Independent Component Analysis (ICA)

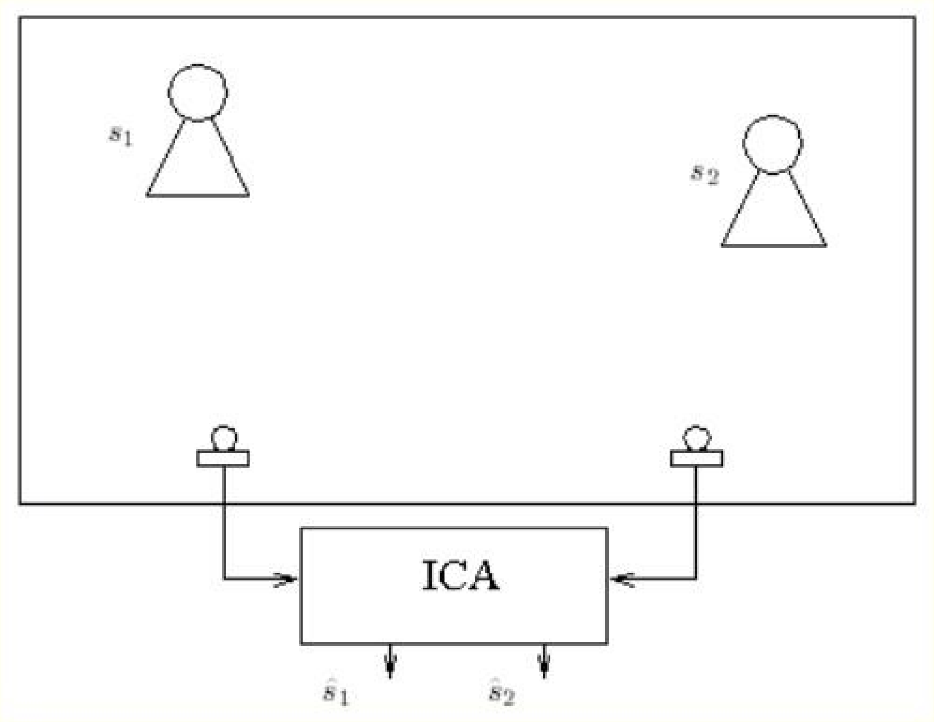
Adopted from:

Independent Component Analysis: A Tutorial

Aapo Hyvärinen and Erkki Oja
Helsinki University of Technology

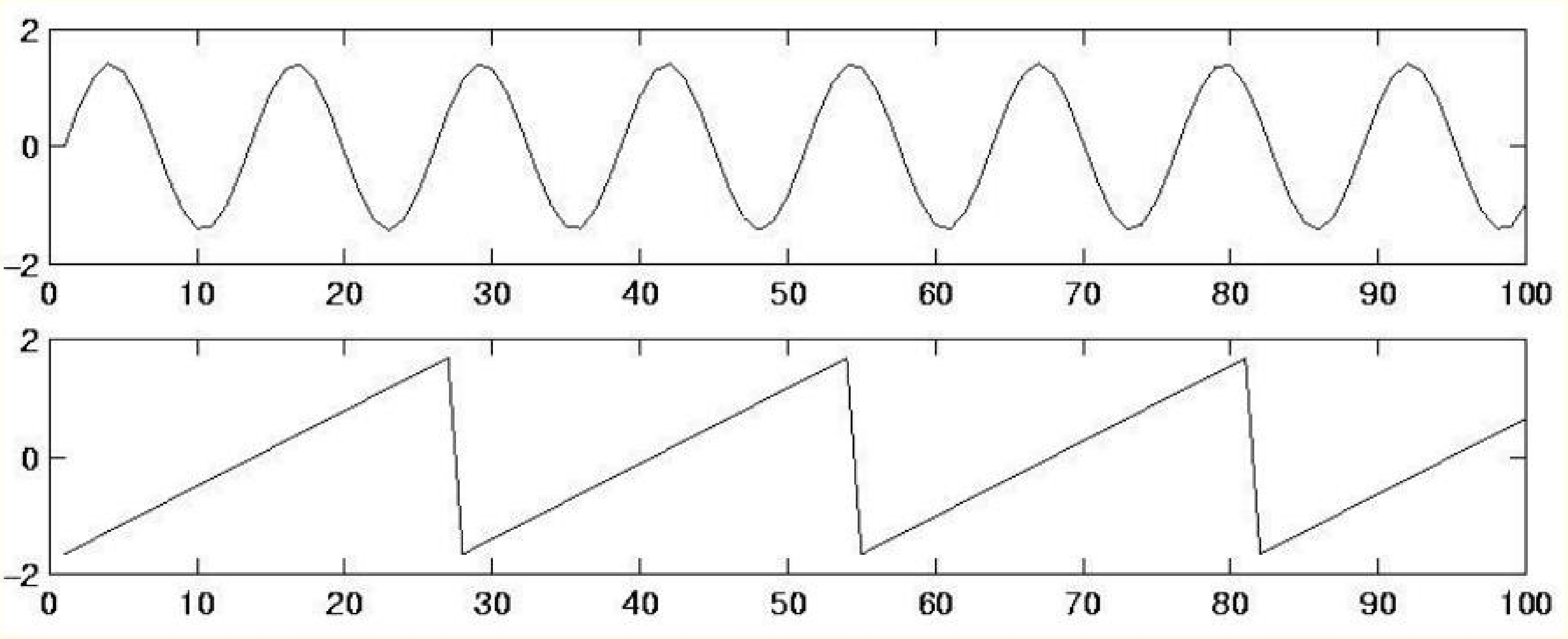
Motivation

- Example: Cocktail-Party-Problem



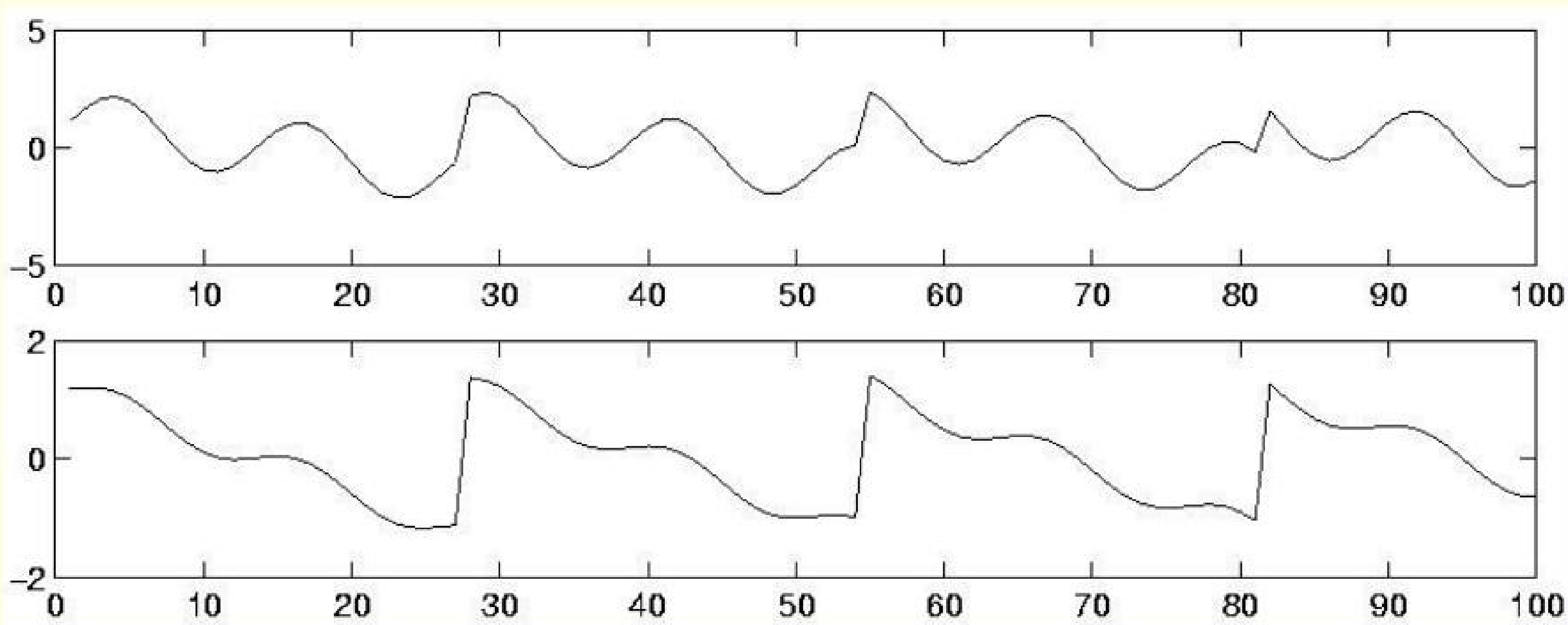
Motivation

- 2 speakers, speaking simultaneously.



Motivation

- 2 microphones in different locations



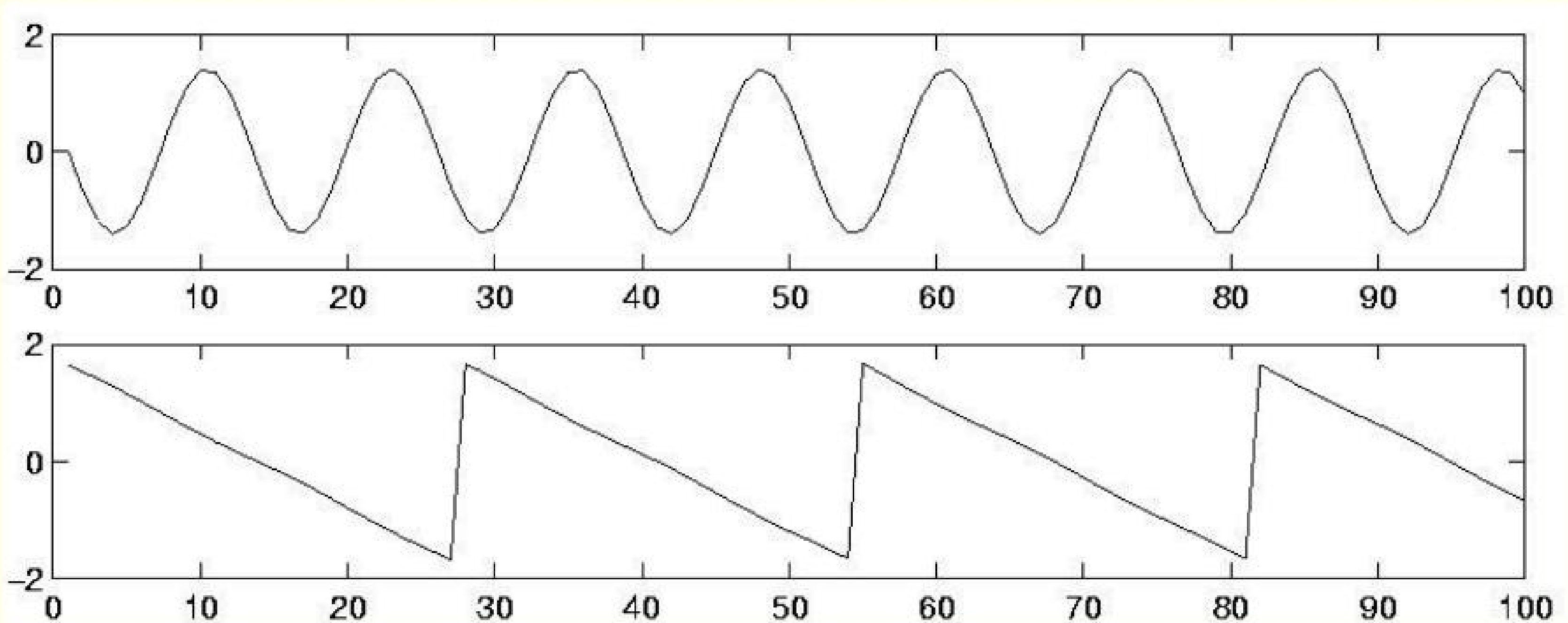
Motivation

$$x_1(t) = a_{11}s_1 + a_{12}s_2$$
$$x_2(t) = a_{21}s_1 + a_{22}s_2$$

a_{ij} ... depends on the distances of the microphones from the speakers

Problem Definition

- Get the original signals out of the recorded ones.



Noise-free ICA model

- Use statistical „latent variables“ system
- Random variable s_k instead of time signal
- $x_j = a_{j1}s_1 + a_{j2}s_2 + .. + a_{jn}s_n$, for all j

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

$$\mathbf{x} = \mathrm{Sum}(\mathbf{a}_i s_i)$$

- $\mathbf{a}_i$ ... basis functions
- s_i ... independent components (IC‘s)

Generative Model

- IC‘s **s** are latent variables => unknown
- Mixing matrix **A** is also unknown
- Task: estimate **A** and **s** using only the observeable random vector **x**

Restrictions

- s_i are statistically independent
  - $p(y_1,y_2) = p(y_1)p(y_2)$
- Non-gaussian distributions
  - Note: if only one IC is gaussian, the estimation is still possible

Solving the ICA model

- Additional assumptions:
  - # of IC's = # of observable mixtures
  - => **A** is square and invertible
- **A** is identifiable => estimate **A**
- Compute $\mathbf{W} = \mathbf{A}^{-1}$
- Obtain IC's from:

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

Ambiguities (I)

- Can‘t determine the variances (energies) of the IC‘s
  - $x = \text{Sum}[(1/C_i)\mathbf{a}_i s_i C_i]$
  - Fix magnitudes of IC‘s assuming unit variance: $E\{s_i^2\} = 1$
  - Only ambiguity of sign remains

Ambiguities (II)

- Can‘t determine the order of the IC‘s
  - Terms can be freely interchanged, because both **s** and **A** are unknown
  - $\mathbf{x} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}$
  - **P** ... permutation matrix

Centering the variables

- Simplifying the algorithm:
  - Assume that both **x** and **s** have zero mean
  - Preprocessing:

$$\mathbf{x} = \mathbf{x}' - E\{\mathbf{x}'\}$$

  - IC's are also zero mean because of:

$$E\{\mathbf{s}\} = \mathbf{A}^{-1}E\{\mathbf{x}\}$$

  - After ICA: add $\mathbf{A}^{-1}E\{\mathbf{x}'\}$ to zero mean IC's

Noisy ICA model

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

- **A** ... mxn mixing matrix
- **s** ... n-dimensional vector of IC's
- **n** ... m-dimensional random noise vector
- Same assumptions as for noise-free model

General ICA model

- Find a linear transformation:

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

- s_i as independent as possible
- Maximize F(**s**) : Measure of independence
- No assumptions on data
- Problem:
  - definition for measure of independence
  - Strict independence is in general impossible

Illustration (I)

- 2 IC's with distribution:

$$p(s_i) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{if } |s_i| \le \sqrt{3} \\ 0 & \text{otherwise} \end{cases}$$

  - zero mean and variance equal to 1
- Joint distribution of IC's:

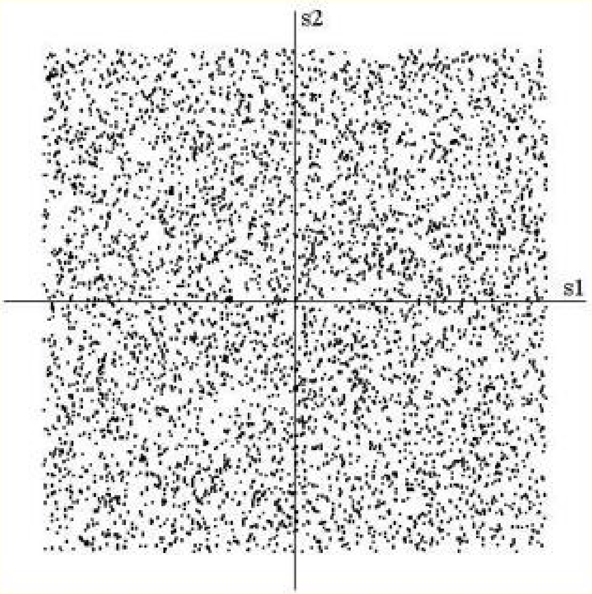
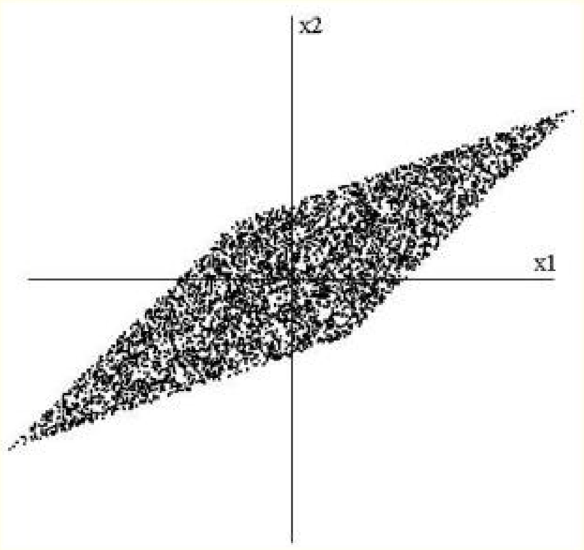


Illustration (II)

- Mixing matrix: $\mathbf{A}_0 = \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}$
- Joint distribution of observed mixtures:



Other Problems

- Blind Source/Signal Separation (BSS)
  - Cocktail Party Problem (another definition)
  - Electroencephalogram
  - Radar
  - Mobile Communication
- Feature extraction
  - Image, Audio, Video, ...representation

Principles of ICA Estimation

- “Nongaussian is independent”: central limit theorem
- Measure of nonguassianity
  - Kurtosis: $\mathrm{kurt}(y) = E\{y^4\} - 3(E\{y^2\})^2$

    (Kurtosis=0 for a gaussian distribution)
  - Negentropy: a gaussian variable has the largest entropy among all random variables of equal variance:

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

Approximations of Negentropy (1)

The classical method of approximating negentropy is using higher-order moments, for example as follows [27]:

$$J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} \operatorname{kurt}(y)^2 \tag{23}$$

The random variable y is assumed to be of zero mean and unit variance. However, the validity of such approximations may be rather limited. In particular, these approximations suffer from the nonrobustness encountered with kurtosis.

Approximations of Negentropy (2)

$$J(y) \approx \sum_{i=1}^{p} k_i [E\{G_i(y)\} - E\{G_i(\nu)\}]^2, \tag{24}$$

where k_i are some positive constants, and ν is a Gaussian variable of zero mean and unit variance (i.e., standardized). The variable y is assumed to be of zero mean and unit variance, and the functions G_i are some nonquadratic functions [18]. Note that even in cases where this approximation is not very accurate, (24) can be used to construct a measure of nongaussianity that is consistent in the sense that it is always non-negative, and equal to zero if y has a Gaussian distribution.

In the case where we use only one nonquadratic function G, the approximation becomes

$$J(y) \propto [E\{G(y)\} - E\{G(\nu)\}]^2 \tag{25}$$

for practically any non-quadratic function G. This is clearly a generalization of the moment-based approximation in (23), if y is symmetric. Indeed, taking $G(y) = y^4$, one then obtains exactly (23), i.e. a kurtosis-based approximation.

But the point here is that by choosing G wisely, one obtains approximations of negentropy that are much better than the one given by (23). In particular, choosing G that does not grow too fast, one obtains more robust estimators. The following choices of G have proved very useful:

$$G_1(u) = \frac{1}{a_1} \log\cosh a_1 u, \quad G_2(u) = -\exp(-u^2/2) \tag{26}$$

where $1 \leq a_1 \leq 2$ is some suitable constant.

The FastICA Algorithm

The FastICA is based on a fixed-point iteration scheme for finding a maximum of the nongaussianity of $\mathbf{w}^T\mathbf{x}$, as measured in (25), see [24, 19]. It can be also derived as an approximative Newton iteration [19]. Denote by g the derivative of the nonquadratic function G used in (25); for example the derivatives of the functions in (26) are:

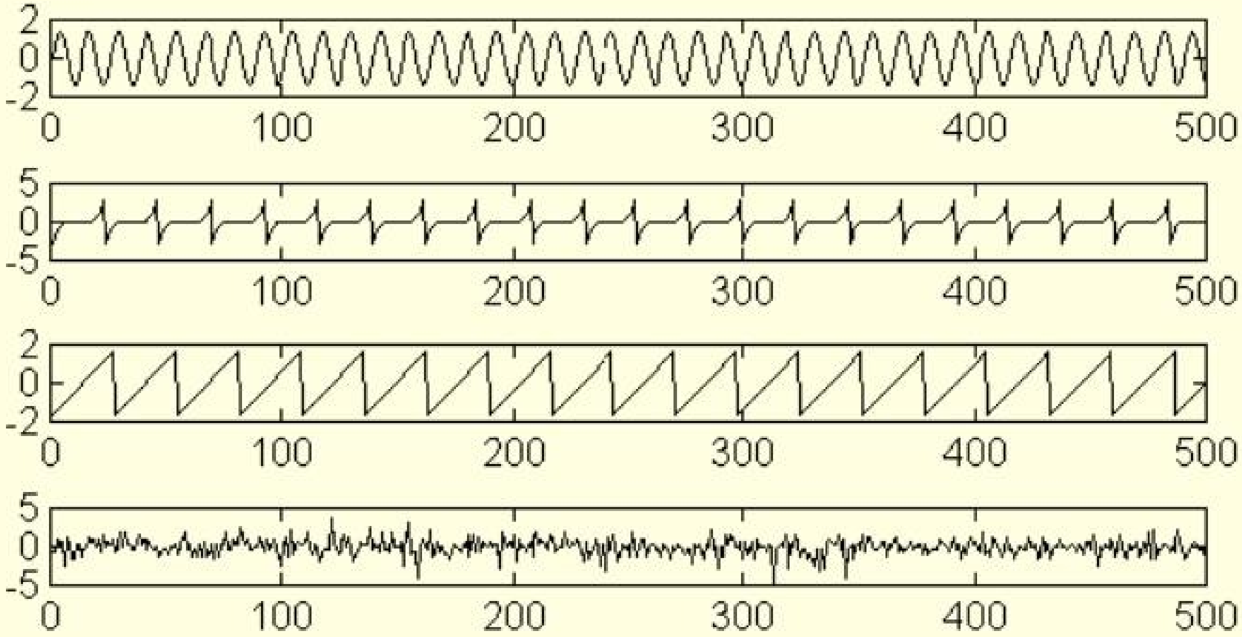
$$g_1(u) = \tanh(a_1 u), \qquad (39)$$
$$g_2(u) = u \exp(-u^2/2)$$

where $1 \leq a_1 \leq 2$ is some suitable constant, often taken as $a_1 = 1$. The basic form of the FastICA algorithm is as follows:

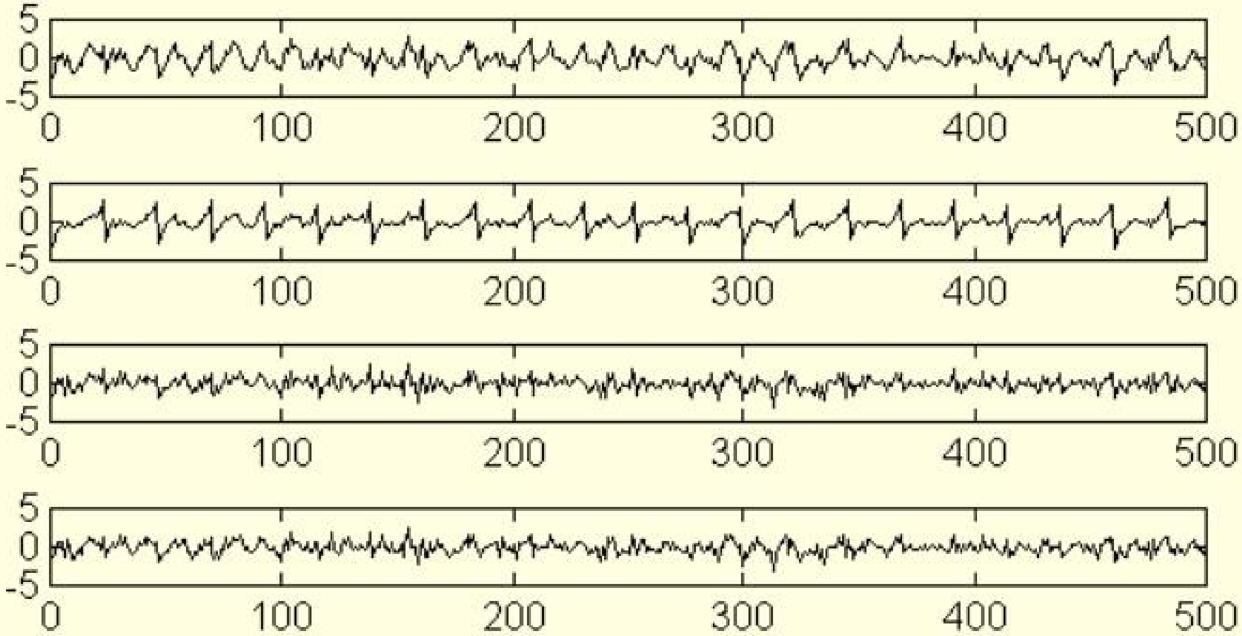
1. Choose an initial (e.g. random) weight vector $\mathbf{w}$.
2. Let $\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$
3. Let $\mathbf{w} = \mathbf{w}^+/\|\mathbf{w}^+\|$
4. If not converged, go back to 2.

Note that convergence means that the old and new values of $\mathbf{w}$ point in the same direction, i.e. their dot-product is (almost) equal to 1. It is not necessary that the vector converges to a single point, since $\mathbf{w}$ and $-\mathbf{w}$ define the same direction. This is again because the independent components can be defined only up to a multiplicative sign. Note also that it is here assumed that the data is prewhitened.

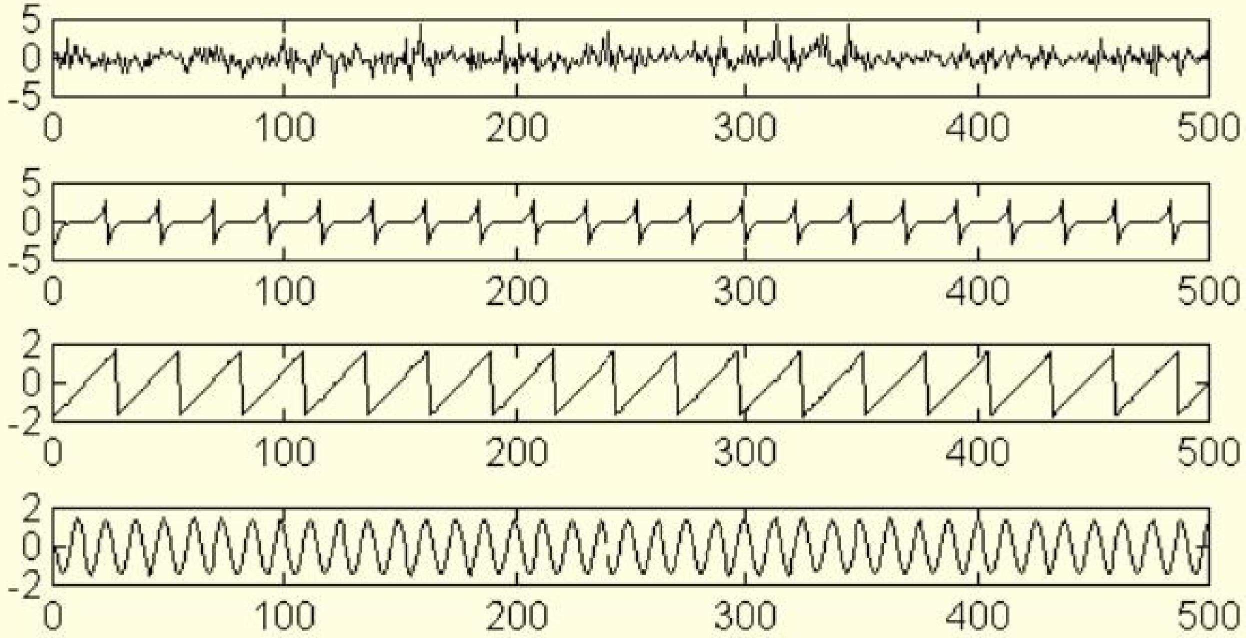
4 Signal BSS demo (original)



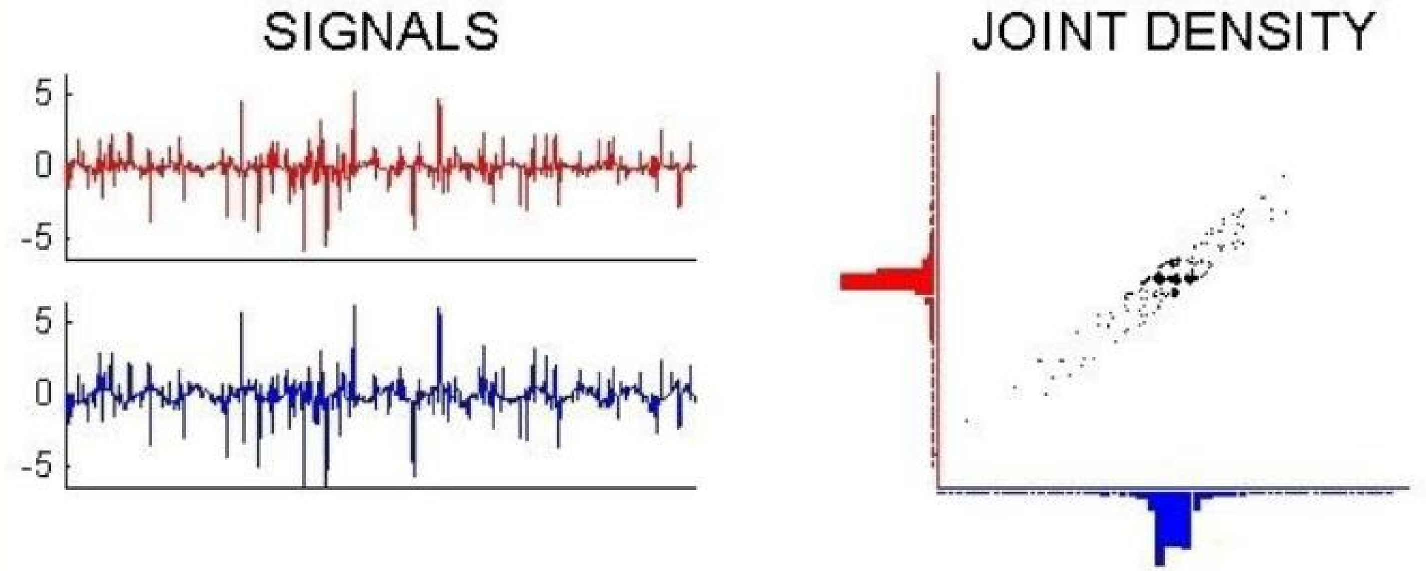
4 Signal BSS demo (Mixtures)



4 Signal BSS demo (ICA)

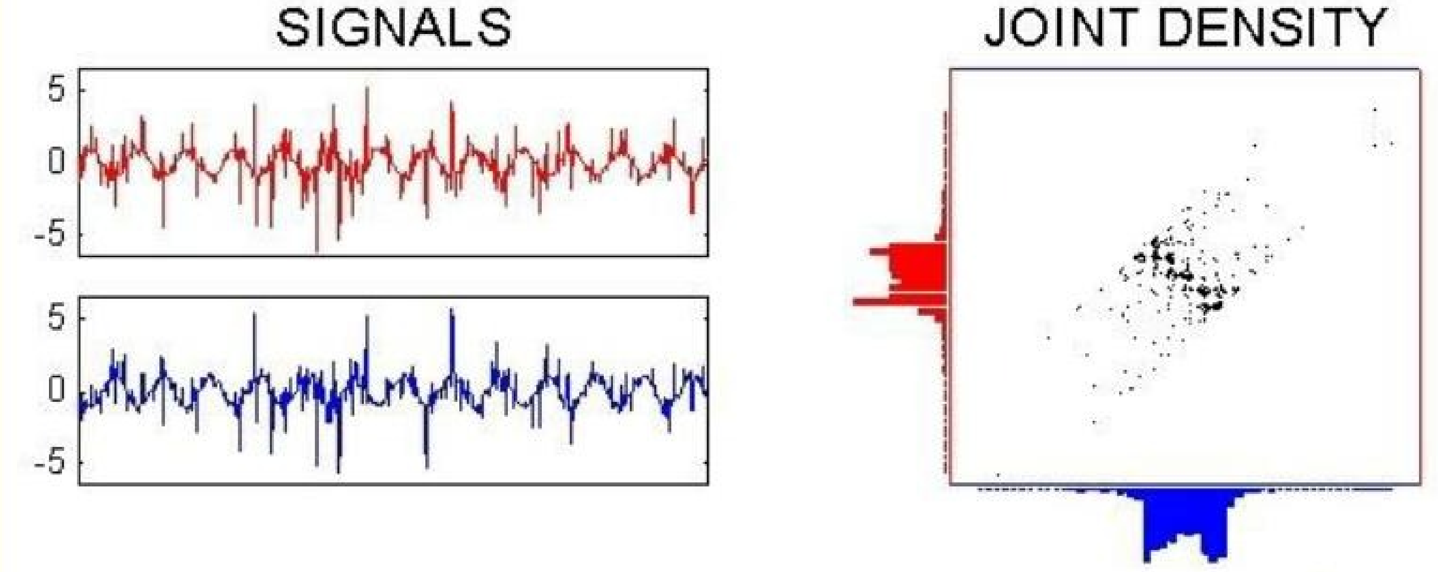


FastICA demo (mixtures)



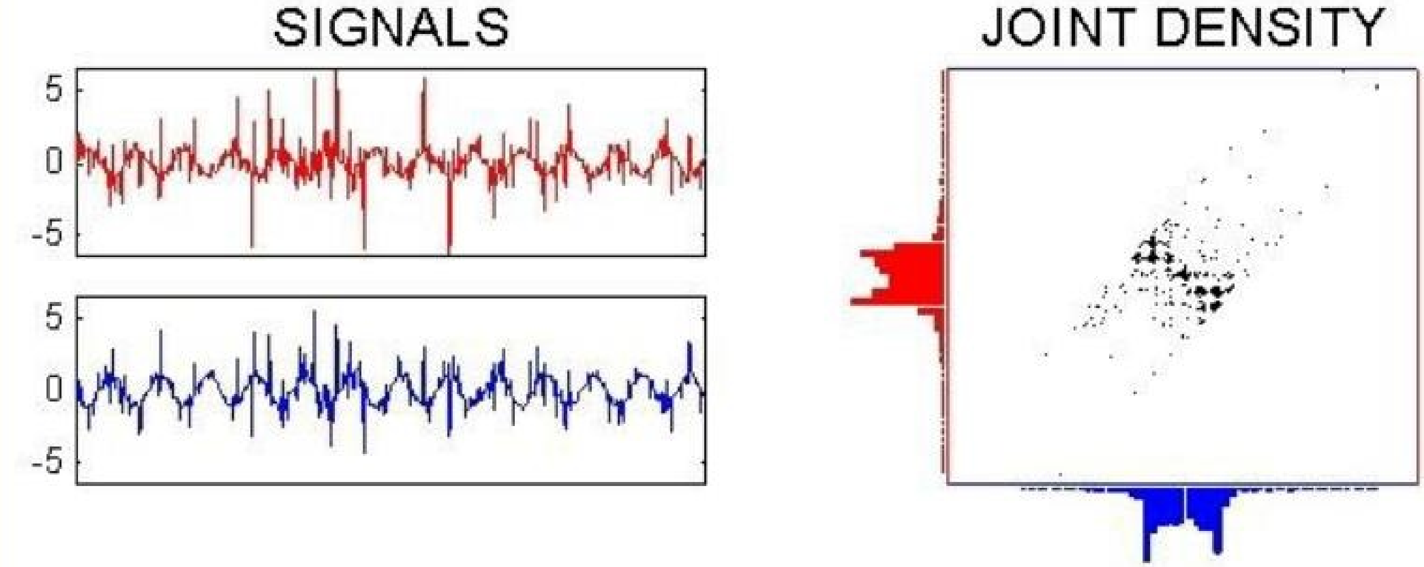
Input signals and density

FastICA demo (whitened)



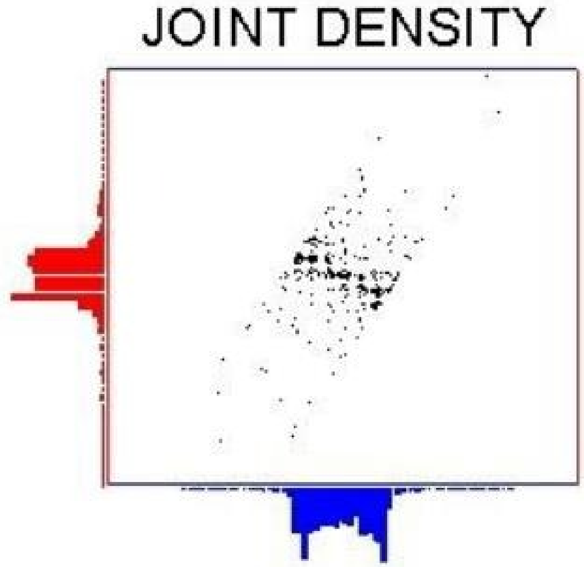
Whitened signals and density

FastICA demo (step 1)



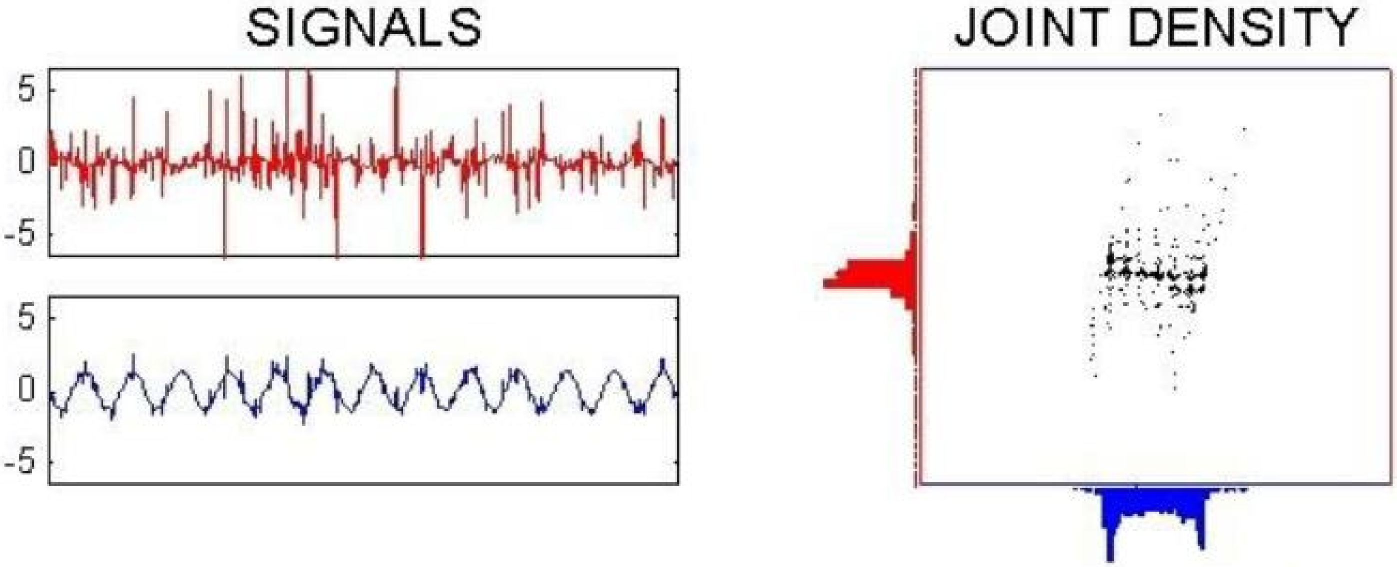
Separated signals after 1 step of FastICA

FastICA demo (step 2)



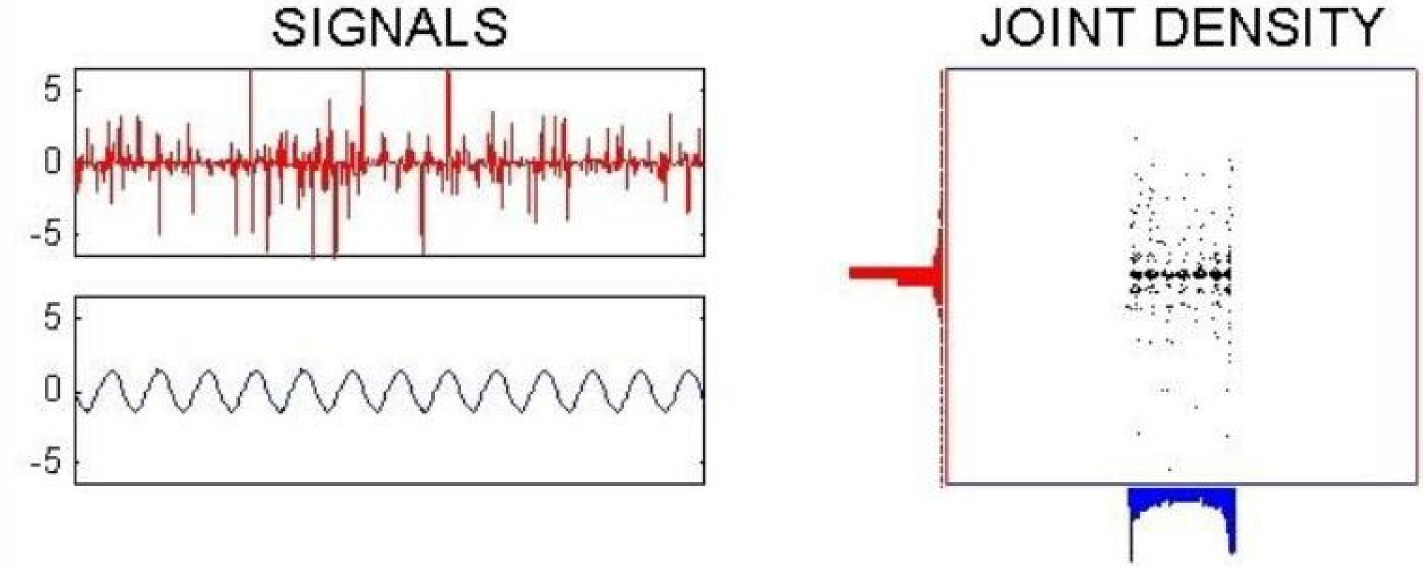
Separated signals after 2 steps of FastICA

FastICA demo (step 3)



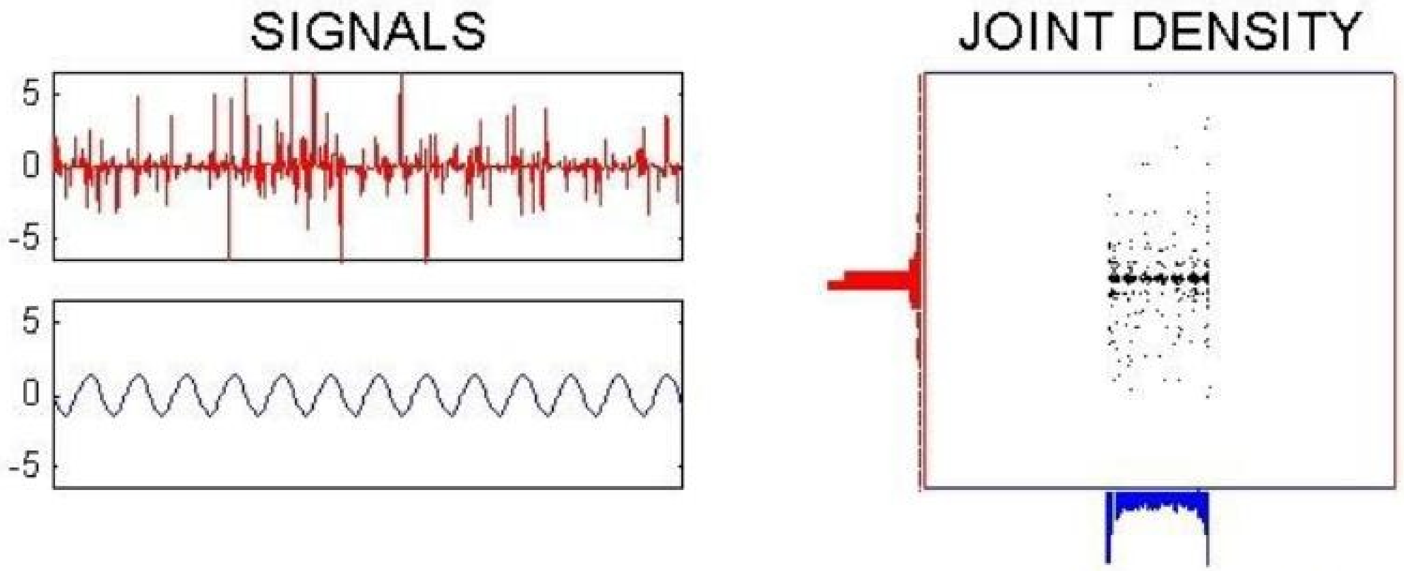
Separated signals after 3 steps of FastICA

FastICA demo (step 4)



Separated signals after 4 steps of FastICA

FastICA demo (step 5 - end)



Separated signals after 5 steps of FastICA

Other Algorithms for BSS

- Temporal Predictability
  - TP of mixture < TP of any source signal
  - Maximize TP to seperate signals
  - Works also on signals with Gaussian PDF
- CoBliSS
  - Works in frequency domain
  - Only using the covariance matrix of the observation
- JADE

Links 1

- Feature extraction (Images, Video)
  - http://hlab.phys.rug.nl/demos/ica/
- Aapo Hyvarinen: ICA (1999)
  - http://www.cis.hut.fi/aapo/papers/NCS99web/node11.html
- ICA demo step-by-step
  - http://www.cis.hut.fi/projects/ica/icademo/
- Lots of links
  - http://sound.media.mit.edu/~paris/ica.html

Links 2

- object-based audio capture demos
  - http://www.media.mit.edu/~westner/sepdemo.html
- Demo for BBS with „CoBliSS“ (wav-files)
  - http://www.esp.ele.tue.nl/onderzoek/daniels/BSS.html
- Tomas Zeman‘s page on BSS research
  - http://ica.fun-thom.misto.cz/page3.html
- Virtual Laboratories in Probability and Statistics
  - http://www.math.uah.edu/stat/index.html

Links 3

- An efficient batch algorithm: JADE
  - http://www-sig.enst.fr/~cardoso/guidesepsou.html
- Dr JV Stone: ICA and Temporal Predictability
  - http://www.shef.ac.uk/~pc1jvs/
- BBS with Degenerate Unmixing Estimation Technique (papers)
  - http://www.princeton.edu/~srickard/bss.html

Links 4

- detailed information for scientists, engineers and industrials about ICA
  - http://www.cnl.salk.edu/~tewon/ica_cnl.html
- FastICA package for matlab
  - http://www.cis.hut.fi/projects/ica/fastica/fp.shtml
- Aapo Hyvärinen
  - http://www.cis.hut.fi/~aapo/
- Erkki Oja
  - http://www.cis.hut.fi/~oja/