A Short Introduction to

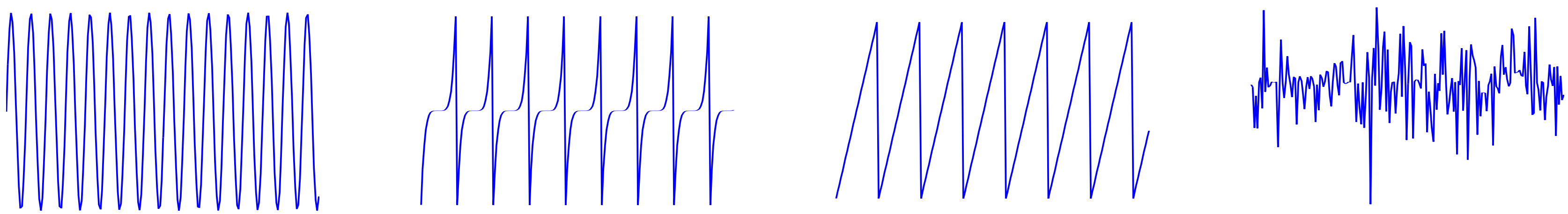
Independent Component Analysis

Aapo Hyvärinen

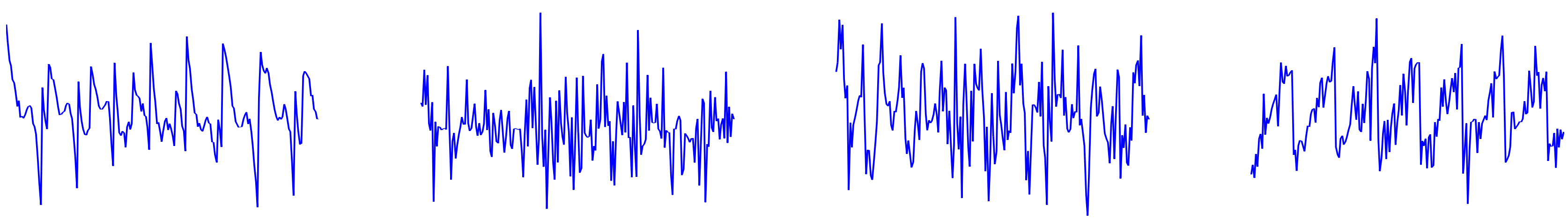
Helsinki Institute for Information Technology and
Depts of Computer Science and Psychology
University of Helsinki

Problem of blind source separation

There is a number of “source signals”:

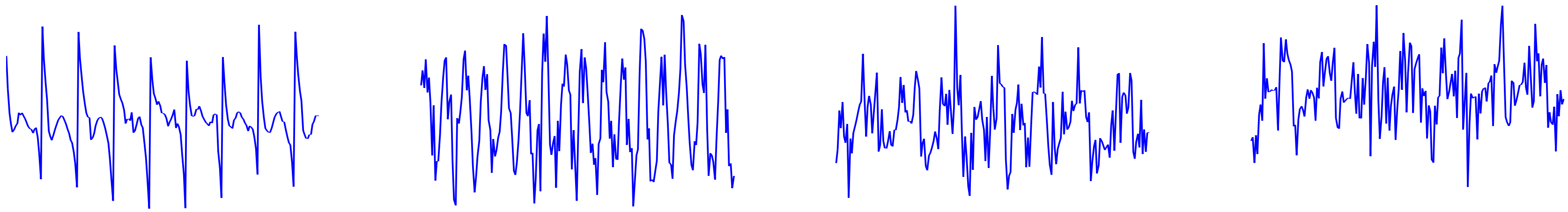


Due to some external circumstances, only linear mixtures of the source signals are observed.



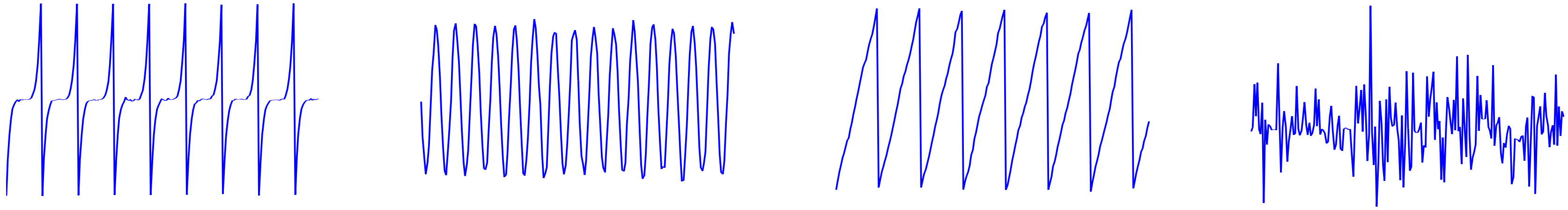
Estimate (separate) original signals!

Principal component analysis does not recover original signals



A solution is possible

Use information on statistical independence to recover:



Independent Component Analysis

(Hérault and Jutten, 1984-1991)

- Observed random variables x_i are modelled as linear sums of hidden variables:

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

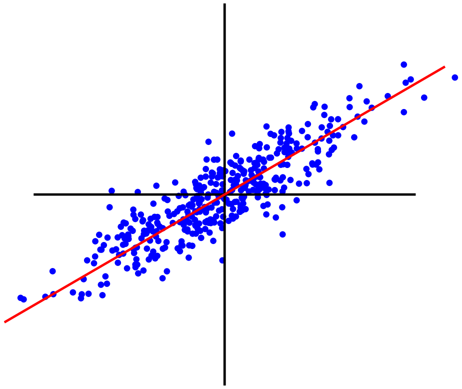
- Mathematical formulation of blind source separation problem
- A form of factor analysis
- Matrix of a_{ij} is constant (factor loadings), called “mixing matrix”.
- The s_i are hidden random factors called “independent components”, or “source signals”
- Problem: Estimate both a_{ij} and s_j, observing only x_i.

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!

Reminder: Principal component analysis

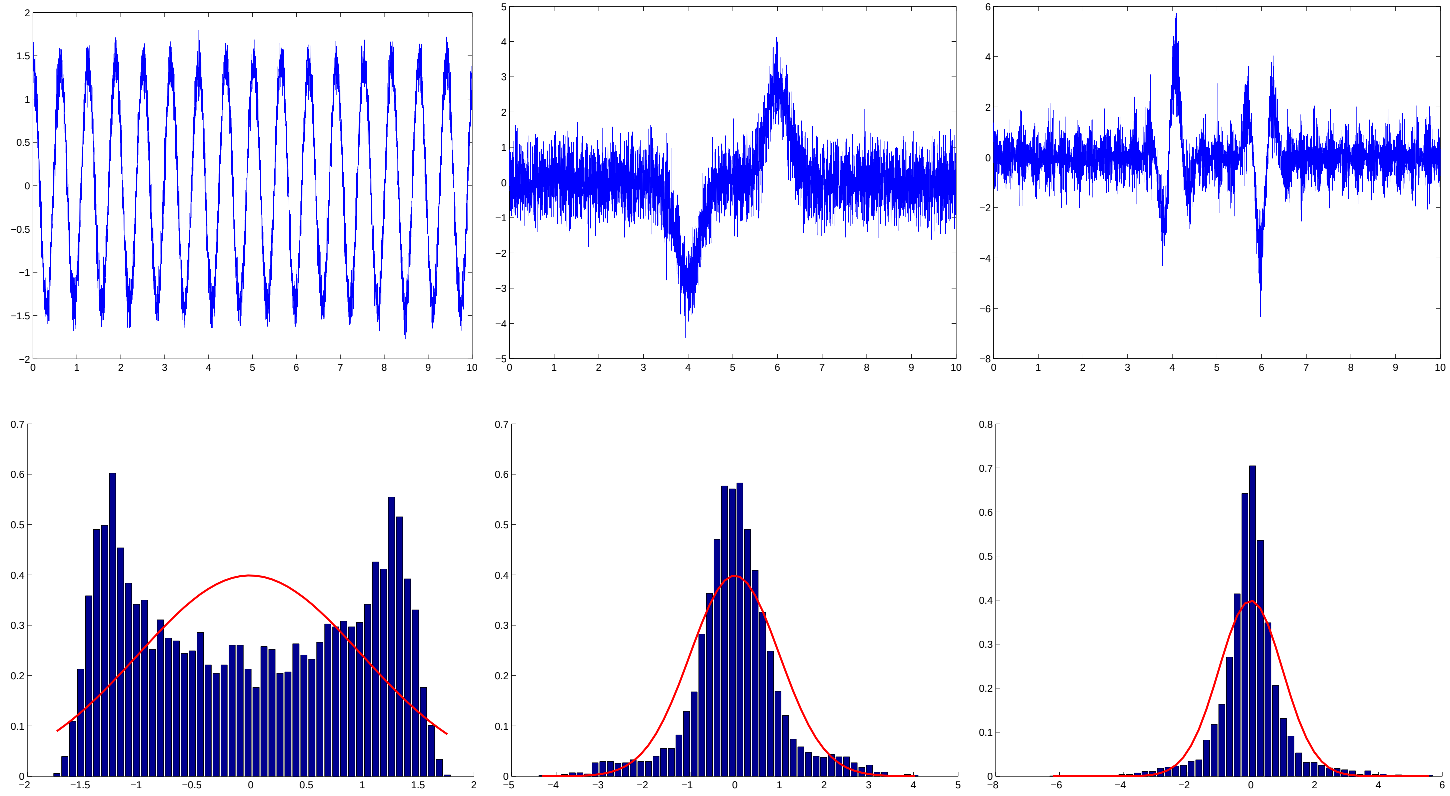
- Basic idea: find directions $\sum_i w_i x_i$ of maximum variance
- We must constrain the norm of $\mathbf{w}$: $\sum_i w_i^2 = 1$, otherwise solution is that w_i are infinite.
- For more than one component, find direction of max var orthogonal to components previously found.
- Classic factor analysis has essentially same idea as in PCA: explain maximal variance with limited number of components



Comparison of ICA, factor analysis and principal component analysis

- ICA is nongaussian FA with no separate noise or specific factors. So many components used that all variance is explained by them.
- No factor rotation left unknown because of identifiability result
- In contrast to FA and PCA, components really give the original source signals or underlying hidden variables
- Catch: only works when components are nongaussian
  - Many "psychological" hidden variables (e.g. "intelligence") may be (practically) gaussian because sum of many independent variables (central limit theorem).
  - But signals measured by sensors are usually quite nongaussian

Some examples of nongaussianity



Why classic methods cannot find original components or sources

- In PCA and FA: find components y_i which are uncorrelated

$$\mathrm{cov}(y_i, y_j) = E\{y_i y_j\} - E\{y_i\}E\{y_j\} = 0 \tag{2}$$

  and maximize explained variance (or variance of components)

- Such methods need only the covariances, $\mathrm{cov}(x_i, x_j)$
- However, there are many different component sets that are uncorrelated, because
  - The number of covariances is $\approx n^2/2$ due to symmetry
  - So, we cannot solve the n^2 factor loadings, not enough information! ("More equations than variables")
- This is why PCA and FA cannot find the underlying components (in general)

Nongaussianity, combined with independence, gives more information

- For independent variables we have

$$E\{h_1(y_1)h_2(y_2)\} - E\{h_1(y_1)\}E\{h_2(y_2)\} = 0. \quad (3)$$

- For nongaussian variables, nonlinear covariances give more information than just covariances.

- This is not true for multivariate gaussian distribution
  - Distribution is completely determined by covariances (and means)
  - Uncorrelated gaussian variables are independent, and their
  - distribution (standardized) is same in all directions (see below)

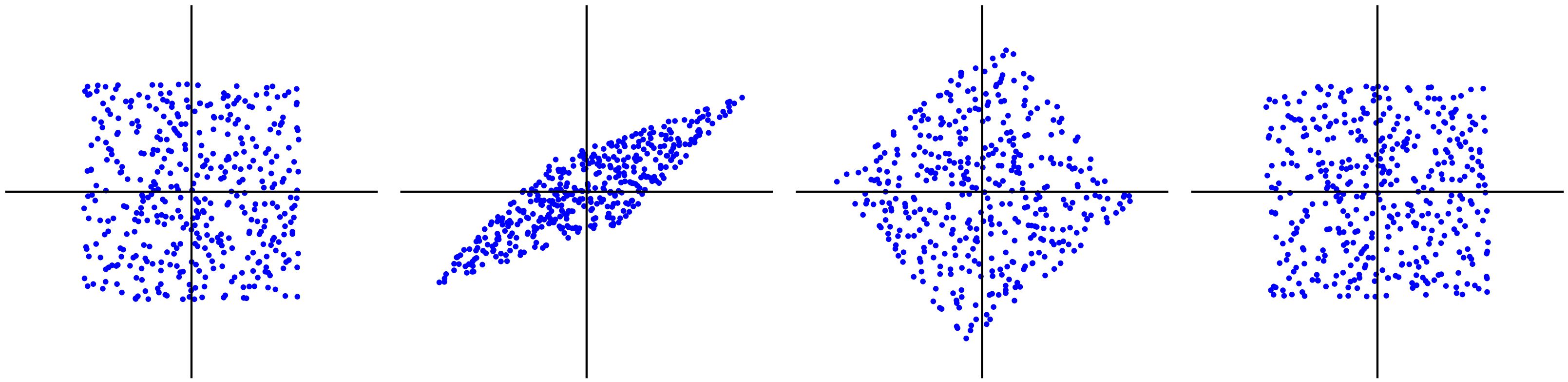
  $\Rightarrow$ ICA model cannot be estimated for gaussian data.

- In practice, simpler to look at properties of linear combinations $\sum_i w_i x_i$. PCA maximizes variance of $\sum_i w_i x_i$, can we do something better? Yes, see below.

Illustration

Two components with uniform distributions:

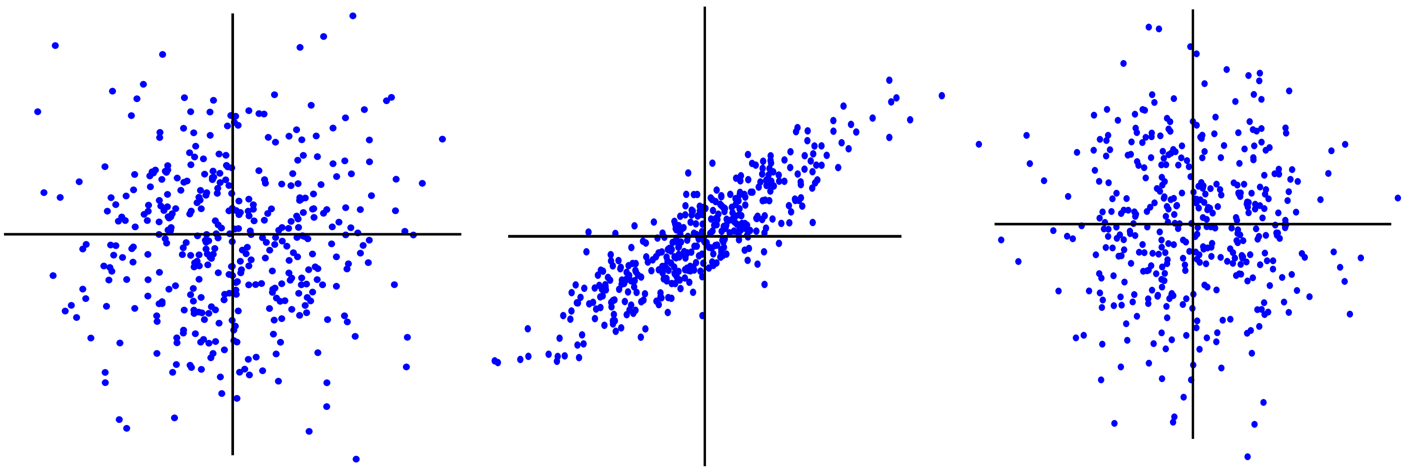
Original components, observed mixtures, PCA, ICA



PCA does not find original coordinates, ICA does!

Illustration of problem with gaussian distributions

Original components, observed mixtures, PCA



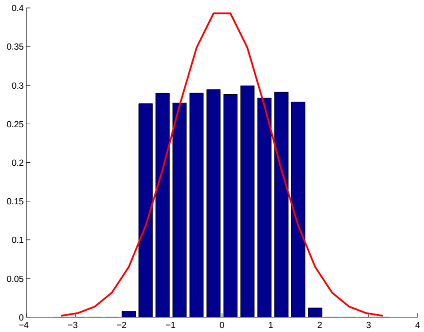
Distribution after PCA is the same as distribution before mixing!

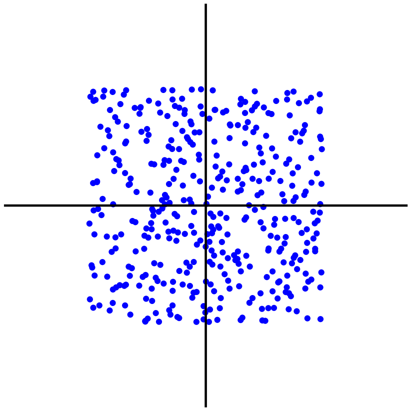
“Factor rotation problem” in classic FA

Basic intuitive principle of ICA estimation

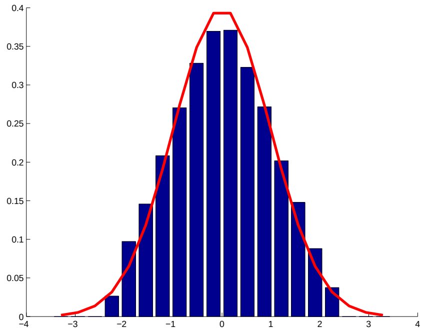
- Inspired the Central Limit Theorem:
  - Average of many independent random variables will have a distribution that is close(r) to gaussian
  - In the limit of an infinite number of random variables, the distribution tends to gaussian
- Consider a linear combination $\sum_i w_i x_i = \sum_i q_i s_i$
- Because of theorem, $\sum_i q_i s_i$ should be more gaussian than s_i.
- *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 $\sum_i w_i x_i$.

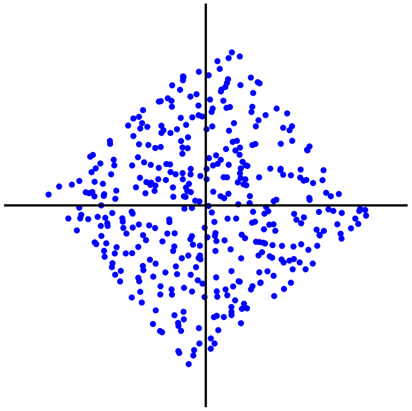
Illustration of changes in nongaussianity





Histogram and scatterplot, original uniform distributions





Histogram and scatterplot, mixtures given by PCA

Development of ICA algorithms

- Nongaussianity measure: Essential ingredient
  - Kurtosis: global consistency, but nonrobust.
  - Differential entropy / likelihood:
    statistically justified, but difficult to compute.
  - Rough approximations of entropy: good compromise.
- Optimization methods
  - Gradient methods (natural gradient, “infomax”)
  - Fast fixed-point algorithm, FastICA (Hyvärinen, 1999)
  - one-by-one estimation vs. estimation of all

Combining ICA with FA/PCA

- In practice, it is useful to combine ICA with classic PCA or FA
  - First, find a small number of factors with PCA or FA
  - Then, perform ICA on those factors
- ICA is then a method of factor rotation
- Very different from varimax etc. which do not use statistical structure, and cannot find original components (in most cases)
- Reduces noise in signals, reduces computation
- (Simplifies algorithms because we can constrain mixing matrix to be orthogonal.)

Preprocessing of data

- Prefiltering possible: ICA model still holds with the same matrix **A**

$$\tilde{x}_i(t) = f(t) * x_i(t) = \sum_{\tau} f(\tau) x_i(t-\tau) \tag{4}$$

$$\Rightarrow \tag{5}$$

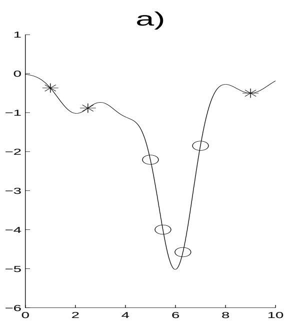
$$\tilde{x}_i(t) = \sum_{j} a_{ij} \tilde{s}_j(t) \tag{6}$$

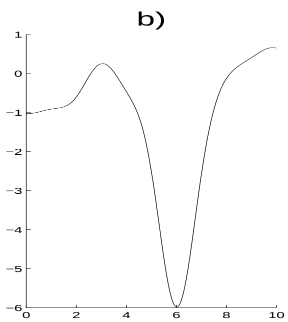
  One can try to find a frequency band in which the source signals are as independent and nongaussian as possible

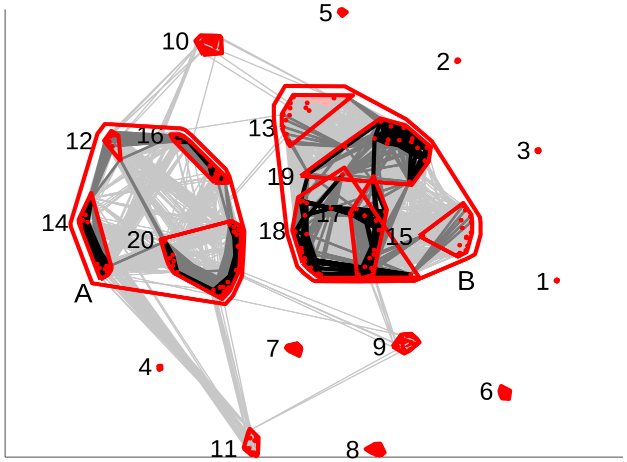
- (And: Dimension reduction by PCA)

Reliability analysis

- Algorithmic reliability: Are there local minima? (see *a)* below)
- Statistical reliability: Is the result just accidental?
  Can be analyzed by bootstrap but this changes local minima *b)*
- We have developed a package *Icasso* that uses computationally intensive methods to visualize and analyze these:

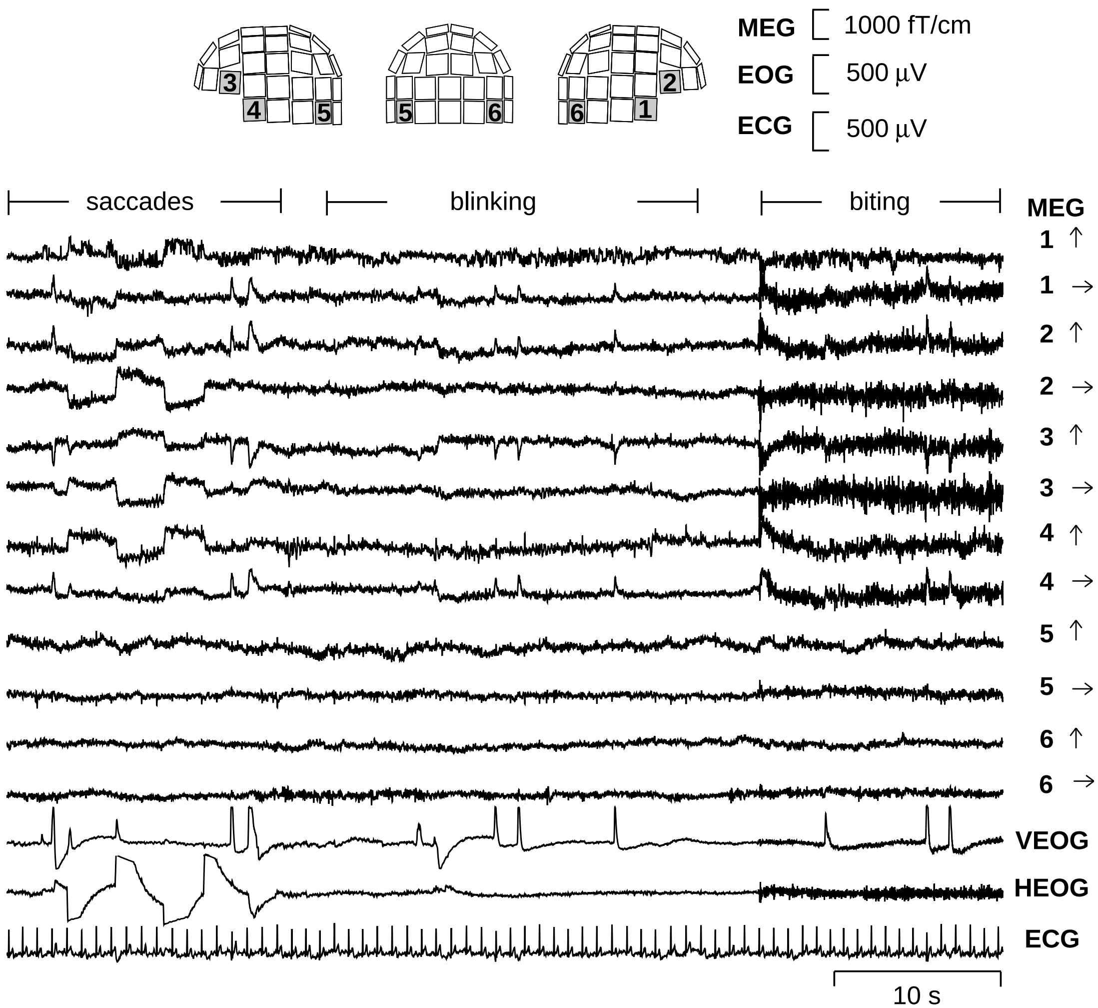




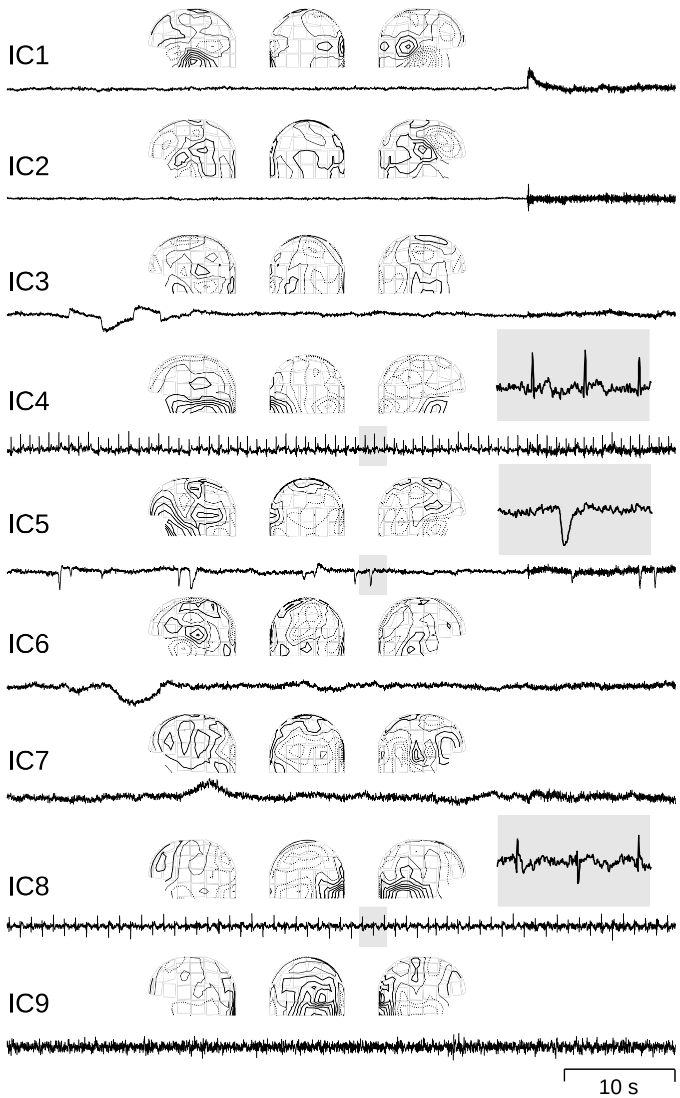


Applications

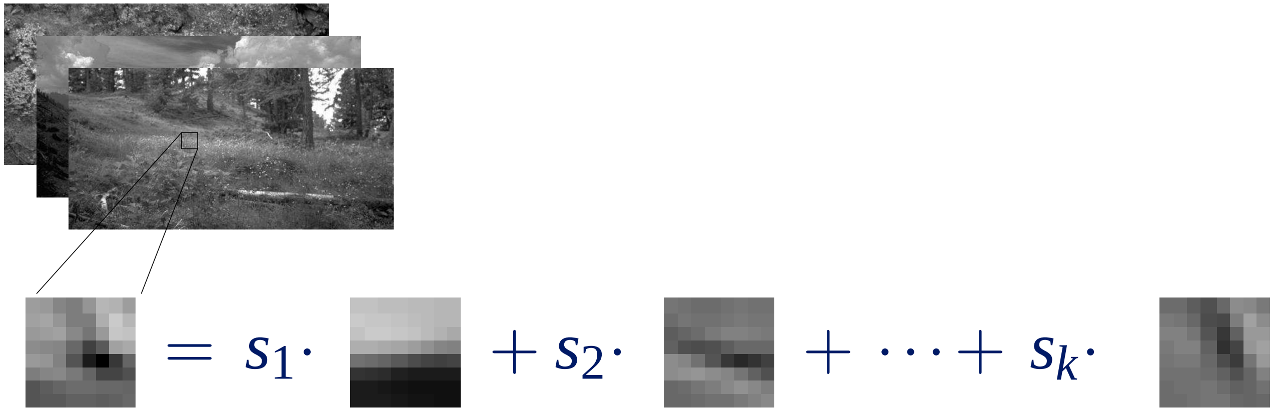
Application to MEG (Vigário et al, 1998)



Independent components of "spontaneous" MEG (Vigário et al, 1998)



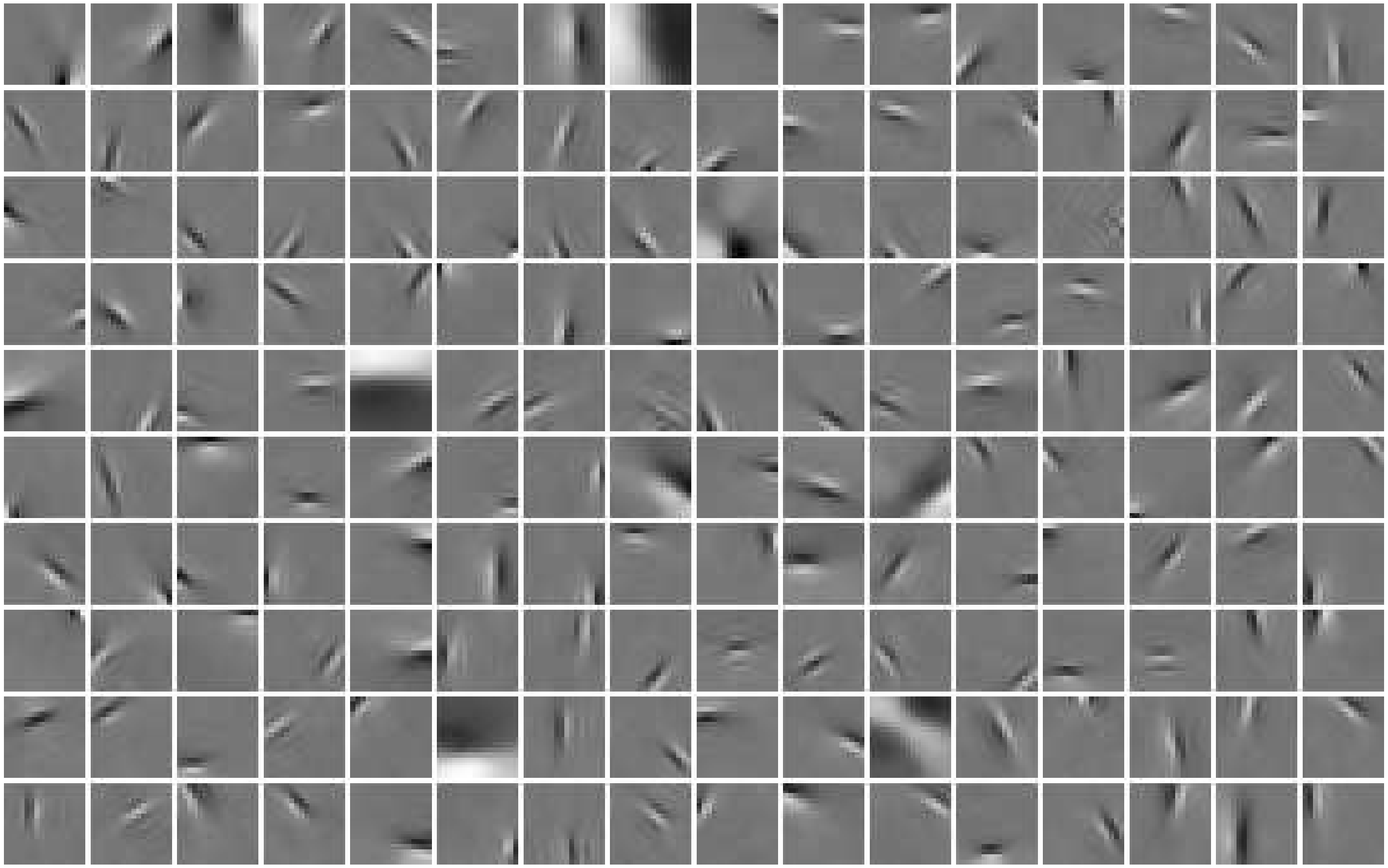
ICA in modelling visual cortex



- Why are the receptive fields in visual cortex the way they are?
- Statistical-ecological approach
  - What is important in a real environment?
  - Natural images have statistical regularities, "explain" receptive fields by statistical properties of natural images
  - ICA gives the "best" features natural images

ICA / sparse coding of natural images

(Olshausen and Field, 1996; Bell and Sejnowski, 1997)

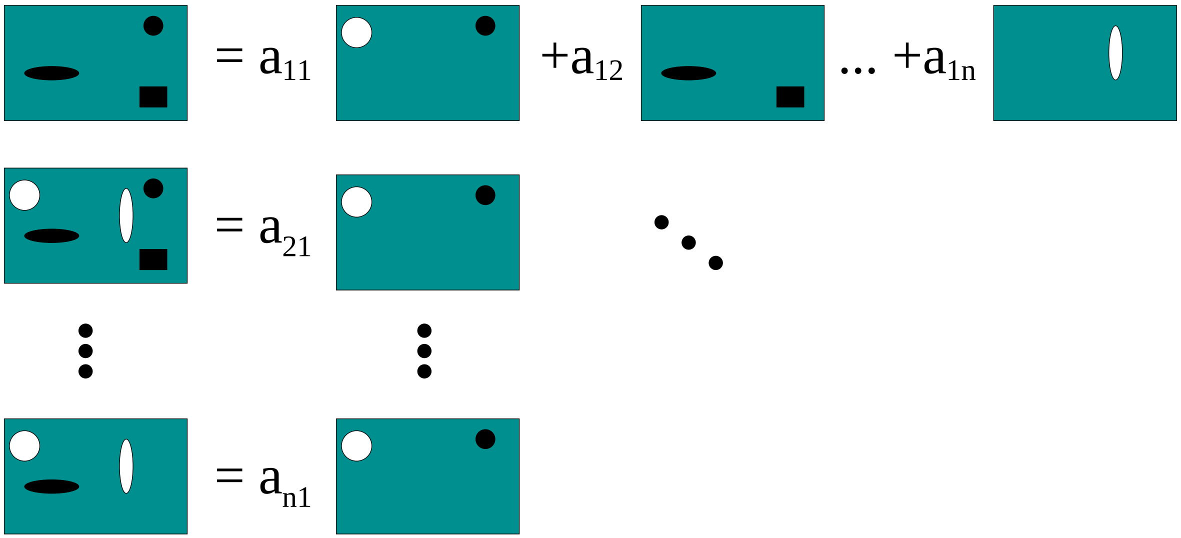


Features similar to receptive fields of simple cells in V1

More theory

ICA of brain images

- Assume we observe several brain images
  - at different time points, or
  - under different imaging conditions
- ICA expresses observed images as linear sums of "source images":



- Reverses the roles of observations and variables

Complication (1): Noisy ICA

- Assume there is (gaussian) sensor noise

$$x_i = \sum_j a_{ij} s_j + n_i \tag{7}$$

- Very difficult problem in general
- But trivial if noise covariance is the same as signal covariance:

$$\mathbf{x} = \mathbf{A}(\mathbf{s} + \mathbf{A}^{-1}\mathbf{n}) = \mathbf{A}\tilde{\mathbf{s}} \tag{8}$$

  the transformed components are independent!
- Or: if noise can be modelled by some components in **s**.
- In practice maybe the best thing to do: reduce noise by time filtering and/or PCA and use ordinary (noise-free) ICA algorithms.

Complication (2): different numbers of components and variables

- In the theoretical analysis, we assume the numbers are equal
- In practice, often we have more variables than components
  - simple solution (1): reduce dimension by PCA
  - simple solution (2): estimate only the k “first” components
- Another very difficult case: Less variables than independent components

Nongaussianity measures: kurtosis

- Problem: how to measure nongaussianity?
- Definition:

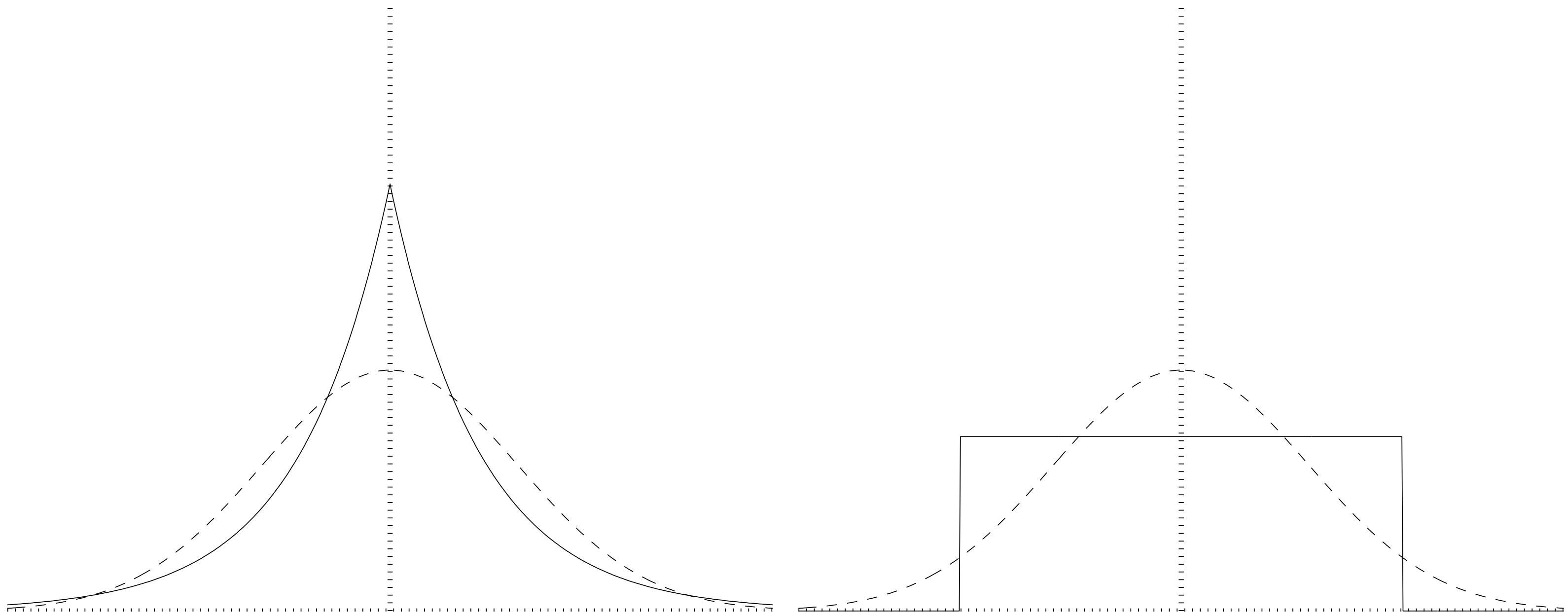
$$\text{kurt}(x) = E\{x^4\} - 3(E\{x^2\})^2 \tag{9}$$

- if variance constrained to unity, essentially 4th moment.
- Simple algebraic properties because it's a cumulant:

$$\text{kurt}(s_1 + s_2) = \text{kurt}(s_1) + \text{kurt}(s_2) \tag{10}$$

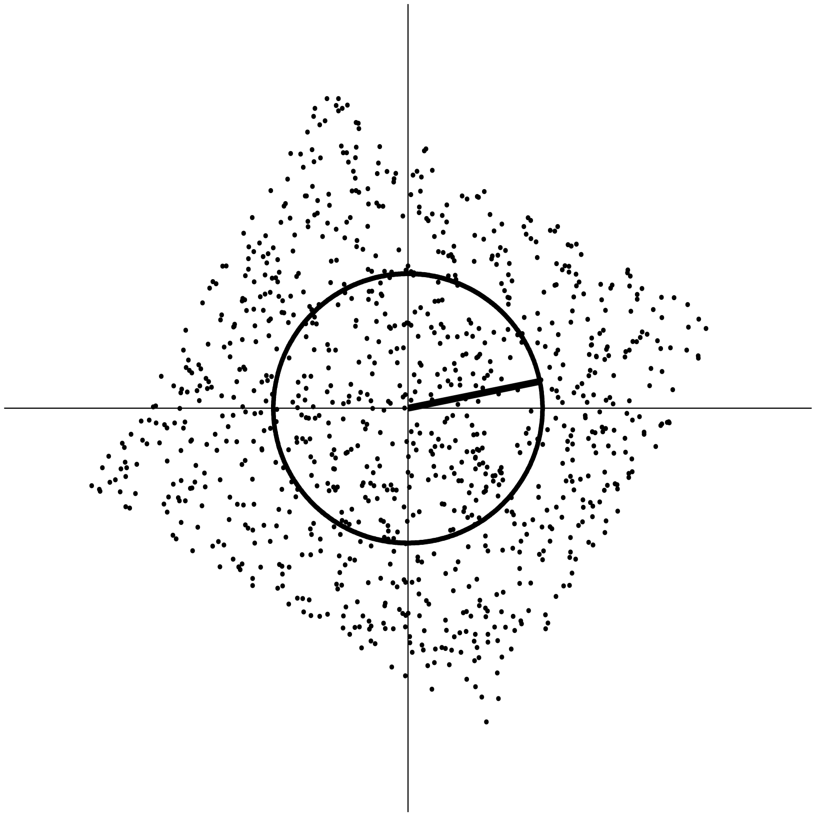
$$\text{kurt}(\alpha s_1) = \alpha^4 \, \text{kurt}(s_1) \tag{11}$$

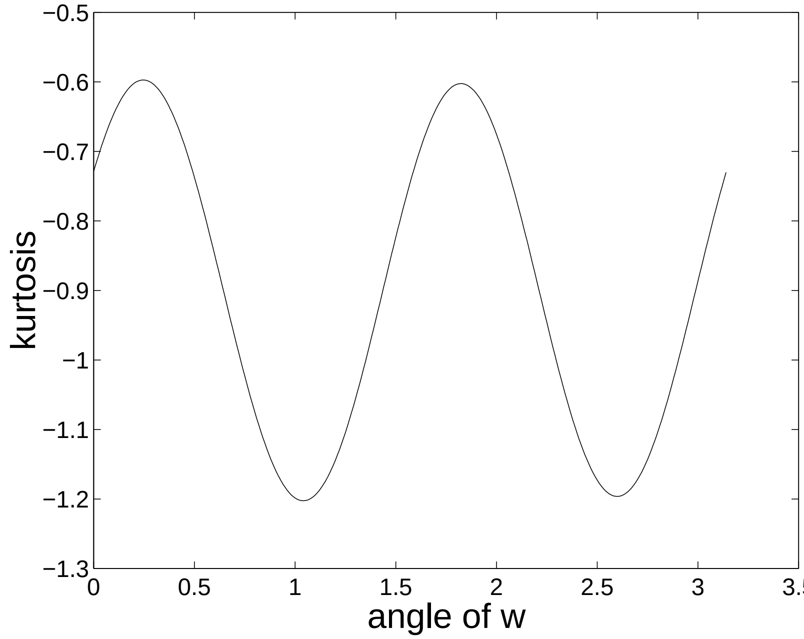
- zero for gaussian RV, non-zero for most nongaussian RV's.
- positive vs. negative kurtosis have typical forms of pdf.
- absolute value a classic measure of nongaussianity



Left: Laplacian pdf, positive kurt (“supergaussian”).

Right: Uniform pdf, negative kurt (“subgaussian”).





Kurtosis is minimized, and its absolute value maximized, in the directions of the independent components.

Why kurtosis is not optimal

- Sensitive to outliers:
  Consider a sample of 1000 values with unit var, and one value equal to 10.
  Kurtosis equals at least $10^4/1000 - 3 = 7$.
- For supergaussian variables, statistical performance not optimal even without outliers.
- Other measures of nongaussianity should be considered.

Differential entropy as nongaussianity measure

- Generalization of ordinary discrete Shannon entropy:

$$H(x) = E\{-\log p(x)\} \tag{12}$$

- for fixed variance, maximized by gaussian distribution.
- often normalized to give negentropy

$$J(x) = H(x_{gauss}) - H(x) \tag{13}$$

- Good statistical properties, but computationally difficult.

Approximation of negentropy

- Approximations of negentropy (Hyvärinen, 1998):

$$J_G(x) = (E\{G(x)\} - E\{G(x_{gauss})\})^2 \tag{14}$$

  where G is a nonquadratic function.

- Generalization of (square of) kurtosis (which is $G(x) = x^4$).
- A good compromise?
  - statistical properties not bad (for suitable choice of G)
  - computationally simple
- Further possibility: Skewness (for nonsymmetric ICs)

Conclusions

- ICA is very simple as a model:
  linear nongaussian latent variables model.
- Solves factor rotation and blind source separation problems,
  if data (components) are nongaussian
- Estimate by maximizing nongaussianity of components.
- Radically different from PCA both in theory and practice.
- Can be applied almost in any field where we have continuous-valued variables, e.g.
  - electro/magnetoencephalograms
  - functional magnetic resonance imaging
  - modelling of vision
  - gene expression data