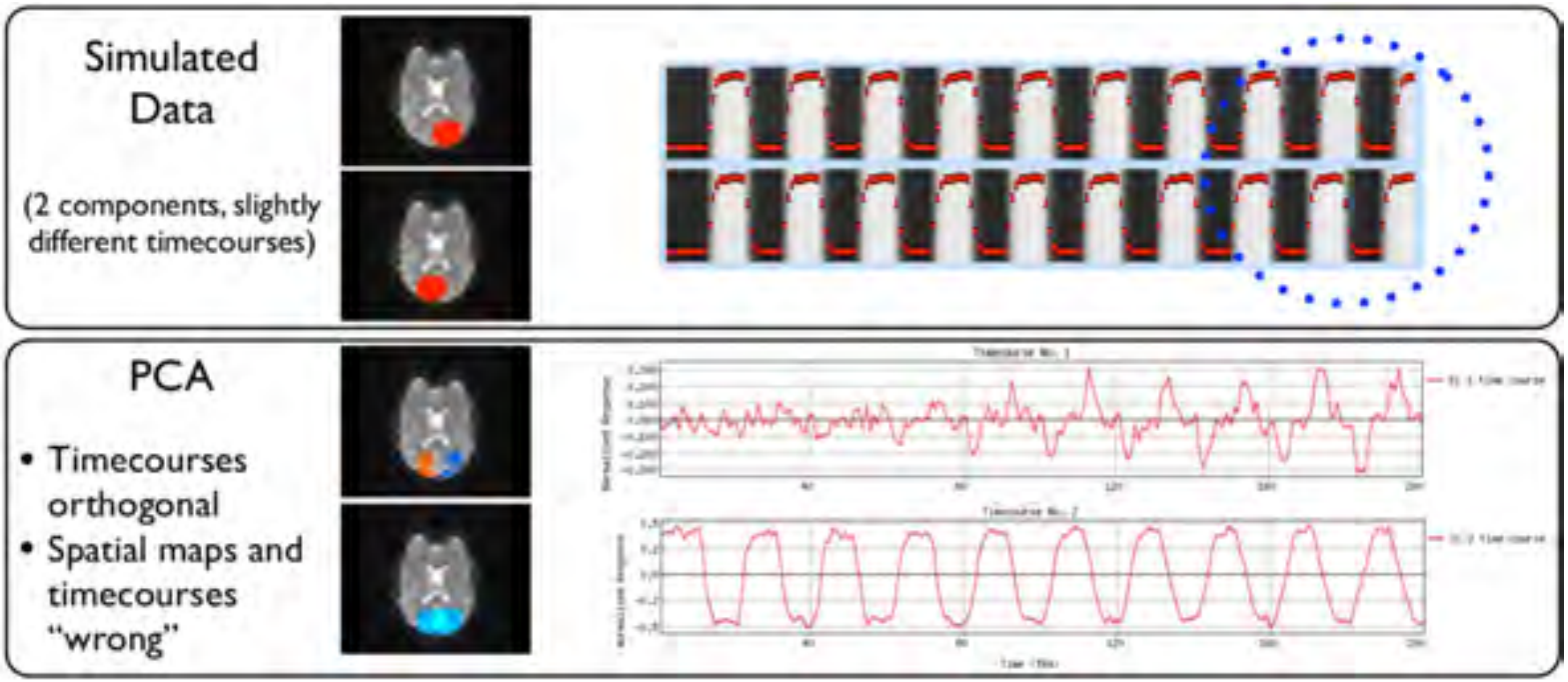


Caution: typical number of voxels is far greater than number of time-points!

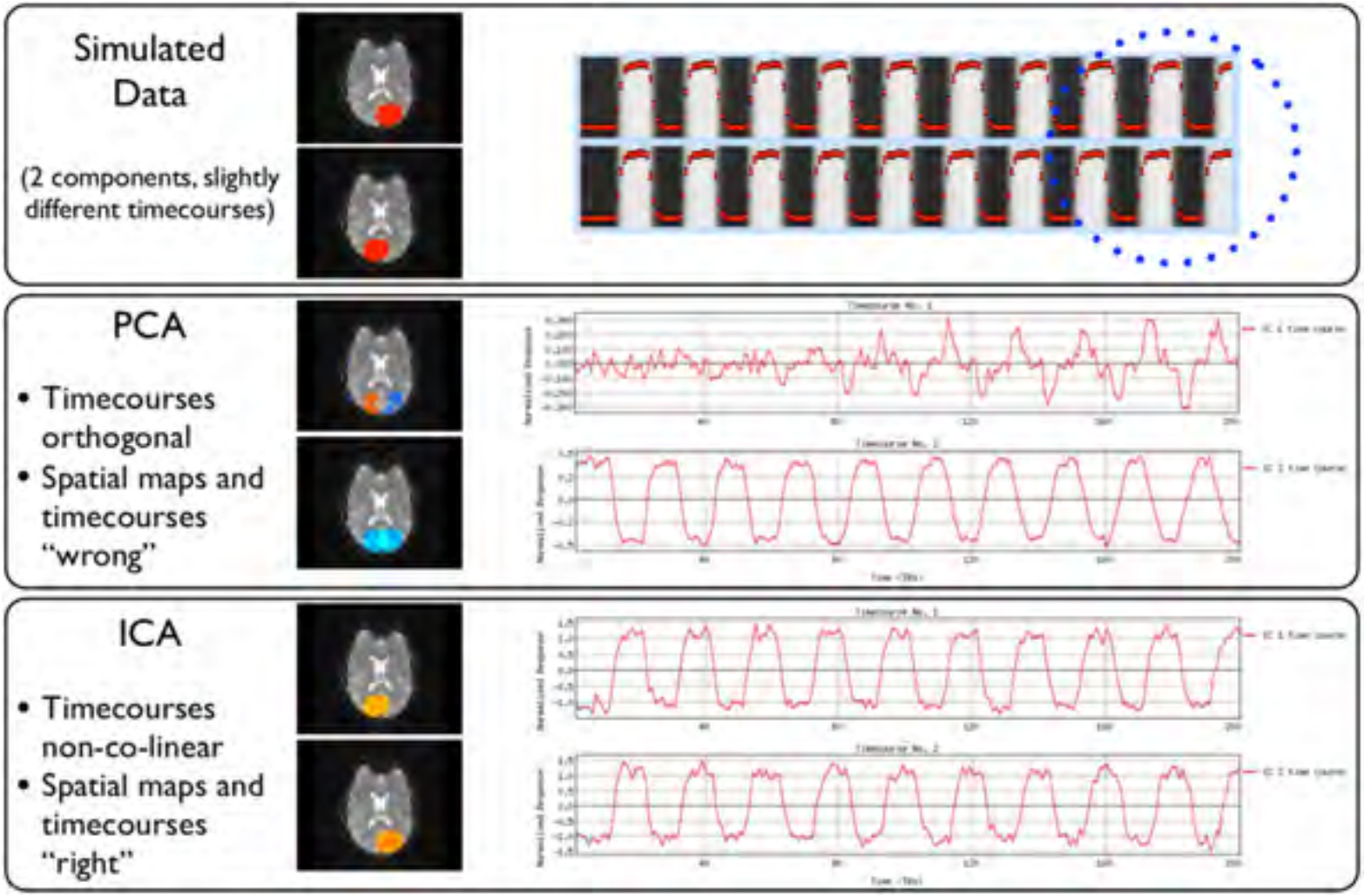


- express observed data as linear combination of spatio-temporal processes
- techniques differ in the way data is represented by components

Comparison PCA and ICA



Orthogonality assumption decomposes phase-shifted time-courses into one time-course (first principal component) and one differential (secondary principal component). Associated spatial maps depict region of commonality and regions of differential.



ICA correctly identifies independent (but non-orthogonal) time-courses and spatial maps.

Points to note

- ICA is used widely in signal processing.
- Functional brain imaging is a popular area of application.
- Spatial ICA, temporal ICA, probabilistic ICA (PICA).
- As all computational methods, ICA works well only when it is appropriate.

The aim of this course was to enable you to judge for yourself which computations methods are, or are not, appropriate for the analysis of your observations!

Moral:

Trust yourself, not authority / literature.