A Short Introduction to

Independent Component Analysis

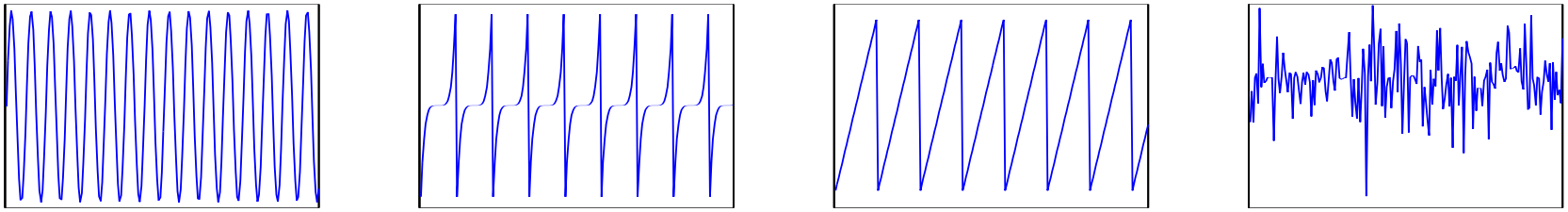
with Some Recent Advances

Aapo Hyvärinen

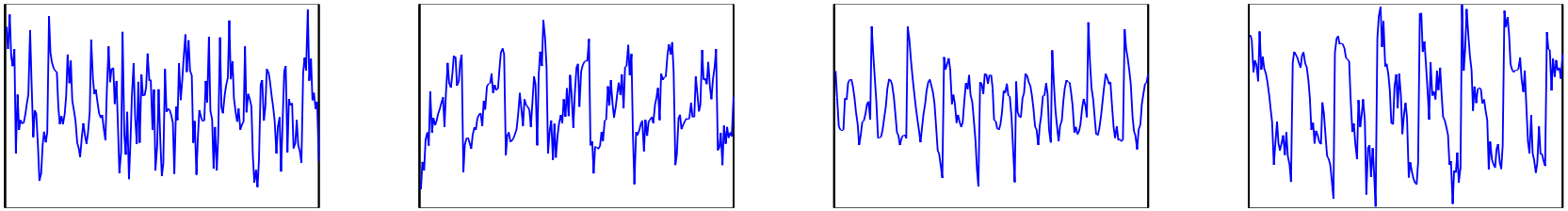
Dept of Computer Science
Dept of Mathematics and Statistics
University of Helsinki

Problem of blind source separation

There is a number of “source signals”:

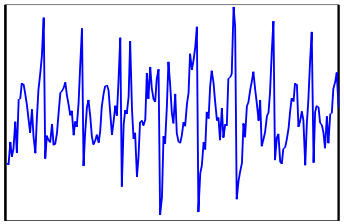
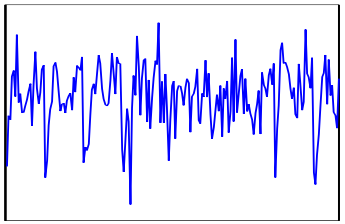
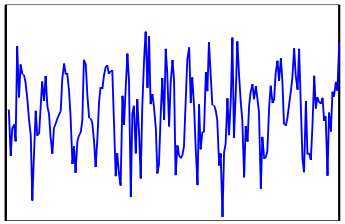
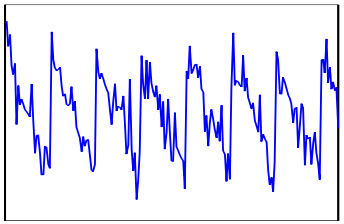


Due to some external circumstances,
only linear mixtures of the source signals are observed:

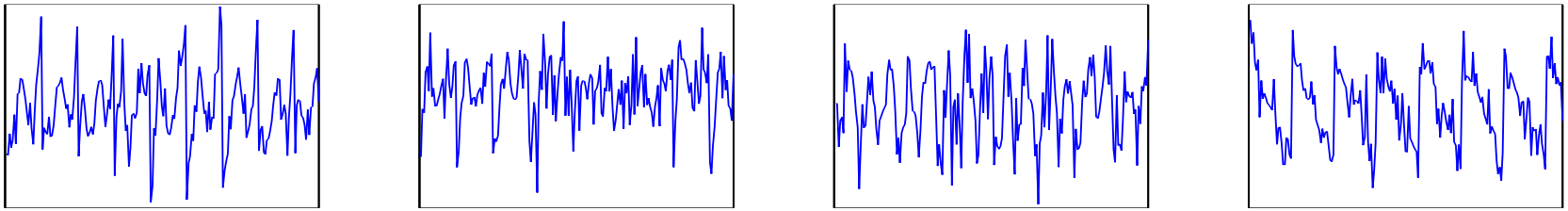


Estimate (separate) original signals!

Principal component analysis does not recover original signals

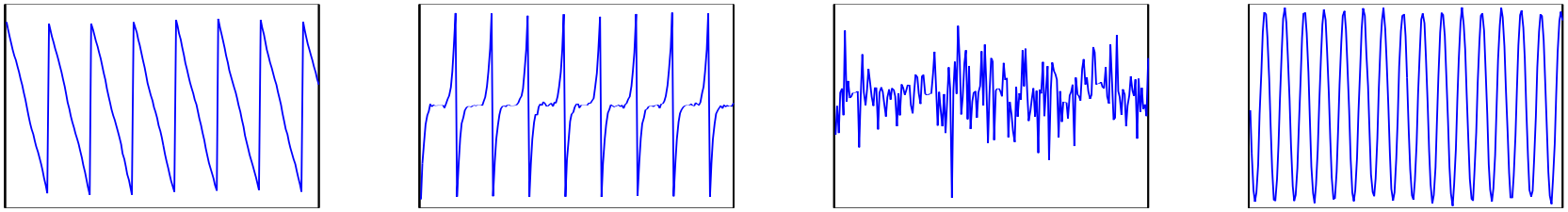





Principal component analysis does not recover original signals



A solution is possible

Use information on statistical independence to recover:



Independent Component Analysis

(Hérault and Jutten, 1984-1991)

- Observed data $x_i(t)$ is modelled using hidden variables $s_i(t)$:

$$x_i(t) = \sum_{j=1}^{m} a_{ij} s_j(t), \qquad i = 1 \ldots n \tag{1}$$

  or as a matrix decomposition

$$\mathbf{X} = \mathbf{AS} \tag{2}$$

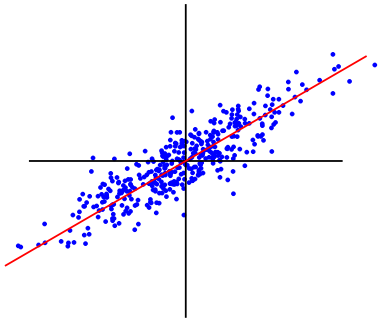
- Matrix of a_{ij} is constant parameter called "mixing matrix"
- Hidden random factors $s_i(t)$ are called "independent components" or "source signals"
- Problem: Estimate both a_{ij} and $s_j(t)$, observing only $x_i(t)$
  - Unsupervised, exploratory approach

When can the ICA model be estimated?

- Must assume:
  - The s_i are mutually statistically independent
  - The s_i are nongaussian (non-normal)
  - (Optional:) Number of independent components is equal to number of observed variables
- Then: mixing matrix and components can be identified (Comon, 1994)
  A very surprising result!

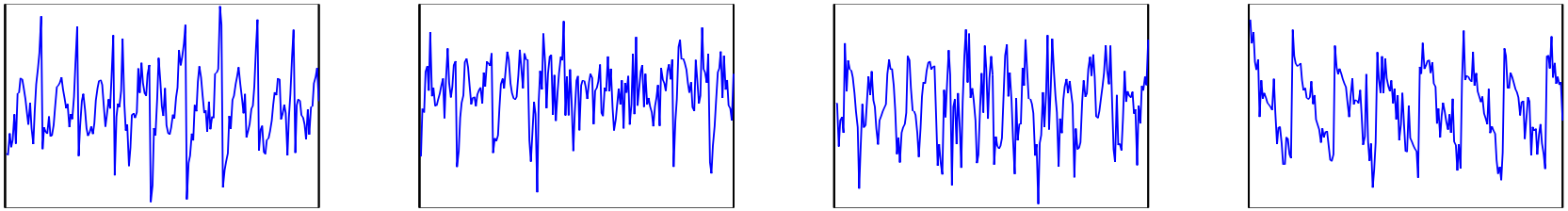
Background: Principal component analysis and factor analysis

- Basic idea: find directions $\sum_i w_i x_i$ of maximum variance



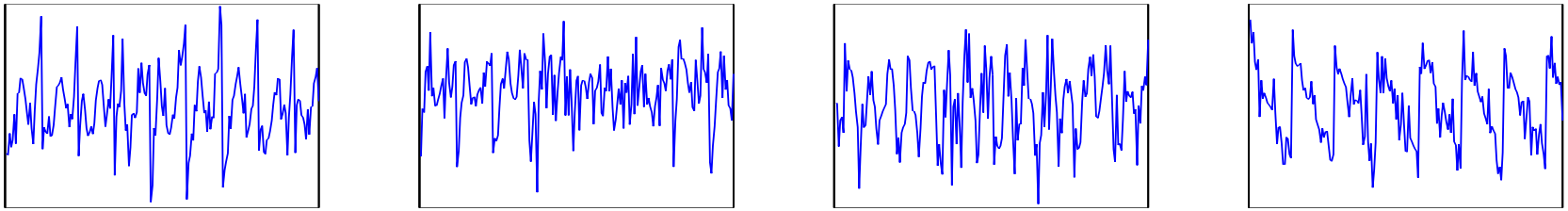
- Goal: explain maximum amount of variance with few components
  - Noise reduction
  - Reduction in computation
  - Easier to interpret (?)
  - (Vain hope: finds original underlying components)

Why cannot PCA or FA find source signals (original components)?



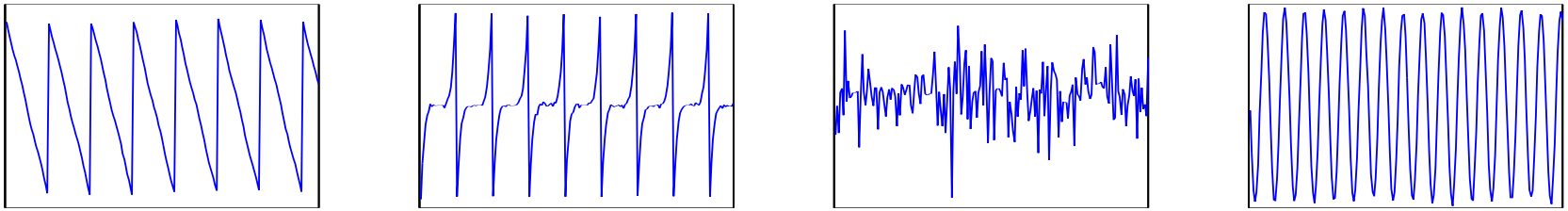
- They use only covariances $\mathrm{cov}(x_i, x_j)$
- Due to symmetry $\mathrm{cov}(x_i, x_j) = \mathrm{cov}(x_j, x_i)$, only $\approx n^2/2$ available
- Mixing matrix has n^2 parameters

Why cannot PCA or FA find source signals (original components)?



- They use only covariances $\mathrm{cov}(x_i, x_j)$
- Due to symmetry $\mathrm{cov}(x_i, x_j) = \mathrm{cov}(x_j, x_i)$, only $\approx n^2/2$ available
- Mixing matrix has n^2 parameters
- So, not enough information in covariances
  - "Factor rotation problem": Cannot distinguish between
    (s_1, s_2) and $(\sqrt{2}s_1 + \sqrt{2}s_2\,,\ \sqrt{2}s_1 - \sqrt{2}s_2)$

ICA uses nongaussianity to find source signals



- Gaussian (normal) distribution is completely determined by covariances
- But: Nongaussian data has structure beyond covariances
  - e.g. $E\{x_i x_j x_k x_l\}$ instead of just $E\{x_i x_j\}$
- Is it reasonable to assume data is nongaussian?
  - Many variables may be gaussian because sum of many independent variables (central limit theorem), e.g. intelligence
  - Fortunately, signals measured by physical sensors are usually quite nongaussian

Some examples of nongaussianity in signals

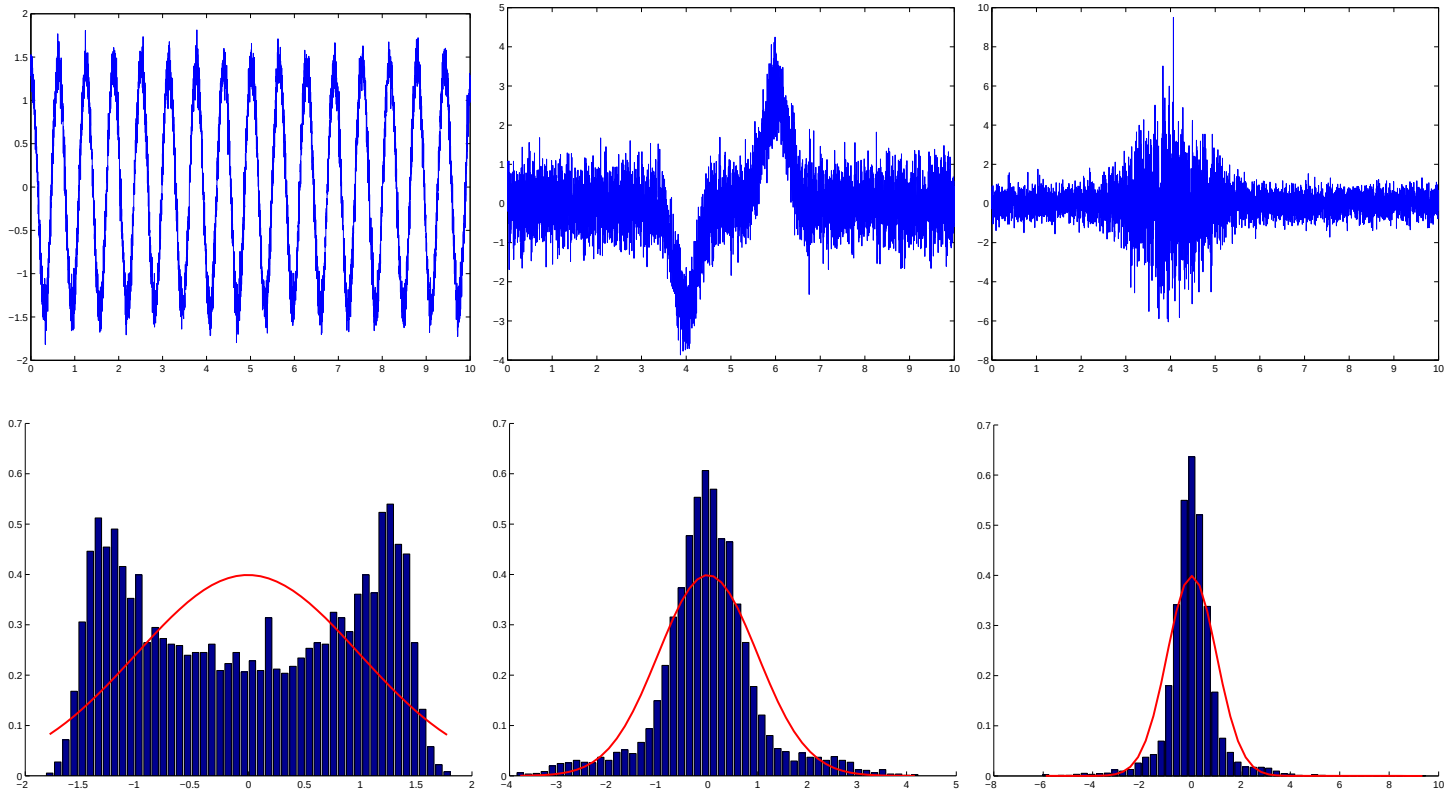
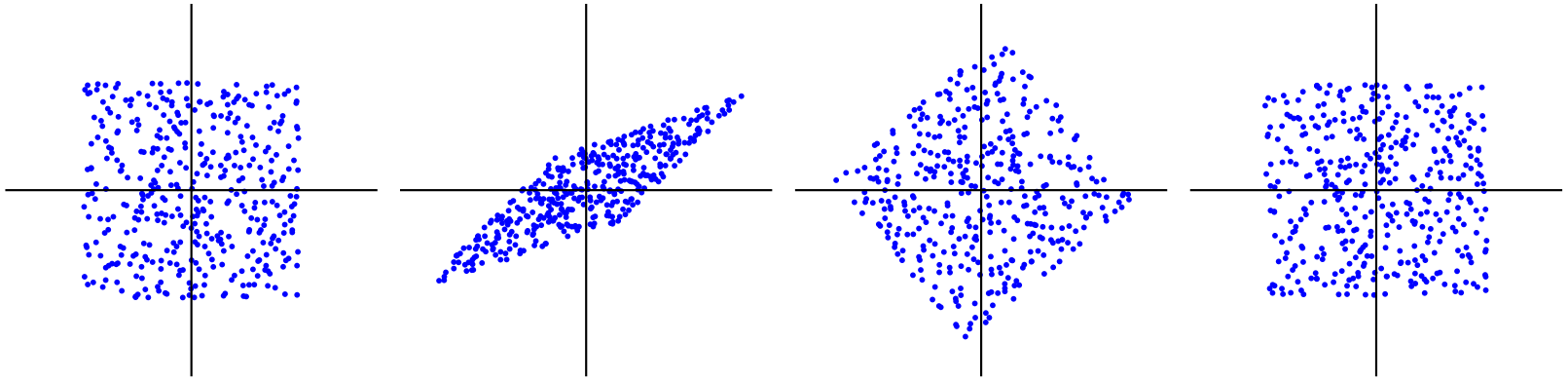


Illustration of PCA vs. ICA with nongaussian data

Two components with uniform distributions:

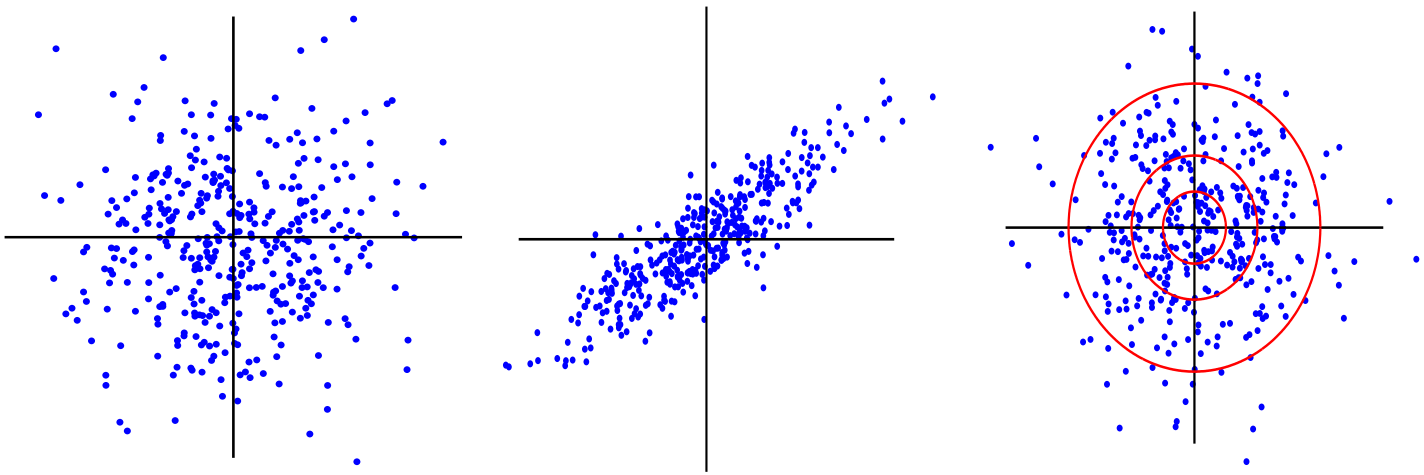
Original components, observed mixtures, PCA, ICA



PCA does not find original coordinates, ICA does!

Illustration of PCA vs. ICA with gaussian data

Original components, observed mixtures, PCA



Distribution after PCA is the same as distribution before mixing!

$\Rightarrow$“Factor rotation problem”

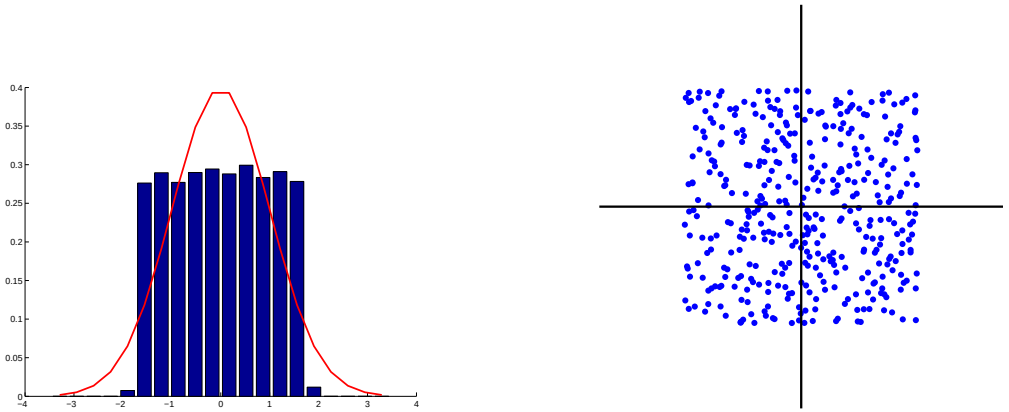
How is nongaussianity used in ICA ?

- Classic Central Limit Theorem:
  - Average of many independent random variables will have a distribution that is close(r) to gaussian
- So, roughly: Any mixture of components will be more gaussian than the components themselves

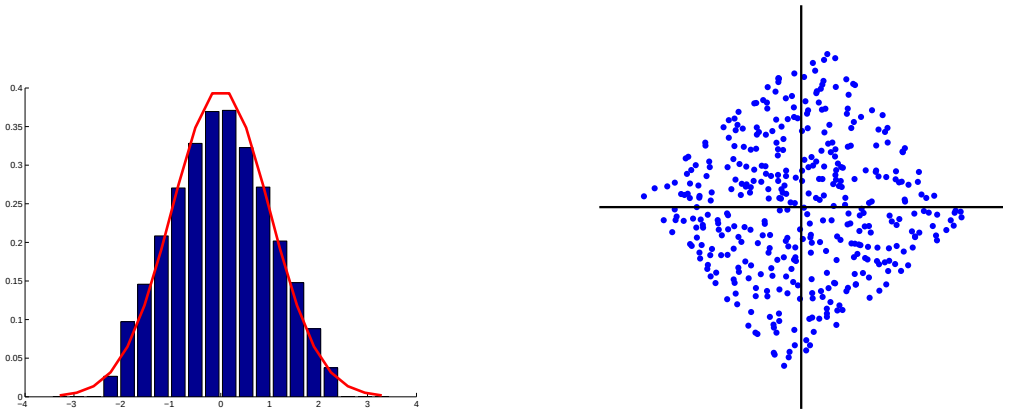
How is nongaussianity used in ICA ?

- Classic Central Limit Theorem:
  - Average of many independent random variables will have a distribution that is close(r) to gaussian
- So, roughly: Any mixture of components will be more gaussian than the components themselves
- Maximizing the nongaussianity of $\sum_i w_i x_i$, we can find s_i.
- Also known as projection pursuit.
- Cf. principal component analysis: maximize variance of $\sum_i w_i x_i$.

Illustration of changes in nongaussianity



Histogram and scatterplot, original uniform distributions

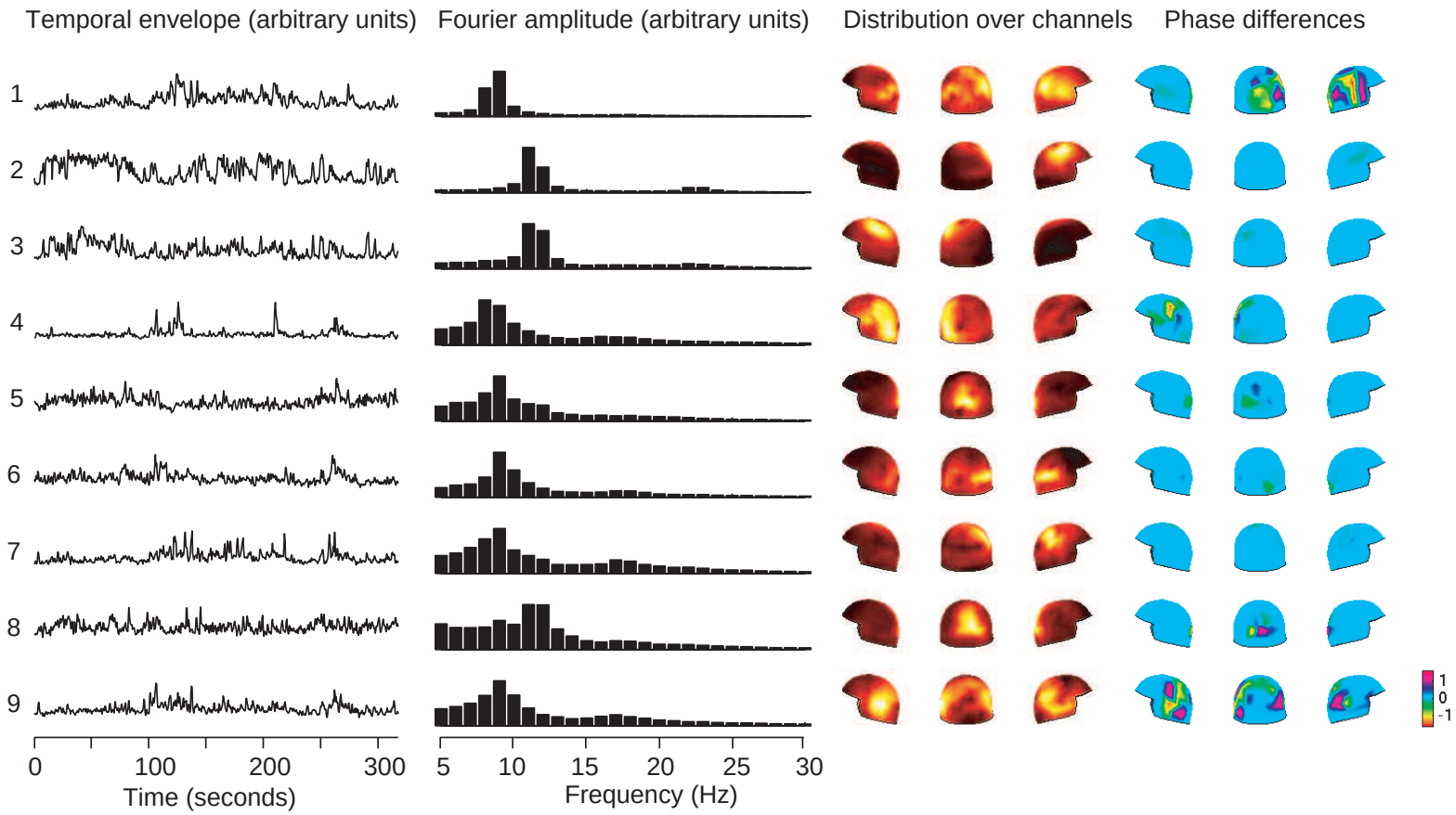


Histogram and scatterplot, mixtures given by PCA

ICA algorithms consist of two ingredients

1. Nongaussianity measure
   - Kurtosis: a classic, but sensitive to outliers.
   - Differential entropy: statistically better, but difficult to compute.
   - Approximations of entropy: a practical compromise.
2. Optimization algorithm
   - Gradient methods: natural gradient, “infomax”
     (Bell, Sejnowski, Amari, Cichocki et al 1994-6)
   - Fixed-point algorithm: FastICA (Hyvärinen, 1999)

Example of separated components in brain signals (MEG)



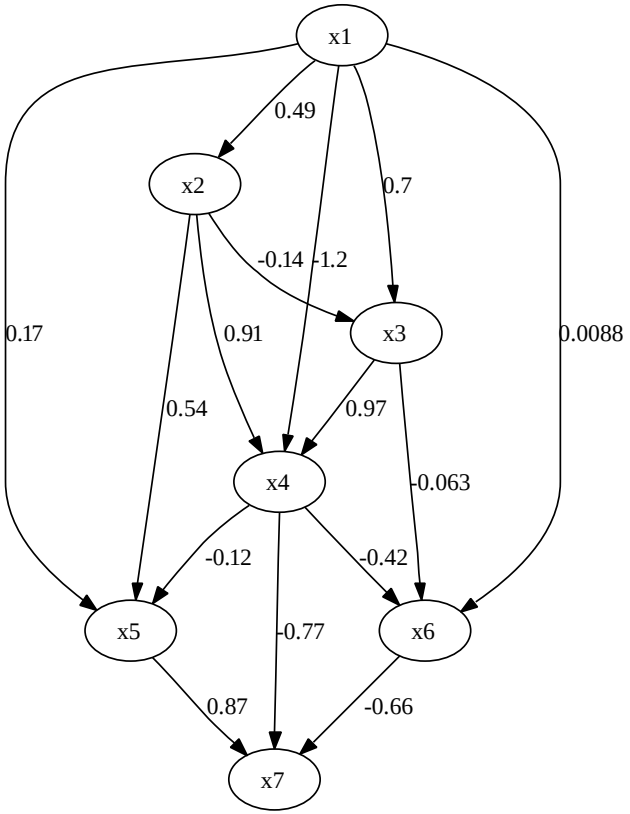
(Hyvärinen, Ramkumar, Parkkonen, Hari, *NeuroImage*, 2010)

Recent Advance 1: Causal analysis

- A structural equation model (SEM) analyses causal connections as

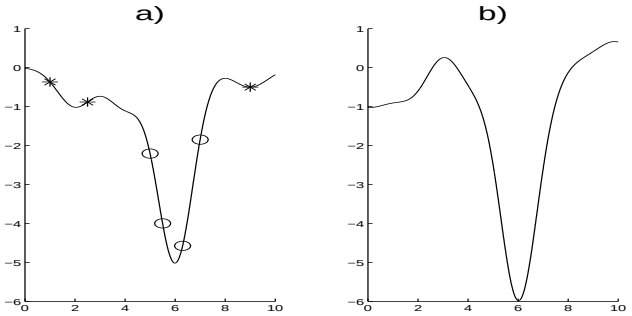
$$x_i = \sum_{j \neq i} b_{ij} x_j + n_i$$

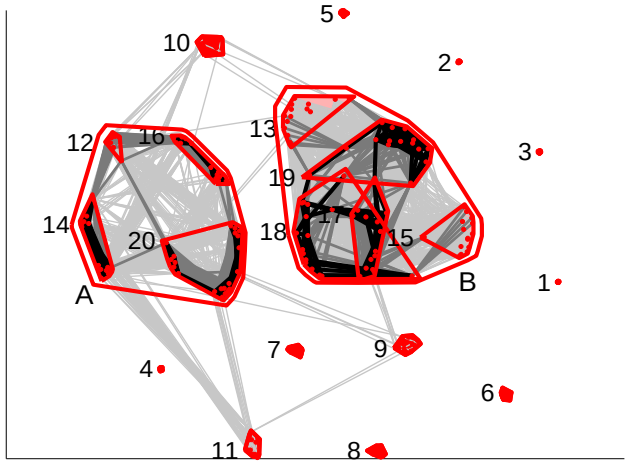
- Cannot be estimated in gaussian case without further assumptions
- Again, nongaussianity solves the problem (Shimizu et al, 2006)



Recent Advance 2: Analysing reliability of components (testing)

- Algorithmic reliability: Are there local minima?
  Can be analysed by rerunning from different initial points *(a)*
- Statistical reliability: Is the result just random/accidental?
  Can be analyzed by bootstrap *(b)*
- Our *Icasso* package (Himberg et al, 2004) visualizes reliability:





- New: A proper testing procedure which gives p-values (Hyvärinen, 2011)

Recent Advance 3: Three-way data ("Group ICA")

- Often ones measures several data matrices $\mathbf{X}_k, k = 1, \ldots, r$, e.g. different conditions, measurement sessions, subjects, etc.
- Each matrix could be modelled $\mathbf{X}_k = \mathbf{A}_k \mathbf{S}_k$
  - but how to connect the results? (Calhoun, 2001)

a) If mixing matrices same for all $\mathbf{X}_k$, use

$$\begin{aligned} \mathcal{X}_1 &= \left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_r \right) \\ &= \mathbf{A} \left(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_r \right) \end{aligned}$$

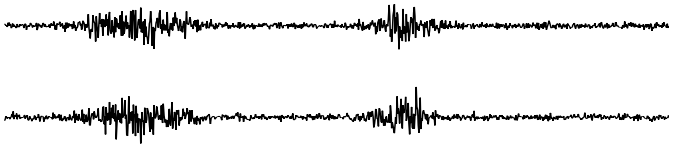
b) If component values same for all $\mathbf{X}_k$, use

$$\mathcal{X}_2 = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_r \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_r \end{pmatrix} \mathbf{S}$$

- If both $\mathbf{A}_k$ and $\mathbf{S}_k$ the same: analyse average data matrix, or combine ICA with PARAFAC (Beckmann and Smith, 2005).

Recent Advance 4: Dependencies between components

- In fact, estimated components are often not independent
  - ICA does not have enough parameters to force independence
- Many authors model correlations of squares, or “simultaneous activity”



Two signals that are uncorrelated but whose squares are correlated.

- On-going work on even linearly correlated components (Sasaki et al, 2011)
- Alternatively, in parallel time series, innovations of VAR model could be independent (Gómez-Herrero et al, 2008)

Recent Advance 5: Better estimation of basic linear mixing

- In case of time signals, we can do ICA on time-frequency decompositions (Pham, 2002; Hyvärinen et al, 2010)
- If the data is by its very nature non-negative, we could impose the same in the model (Hoyer, 2004)
  - Zero must have some special meaning as a baseline
  - E.g. Fourier spectra
- More precise modelling of nongaussianity of components could also improve estimation.

Summary

- ICA is a very simple model: Simplicity implies wide applicability.
- A nongaussian alternative to PCA or factor analysis.
- Finds a linear decomposition by maximizing nongaussianity of the components.
  - These (hopefully) correspond to the original sources
- Recent advances:
  - Causal analysis, or structural equation modelling, using ICA
  - Testing of independent components for statistical significance
  - Group ICA, i.e. ICA on three-way data
  - Modelling dependencies between components
  - Imporovements in estimating the basic linear mixing model
- “Nongaussianity is beautiful” !?