Introduction to Independent Component Analysis

Barnabás Póczos

University of Alberta

Nov 26, 2009

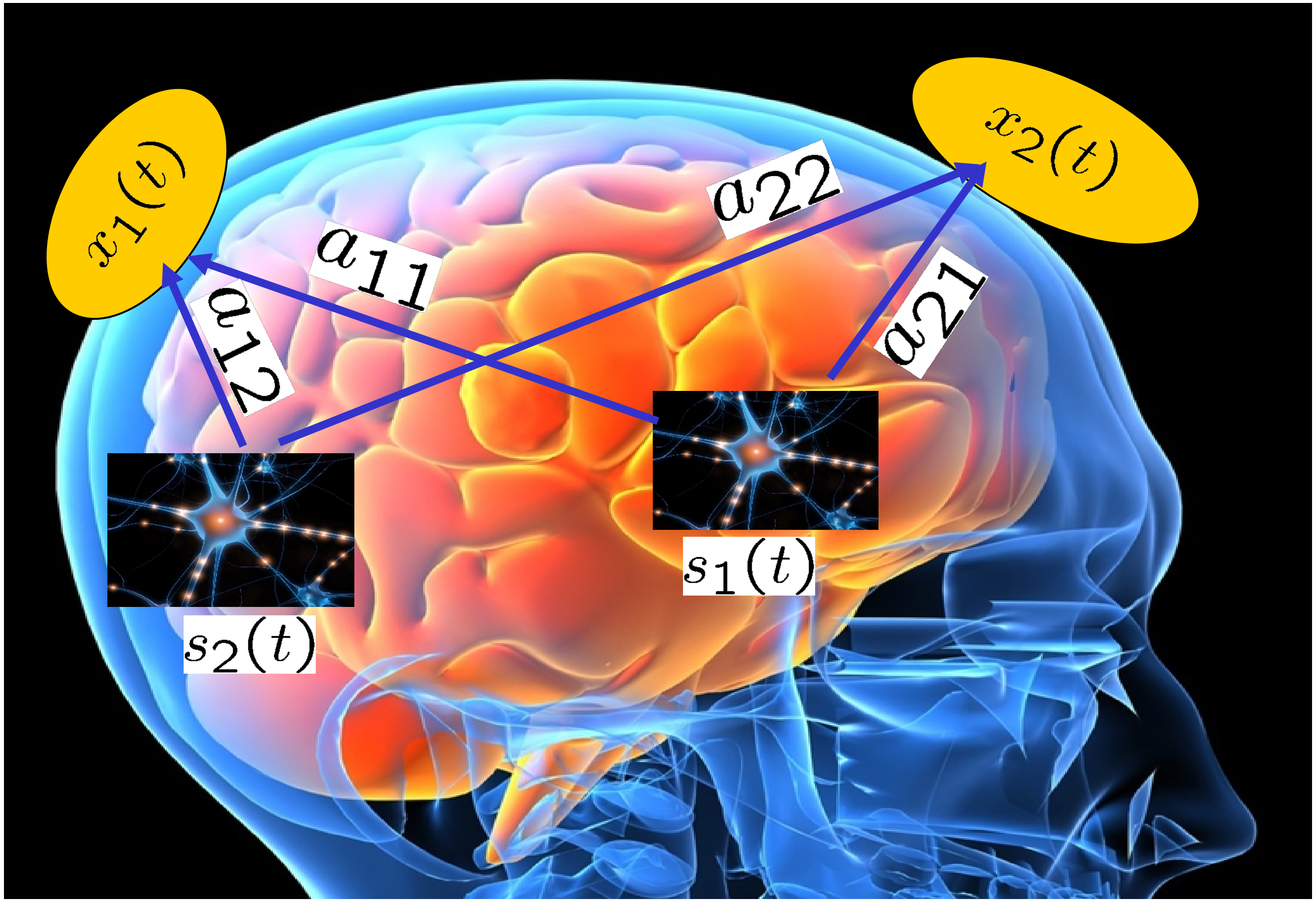
Contents

- Independent Component Analysis
  - ICA model
  - ICA applications
  - ICA generalizations
  - ICA theory
- Independent Subspace Analysis
  - ISA model
  - ISA theory
  - ISA results

Independent Component Analysis

$$
\begin{aligned}
x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\
x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t)
\end{aligned}
$$

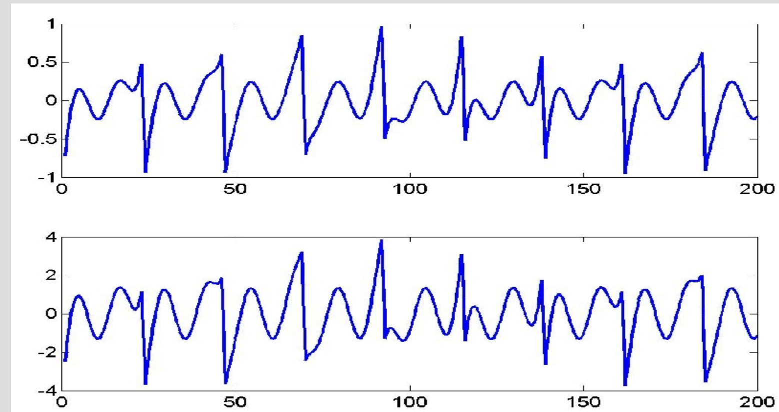
Goal: Estimate $\{s_i(t)\}$,
(and also $\{a_{ij}\}$)



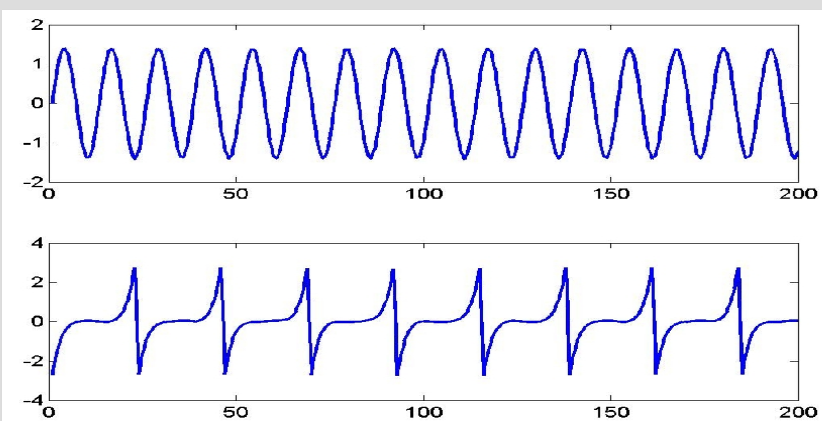
Independent Component Analysis

$$
\begin{aligned}
x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\
x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t)
\end{aligned}
$$

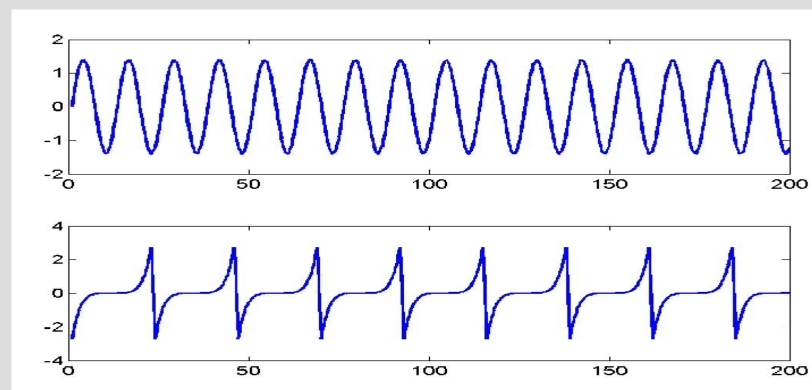
Model



Observations (Mixtures)



ICA estimated signals



original signals

Independent Component Analysis

Model

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned}$$

We observe

$$\begin{pmatrix} x_1(1) \\ x_2(1) \end{pmatrix}, \begin{pmatrix} x_1(2) \\ x_2(2) \end{pmatrix}, \ldots, \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$$

We want

$$\begin{pmatrix} s_1(1) \\ s_2(1) \end{pmatrix}, \begin{pmatrix} s_1(2) \\ s_2(2) \end{pmatrix}, \ldots, \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix}$$

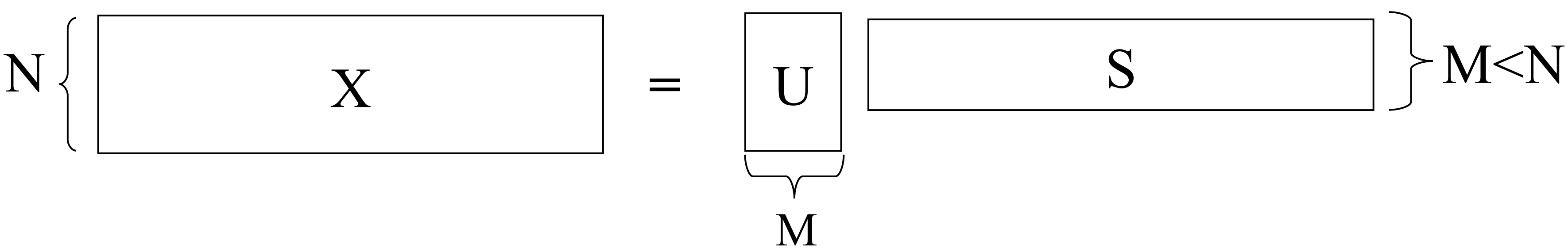
But we don't know $\{a_{ij}\}$, nor $\{s_i(t)\}$

Goal: Estimate $\{s_i(t)\}$, (and also $\{a_{ij}\}$)

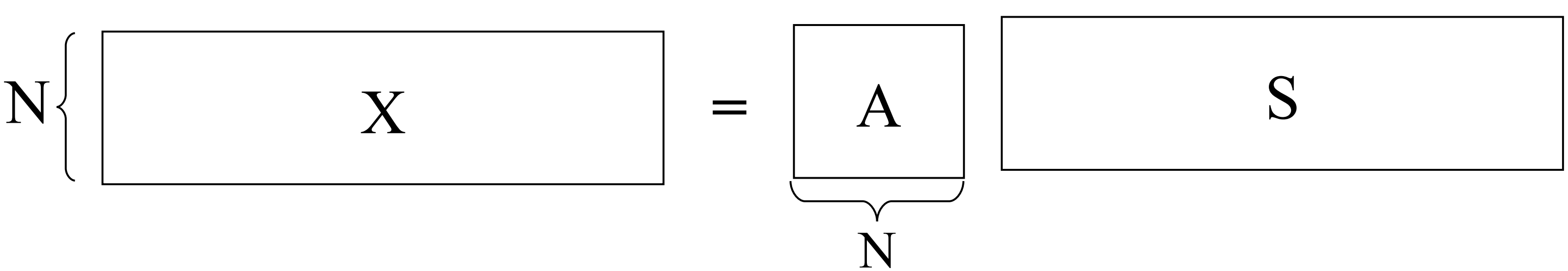
ICA vs PCA, **Similarities**

- Perform linear transformations
- Matrix factorization

PCA: ***low rank*** matrix factorization for ***compression***



ICA: ***full rank*** matrix factorization to ***remove dependency*** between the rows



ICA vs PCA, **Differences**

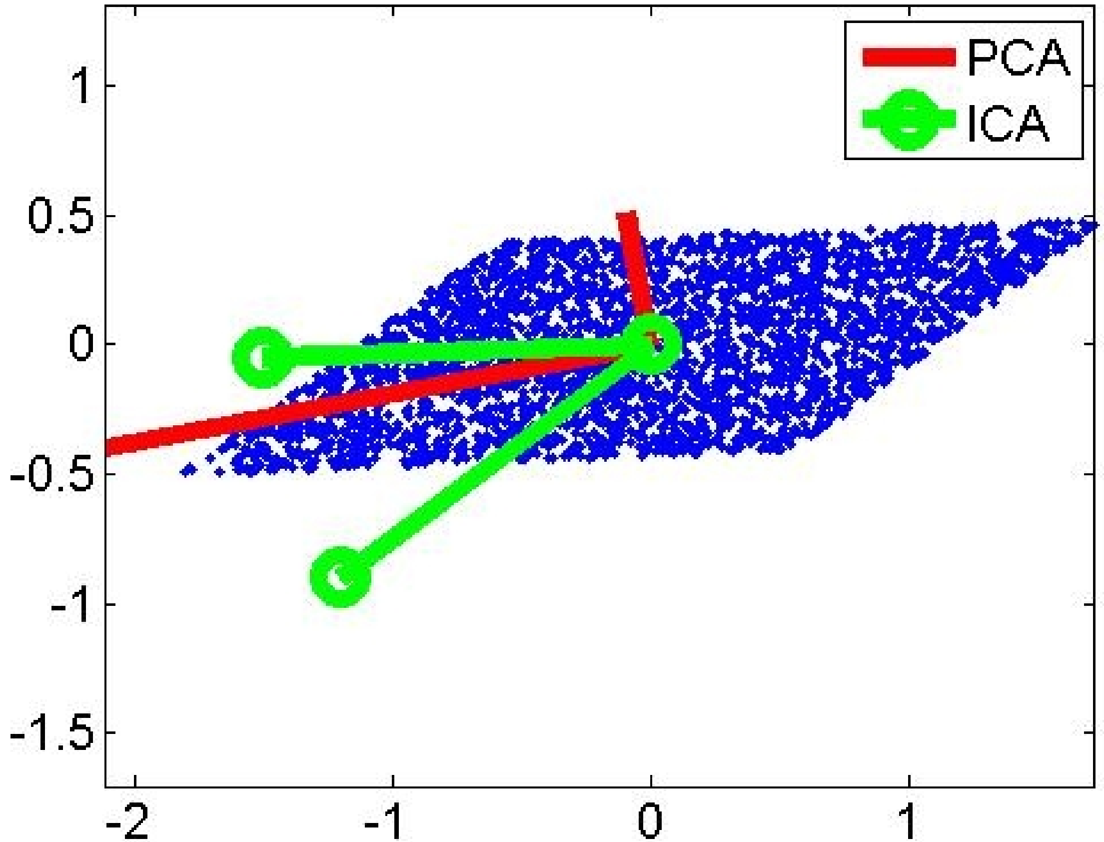
- PCA: $\mathbf{X=US,\ U^TU=I}$
- ICA: $\mathbf{X=AS}$

- PCA **does** compression
  - M<N
- ICA does **not** do compression
  - same # of features (M=N)

- PCA just removes correlations, **not** higher order dependence
- ICA removes correlations, **and** higher order dependence

- PCA: some components are **more important** than others (based on eigenvalues)
- ICA: components are **equally important**

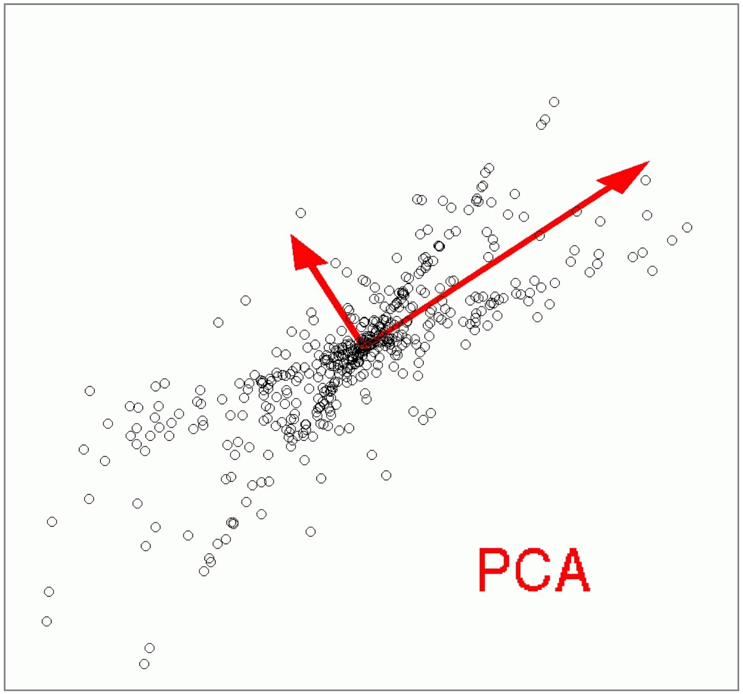
ICA vs PCA

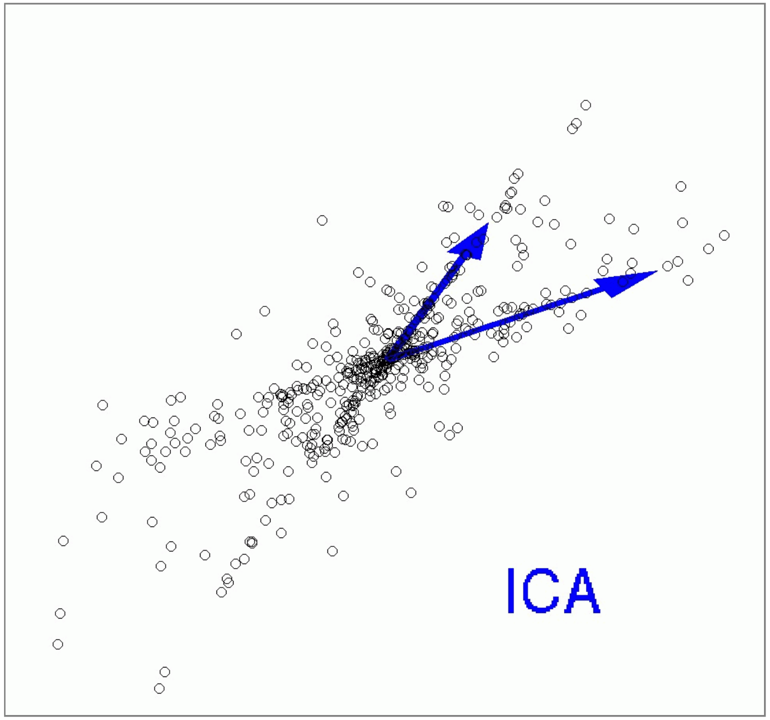


Note

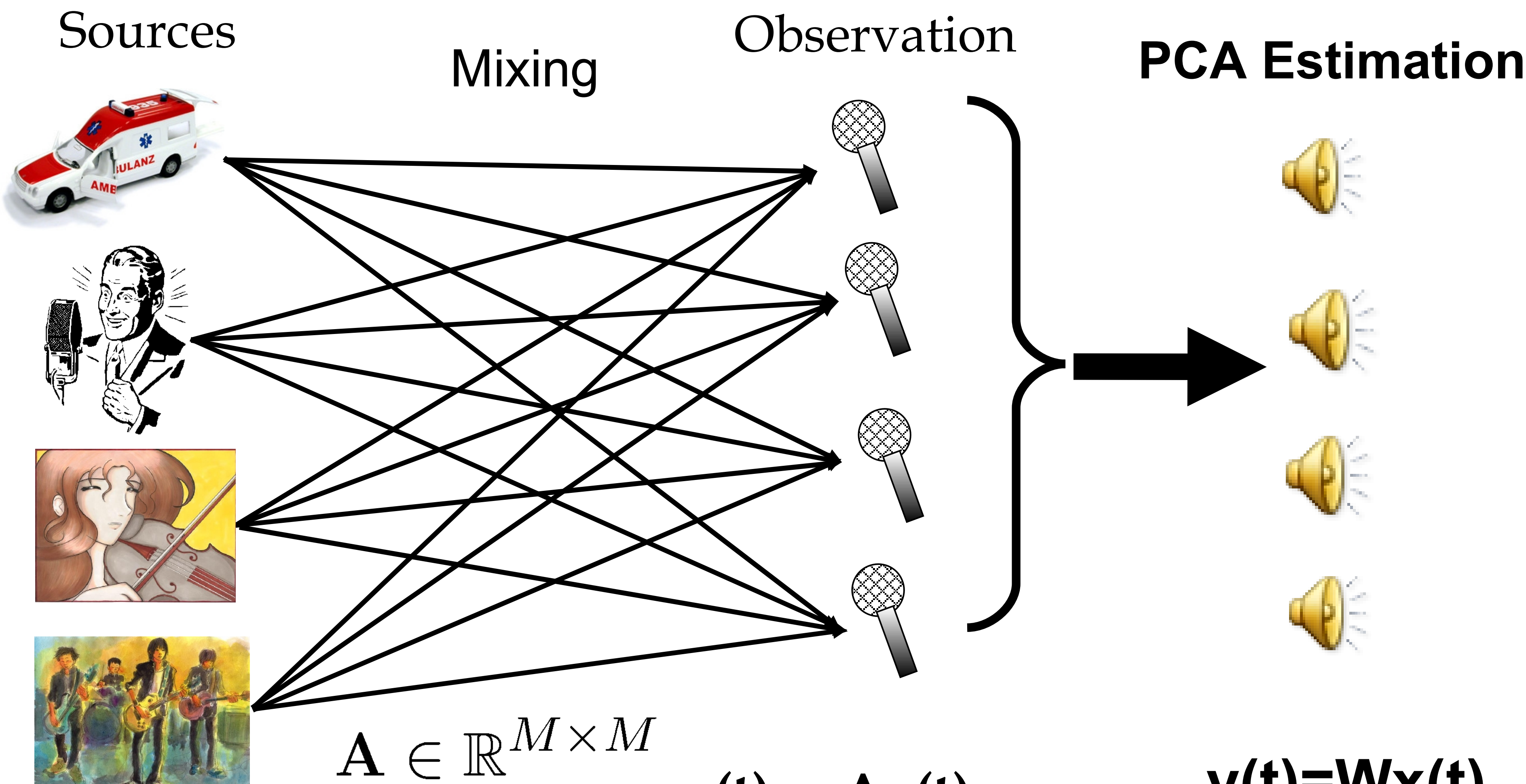
- PCA vectors are orthogonal
- ICA vectors are **not** orthogonal

PCA vs ICA





The Cocktail Party Problem
SOLVING WITH PCA



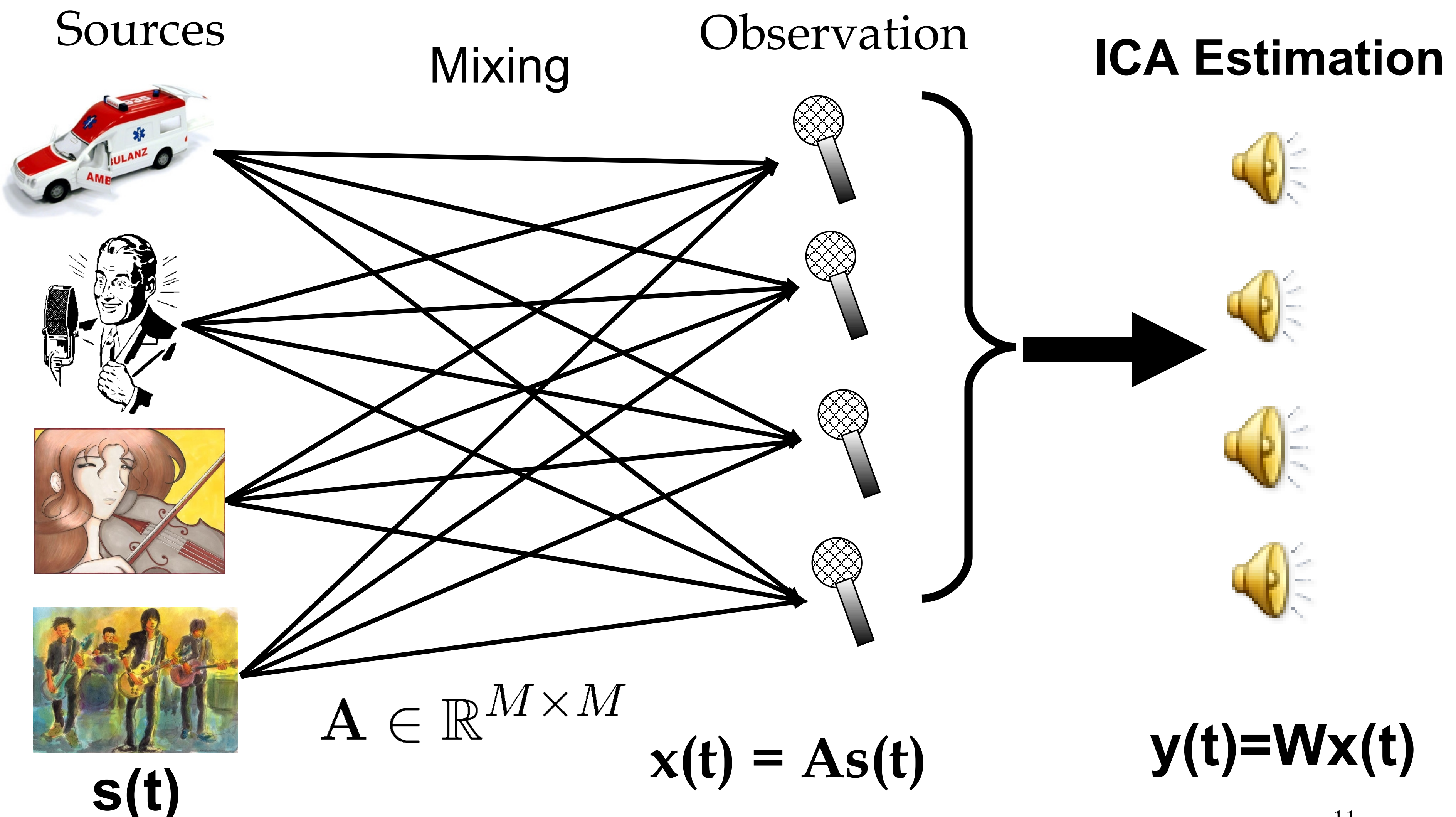
$\mathbf{A} \in \mathbb{R}^{M \times M}$

s(t)

x(t) = As(t)

y(t)=Wx(t)

The Cocktail Party Problem
SOLVING WITH ICA



Some ICA Applications

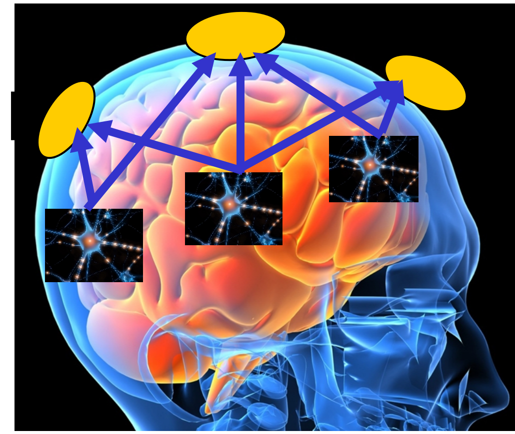
STATIC

- Image denoising
- Microarray data processing
- Decomposing the spectra of galaxies
- Face recognition
- Facial expression recognition
- Feature extraction
- Clustering
- Classification

TEMPORAL

- Medical signal processing – fMRI, ECG, EEG
- Brain Computer Interfaces
- Modeling of the hippocampus, place cells
- Modeling of the visual cortex
- Time series analysis
- Financial applications
- Blind deconvolution

ICA Application, Removing Artifacts from EEG



- EEG ~ *Neural cocktail party*
- Severe ***contamination*** of EEG activity
  - eye movements
  - blinks
  - muscle
  - heart, ECG artifact
  - vessel pulse
  - electrode noise
  - line noise, alternating current (60 Hz)

- ICA can improve signal
  - effectively ***detect, separate and remove*** activity in EEG records from a wide variety of artifactual sources. (Jung, Makeig, Bell, and Sejnowski)

- ICA weights help find **location** of sources

ICA decomposition

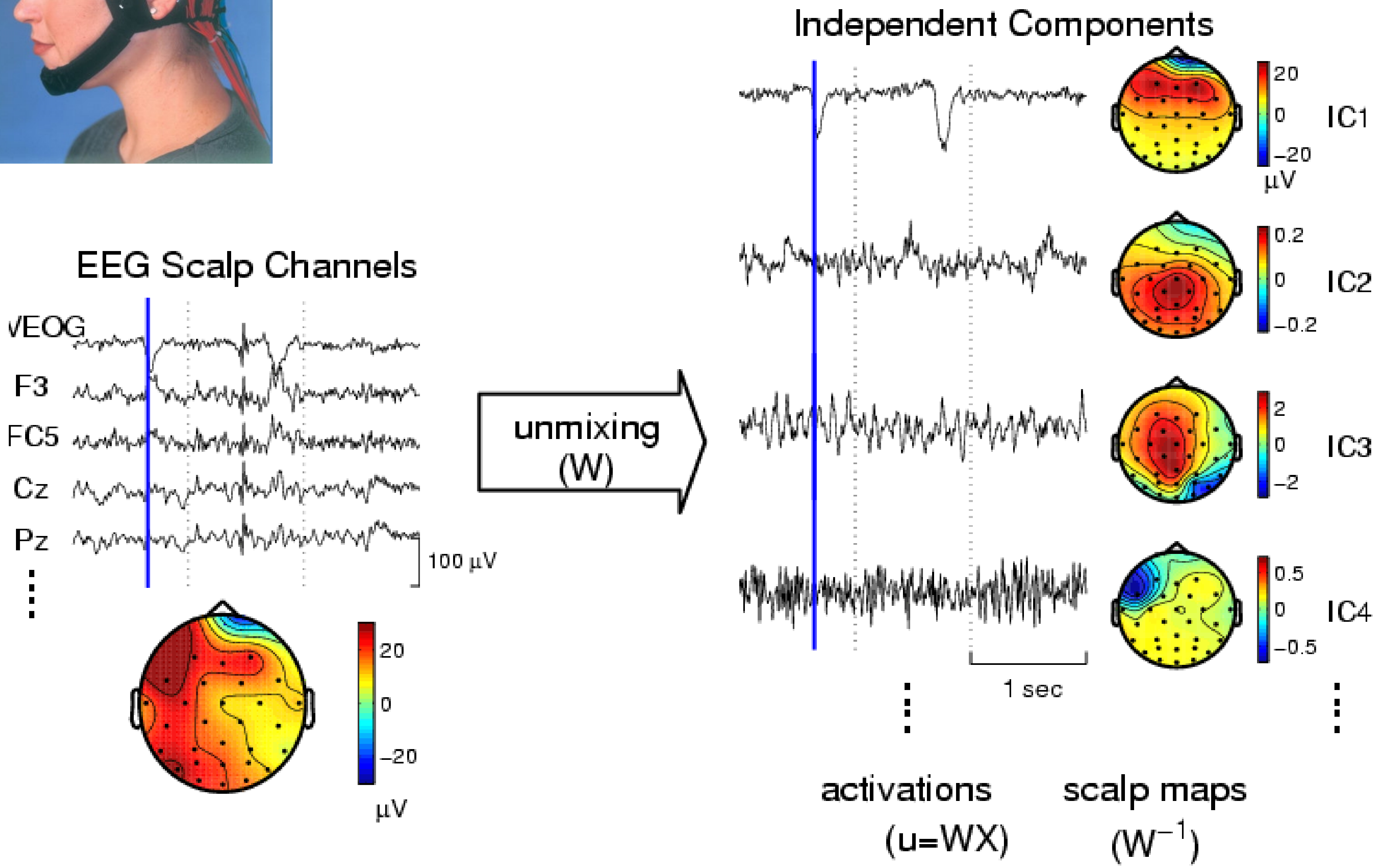


Fig from Jung

Summed Projection of Selected Components

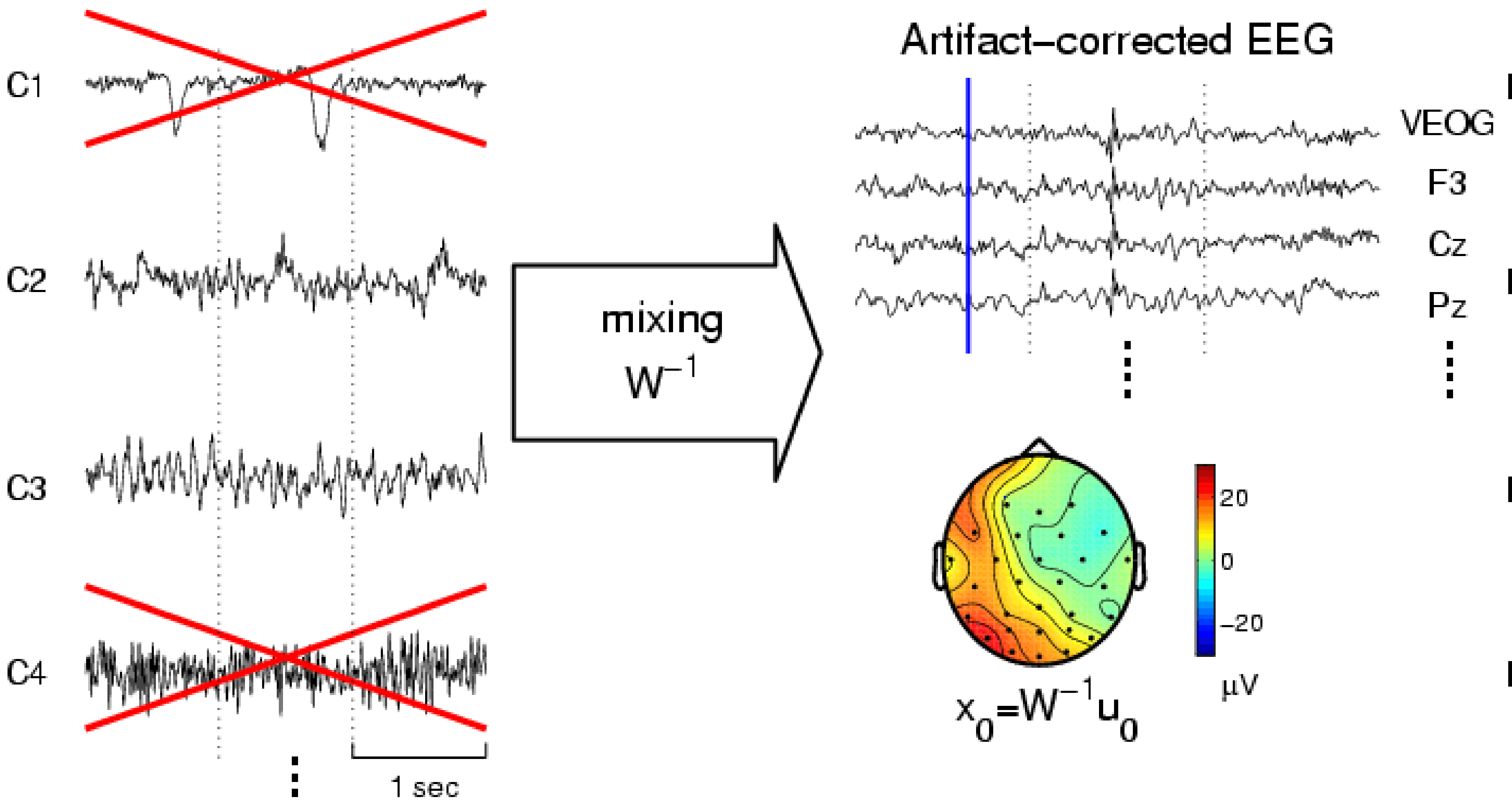


Fig from Jung

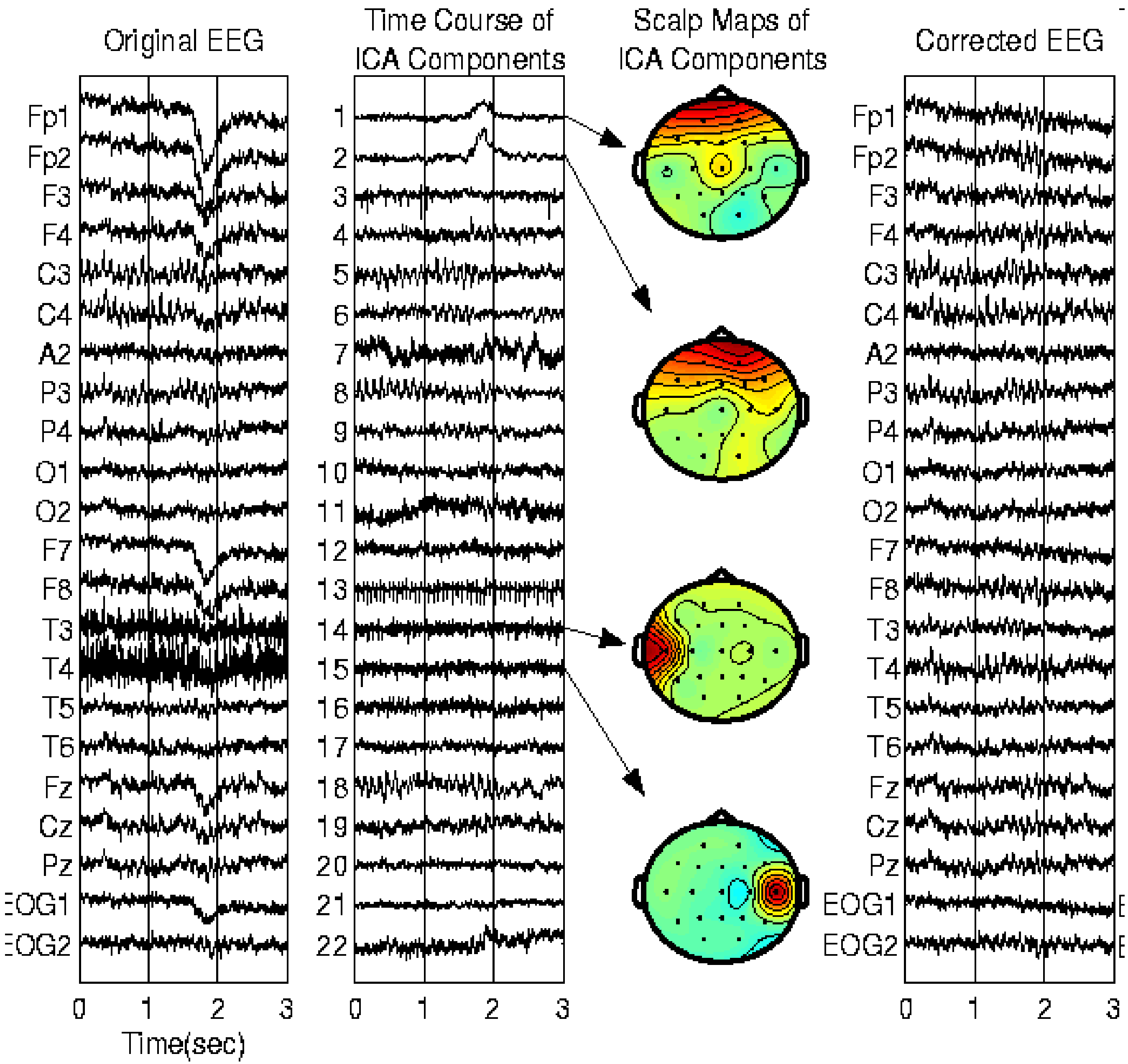
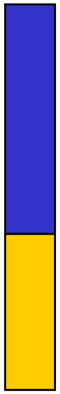
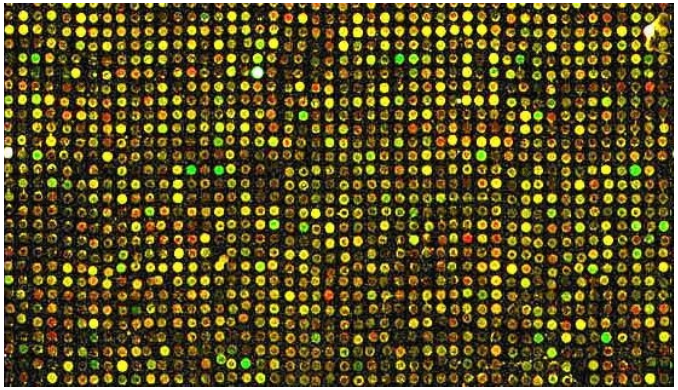


Fig from Jung

PCA+ICA for Microarray data processing



labels

$\mathbf{X}^T =$  $= \mathbf{A} * \mathbf{S}$

$\mathbf{a}_k$ $\mathbf{s}_k$

$\mathbf{X}^T \in R^{M \times N}$

M = number of experiments

N = number of genes

PCA alone can estimate **US** only $\Rightarrow$ doesn't work

Assumption:

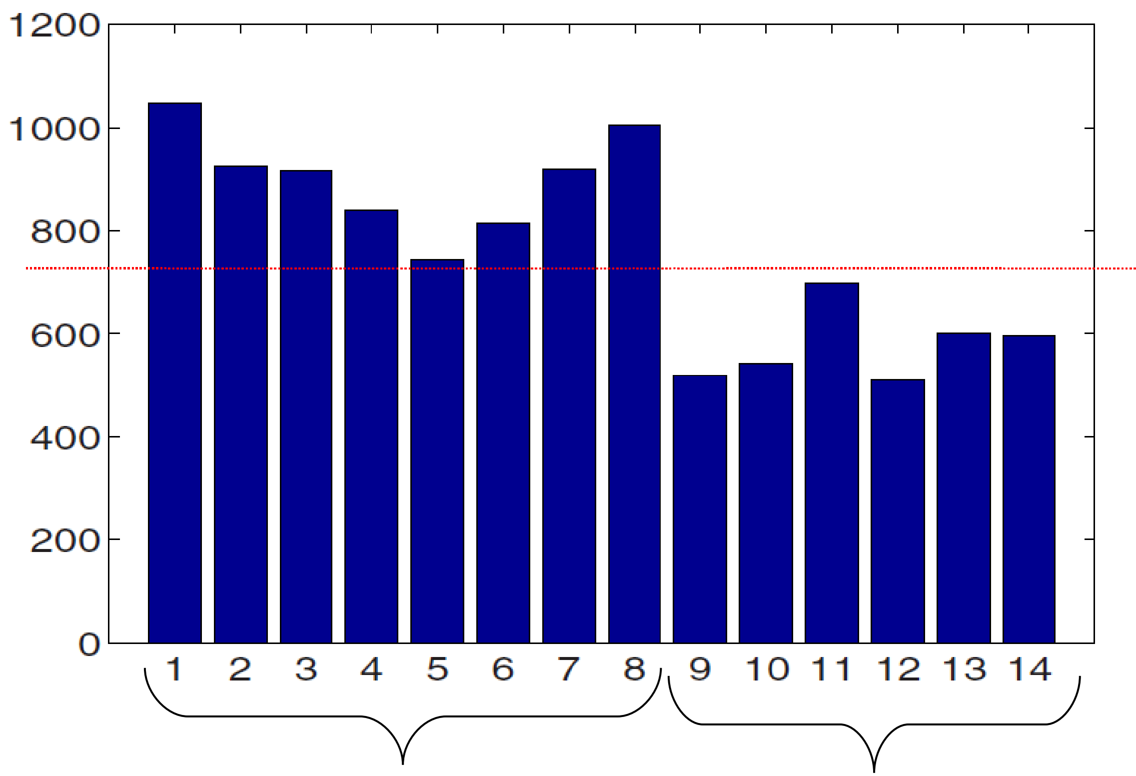
- each experiment is a mixture of **independent expression modes** ($\mathbf{s}_1, \ldots \mathbf{s}_K$).
- some of these modes (e.g. $\mathbf{s}_k$) can be related to the difference between the classes.
  - $\rightarrow$ $\mathbf{a}_k$ correlates with the class labels

ICA for Microarray data processing
(Schachtner et al, ICA07)

Breast Cancer Data set

M=14 patients
N=22283 genes
2 classes

9th column of **A**:



Class 1, weak metastasis

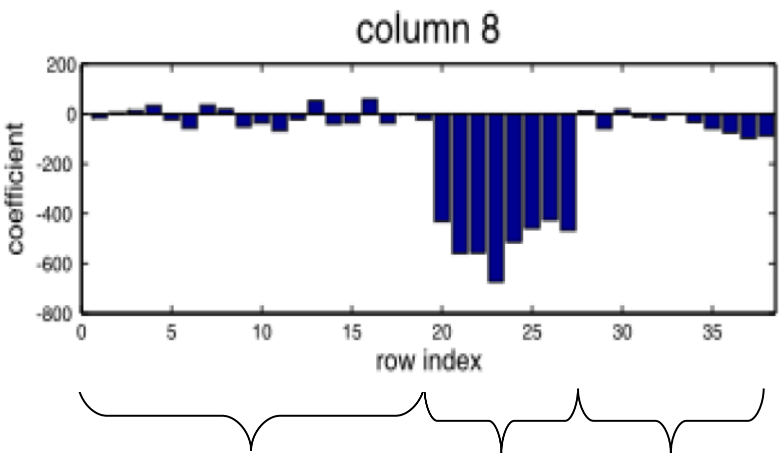
Class 2, strong metastasis

|Corr($\mathbf{a}_9$, $\mathbf{d}$)|=0.89, where **d** is the vector of class labels

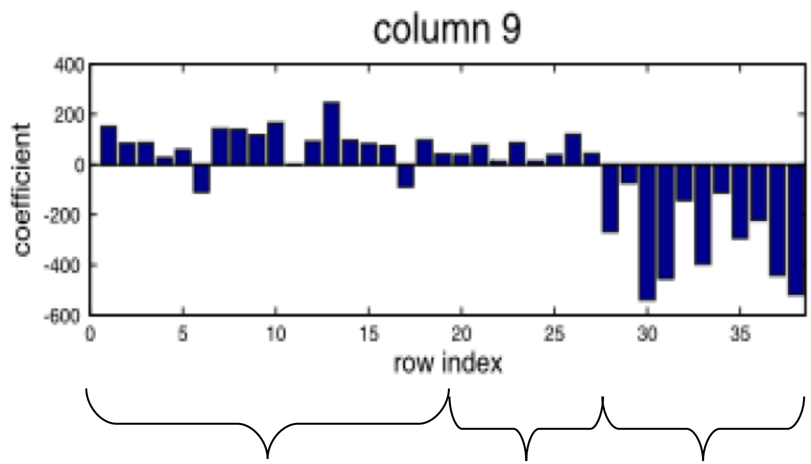
ICA for Microarray data processing

(Schachtner et al, ICA07)

Leukemia Data set M=38 Patients
N=5000 genes
3 classes: ALL-B, ALL-T, AML



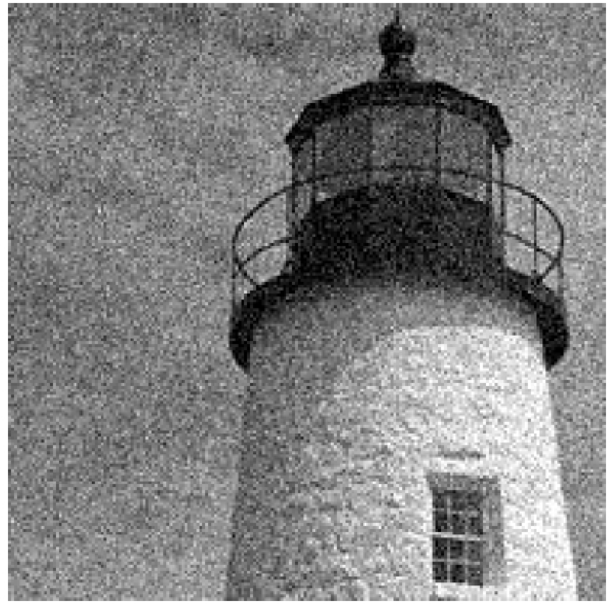
ALL-B ALL-T AML



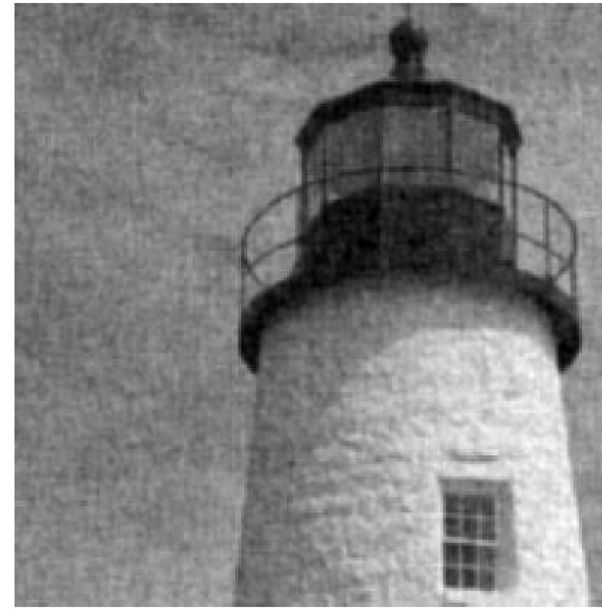
ALL-B ALL-T AML

ICA for Image Denoising
(Hoyer, Hyvarinen)

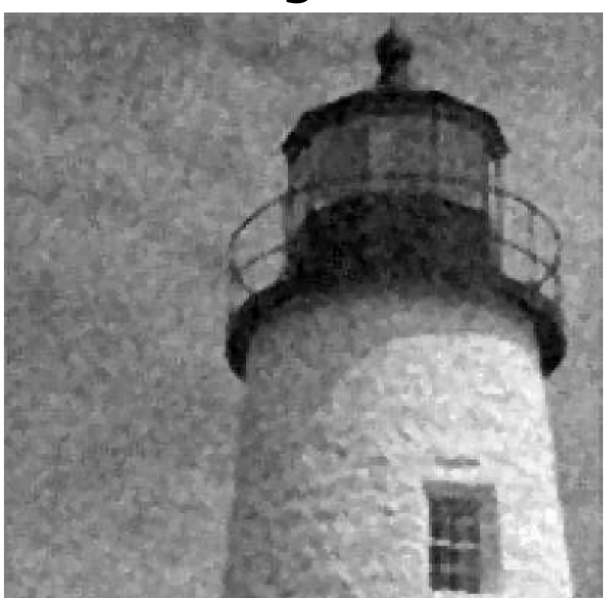
original



noisy

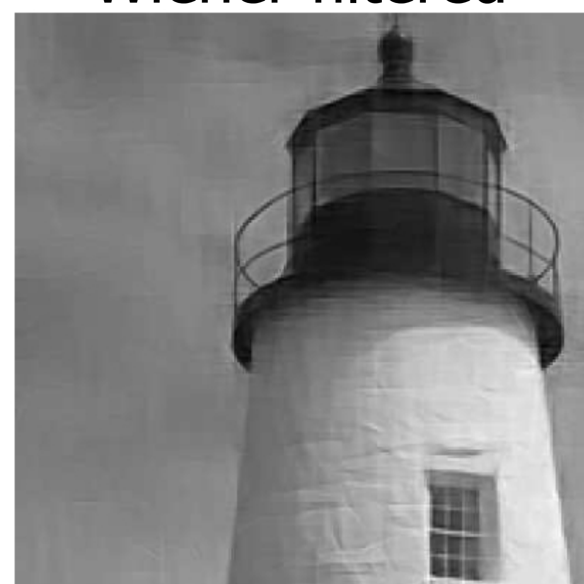


Wiener filtered



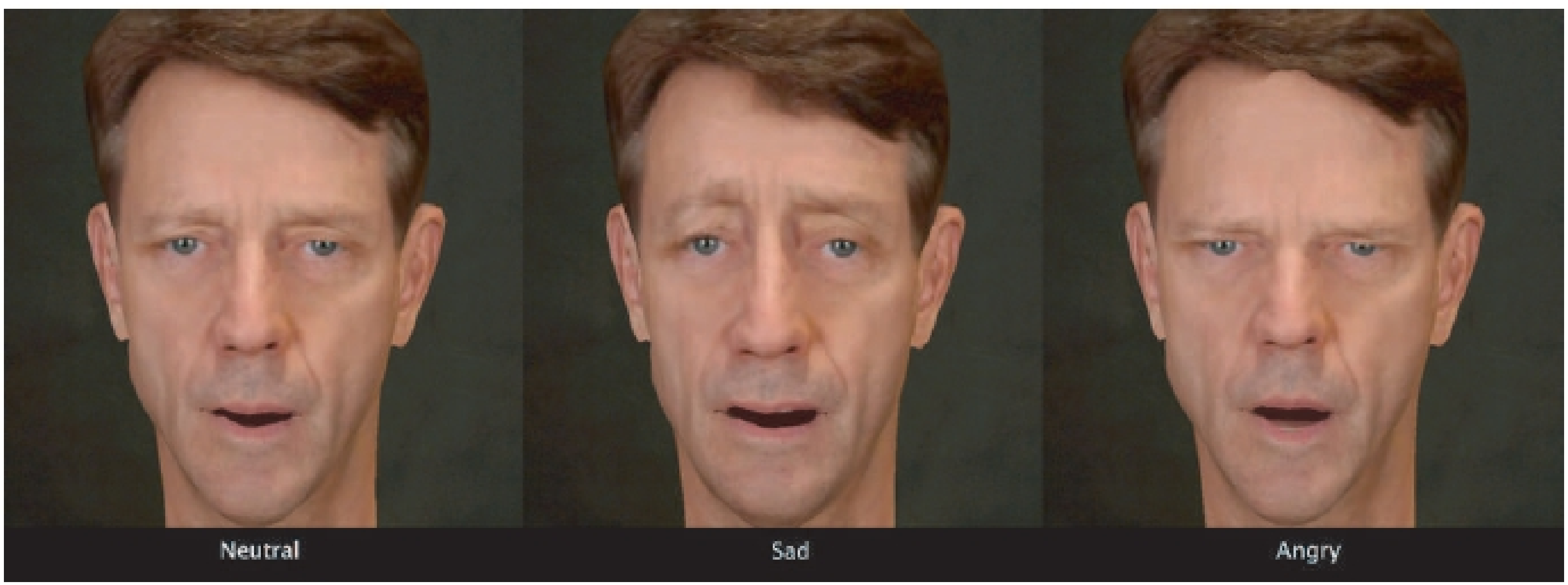
median filtered

ICA denoised

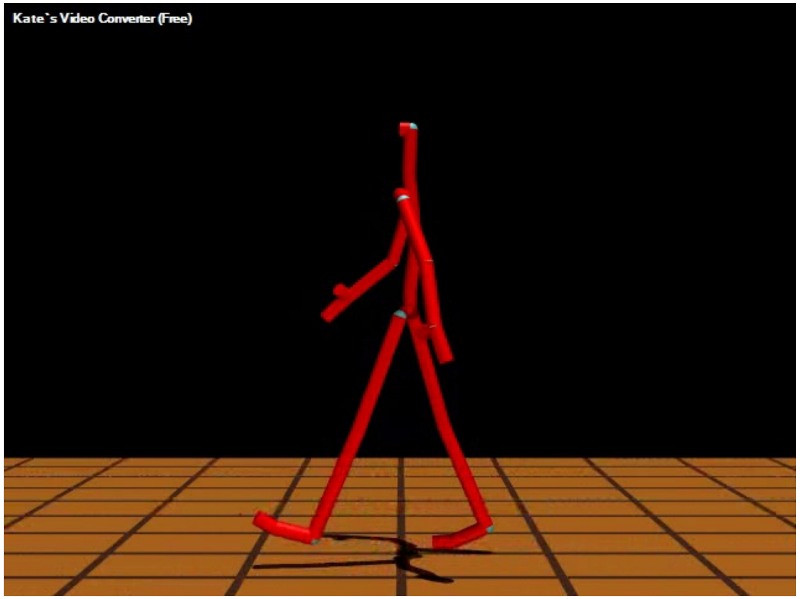


ICA for Motion Style Components
(Mori & Hoshino 2002, Shapiro et al 2006, Cao et al 2003)

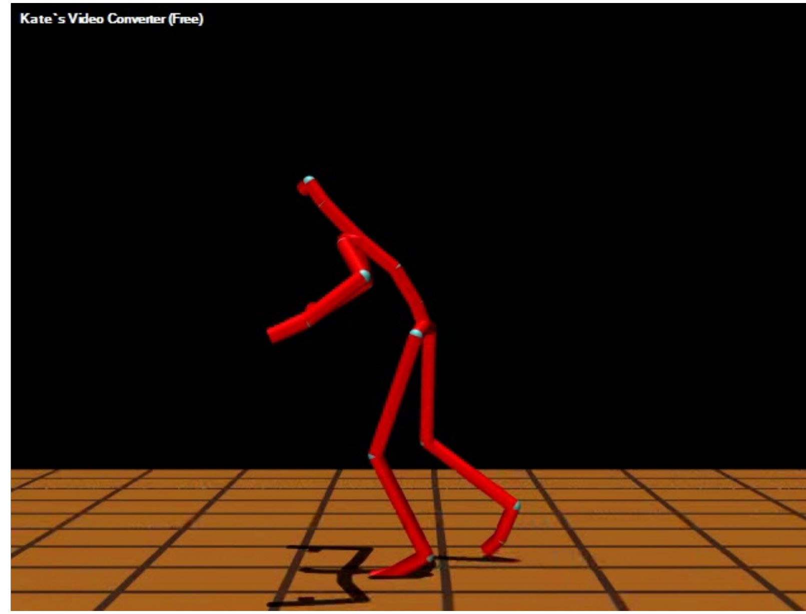
- Method for analysis and synthesis of human motion from motion captured data
- Provides perceptually meaningful components
- 109 markers, 327 parameters
  $\Rightarrow$ 6 independent components (emotion, content,…)



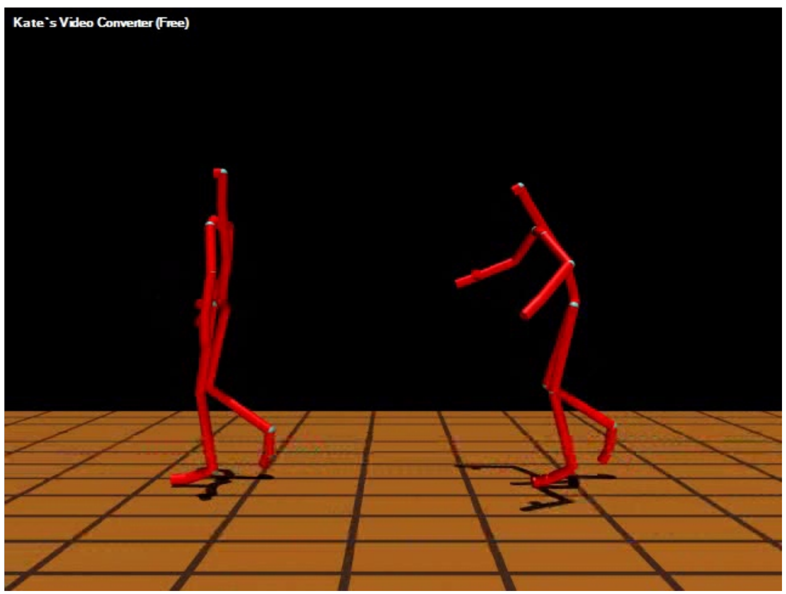

Neutral
%33 Sad + %67 Neutral
%67 Sad + %33 Neutral
Sad
Neutral
%33 Sad + %67 Neutral
%67 Sad + %33 Neutral
Sad



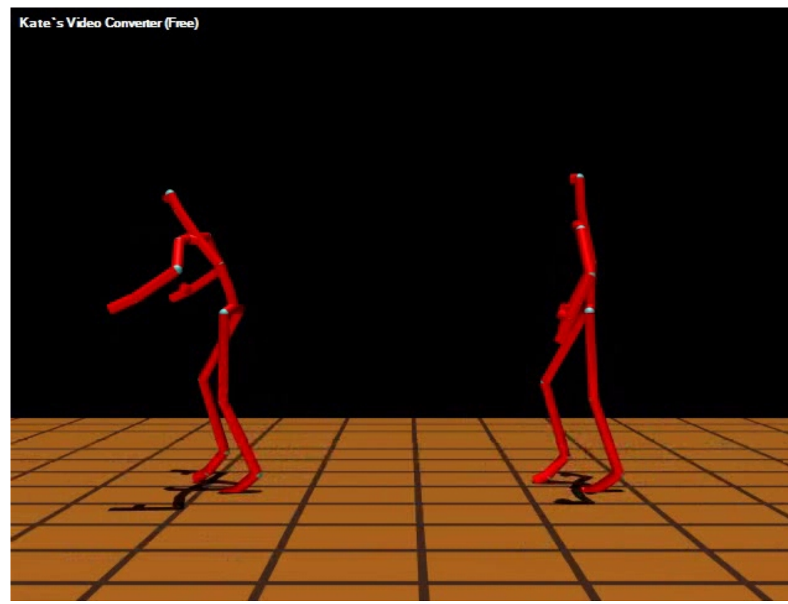
walk



sneaky

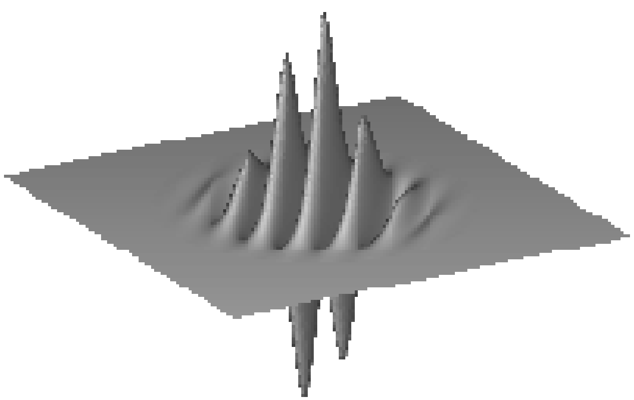


walk with sneaky

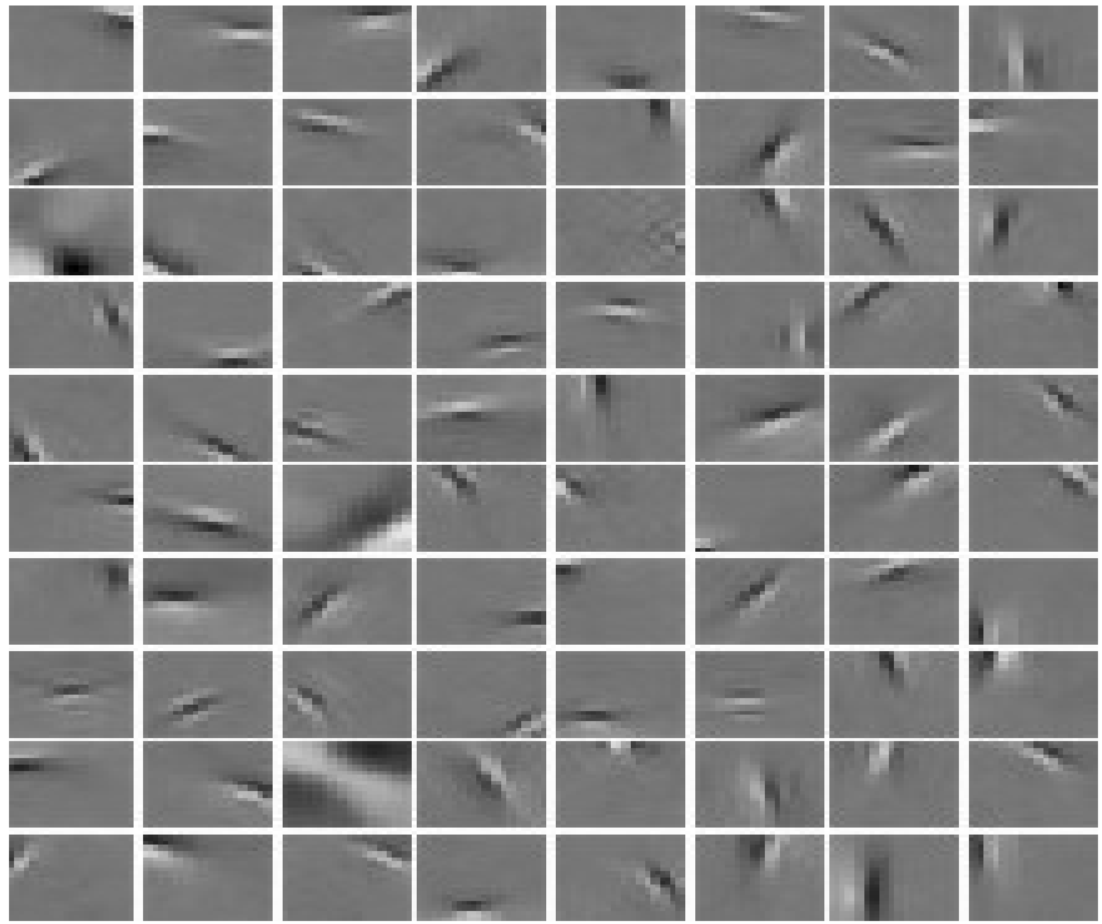


sneaky with walk

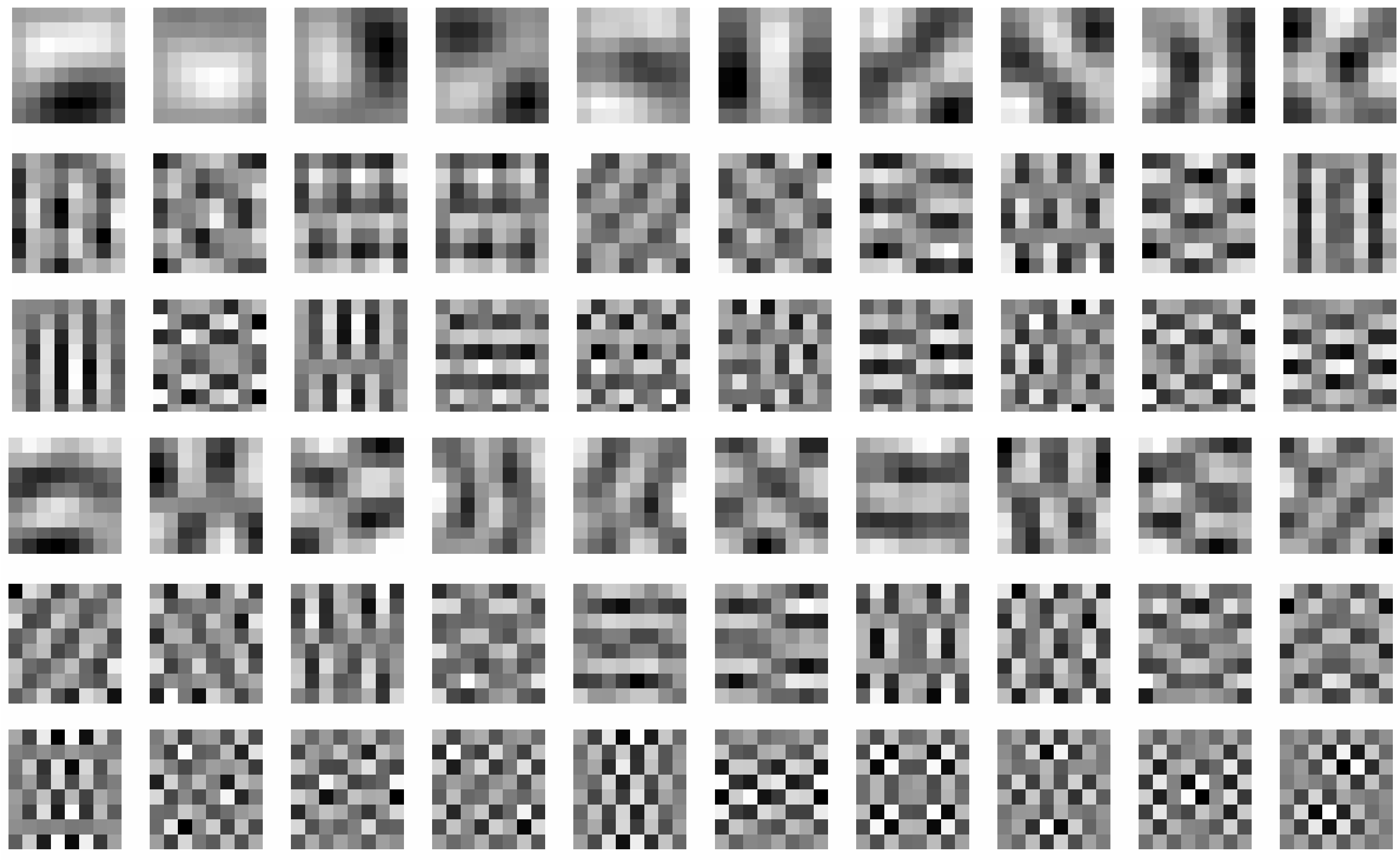
ICA basis vectors extracted from natural images



Gabor wavelets,
edge detection,
receptive fields of V1 cells...



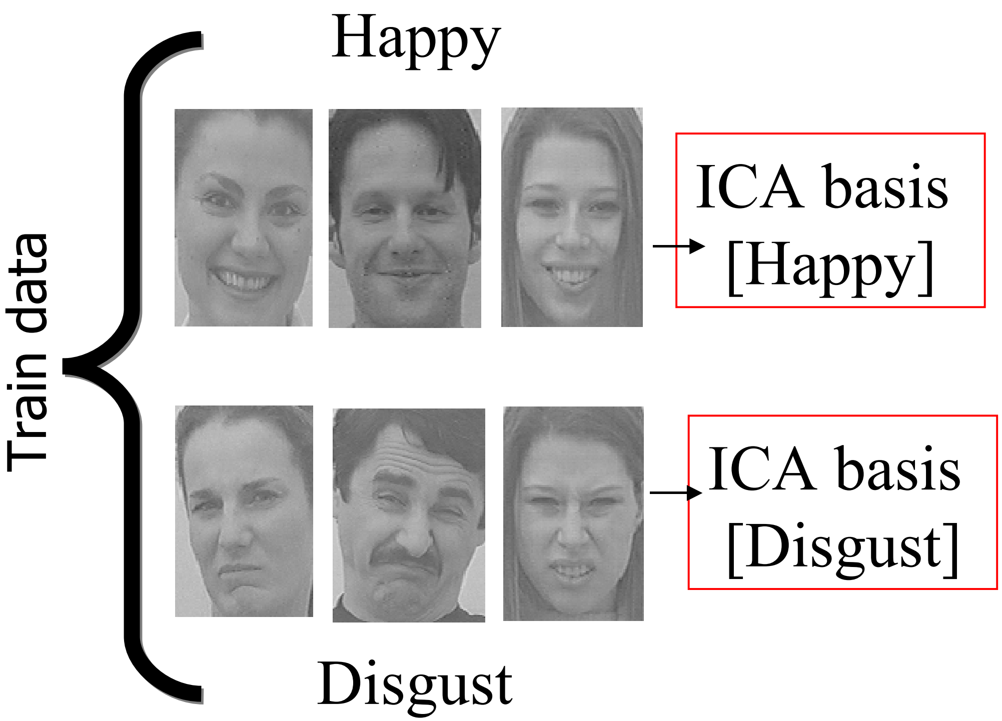
PCA basis vectors extracted from natural images

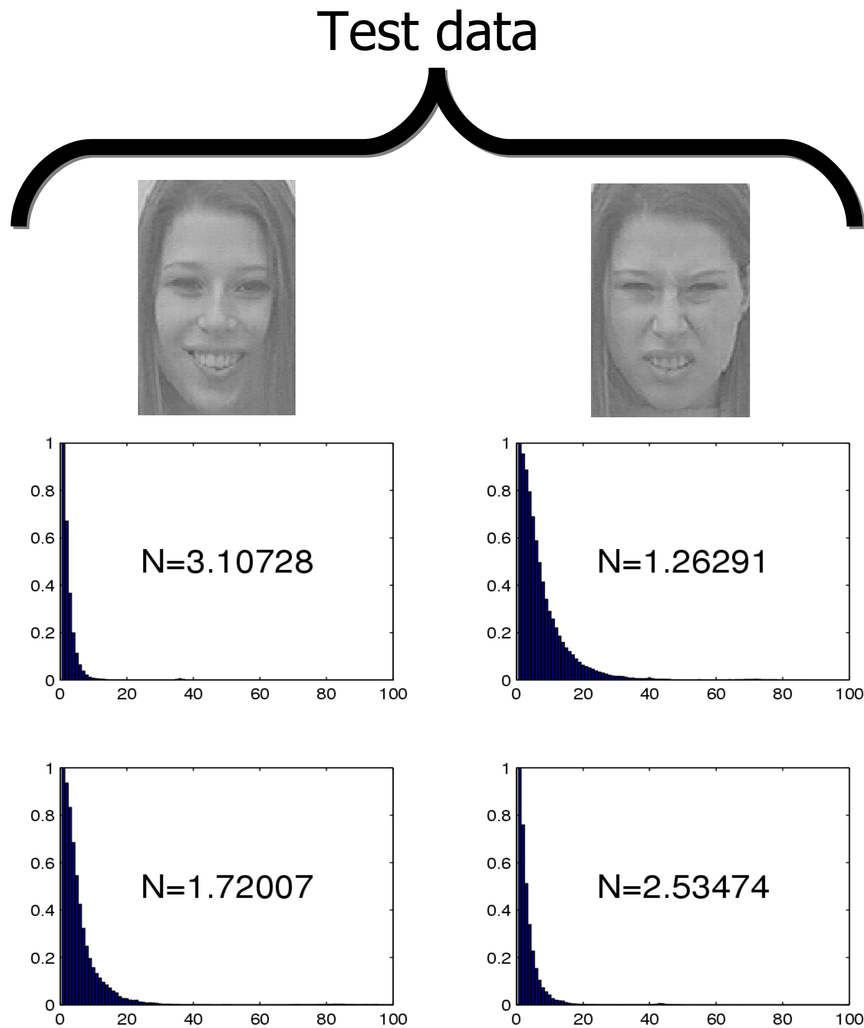


Using ICA for classification

Activity distributions of

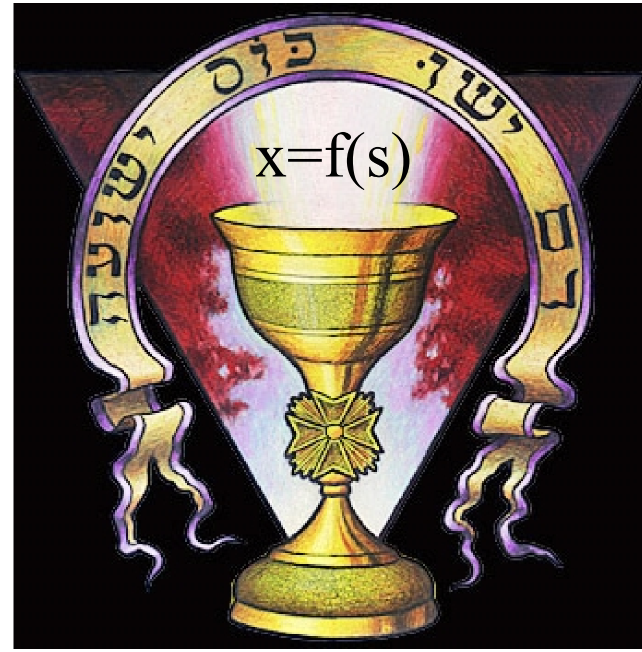
- within-category test images are much narrower
- off-category is closer to the Gaussian distribution





ICA Generalizations

- Independent Subspace Analysis
- Multilinear ICA
- Blind Source Deconvolution
- Blind SubSpace Deconvolution
- Nonnegative ICA
- Sparse Component Analysis
- Slow Component Analysis
- Noisy ICA
- Undercomplete, Overcomplete ICA
- Varying mixing matrix
- Online ICA
- (Post) Nonlinear ICA



The Holy Grail

ICA Theory

Basic terms, definitions

- uncorrelated and independent variables
- entropy, joint entropy, neg_entropy
- mutual information
- Kullback-Leibler divergence

Statistical (in)dependence

Definition:

Y_1, Y_2 are independent $\Leftrightarrow p(y_1, y_2) \ = \ p(y_1)\, p(y_2)$

Lemma:

Let h_1, h_2 be arbitary functions.
Y_1, Y_2 are independent $\Rightarrow$

$$\mathbb{E}[h_1(Y_1)\ h_2(Y_2)] \ = \ \mathbb{E}[h_1(Y_1)]\, \mathbb{E}[h_2(Y_2)]$$

Proof: Homework

Correlation

Definition:

$$corr(Y_1, Y_2) = \frac{\mathbb{E}\Big[(Y_1 - \mathbb{E}[Y_1])\ (Y_2 - \mathbb{E}[Y_2])\Big]}{var(Y_1)^{1/2}\ var(Y_2)^{1/2}}$$

Lemma:

$$corr(Y_1, Y_2) = 0 \Leftrightarrow \mathbb{E}[Y_1 Y_2] = \mathbb{E}[Y_1]\,\mathbb{E}[Y_2]$$

Proof: Homework

Lemma:

Y_1, Y_2 are independent $\underset{\not\Leftarrow}{\Rightarrow}$ Y_1, Y_2 are uncorrelated

Proof: Homework

Lemma: If (Y_1, Y_2) are jointly Gaussian, then Y_1, Y_2 are independent $\Leftrightarrow$ Y_1, Y_2 are uncorrelated

Proof: Homework

Mutual Information, Entropy

Definition (Mutual Information)

$$0 \le I(Y_1,\dots,Y_M) \doteq \int p(y_1,\dots,y_M) \log \frac{p(y_1,\dots,y_M)}{p(y_1)\dots p(y_M)} d\mathbf{y}$$

$$= KL(p(y_1,\dots,y_M) \| p(y_1)\dots p(y_M))$$

$$= \sum_{i=1}^{M} H(Y_i) - H(Y_1,\dots,Y_M)$$

Definition (Shannon entropy)

$$H(\mathbf{Y}) \doteq H(Y_1,\dots,Y_m) \doteq -\int p(y_1,\dots,y_m) \log p(y_1,\dots,y_m) d\mathbf{y}.$$

Definition (KL divergence)

$$0 \le KL(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

Solving the ICA problem with i.i.d. sources

ICA problem: $\mathbf{x} = \mathbf{A}\mathbf{s}$, $\mathbf{s} = [s_1; \ldots; s_M]$ are jointly independent.

Ambiguity:

$\mathbf{s} = [s_1; \ldots; s_M]$ sources can be recovered only up to

sign, scale and permutation.

Proof:

- $\mathbf{P}$ = arbitrary permutation matrix,
- $\mathbf{\Lambda}$ = arbitrary diagonal scaling matrix.

$$\Rightarrow \mathbf{x} = [\mathbf{A}\mathbf{P}^{-1}\mathbf{\Lambda}^{-1}][\mathbf{\Lambda}\mathbf{P}\mathbf{s}]$$

Solving the ICA problem with i.i.d. sources

Lemma:

We can assume that $E[\mathbf{s}] = 0$.

Proof:

Removing the mean does not change the mixing matrix.

$$\mathbf{x} - E[\mathbf{x}] = \mathbf{A}(\mathbf{s} - E[\mathbf{s}]).$$

In what follows we assume that $E[\mathbf{s}\mathbf{s}^T] = \mathbf{I}_M$, $E[\mathbf{s}] = 0$.

Whitening

Let $\mathbf{A} \in \mathbb{R}^{N \times M}$ with full rank, $N \geq M$, and $\mathbf{x} = \mathbf{A}\mathbf{s}$

Theorem (Whitening)

$\exists\ \mathbf{Q} \in \mathbb{R}^{M \times N}$ such that

$$\text{if } \mathbf{x}^* \doteq \mathbf{Q}\mathbf{x} = \mathbf{Q}\mathbf{A}\mathbf{s} = \mathbf{A}^*\mathbf{s},\ \mathbf{A}^* \doteq \mathbf{Q}\mathbf{A}$$

$$\Rightarrow E[\mathbf{x}^*\mathbf{x}^{*T}] = \mathbf{I}_M, \text{ and } \mathbf{A}^*\mathbf{A}^{*T} = \mathbf{I}_M.$$

Definitions

- $\mathbf{x}^* = \mathbf{Q}\mathbf{x}$ transformation is the *whitening* transformation.
- $\mathbf{Q}$ is the *whitening matrix*
- $\boxed{\mathbf{x}^* \doteq \mathbf{A}^*\mathbf{s}}$ is the *whitened* ICA task.

Note After whitening we need only to consider **orthogonal matrices** for (de)mixing. ($\mathbf{A}^*$ is orthogonal)

Proof of the whitening theorem

We can use PCA for whitening!

- Let $\boldsymbol{\Sigma} \doteq cov(\mathbf{x}) = E[\mathbf{x}\mathbf{x}^T] = \mathbf{A}E[\mathbf{s}\mathbf{s}^T]\mathbf{A}^T = \mathbf{A}\mathbf{A}^T$.

- Do **PCA**: $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$, $rank(\boldsymbol{\Sigma}) = M$,

  $\Rightarrow \boldsymbol{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T$,

  where $\mathbf{U} \in \mathbb{R}^{N \times M}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_M$, **Prinicipal vectors**

  $\mathbf{D} \in \mathbb{R}^{M \times M}$, diagonal with rank M. **Prinicipal values**

- Let $\mathbf{Q} \doteq \mathbf{D}^{-1/2}\mathbf{U}^T \in \mathbb{R}^{M \times N}$ *whitening matrix*
- Let $\mathbf{A}^* \doteq \mathbf{Q}\mathbf{A}$
- $\mathbf{x}^* \doteq \mathbf{Q}\mathbf{x} = \mathbf{Q}\mathbf{A}\mathbf{s} = \mathbf{A}^*\mathbf{s}$ is our new (*whitened*) ICA task.

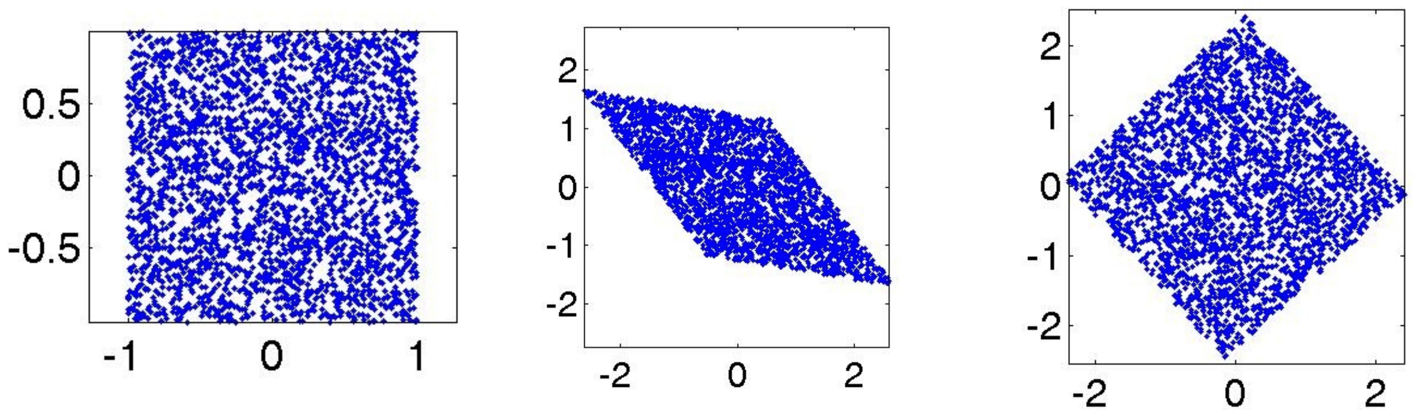
$\Rightarrow \boxed{E[\mathbf{x}^*\mathbf{x}^{*T}] = \mathbf{I}_M, \text{ and } \mathbf{A}^*\mathbf{A}^{*T} = \mathbf{I}_M.}$

Whitening solves half of the ICA problem

Note:

The number of free parameters of an N by N orthogonal matrix is (N-1)(N-2)/2.

$\Rightarrow$ whitening solves **half** of the ICA problem



original mixed whitened

After whitening it is enough to consider **orthogonal matrices** only for separation.

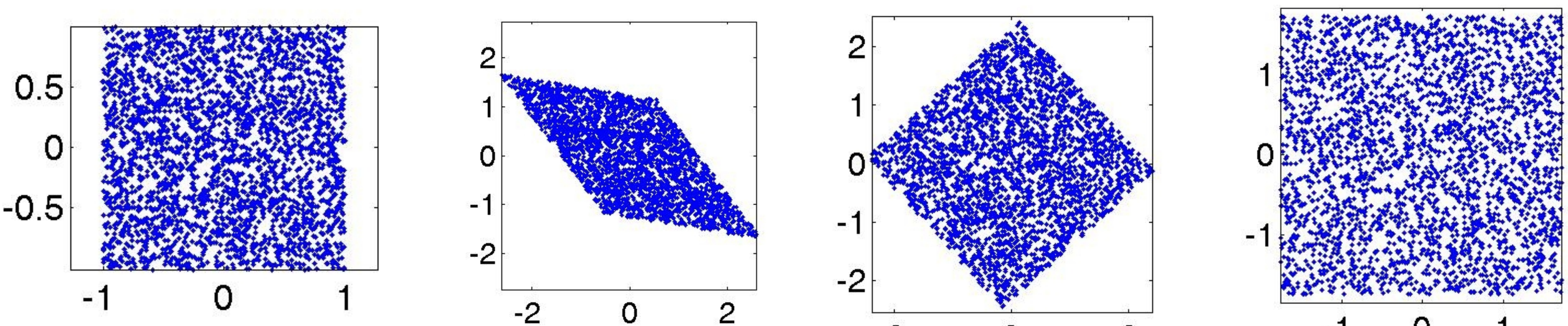
Solving ICA

ICA task: Given **x**,

- find **y** (the estimation of **s**),
- find **W** (the estimation of $\mathbf{A}^{-1}$)

ICA solution: **y=Wx**

- Remove mean, E[**x**]=0
- Whitening, E[$\mathbf{x}\mathbf{x}^T$]=**I**
- Find an orthogonal **W** optimizing an objective function
  - Sequence of 2-d Jacobi (Givens) rotations



original
mixed
whitened
rotated (demixed)

Optimization Using Jacobi Rotation Matrices

$$\mathbf{G}(p,q,\theta) \doteq \begin{array}{cc} \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\theta) & \cdots & -\sin(\theta) & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \sin(\theta) & \cdots & \cos(\theta) & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} & \begin{array}{l} \\ \\ \leftarrow \mathrm{p} \\ \\ \leftarrow \mathrm{q} \\ \\ \\ \end{array} \\ \begin{array}{ccccccc} & & \uparrow & & \uparrow & & \\ & & \mathrm{p} & & \mathrm{q} & & \end{array} & \end{array} \in \mathbf{R}^{M \times M}$$

$$\mathit{Observation}: \mathbf{x} = \mathbf{A}\mathbf{s}$$

$$\mathit{Estimation}: \mathbf{y} = \mathbf{W}\mathbf{x}$$

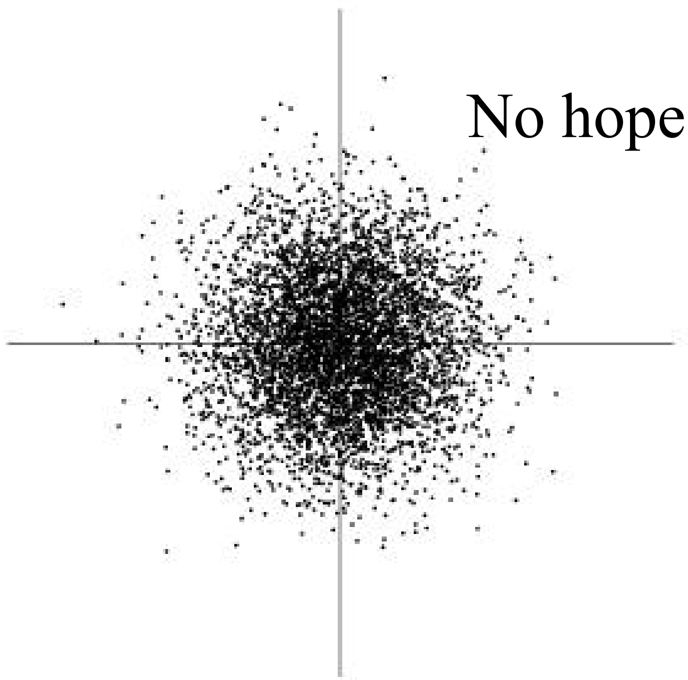
$$\mathbf{W} = \arg\min_{\tilde{\mathbf{W}} \in \mathcal{W}} J(\tilde{\mathbf{W}}\mathbf{x}),$$

$$\text{where } \mathcal{W} = \{\mathbf{W} \mid \mathbf{W} = \prod_i G(p_i, q_i, \theta_i)\}$$

Gaussian sources are problematic

The Gaussian distribution is ***spherically symmetric***.

Mixing it with an orthogonal matrix… produces the same distribution...



No hope for recovery... ☹

However, this is the only ‘nice’ distribution that we cannot recover! ☺

ICA Cost Functions

Let $\mathbf{y} \doteq \mathbf{W}\mathbf{x}$, $\mathbf{y} = [y_1; \ldots; y_M]$, and let us measure the dependence using Shannon's mututal information:

$$J_{ICA_1}(\mathbf{W}) \doteq I(y_1, \ldots, y_M) \doteq \int p(y_1, \ldots, y_M) \log \frac{p(y_1, \ldots, y_M)}{p(y_1) \ldots p(y_M)} d\mathbf{y},$$

Let $H(\mathbf{y}) \doteq H(y_1, \ldots, y_m) \doteq -\int p(y_1, \ldots, y_m) \log p(y_1, \ldots, y_m) d\mathbf{y}$.

$H(\mathbf{W}\mathbf{x}) = H(\mathbf{x}) + \log|\det \mathbf{W}|$, thus

$$\begin{aligned} I(y_1, \ldots, y_M) &= \int p(y_1, \ldots, y_M) \log \frac{p(y_1, \ldots, y_M)}{p(y_1) \ldots p(y_M)} \\ &= -H(y_1, \ldots, y_M) + H(y_1) + \ldots + H(y_M) \\ &= -H(x_1, \ldots, x_M) - \log|\det \mathbf{W}| + H(y_1) + \ldots + H(y_M). \end{aligned}$$

$H(x_1, \ldots, x_M)$ is constant, $\log|\det \mathbf{W}| = 0$, thus

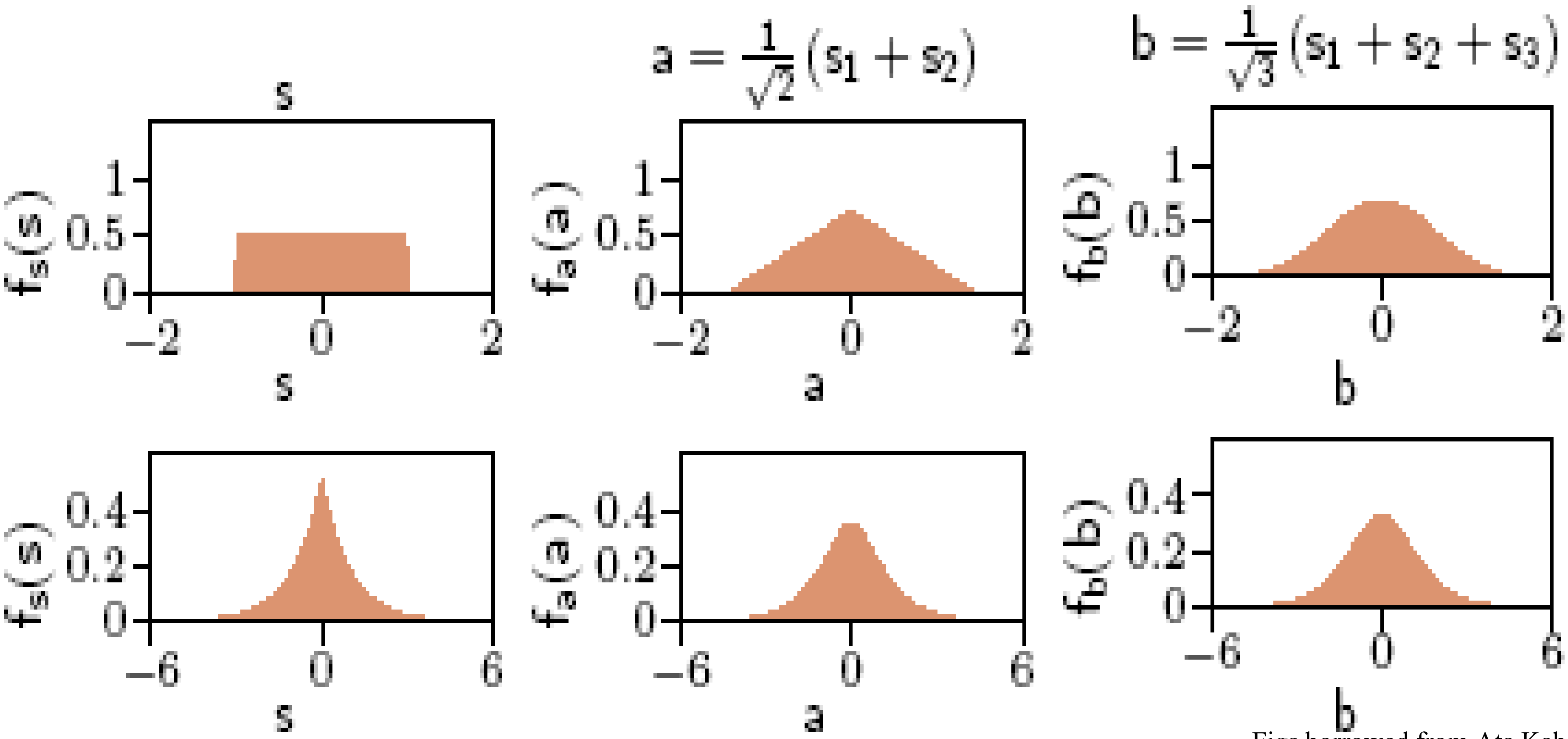
$$J_{ICA_2}(\mathbf{W}) \doteq H(y_1) + \ldots + H(y_M)$$ $\Rightarrow$ go away from normal distribution

Central Limit Theorem

The sum of independent variables converges to the normal distribution

⇒ For separation go far away from the normal distribution

⇒ **Negentropy, |kurtozis| maximization**



Figs borrowed from Ata Kaban

Algorithms

There are more than 100 different ICA algorithms…

- Mutual information (MI) estimation
  - Kernel-ICA [Bach & Jordan, 2002]
- Entropy, negentropy estimation
  - Infomax ICA [Bell & Sejnowski 1995]
  - RADICAL [Learned-Miller & Fisher, 2003]
  - FastICA [Hyvarinen, 1999]
  - [Girolami & Fyfe 1997]
- ML estimation
  - KDICA [Chen, 2006]
  - EM-ICA [Welling]
  - [MacKay 1996; Pearlmutter & Parra 1996; Cardoso 1997]
- Higher order moments, cumulants based methods
  - JADE [Cardoso, 1993]
- Nonlinear correlation based methods
  - [Jutten and Herault, 1991]

ICA ALGORITHMS

David J.C. MacKay (97)

Maximum Likelihood ICA Algorithm

- simplest approach
- requires knowing densities of hidden sources $\{f_i\}$

rows of **W**

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t),\ \mathbf{s}(t) = \mathbf{W}\mathbf{x}(t),\ \text{where } \mathbf{A}^{-1} = \mathbf{W} = [\mathbf{w}_1; \ldots ; \mathbf{w}_M] \in \mathbb{R}^{M \times M}$$

$$L = \sum_{t=1}^{T} \log p(\mathbf{x}(t)) = \sum_{t=1}^{T} \log p(\mathbf{A}\mathbf{s}(t)) = \sum_{t=1}^{T} \log |\mathbf{W}| + \log p(\mathbf{s}(t))$$

$$= T \log |\mathbf{W}| + \sum_{t=1}^{T} \sum_{i=1}^{M} \log f_i(\underbrace{\mathbf{w}_i \mathbf{x}(t)}_{s_i(t)})$$

$$\Rightarrow \max_{\mathbf{W}} L \Rightarrow \frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial w_{ij}} T \log |\mathbf{W}| + \frac{\partial L}{\partial w_{ij}} \sum_{t=1}^{T} \sum_{k=1}^{M} \log f_k(\underbrace{\mathbf{w}_k \mathbf{x}(t)}_{s_k(t)})$$

$$\Rightarrow \frac{\partial L}{\partial w_{ij}} \approx T(\mathbf{W}^T)^{-1}_{ij} + \sum_{t=1}^{T} \frac{f_i{}'(s_i(t))}{f_i(s_i(t))} x_j(t)$$

$$\Rightarrow \boxed{\Delta \mathbf{W} \propto [\mathbf{W}^T]^{-1} + \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{W}\mathbf{x}(t))\mathbf{x}^T(t),}\ \text{where } g_i = f_i'/f_i$$

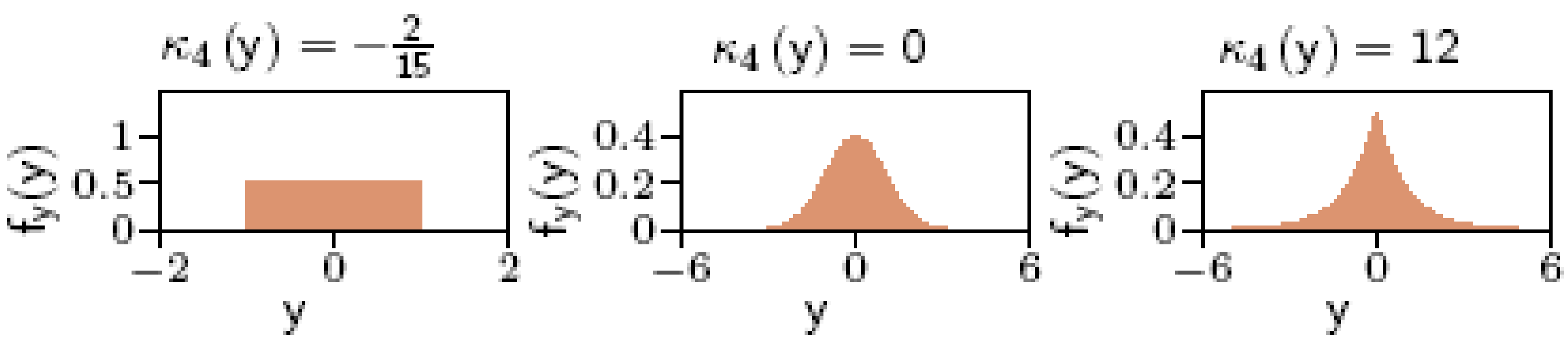
ICA algorithm based on Kurtosis maximization

Kurtosis = 4^{th} order cumulant

Measures

- the distance from normality
- the degree of peakedness

- $\kappa_4(y) = \mathrm{E}\{y^4\} - \underbrace{3\left(\mathrm{E}\{y^2\}\right)^2}_{= 3 \text{ if } \mathrm{E}\{y\} = 0 \text{ and whitened}}$



The Fast ICA algorithm (Hyvarinen)

Probably the most famous ICA algorithm

- Given whitened data $\mathbf{z}$
- Estimate the 1^{st} ICA component:

  $\star$ $y = \mathbf{w}^T\mathbf{z}$, $\|\mathbf{w}\| = 1$, $\Leftarrow \mathbf{w} = 1^{st}$ row of $\mathbf{W}$

  $\star$ maximize kurtosis $f(\mathbf{w}) \doteq \kappa_4(y) \doteq \mathbb{E}[y^4]$
  with constraint $h(\mathbf{w}) = \|\mathbf{w}\|^2 - 1 = 0$

  $\star$ At optimum $f'(\mathbf{w}) + \lambda h'(\mathbf{w}) = \mathbf{0}^T$
  $\Rightarrow 4\mathbb{E}[(\mathbf{w}^T\mathbf{z})^3\mathbf{z}] + 2\lambda\mathbf{w} = 0$

  $\star$ After calculating λ we arrive at the following iteration:

  Let $\mathbf{w}_1$ be the fix pont of:

  $$\tilde{\mathbf{w}}(k+1) = \mathbb{E}[(\mathbf{w}(k)^T\mathbf{z})^3\mathbf{z}] - 3\mathbf{w}(k)$$
  $$\mathbf{w}(k+1) = \frac{\tilde{\mathbf{w}}(k+1)}{\|\tilde{\mathbf{w}}(k+1)\|}$$

- Estimate the 2^{nd} ICA component similarly
  using the $\mathbf{w} \perp \mathbf{w}_1$ additional constraint... and so on ...

Dependence Estimation Using Kernel Methods

The Kernel ICA Algorithm

Kernel covariance (KC)

A. Gretton, R. Herbrich, A. Smola, F. Bach, M. Jordan

Let $\mathbf{x} \in \mathbb{R}^{d_x}$, $\mathbf{y} \in \mathbb{R}^{d_y}$ stochastic variables.

We want to measure their dependence.

$$J_{KC} \doteq \sup_{\substack{f \in \mathcal{F}^x, g \in \mathcal{F}^y \\ \|f\| \le 1, \|g\| \le 1}} |E\{[f(\mathbf{x}) - Ef(\mathbf{x})][g(\mathbf{y}) - Eg(\mathbf{y})]\}|$$

$$J_{KC}^{emp} \doteq \sup_{\substack{f \in \mathcal{F}^x, g \in \mathcal{F}^y \\ \|f\| \le 1, \|g\| \le 1}} |\frac{1}{m} \sum_{l=1}^{m} \{[f(\mathbf{x}_l) - \frac{1}{m} \sum_{j=1}^{m} f(\mathbf{x}_j)][g(\mathbf{y}_l) - \frac{1}{m} \sum_{j=1}^{m} g(\mathbf{y}_j)]\}|$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m$, and $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are m pieces of i.i.d. samples and $\mathcal{F}^x$, $\mathcal{F}^y$ are sets of real valued functions.

The calculation of the supremum over function sets is extremely difficult. Reproducing Kernel Hilbert Spaces make it easier.

RKHS construction for x, y stochastic variables.

Let $K^x(\cdot,\cdot) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$, $K^y(\cdot,\cdot) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ kernel functions.

These kernels define the following RKHS:

$$\mathcal{F}^x \doteq \{f : f = \sum_{j=1}^{\infty} \Psi_j \Phi_j^x(\cdot), \ \sum_{j=1}^{\infty} \frac{\Psi_j^2}{\lambda_j^x} < \infty\},$$

$$\mathcal{F}^y \doteq \{f : f = \sum_{j=1}^{\infty} \Psi_j \Phi_j^y(\cdot), \ \sum_{j=1}^{\infty} \frac{\Psi_j^2}{\lambda_j^y} < \infty\},$$

where $\Phi_j^x(\cdot)$, $\Phi_j^y(\cdot)$, λ_j^x, λ_j^y are eigenfunctions and eigenvalues corresponding to the $K^x(\cdot,\cdot)$, $K^y(\cdot,\cdot)$ Hilbert spaces.

The Representer Theorem

Theorem:

$$\left.\begin{array}{r} k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \text{Mercer kernel on } \mathcal{X} \\ \boldsymbol{z} = (x_1, y_1), \ldots, (x_m, y_m) \in (\mathcal{X} \times \mathcal{Y})^m \text{ training sample} \\ g_{emp} : (\mathcal{X} \times \mathcal{Y} \times \mathbb{R})^m \to \mathbb{R} \cup \{\infty\} \\ g_{reg} : \mathbb{R} \to [0, \infty) \text{ strictly increasing function} \\ \mathcal{F} : \text{ RKHS induced by } k(\cdot,\cdot) \end{array}\right\} \Rightarrow$$

$$\Rightarrow f^* = \arg\min_{f \in \mathcal{F}} R_{reg}[f, \boldsymbol{z}]$$

$$\doteq \arg\min_{f \in \mathcal{F}} \underbrace{g_{emp}[(x_i, y_i, f(x_i))_{i \in \{1 \ldots m\}}]}_{1^{st}\text{ term, empirical loss}} + \underbrace{g_{reg}(\|f\|)}_{2^{nd}\text{ term, regularization}}$$

admits the following representation:

$$f^*(\cdot) = \sum_{i=1}^{m} c_i k(x_i, \cdot), \quad c = (c_1, \ldots, c_m) \in \mathbb{R}^m$$

Kernel covariance (KC)

Yay! We can use the representer theorem for our problem ☺

The optimal f, g can be found in these forms:

$$f^*(\cdot) = \sum_{i=1}^{m} c_i k(x_i, \cdot), \quad \mathbf{c} = (c_1, \dots, c_m) \in \mathbb{R}^m$$
$$g^*(\cdot) = \sum_{i=1}^{m} d_i k(x_i, \cdot), \quad \mathbf{d} = (d_1, \dots, d_m) \in \mathbb{R}^m$$

Kernel covariance (KC)

$f(\mathbf{x}) = \langle f, K^x(\cdot,\mathbf{x})\rangle_{\mathcal{F}^x}$ and $f(\cdot) = \sum_{j=1}^{m} c_j K^x(\cdot,\mathbf{x}_j) + f^{\perp}(\cdot)$, thus

$$f(\mathbf{x}_i) \;=\; \langle f, K^x(\cdot,\mathbf{x}_i)\rangle_{\mathcal{F}^x} = \langle \sum_{j=1}^{m} c_j K^x(\cdot,\mathbf{x}_j) + f^{\perp}(\cdot), K^x(\cdot,\mathbf{x}_i)\rangle_{\mathcal{F}^x} = \sum_{j=1}^{m} c_j K^x(\mathbf{x}_j,\mathbf{x}_i).$$

$\Longrightarrow$

$$[f(\mathbf{x}_1) - \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i), \ldots, f(\mathbf{x}_m) - \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i)] = \mathbf{c}^T\widetilde{\mathbf{K}}^x$$

$$[g(\mathbf{y}_1) - \frac{1}{m}\sum_{i=1}^{m} g(\mathbf{y}_i), \ldots, g(\mathbf{y}_m) - \frac{1}{m}\sum_{i=1}^{m} g(\mathbf{y}_i)] = \mathbf{d}^T\widetilde{\mathbf{K}}^y$$

Where $\mathbf{K}^x = \{K(\mathbf{x}_i,\mathbf{x}_j)\}_{i,j}$, $\mathbf{H} \doteq \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$, and $\widetilde{\mathbf{K}}^x \doteq \mathbf{H}\mathbf{K}^x\mathbf{H}$

Thus, for the estimation of J_{KC}^{emp} we have to calculate the maximum of $\mathbf{c}^T\widetilde{\mathbf{K}}^x\widetilde{\mathbf{K}}^y\mathbf{d}$ over $\mathbf{c},\mathbf{d} \in \mathbb{R}^n$ subject to $\mathbf{c}^T\widetilde{\mathbf{K}}^x\mathbf{c} = 1$, $\mathbf{d}^T\widetilde{\mathbf{K}}^y\mathbf{d} = 1$.

Amari Error for Measuring the Performance

- Measures how close a square matrix is to a permutation matrix

$$\mathbf{B} = \mathbf{W}\mathbf{A}$$

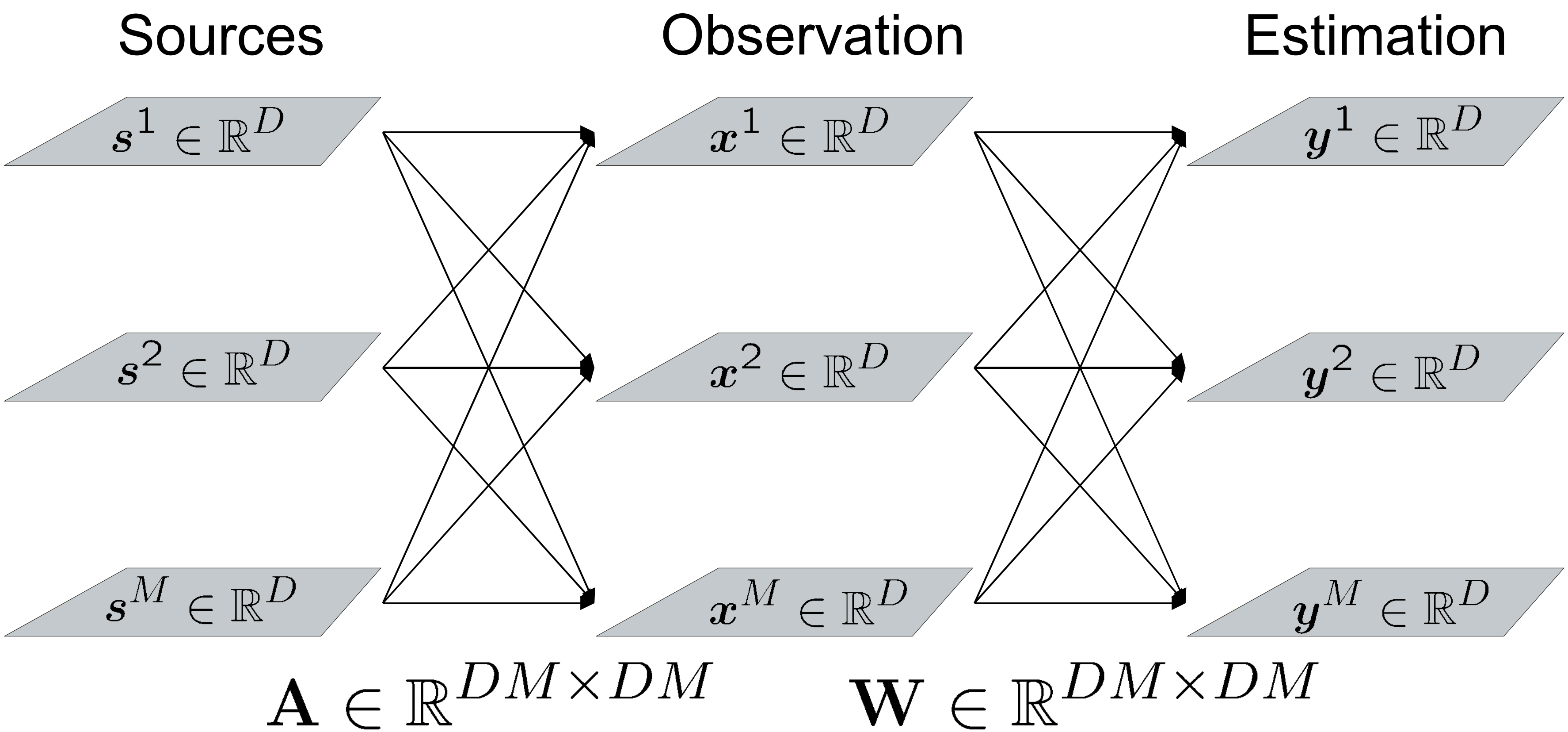
demixing mixing

$$r(\mathbf{B}) = \frac{1}{2M(M-1)} \sum_{i=1}^{M} \left(\frac{\sum_{j=1}^{M} |b_{ij}|}{\max_j |b_{ij}|} - 1 \right) + \frac{1}{2M(M-1)} \sum_{j=1}^{M} \left(\frac{\sum_{i=1}^{M} |b_{ij}|}{\max_i |b_{ij}|} - 1 \right)$$

$$r(\mathbf{B}) \in [0,1], \quad r(\mathbf{B}) = 0 \Leftrightarrow \mathbf{B} \text{ is a permutation matrix}$$

Independent Subspace Analysis

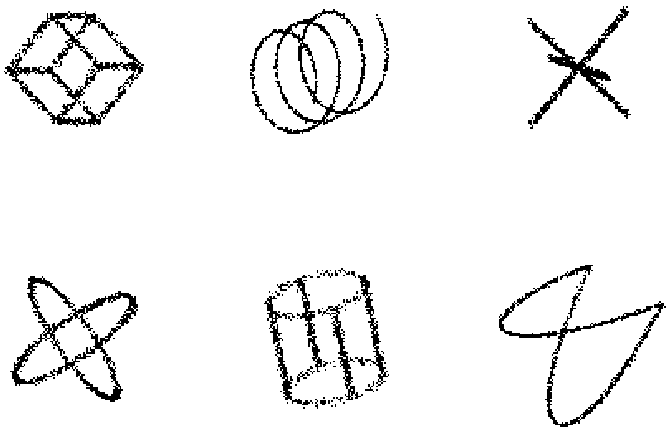
Independent Subspace Analysis (ISA, The Woodstock Problem)



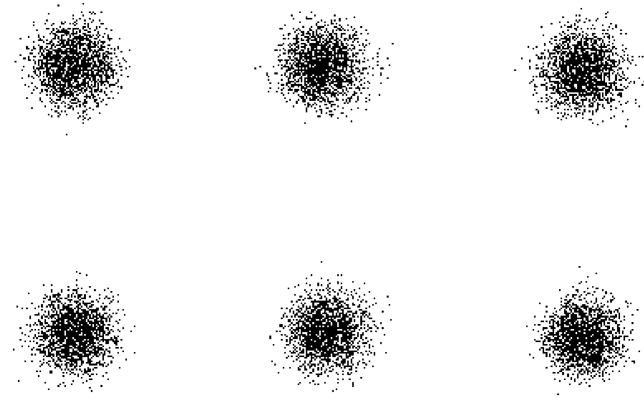
Find $\mathbf{W}$, recover $\mathbf{W}\boldsymbol{x}$

Independent Subspace Analysis

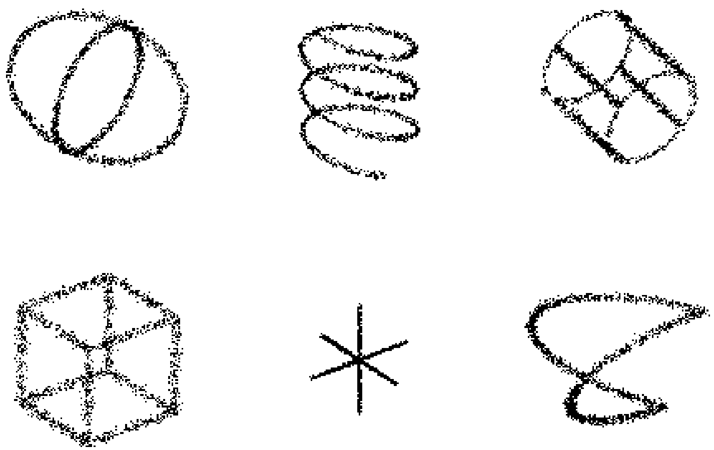
Original



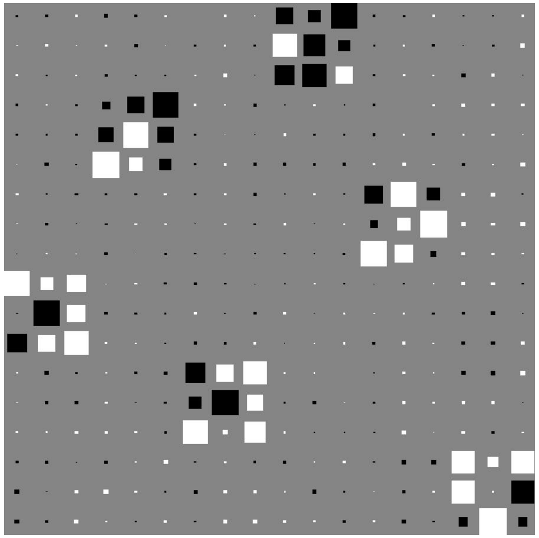
Mixed



Separated



Hinton diagram



ISA Cost Functions

Mutual Information: $I(\mathbf{y}^1, \ldots, \mathbf{y}^m) = \int \log \frac{p(\mathbf{y})}{p(\mathbf{y}^1) \cdots p(\mathbf{y}^m)} d\mathbf{y}$

Shannon-entropy: $H(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y}$

Assume $\mathbf{y} = \mathbf{W}\mathbf{x}$. Then

$$
\begin{aligned}
H(\mathbf{y}) &= H(\mathbf{y}^1, \ldots, \mathbf{y}^m) = H(\mathbf{W}\mathbf{x}) = H(\mathbf{x}) + \log |\mathbf{W}| \\
I(\mathbf{y}^1, \ldots, \mathbf{y}^m) &= -H(\mathbf{x}) - \log |\mathbf{W}| + \sum_{i=1}^{m} H(\mathbf{y}^i) \\
I(\mathbf{y}^1, \ldots, \mathbf{y}^m) &= -H(\mathbf{y}^1, \ldots, \mathbf{y}^m) + \sum_{j=1}^{m} \sum_{i=1}^{d} H(y_i^j) - \sum_{j=1}^{m} I(y_1^j, \ldots, y_d^j) \\
H(\mathbf{y}^j) &= H(y_1^j, \ldots, y_d^j) = \sum_{i=1}^{d} H(y_i^j) - I(y_1^j, \ldots, y_1^j)
\end{aligned}
$$

and we get the following ISA cost functions:

ISA Cost Functions

$$J_{ISA_1}(\mathbf{W}) \doteq I(\mathbf{y}^1, \ldots, \mathbf{y}^m)$$

$$J_{ISA_2}(\mathbf{W}) \doteq H(\mathbf{y}^1) + \ldots + H(\mathbf{y}^m)$$

$$J_{ISA_3}(\mathbf{W}) \doteq \sum_{j=1}^{m} \sum_{i=1}^{d} H(y_i^j) - \sum_{j=1}^{m} I(y_1^j, \ldots, y_d^j)$$

$$J_{ISA_4}(\mathbf{W}) \doteq I(y_1^1, \ldots, y_d^m) - \sum_{j=1}^{m} I(y_1^j, \ldots, y_d^j)$$

Multidimensional Entropy Estimation

Multi-dimensional Entropy Estimations, Method of Kozahenko and Leonenko

Let $\{\mathbf{z}(1), \ldots, \mathbf{z}(n)\}$ denote n i.i.d. samples drawn from the distribution of $\mathbf{z} \in \mathbf{R}^d$.

Let $\mathcal{N}_{1,j}$ be the nearest neighbour of $\mathbf{z}(j)$ in the sample set.

Then the nearest neighbor entropy estimation:

$$\widehat{H}(\mathbf{z}) = \frac{1}{n} \sum_{j=1}^{n} \log(n \|\mathcal{N}_{1,j} - \mathbf{z}(j)\|) + \ln(2) + C_E,$$

where $C_E = -\int_0^\infty e^{-t} \ln(t) dt$ is the Euler-constant.

This estimation is means-square consistent, but not robust.
Let us try to use more neighbors!

Multi-dimensional Rényi's Entropy Estimations

Let us apply Rényi's-entropy for estimating the Shannon-entropy:

$$H_\alpha = \frac{1}{1-\alpha} \log \int f^\alpha(\mathbf{z})\, d\mathbf{z}$$

$$\lim_{\alpha \to 1} H_\alpha = -\int f(\mathbf{z}) \log f(\mathbf{z})\, d\mathbf{z}$$

Let us use
- ***K-nearest neighbors***
- ***minimum spanning trees***

for estimating the multi-dimensional Rényi's entropy.
(It could be much more general…)

Beardwood - Halton - Hammersley Theorem for kNN graphs

Let $\{\mathbf{z}(1),\ldots,\mathbf{z}(n)\}$ denote n i.i.d. samples drawn from the distribution of $\mathbf{z} \in \mathbf{R}^d$.

Let $\mathcal{N}_{k,j}$ be the k nearest neighbours of $\mathbf{z}(j)$ in the sample set.

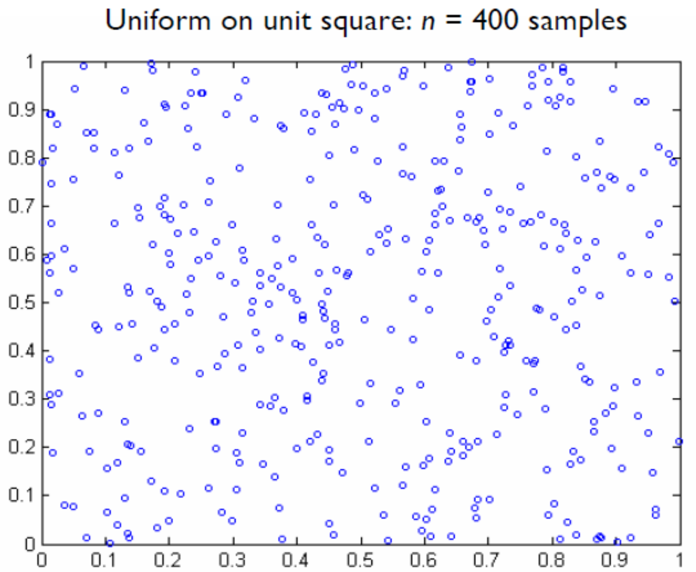
Let $\gamma = d - d\alpha$, then

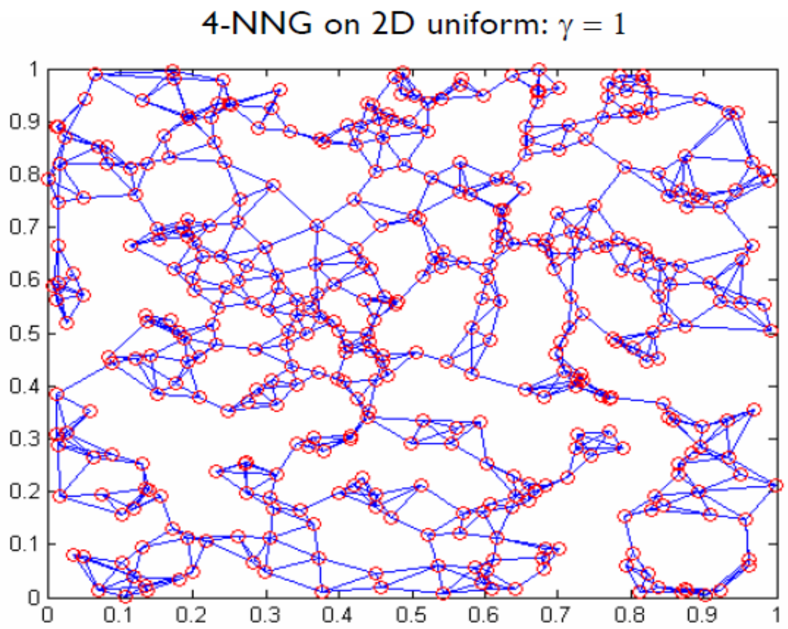
$$\frac{1}{1-\alpha} \log\left(\frac{1}{kn^\alpha} \sum_{j=1}^{n} \sum_{\mathbf{v}\in\mathcal{N}_{k,j}} \|\mathbf{v}-\mathbf{z}(j)\|^\gamma\right) \to H_\alpha(\mathbf{z}) + c,$$

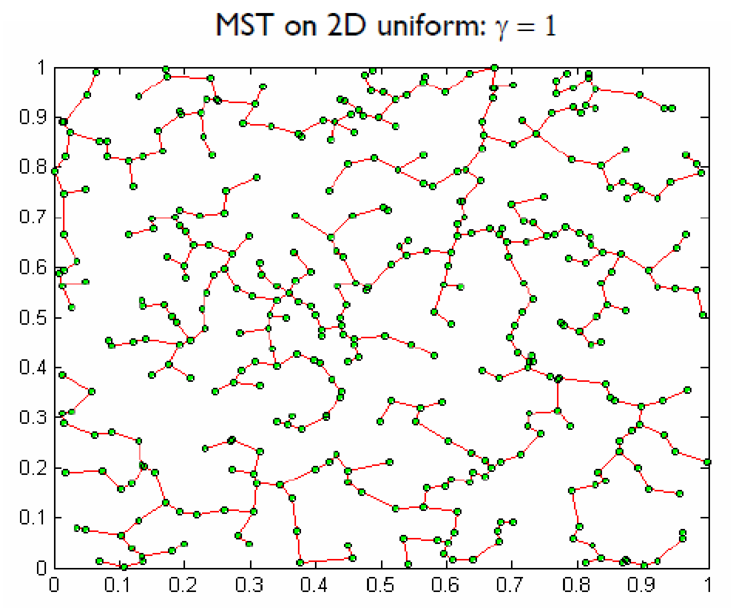
as $n \to \infty$

Lots of other graphs, e.g. MST, TSP, minimal matching, Steiner graph…etc could be used as well.

Examples
(J. A. Costa and A. O. Hero)

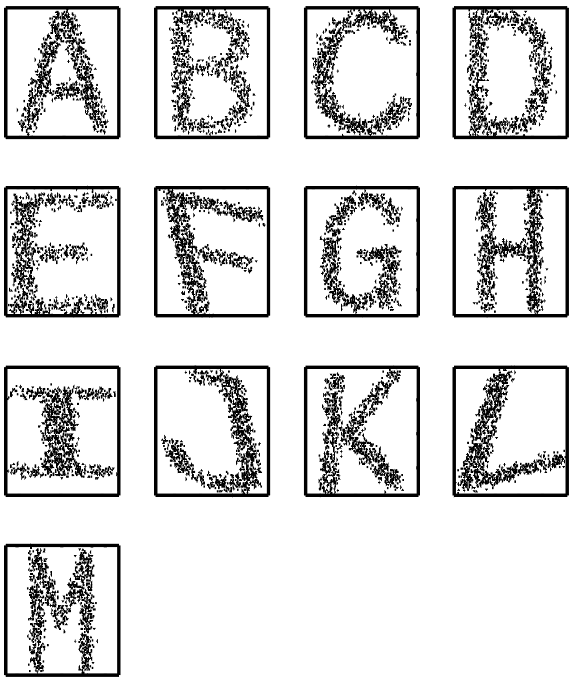




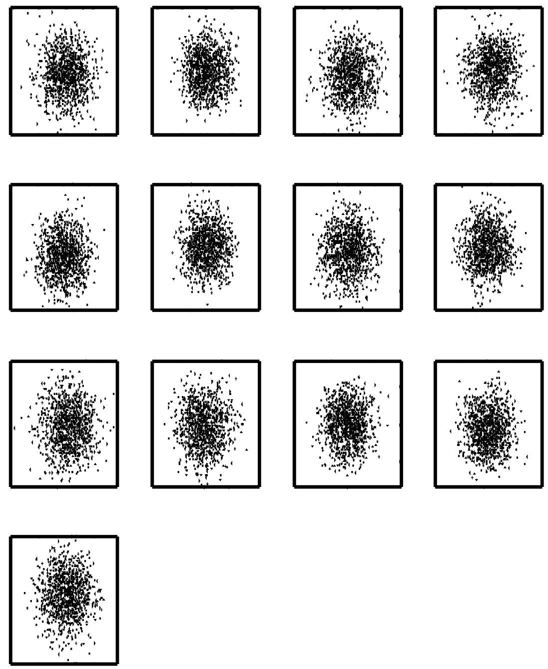


Independent Subspace Analysis Results

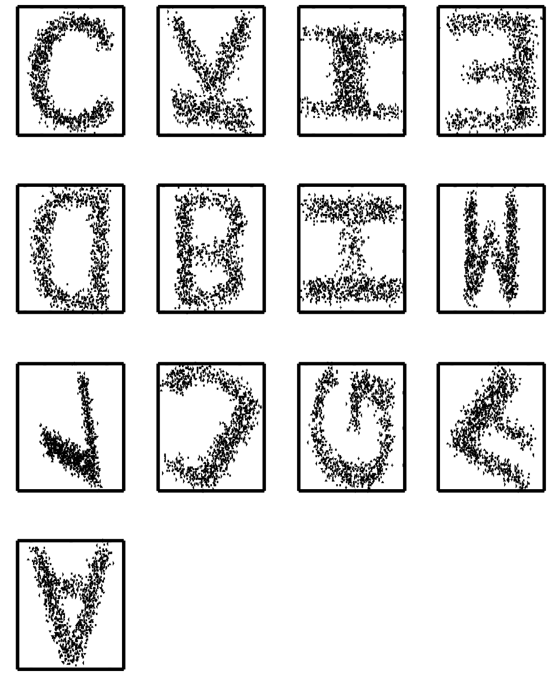
Numerical Simulations
2D Letters (i.i.d.)



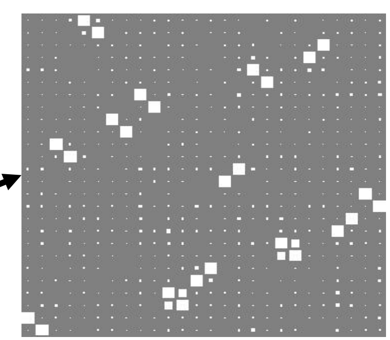
Sources



Observation

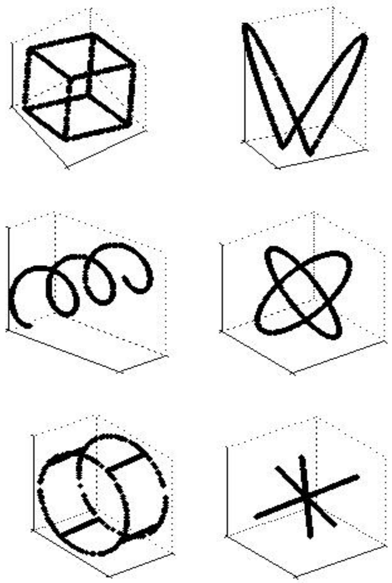


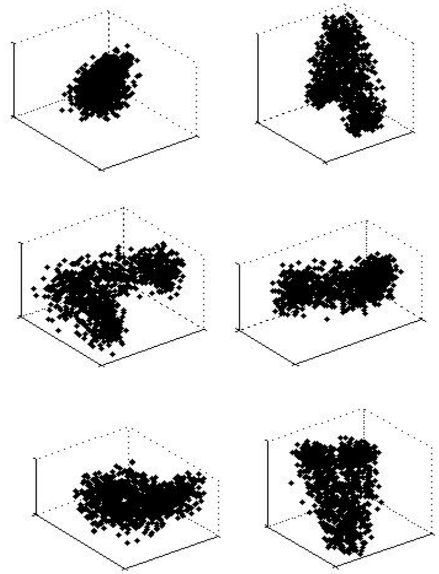
Estimated sources

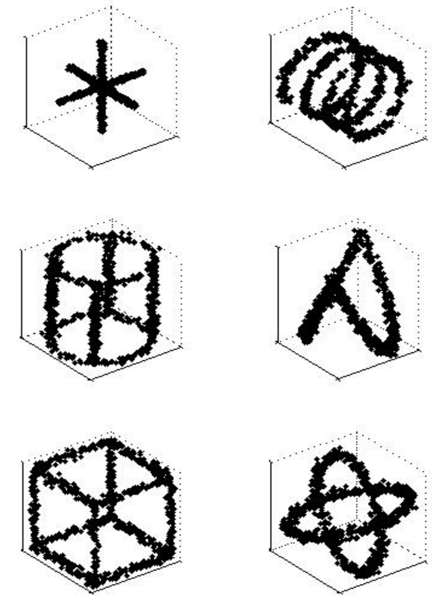


Performance matrix

Numerical Simulations
3D Curves (i.i.d.)





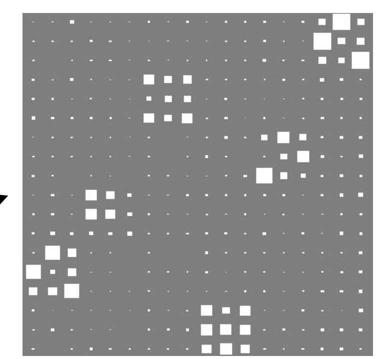


Sources

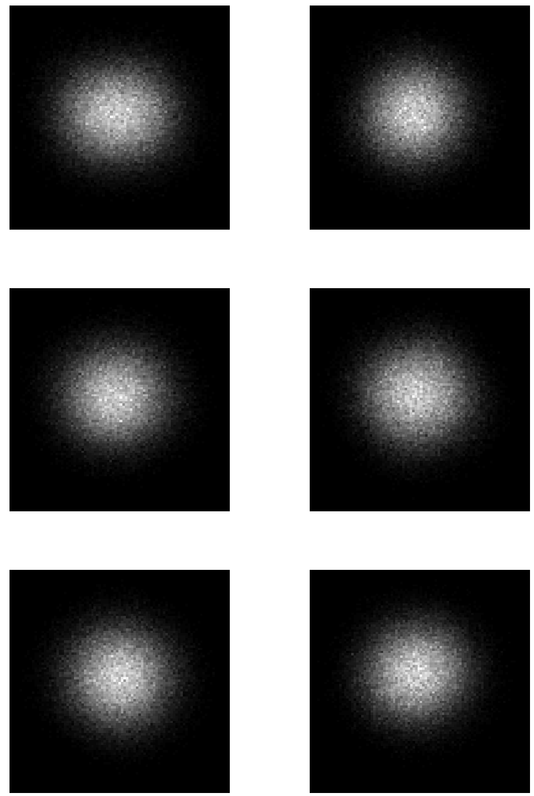
Observation

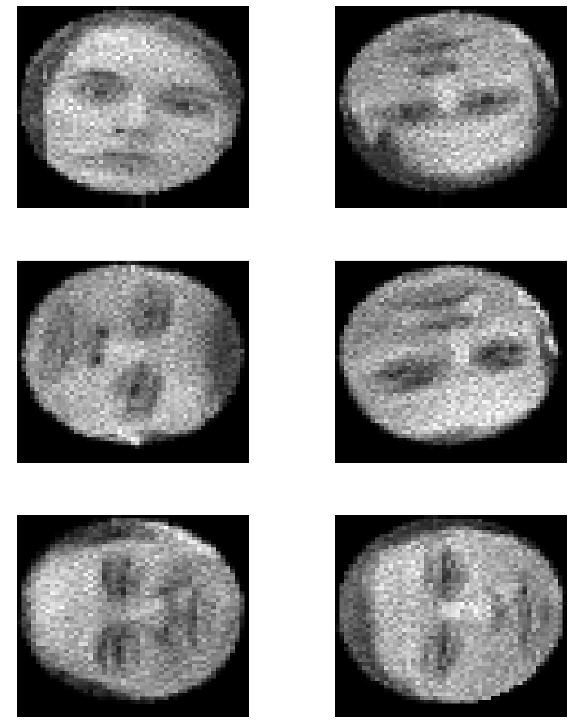
Estimated sources

Performance matrix



Numerical Simulations
Facial images (i.i.d.)



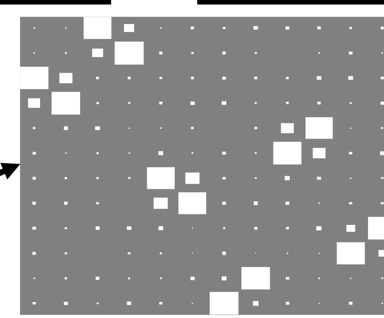


Sources

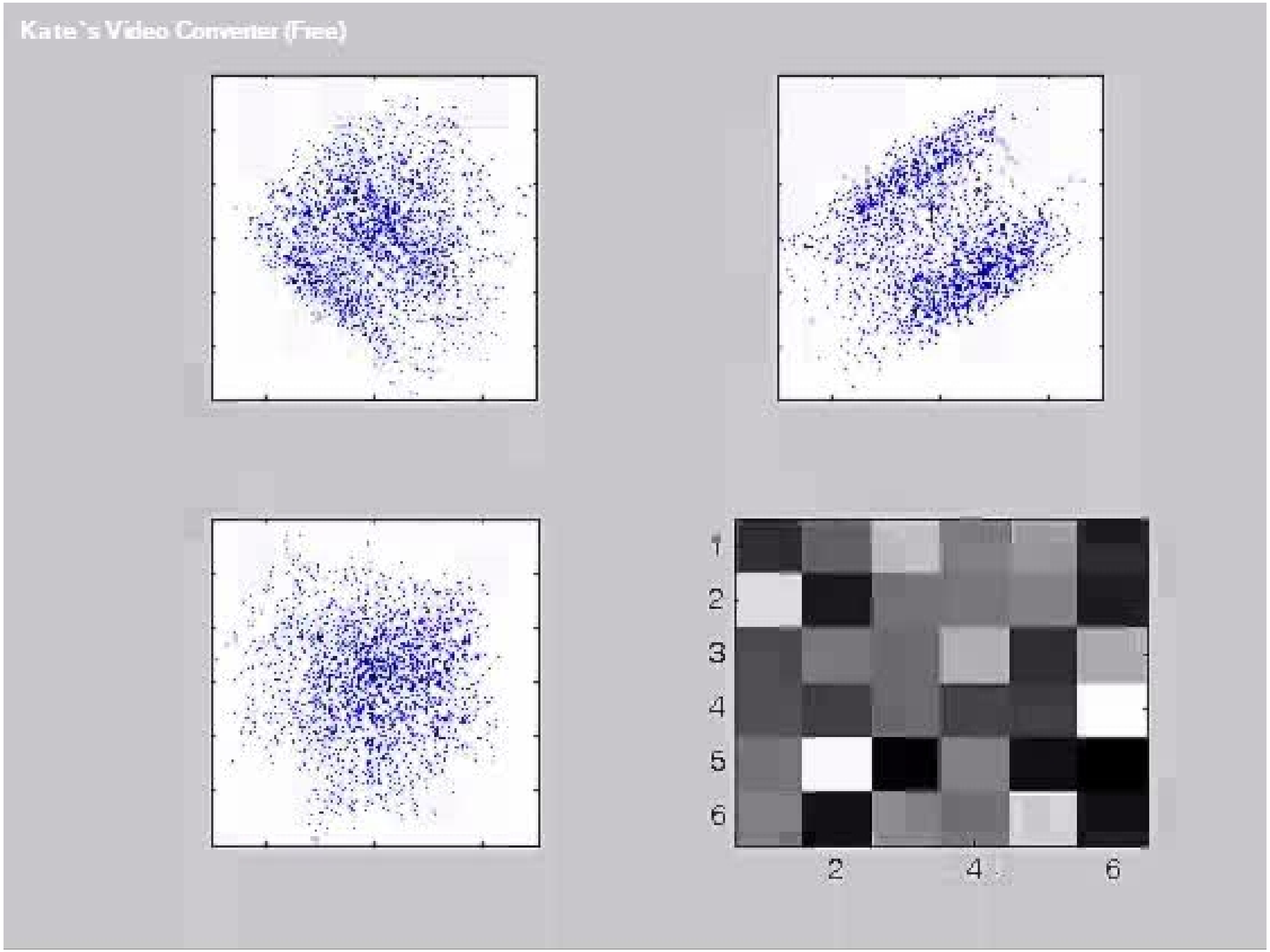
Observation

Estimated sources

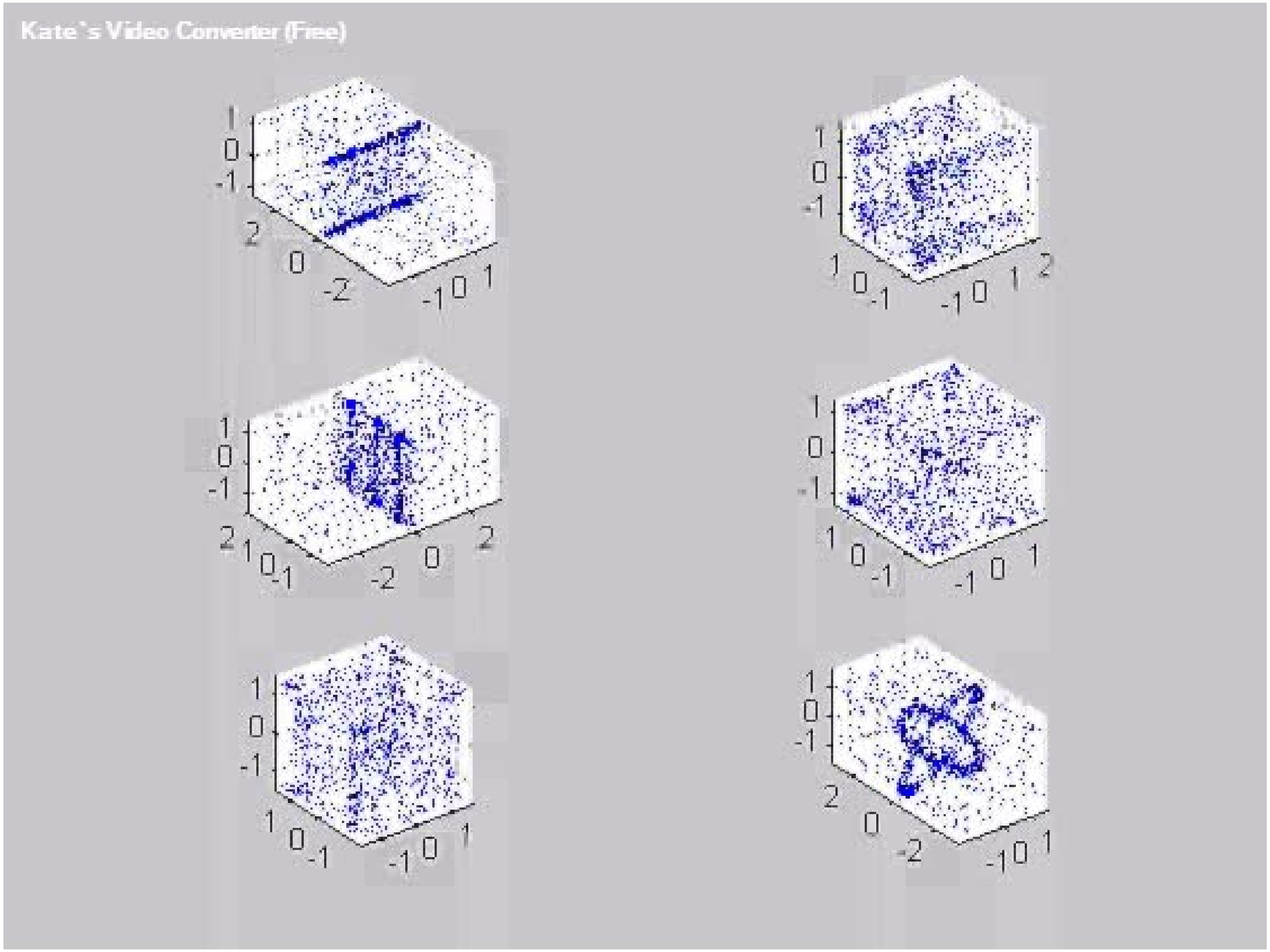
Performance matrix



ISA 2D



ISA 3D after ICA preprocessing



Thanks for the Attention! ☺