Linear Models: FA, ICA, NFA

Shikui Tu

Department of Computer Science and Engineering, Shanghai Jiao Tong University

2021-04-27

Outline

- **Preliminary on probability and statistics**
- Independent Component Analysis (ICA)
- Independent FA (IFA), Non-Gaussian FA (NFA)
- Recent papers related to PCA/ICA/GMM

What is independence?

- The variables y_1 and y_2 are said to be independent, if information on the value of y_1 dose not give any information on the value of y_2, and vice versa.

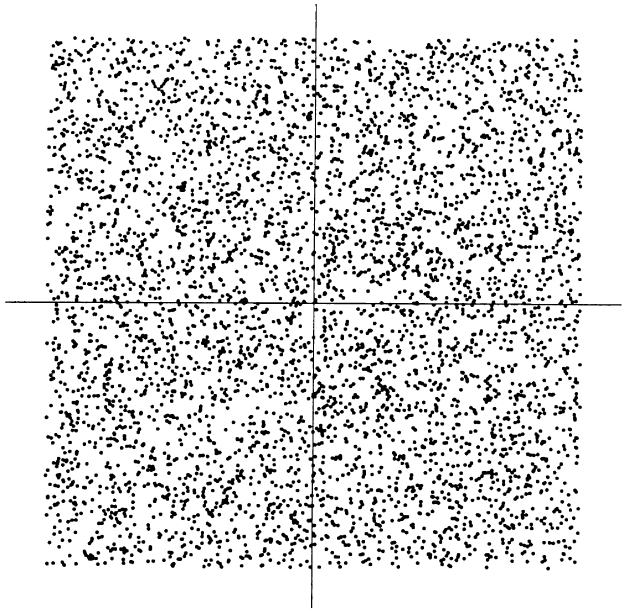


Fig. 5. The joint distribution of the independent components s_1 and s_2 with uniform distributions. Horizontal axis: s_1, vertical axis: s_2.

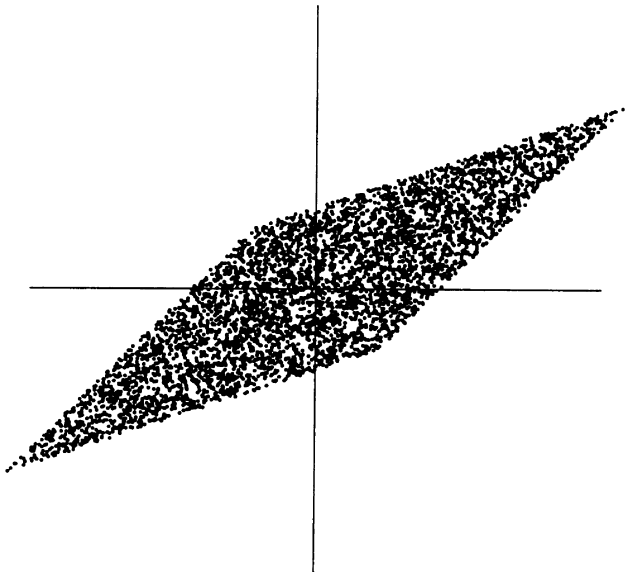


Fig. 6. The joint distribution of the observed mixtures x_1 and x_2. Horizontal axis: x_1, vertical axis: x_2.

What is independence?

Joint distribution is factorizable:

$$p(y_1, y_2) = p_1(y_1)p_2(y_2).$$

Which, for any functions h_1, h_2, implies

$$E\{h_1(y_1)h_2(y_2)\} = E\{h_1(y_1)\}E\{h_2(y_2)\}$$

$$E\{h_1(y_1)h_2(y_2)\} = \int\int h_1(y_1)h_2(y_2)p(y_1, y_2)\mathrm{d}y_1\ \mathrm{d}y_2$$

$$= \int\int h_1(y_1)p_1(y_1)h_2(y_2)p_2(y_2)\mathrm{d}y_1\ \mathrm{d}y_2$$

$$= \int h_1(y_1)p_1(y_1)\mathrm{d}y_1 \int h_2(y_2)p_2(y_2)\mathrm{d}y_2$$

$$= E\{h_1(y_1)\}E\{h_2(y_2)\}.$$

Uncorrelated variables are only partly independent

Uncorrelation: $E\{y_1 y_2\} - E\{y_1\}E\{y_2\} = 0$

Uncorrelation does not imply independence.

$$P\big((y_1, y_2) = (0, +1)\big) = \frac{1}{4}$$

$$P\big((y_1, y_2) = (0, -1)\big) = \frac{1}{4}$$

$$P\big((y_1, y_2) = (+1, 0)\big) = \frac{1}{4}$$

$$P\big((y_1, y_2) = (-1, 0)\big) = \frac{1}{4}$$

$$E\{y_1^2 y_2^2\} = 0 \neq \frac{1}{4} = E\{y_1^2\}E\{y_2^2\}$$

Conditional independence

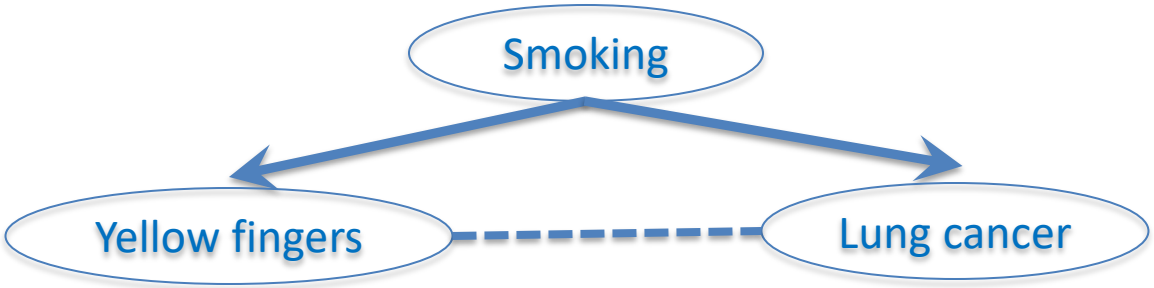
- Definition

$$p(X, Y|Z) = p(X|Z)p(Y|Z)$$

- *X* and *Y* are conditionally independent given *Z*: If *Z* is known, *Y* is not useful when modeling/predicting *X*.

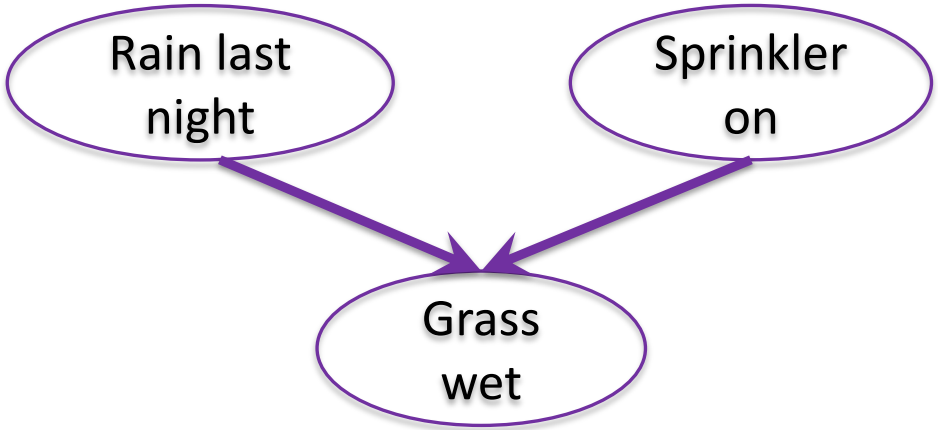
Ways to Produce Dependence

- Common cause



- Causal relation

- Conditional dependence given common effect



Some properties of Gaussian

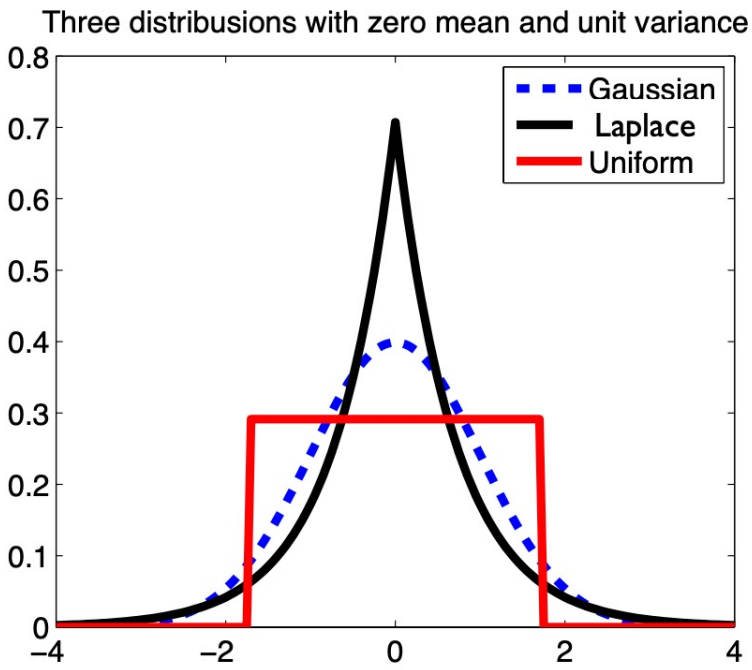
- Uncorrelatedness implies independence
- Approximately holds in many cases because of **central limit theorem** (CLT):

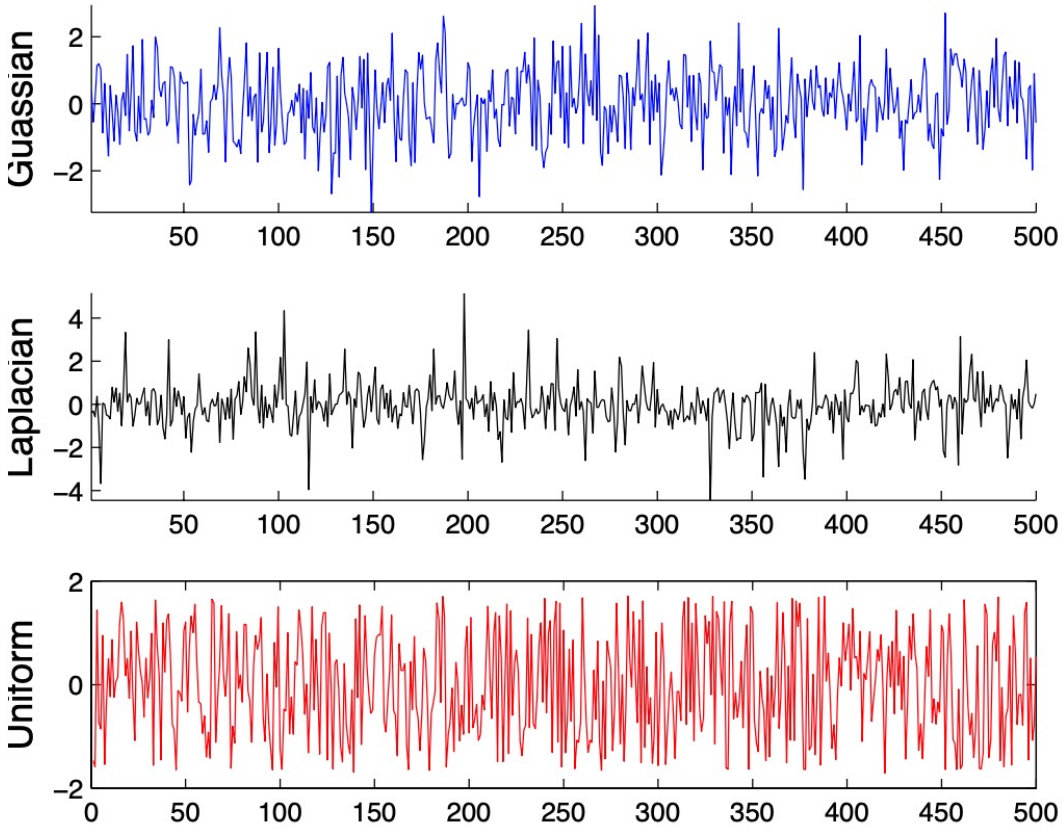
  Under some conditions, $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to a Gaussian distribution for independent X_i with finite mean and variance

- Cramér's decomposition theorem

  Let a random variable ξ be normally distributed and admits a decomposition as a sum $\xi=\xi_1+\xi_2$ of two independent random variables. Then the summands ξ_1 and ξ_2 are normally distributed as well.

Sub-Gaussian and Super-Gaussian





Measure the independence of two variables

- Natural measure of statistical dependence: mutual information

$$I(X;Y) = \sum_{y} \sum_{x} P(x,y) \log \left(\frac{P(x,y)}{P(x)\,P(y)} \right),$$

$$I(X;Y) = \int \int p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) dx\,dy,$$

- Non-negative; is zero iff *X* and *Y* are independent

Outline

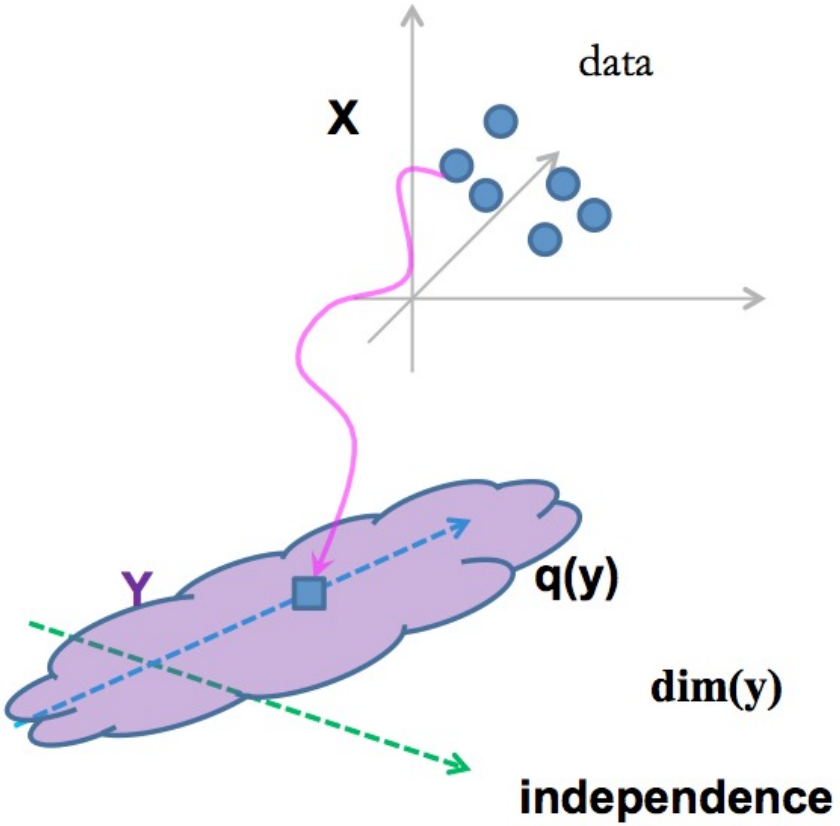
- Preliminary on probability and statistics

- **Independent Component Analysis (ICA)**

- Independent FA (IFA), Non-Gaussian FA (NFA)

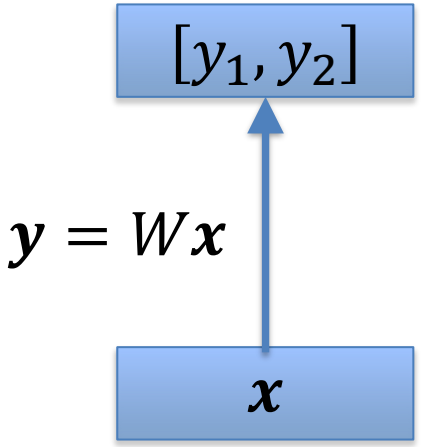
- Recent papers related to PCA/ICA/GMM

Independent Component Analysis (ICA)



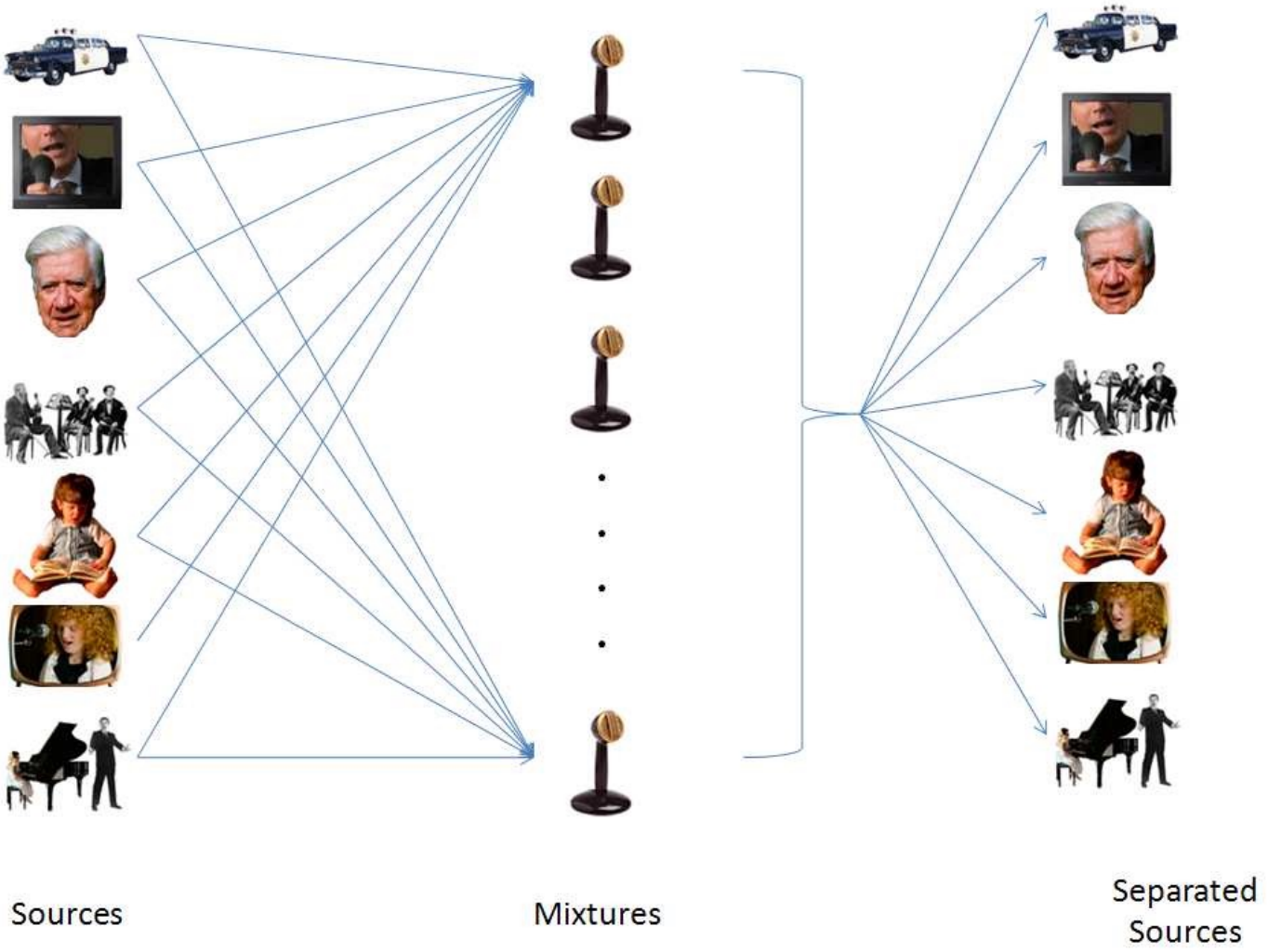
Find W such that

$$p(\boldsymbol{y}) = p(\boldsymbol{y_1}) \cdots p(\boldsymbol{y_m})$$



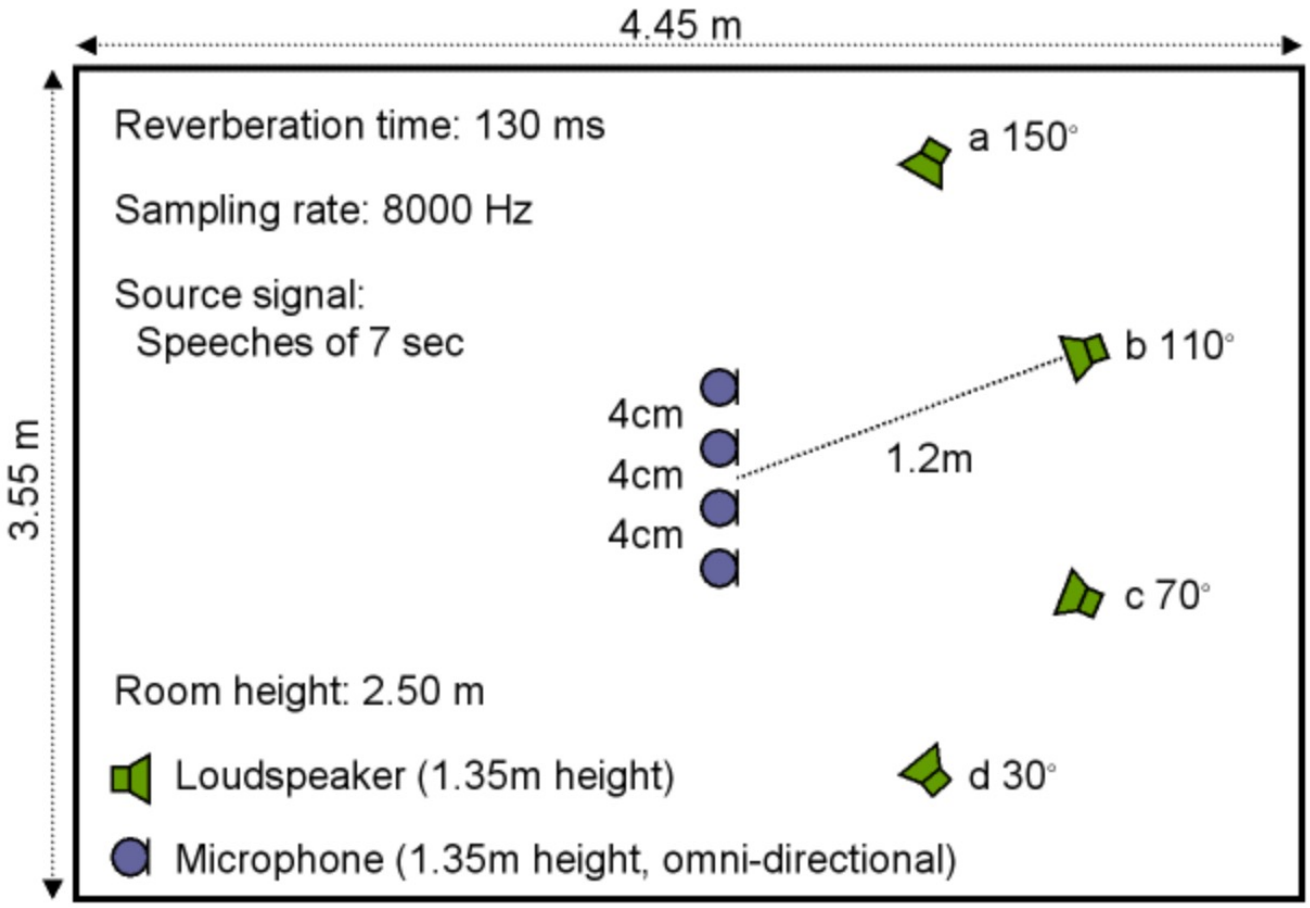
Blind Source Separation (BSS)

Cocktail party problem



Demo: http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html

BSS Sound Demonstration



Source signals:

a b c

Observed signals:

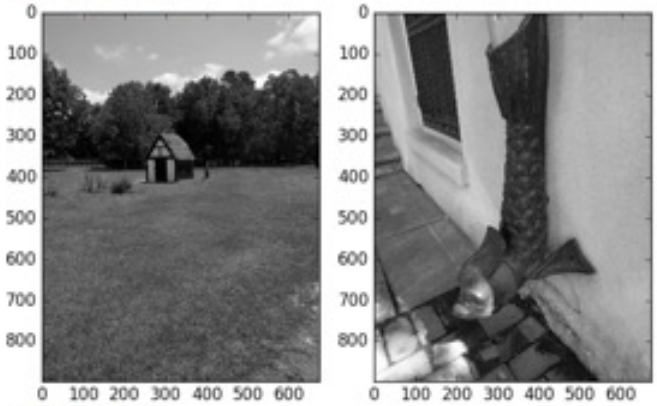
x1 x2 x3

Separated signals:

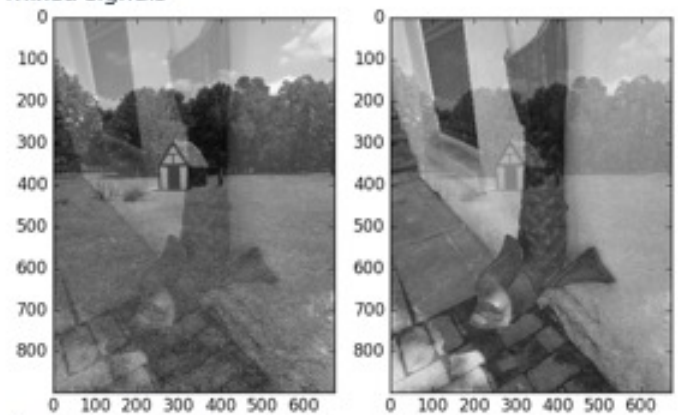
y1 y2 y3

http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html

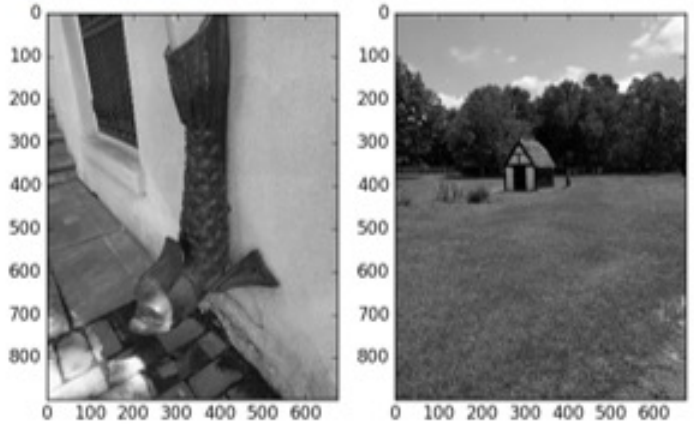
Original Signals



Mixed Signals



Separated signals

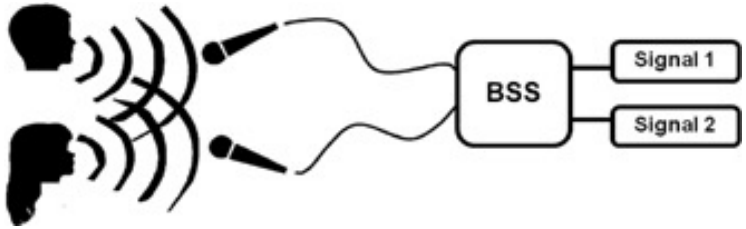


BSS: problem definition

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

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

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



s_1 and s_2

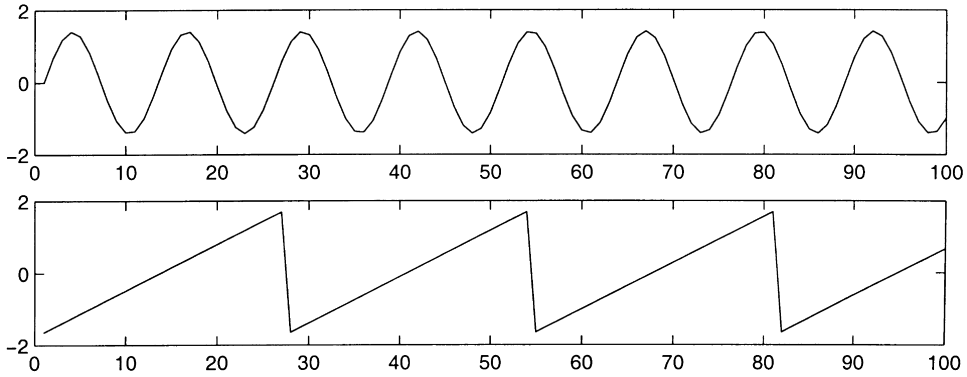


Fig. 1. The original signals.

x_1 and x_2

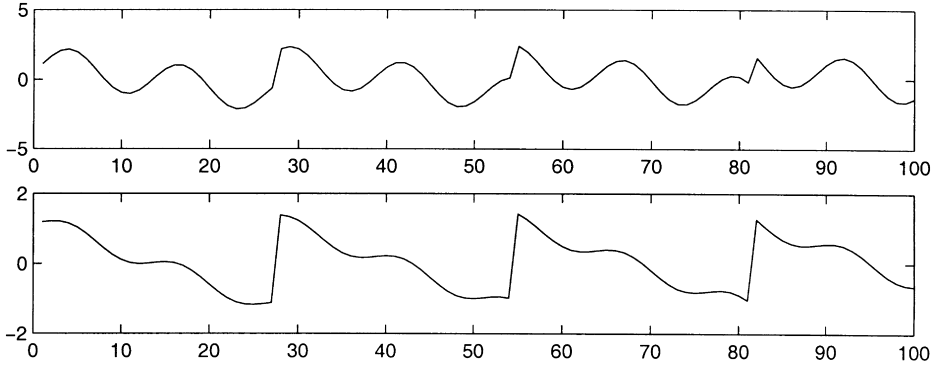


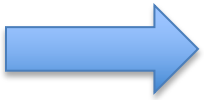
Fig. 2. The observed mixtures of the source signals in Fig. 1.

Task: Find W, and compute $\boldsymbol{s} = W\boldsymbol{x}$

Indeterminacies of ICA

- The variances (energies) of the independent components.

$$\boldsymbol{x} = A\boldsymbol{s} = \sum_{i=1}^{n} \boldsymbol{a} s_i = \sum_{i=1}^{n} (\boldsymbol{a} \cdot \lambda_i^{-1})(\lambda_i s_i)$$



Assume $E[s_i^2] = 1$

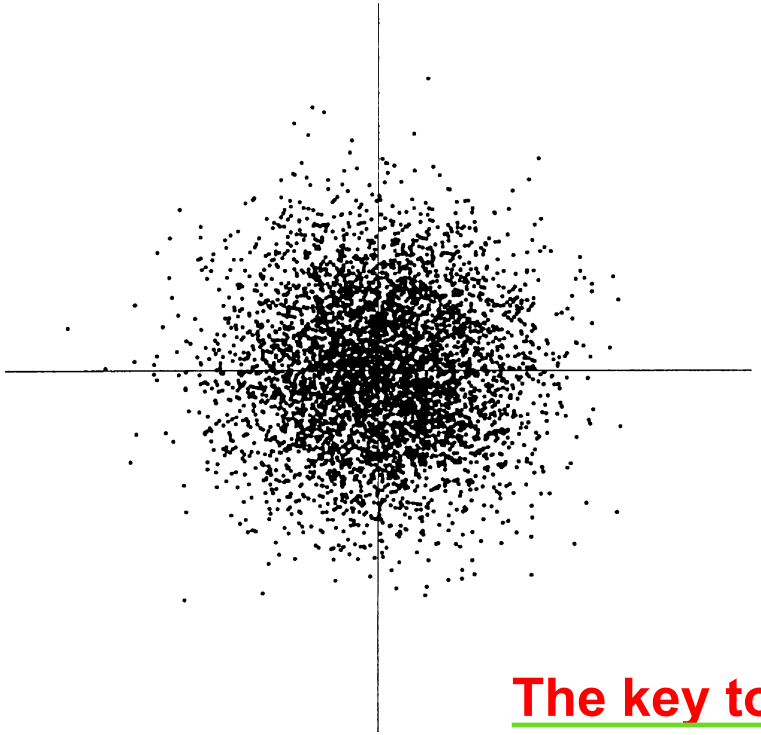
- The order of the independent components

$$\boldsymbol{x} = A\boldsymbol{s} = (AP^{-1})(P\boldsymbol{s})$$

P is a permutation matrix

At most one Gaussian variable is allowed in ICA

$$p(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right)$$



- Completely **symmetric**, no information on the directions of the columns of $\boldsymbol{A}$
- Any **orthogonal transformation** of Gaussian (x_1, x_2) has exactly the same distribution as (x_1, x_2).

The key to estimating ICA is non-Gaussianity.

Fig. 7. The multivariate distribution of two independent Gaussian variables.

Principles of ICA estimation

- "Non-Gaussian is independent"
  - The Central Limit Theorem, a classical result in probability theory, tells that the distribution of a sum of independent random variables tends toward a Gaussian distribution, under certain conditions.
  - Thus, a sum of two independent random variables usually has a distribution that is closer to Gaussian than any of the two original random variables.

$$y = \boldsymbol{w}^T \boldsymbol{x} = \boldsymbol{w}^T A \boldsymbol{s} = (\boldsymbol{w}^T A)\boldsymbol{s} = \boldsymbol{z}^T \boldsymbol{s}$$

$\boldsymbol{z}^T \boldsymbol{s}$ is more Gaussian than any of the s_i (Assume s_i is i.i.d.)

Therefore, we could take w that maximizes the non-Gaussianity.

Measures of non-Gaussian

- Kurtosis or the fourth order cumulant

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

$$= E\{y^4\} - 3, \text{ if } E\{y^2\} = 1 \text{ for unit variance.}$$

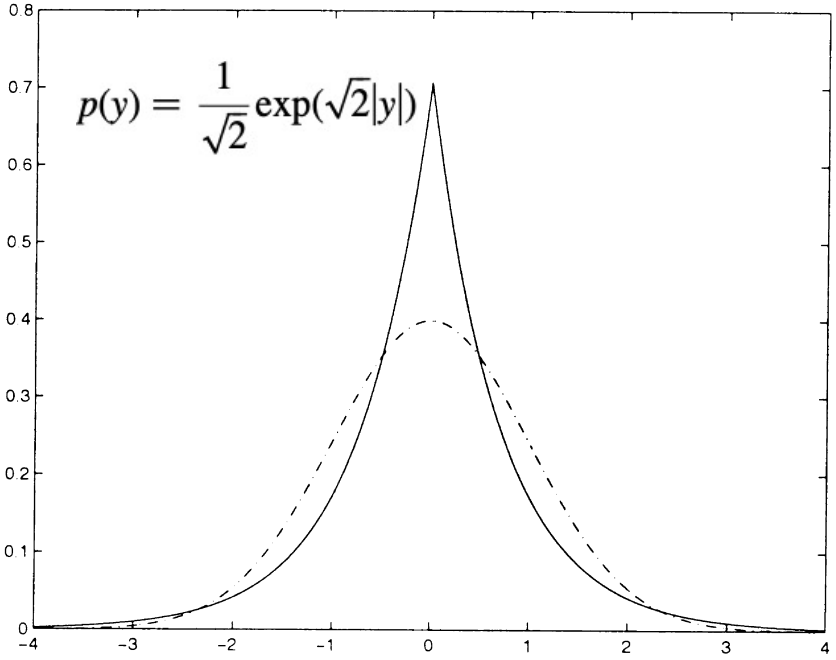


Fig. 8. The density function of the Laplace distribution, which is a typical super-Gaussian distribution. For comparison, the Gaussian density is given by a dashed line. Both densities are normalized to unit variance.

$\text{kurt}(y) < 0$, sub-Gaussian

$\text{kurt}(y) > 0$, super-Gaussian

Measures of non-Gaussian

A Gaussian variable has the largest entropy among all random variables of equal variance over the entire real axis.

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

Where $H(\boldsymbol{y})$ is entropy defined by

$$H(Y) = -\sum_i P(Y = a_i)\log P(Y = a_i) \qquad \text{discrete}$$

$$H(\mathbf{y}) = -\int f(\mathbf{y})\log f(\mathbf{y})\mathrm{d}\mathbf{y} \qquad \text{continuous}$$

The more "random", i.e. unpredictable and unstructured the variable is, the larger its entropy.

Negentropy is in some sense the optimal estimator of non-Gaussianity, as far as statistical properties are concerned. The problem in using negentropy is, however, that it is computationally very difficult due to estimation of pdf.

Measures of non-Gaussian

- Mutual information

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

Actually it is KL divergence between $P(y_1, \ldots, y_m)$ and $P(y_1) \cdots P(y_m)$

There is a fundamental relation between mutual information and negentropy:

$$I(y_1, y_2, \ldots, y_n) = C - \sum_i J(y_i).$$

http://fourier.eng.hmc.edu/e161/lectures/ica/node5.html

FastICA algorithm [Hyvarinen 1999]

- FastICA learning finds a unit vector $\boldsymbol{w}$ such that the projection $y = \boldsymbol{w}^T \boldsymbol{x}$ maximize non-Gaussian via an approximation of negentropy:

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

ν is a zero-mean unit-variance Gaussian

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

Denote by g the derivative of the above non-quadratic function G:

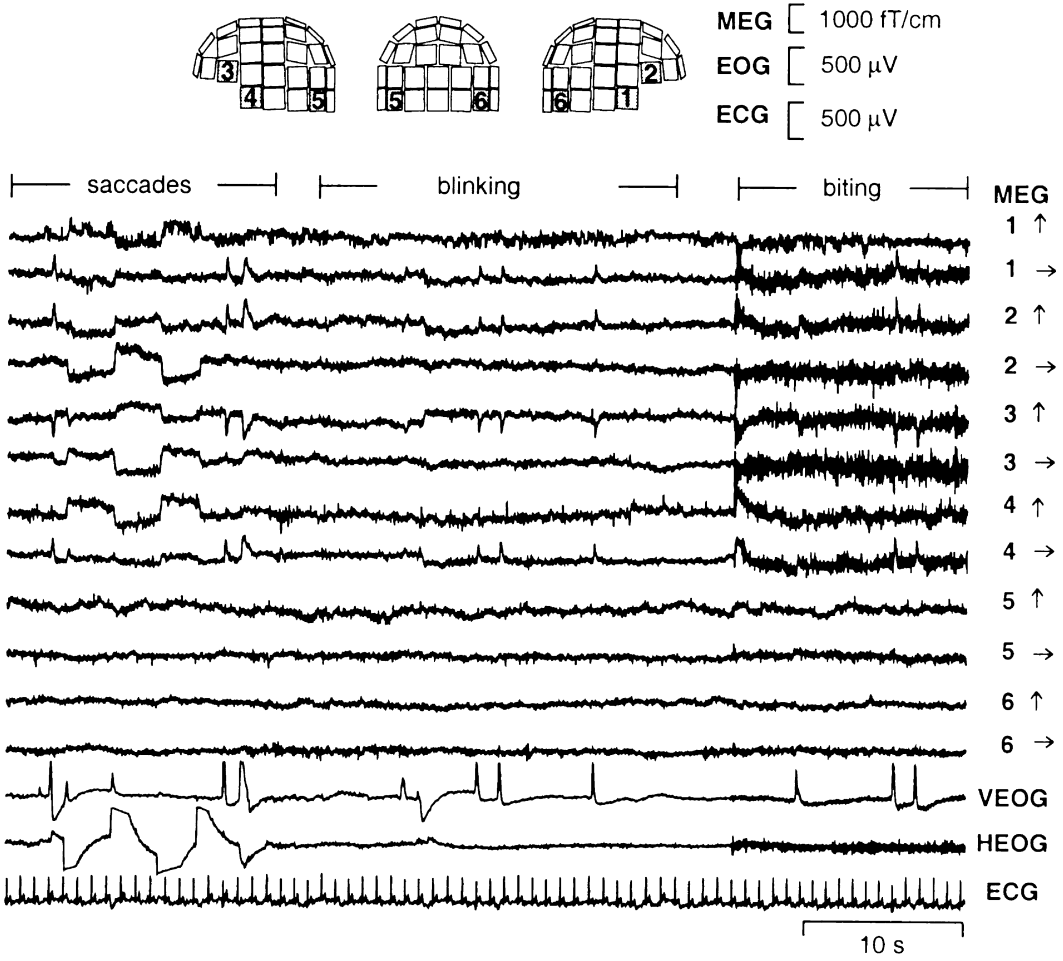
$$g_1(u) = \tanh(a_1 u), \qquad g_2(u) = u \exp(-u^2/2)$$

The basic form of FastICA algorithm:

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.

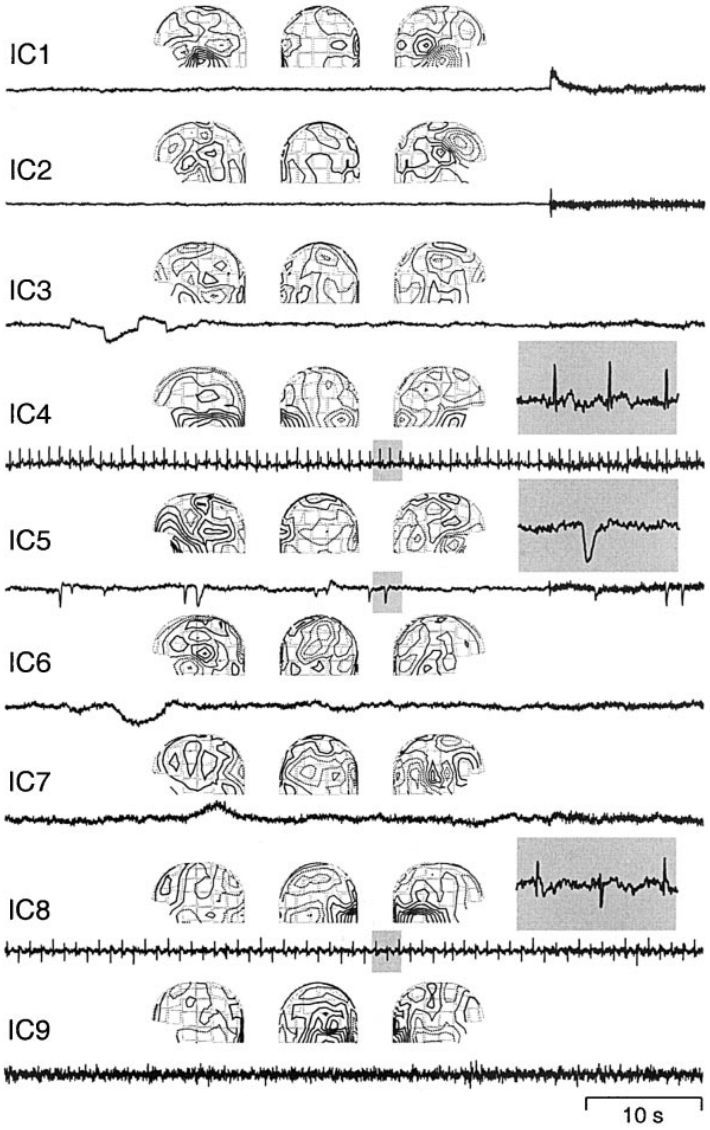
$$\cosh x = \frac{e^x + e^{-x}}{2} = \frac{e^{2x}+1}{2e^x} = \frac{1+e^{-2x}}{2e^{-x}}. \quad \tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x}-1}{e^{2x}+1} = \frac{1-e^{-2x}}{1+e^{-2x}}.$$

Separation of artifacts in Magnetoencephalography(MEG) data



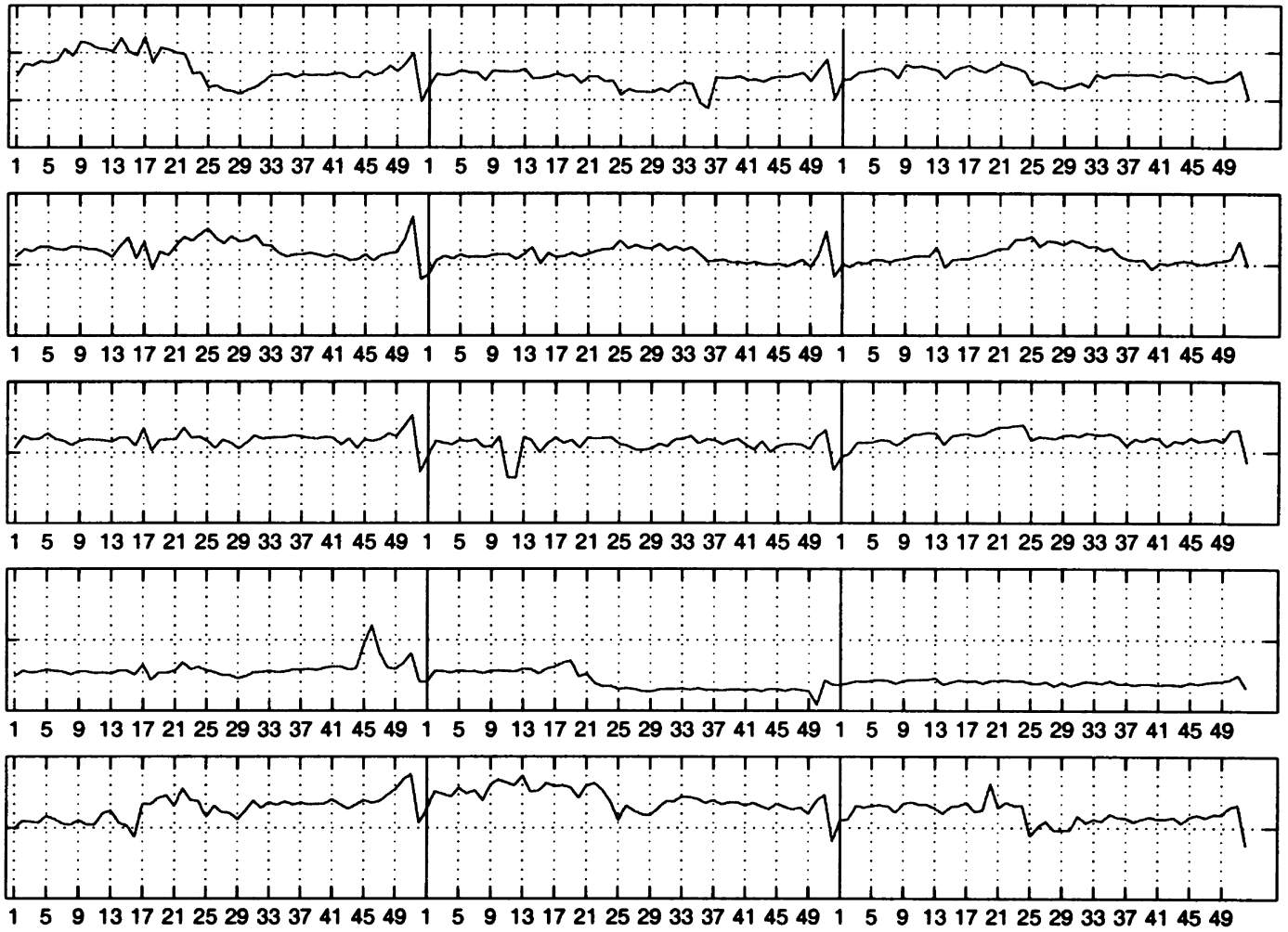
Samples of MEG signals, showing artifacts produced by blinking, saccades, biting and cardiac cycle. For each of the six positions shown, the two orthogonal directions of the sensors are plotted.

Separation of artifacts in Magnetoencephalography(MEG) data



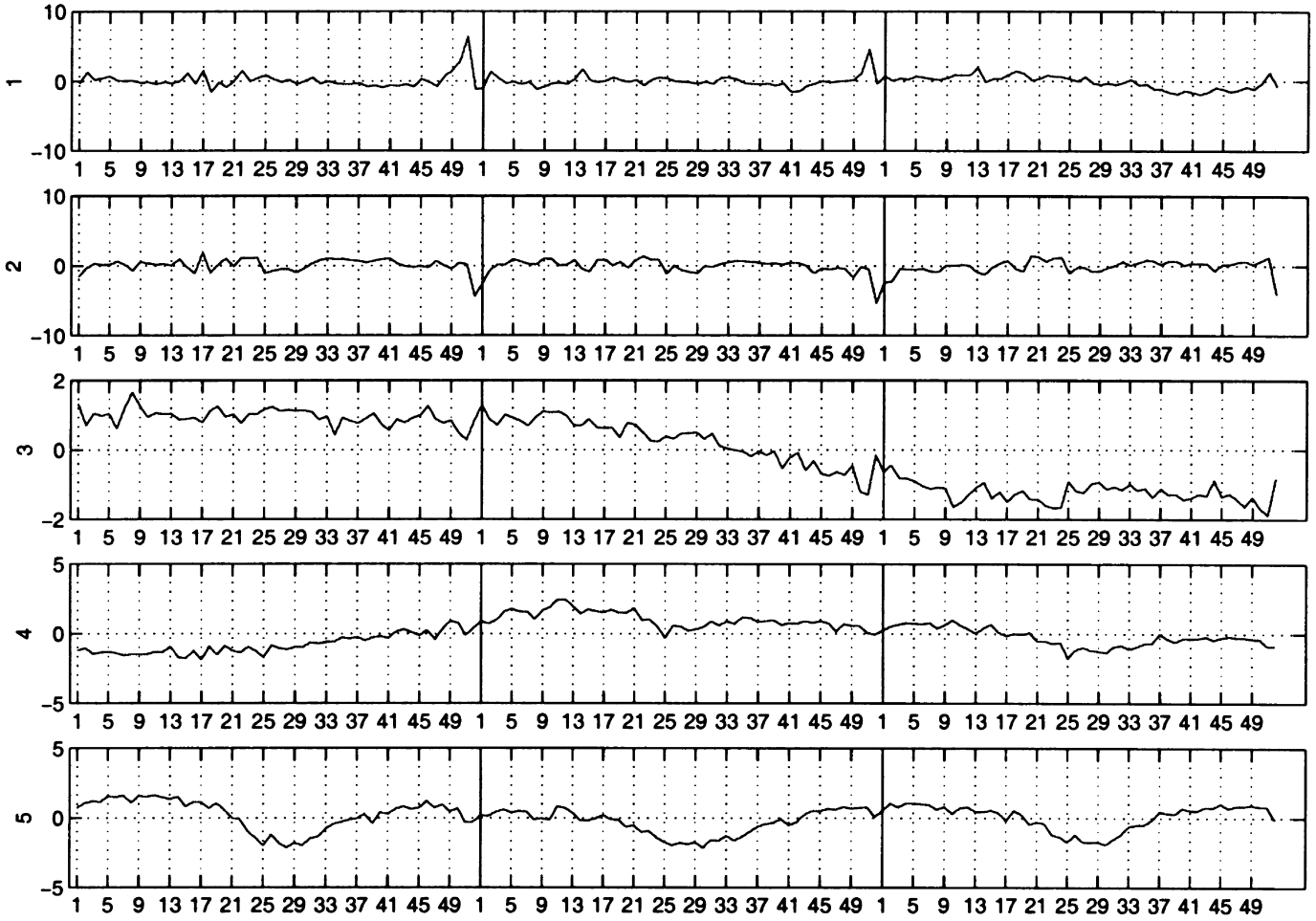
Nine independent components found from the MEG data. For each component the left, back and right views of the field patterns generated by these components are shown—full line stands for magnetic flux coming out from the head, and dotted line the flux inwards.

Finding hidden factors in financial data



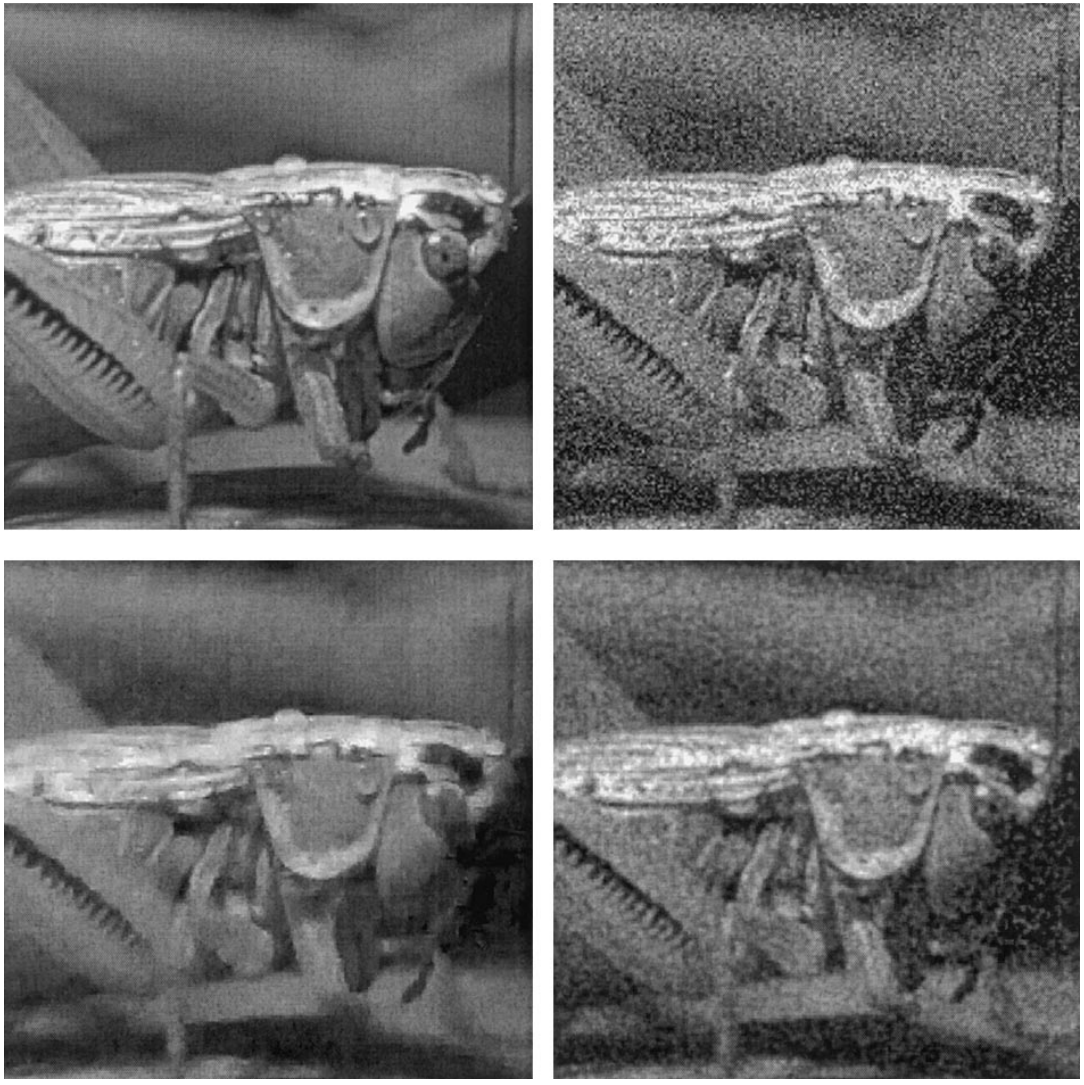
Five samples of **the original cashflow time series** (mean removed, normalized to unit standard deviation).

Finding hidden factors in financial data



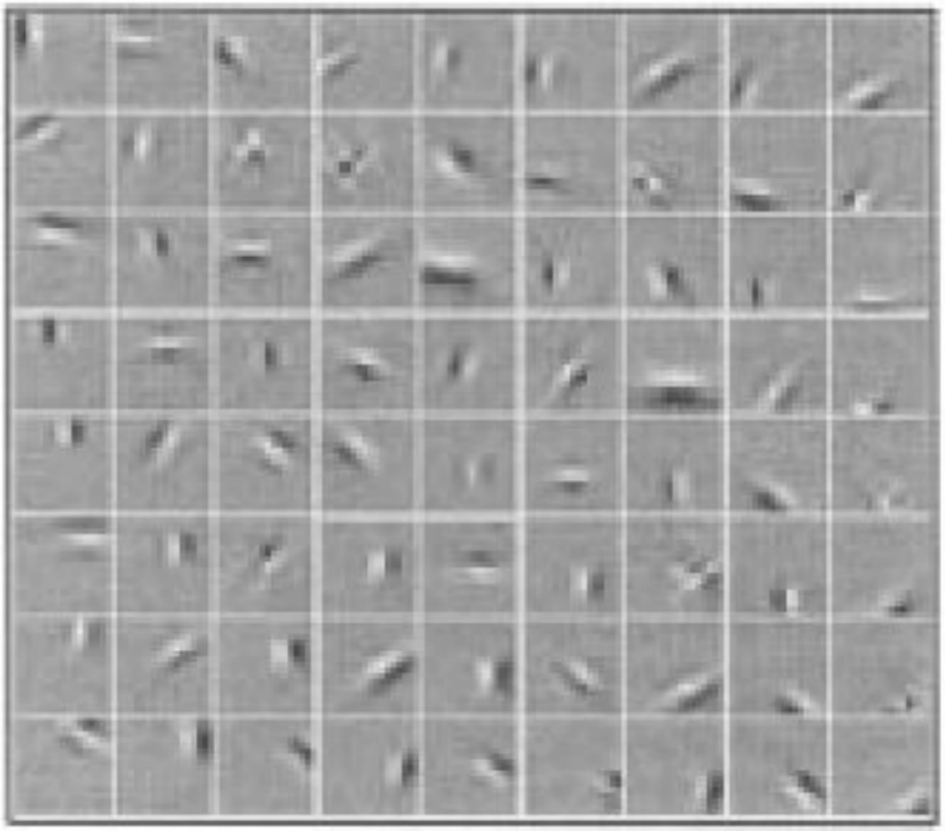
Five independent components or fundamental factors found from the cashflow data.

Reducing noise in natural images



An experiment in denoising. Upper left: original image. **Upper right: original image** corrupted with noise; the noise level is 50%. **Lower left: the recovered image** after applying sparse code shrinkage. Lower right: for comparison, a wiener filtered image.

ICA decomposition of natural images



- A selection of 144 basis functions (columns of W^{-1}) obtained from training on patches of 12- by-12 pixels from pictures of natural scenes.
- Components are similar to Gabor filters (oriented edge detectors)

T.-P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T.-W. Lee, and T. J. Sejnowski, “Imaging Brain Dynamics Using Independent Component Analysis,” Proc IEEE Inst Electr Electron Eng, vol. 89, no. 7, pp. 1107–1122, Jul. 2001, doi: 10.1109/5.939827.

Discussion

- ICA *vs.* PCA: differences and relationship?

Thank you!