



# Independent Component Analysis and Its Applications

---

By Qing Xue, 10/15/2004



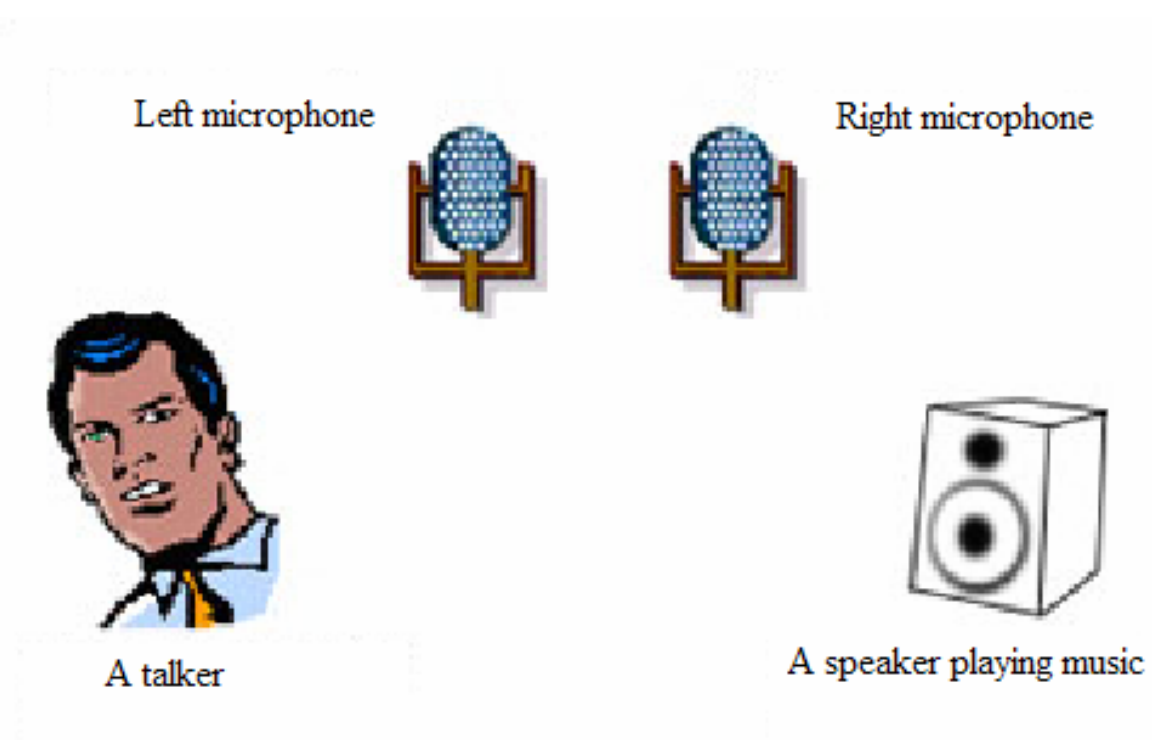
# Outline

---

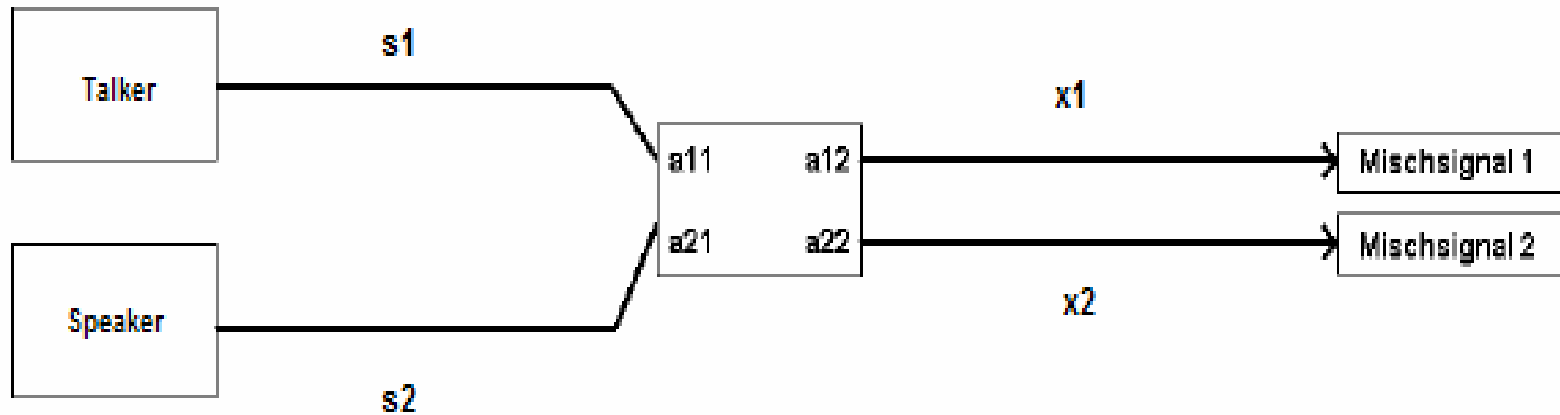
- Motivation of ICA
- Applications of ICA
- Principles of ICA estimation
- Algorithms for ICA
- Extensions of basic ICA framework

# Motivation of ICA

- Cocktail-party problem



# Motivation of ICA



$$\mathbf{s} = (s_1 ; s_2)$$

$$\mathbf{x} = (x_1 ; x_2)$$

$$\mathbf{A} = (a_{11} \ a_{12} ; a_{21} \ a_{22})$$

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



# Motivation of ICA

---

- The problem:

How to separate the voice from the music using recordings of several microphones in the room.

- Solution:

- $a_{ij}$  are known: the linear equation can be solved by classical methods.
- $a_{ij}$  are unknown: ICA can be used to estimate the  $a_{ij}$  and separate the original source signals from their mixtures.

# Applications of ICA

- Blind source separation (classical application of ICA)
  - Cocktail party problem

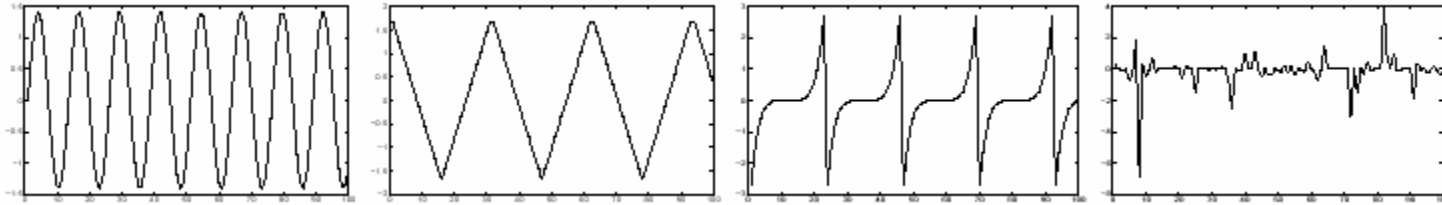


Fig1: An illustration of blind source separation. This figure shows four source signals or independent components.



# Applications of ICA

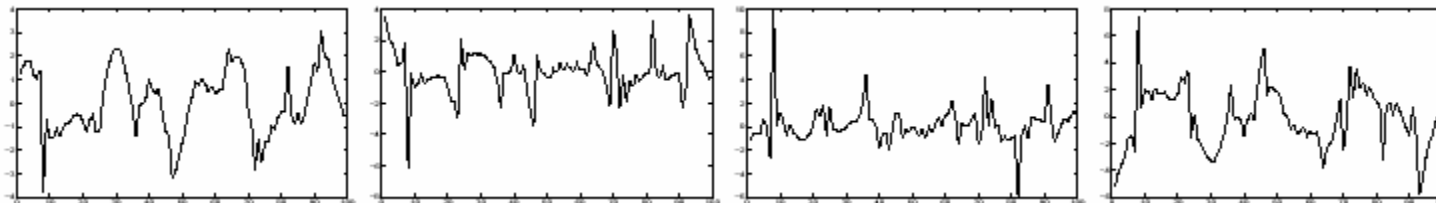


Fig2: Due to some external circumstances, only linear mixtures of the source signals in Fig1 can be observed.

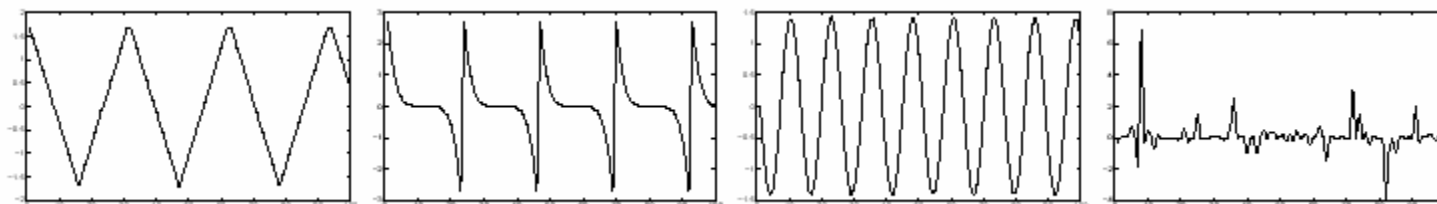


Fig3: The estimates of the source signals using only the linear mixtures in Fig2. Estimates are accurate up to multiplying factors.



# Applications of ICA

---

## Separation of artifacts in MEG data

- Magnetoencephalography (MEG):
  - The measurement of the magnetic activity of the brain.
  - Provides very good temporal resolution and moderate spatial resolution.
- Problem when using a MEG record:
  - Extracting the essential features of the neuromagnetic signals in the presence of artifacts.
- An ICA approach to separate brain activity from artifacts

# Applications of ICA

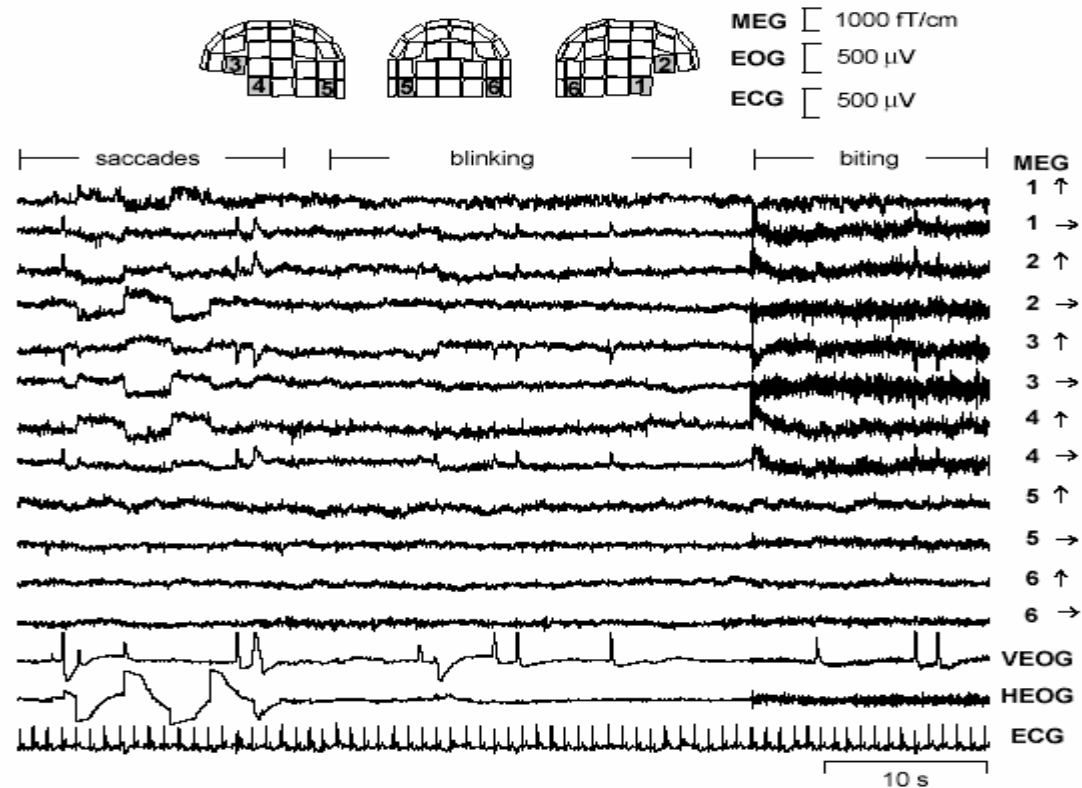


Fig4: Samples of MEG signals, showing artifacts produced by blinking, saccades, biting and cardiac cycle. For each of the 6 positions shown, the two orthogonal directions of the sensors are plotted.

# Applications of ICA

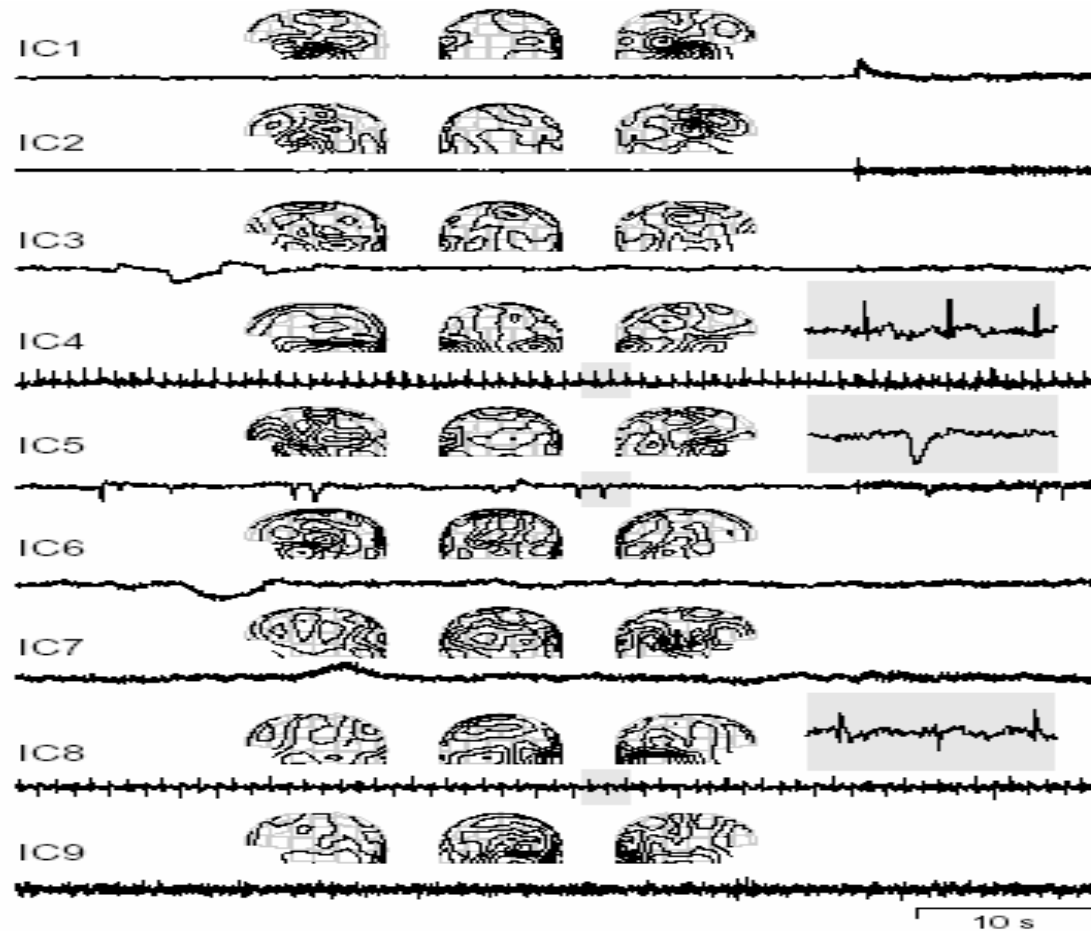


Fig5: Nine independent components found from the MEG data



# Application of ICA

---

- Finding hidden factors in financial data

Financial data with parallel time series (e.g. currency exchange rates, daily returns of stocks)

- May have some common underlying factors.
- ICA decomposes the data into independent components to give an insight to the structure of the data set.



# Application of ICA

---

- Problem:
  - Suppose a cashflow of several stores belonging to the same retail chain
  - Find the fundamental factors common to all stores that affect the cashflow data.
- Input data
  - The weekly cash flow in 40 stores in the same retail chain;
  - The cash flow measurements cover 140 weeks.
  - Pre-whitened to project the original signal vectors to the subspace spanned by their first five principal components.

# Applications of ICA

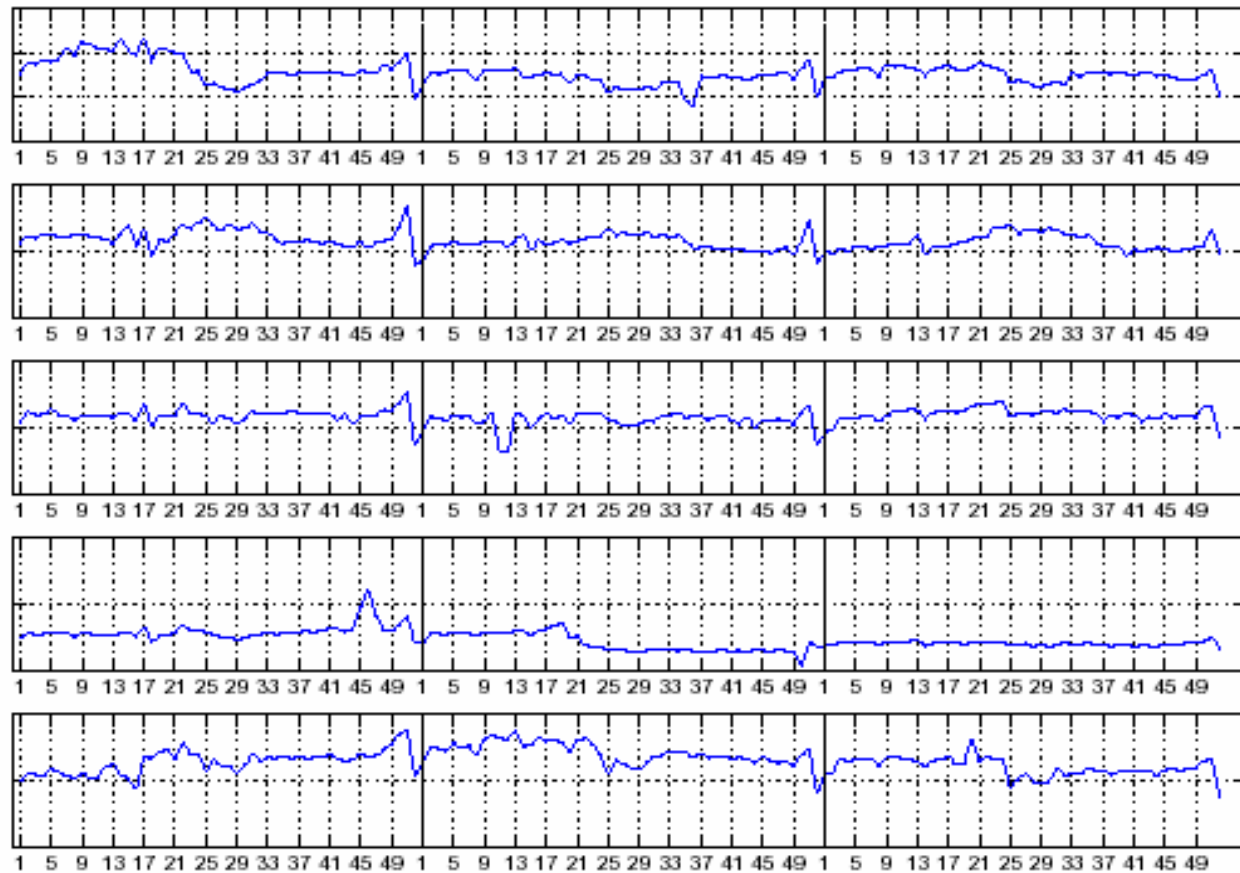


Fig6: Five samples of the original cashflow time series. Horizontal axis: time in weeks.

# Applications of ICA

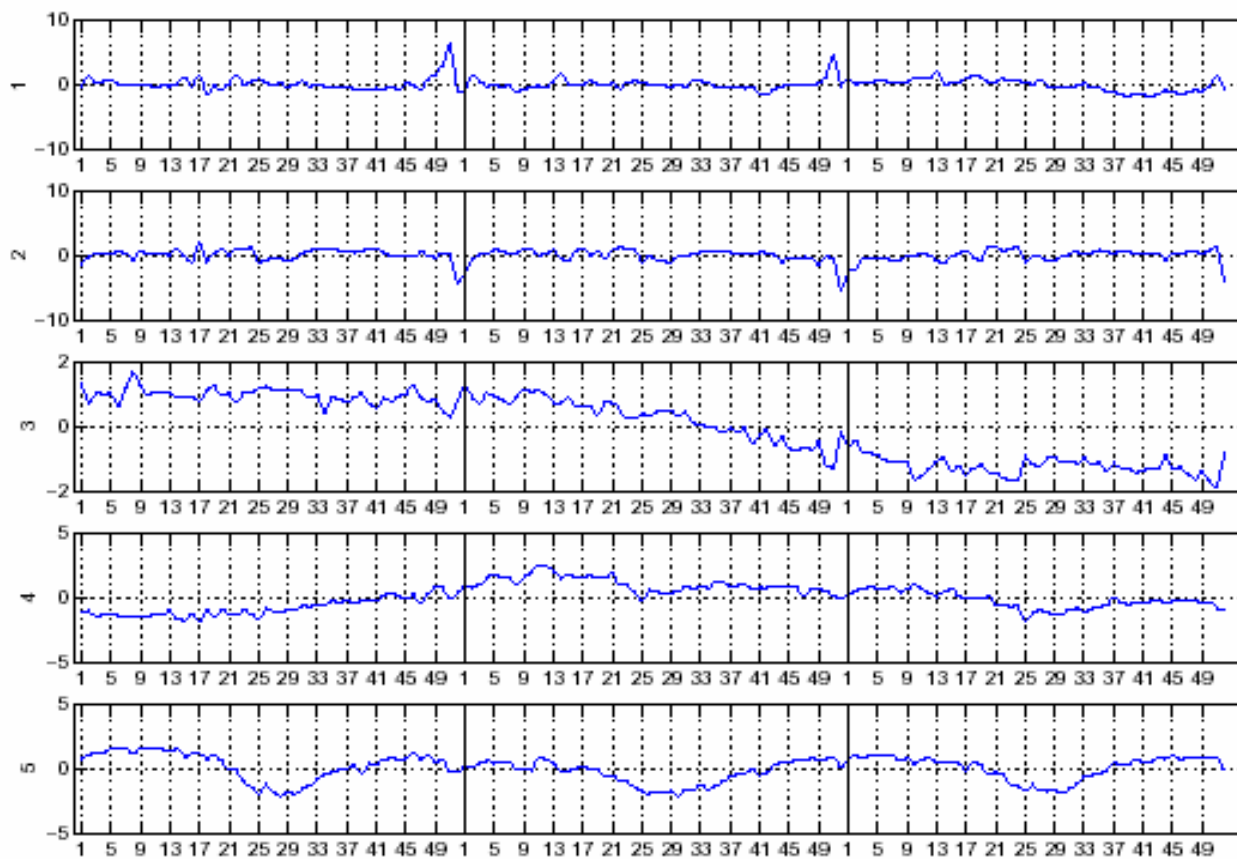


Fig7: Five independent components or fundamental factors found from the cashflow data



# Application of ICA

---

## ■ Interpretations of the factors

- Factor 1 and 2 follow the sudden changes caused by holidays, the most prominent is the Christmas time.
- Factor 3 could represent a still slower variation, such as a trend
- Factor 4 might follow the relative competitive position of the retail chain with respect to its competitors.
- Factor 5 reflects the slower seasonal variation, with the summer holidays clearly visible.



# Application of ICA

---

- Feature extraction

The columns of  $\mathbf{A}$  represent features, and  $s_i$  is the coefficient of the  $i$ -th feature in an observed data vector  $\mathbf{x}$ . ICA is used to find features that are as independent as possible.



# Linear Representation of Multivariate Data

---

- A suitable representation of the multivariate data can facilitate the subsequent analysis of the data.
- Linear transformation of the original data: computational and conceptual simplicity. The problem can be rephrased as

$$\begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{pmatrix} = W \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_m(t) \end{pmatrix} \Rightarrow y = Wx$$

$\mathbf{x}$ : random vector whose elements are the mixtures  $x_1, \dots, x_m$ ,

$\mathbf{y}$ : random vector with elements  $y_1, \dots, y_n$ .

$W$ : the matrix to be determined by the statistical properties components  $y_i$ .



# Dimension Reduction Methods

---

- Principal component analysis (PCA)
  - Choose  $\mathbf{W}$  to limit the number of components  $y_i$  to be quite small.
  - To determine  $\mathbf{W}$  so that each component contains as much information on the data as possible.



# Independence as a Guiding Principle

---

- Determine  $\mathbf{W}$  by finding statistically independent components of  $y_i$
- Any one of these components gives no information on the other ones.
- The starting point of ICA  
Find statistically independent components in the general case where the data is nongaussian.



# Definition of Linear ICA

---

- Definition

- Given a set of observations of random variables  $(x_1(t), x_2(t), \dots, x_n(t))$  ( $t$  is the time or sample index)
- Assume that they are generated by a linear mixture of independent components  $(s_1(t), s_2(t), \dots, s_m(t))$

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix} = A \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_m(t) \end{pmatrix}$$

- ICA consists of estimating both the unknown matrix  $\mathbf{A}$  and the  $s_i(t)$ , only by observe the  $x_i(t)$ .



# Definition of Linear ICA

---

- Alternative definition

Find a linear transformation given by a matrix  $\mathbf{W}$ , so that the random variables  $y_i$ ,  $i=1, \dots, n$  are as independent as possible.

$$\begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{pmatrix} = \mathbf{W} \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_m(t) \end{pmatrix}$$

Each  $x_i(t)$  is a sample on one observed random variable.



# What is Independence?

---

## ■ Statistical independence

- Independence between two random variables  $y_1$  and  $y_2$

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

$p(y_1, y_2)$ : is the joint pdf of  $y_1$  and  $y_2$ ;

$p_1(y_1)$  and  $p_2(y_2)$  are the marginal pdf of  $y_1$  and  $y_2$  respectively.

- If  $y_1$  and  $y_2$  are independent

$$E\{g_1(y_1)g_2(y_2)\} - E\{g_1(y_1)\}E\{g_2(y_2)\} = 0$$

for any two functions  $g_1$  and  $g_2$ .



# Uncorrelatedness

---

- Uncorrelated variables are only partly independent
  - Uncorrelatedness: a weaker form independence is if  $y_1$  and  $y_2$  are uncorrelated

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

- Independence implies uncorrelatedness  
ICA methods constrain the estimation procedure to give uncorrelated estimates of the independent components, which can reduce the number of the free parameters.



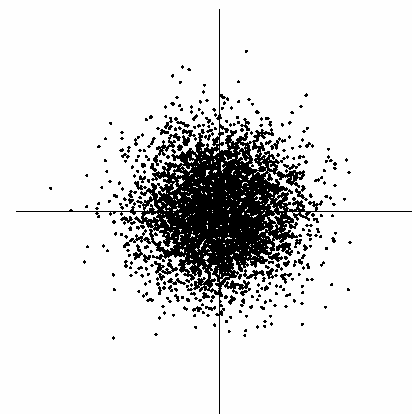
# Gaussian Variables are Forbidden

---

- Fundamental restriction in ICA:

Independent components must be nongaussian for ICA to be possible.

- $s_1$  and  $s_2$  have a gaussian distribution
- Mixing matrix  $A$  is orthogonal
- Thus  $x_1$  and  $x_2$  are gaussian
- Joint distribution of  $x_1$  and  $x_2$  is completely symmetric and  $A$  can not be estimated.



Joint Distribution of two independent gaussian variables



# ICA by Maximization of Nongaussianity

---

- Nongaussian is independent
  - Central limit theorem
  - Sums of nongaussian random variables are closer to gaussian than the original ones.
  - A linear combination of the observed mixture variables will be maximally nongaussian if it equals one of the independent components.



# Measures of Nongaussianity

---

- Kurtosis (fourth-order cumulant)
  - $\text{kurt}(y) = E\{y^4\} - 3(E\{y^2\})^2$
  - Zero for a gaussian random variable
  - Negative for subgaussian random variables (flat pdf)
  - Positive for supergaussian random variables (spiky pdf with heavy tails)
  - Nongaussianity: absolute value of kurtosis
  - Start from some weight vector  $\mathbf{w}$  compute the direction in which the kurtosis of  $y = \mathbf{w}^T x$  is growing or decreasing most strongly and find a new  $\mathbf{w}$  using gradient methods.
  - Sensitive to outliers from the sample.



# Measures of Nongaussianity

---

- Negentropy

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

- $H(\mathbf{y})$  is the differential entropy  $H$  of a random vector  $\mathbf{y}$  with density  $f(\mathbf{y})$ ;  $\mathbf{y}_{gauss}$  is a gaussian random variable of the same covariance matrix as  $\mathbf{y}$ .

- Computationally very difficult: requires an estimate (possibly nonparametric) of the pfd.
  - The optimal estimator of nongaussianity concerning statistical properties.



# Measures of Nongaussianity

---

- Approximations of negentropy
  - $G$  is a practical and non-quadratic function,  $c$  is a constant and  $v$  is a Gaussian variable of zero mean and unit variance.

$$J(y) \approx c[E\{G(y)\} - E\{G(v)\}]^2$$

- Choosing a  $G$  that does not grow too fast help obtaining more robust estimators.
- Give a very good compromise between the properties of Kurtosis and Negentropy
- Fast to compute; have appealing statistical properties especially robustness.



# ICA by Maximum Likelihood Estimation

---

- Formulate the likelihood in the ICA model

Denoting  $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_n)^T$

$f_i$  are the density functions of  $s_i$  (here assumed to be known)

log-likelihood takes the form

$$L = \sum_{t=1}^T \sum_{i=1}^n \log f_i(\mathbf{w}_i^T x(t)) + T \log |\det W|$$

- Estimate the model by maximum likelihood methods
- Require the knowledge of the probability densities of the independent components
- Be very sensitive to outliers.



# ICA by Minimization of Mutual Information

---

- A natural measure of the dependence between random variables: mutual information  $I$  between  $m$  random variables  $y_i$  is defined as

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

- $I$  is always non-negative and is zero iff  $y_i$  are statistically independent.
- For an invertible linear transformation  $\mathbf{y} = \mathbf{W}\mathbf{x}$ :

$$I(y_1, y_2, \dots, y_m) = \sum_i H(y_i) - H(x) - \log |\det W|$$

- Choose  $W$  which minimizes the mutual information between the components  $y_i$  is a natural way of estimating the ICA model.
- Restricted by the probability density of independent components



# Algorithms for ICA

---

- Introduction

- Choose one of the principles of estimation for ICA
- Practical algorithms to optimized the objective function corresponding to the principle.
- The statistical properties of the ICA method depend only on the objective functions used.



# Preprocessing of the Data

---

## Preprocessing:

Makes the problem of ICA estimation simpler and better conditioned.

- Centering: center  $\mathbf{x}$  by subtract its mean vector
  - Centering implies that  $\mathbf{s}$  is zero-mean as well.
  - Centering the observed  $\mathbf{x}$  is solely to simplify ICA algorithms
  - After estimating the mixing matrix with the centered data, mean vector of  $\mathbf{s}$ , which is  $WE\{\mathbf{x}\}$ , can be added back.



# Preprocessing of the Data

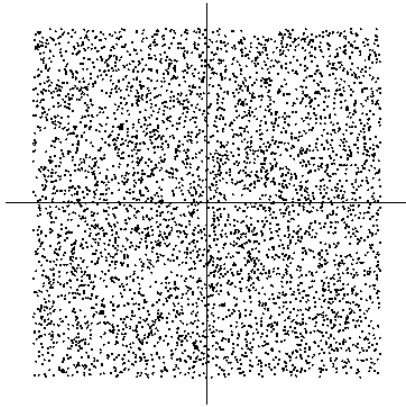
---

- **Whitening**

- Transform the observed  $\mathbf{x}$  linearly to obtain a white one, whose components are uncorrelated with unity variances.
- Transforms the mixing matrix into an orthogonal new one.
- Reduces the number of parameters to be estimated (from  $n^2$  to  $n(n-1)/2$ ).
- May also reduce the dimension of the data (e.g. discard those with very small eigenvalue in EVD).

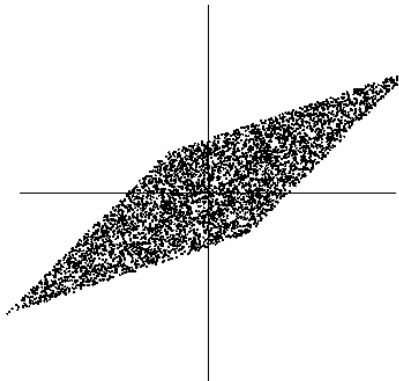
# Preprocessing of the Data

## ■ Example of whitening



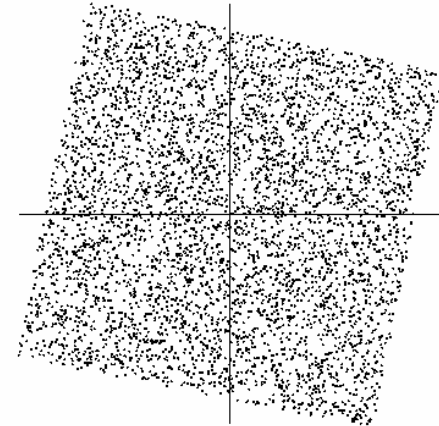
The joint distribution of the independent components  $s_1$  and  $s_2$  with uniform distributions.

Horizontal axis:  $s_1$ , vertical axis:  $s_2$



The joint distribution of observed mixtures  $x_1$  and  $x_2$ .

Horizontal axis:  $x_1$  vertical axis:  $x_2$



The joint distribution of the whitened mixtures

The whitened mixtures is clearly a rotated version of the original data. Estimation is left to be a single angle that gives the rotation.



# Algorithms for ICA

---

- FastICA algorithm

- In many practical situations the convergence of adaptive algorithms based on gradient descent is often slow.
- Batch (block) algorithms based on fixed-point iteration are remedies for this problem.
- The FastICA algorithm is based on a fixed-point iteration scheme for finding a maximum of the nongaussianity of  $\mathbf{w}^T \mathbf{x}$ , measured in negentropy  $J(y)=[E\{G(y)\}-E\{G(v)\}]^2$ , where  $G$  is any non-quadratic function, and  $v$  is a standardized Gaussian variable.



# Algorithms for ICA

---

- One-unit FastICA algorithm :

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

Convergence means the old and new values of  $\mathbf{w}$  point in the same direction (dot product is equal to 1).

Considering the computations, the averages can be estimated using a smaller sample. The sample points should be chosen separately at every iteration. If the convergence is not satisfactory, one may increase the sample size.



# Algorithms for ICA

---

- FastICA algorithm for several units
  - To estimate several independent components, the one-unit FastICA is run several times to obtain vectors  $\mathbf{w}_1, \dots, \mathbf{w}_n$ .
  - To prevent different vectors from converging to the same maximum, a decorrelation step is added after each iteration.



# Algorithms for ICA

---

- Properties of FastICA
  - The convergence is cubic instead of linear convergence for ordinary ICA algorithms based on gradient descent method.
  - Finds directly independent components of any non-Gaussian distribution in contrast to many algorithms, where some estimate of the pdf has to be first available.
  - The independent components can be estimated one by one. This decreases the computational load in cases where only some of the independent components need to be estimated.



# Extensions of Basic ICA Framework

---

- Estimation of the noisy ICA model
  - In practice it may be unrealistic to assume that the data could be divided into signals and noise in any meaningful way
  - Noisy ICA model gives estimation of ordinary latent variable.
- Estimates the model with overcomplete bases  
More independent components than observed mixtures



# Extensions of Basic ICA Framework

---

- Tailor ICA methods to a given practical application

Example: estimate a ICA-similar-model in which the components are not necessarily all independent.

It would be interesting to formulate the exact conditions that enable the such estimation.

- Avoiding and detecting overlearning

Overlearning occurs if there are not enough samples in the data or there is a considerable amount of noise present. This results in the generation of spike-like signals.



# Extensions of Basic ICA Framework

---

- Integrating time-correlations in ICA methods

If  $\mathbf{x}(t)$  come from a stochastic process, instead of being a sample of a random variable, blind source separation can also be accomplished by using time-correlations.

- Make together the blind source separation and some kind of blind deconvolution consider particular phenomena
  - There may be time delays of signals propagation
  - Echo in blind source separation



# Reference

---

- **Survey on Independent Component Analysis**, Aapo Hyvarinen, Helsinki University of Technology, Laboratory of Computer and Information Science.
- **Independent Component Analysis: a Tutorial**, Aapo Hyvarinen, Helsinki University of Technology, Laboratory of Computer and Information Science.
- **Independent Component Analysis Introduction**, A. Hyvärinen, J. Karhunen, E. Oja, 2001 John Wiley & Sons.
- **A Unifying Information-theoretic Framework for Independent Component Analysis**, Te-Won Lee, Mark Girolami, Anthony J. Bell and Terrence J. Sejnowski, International Journal on Mathematical and Computer Models ,1999