Course 495: Advanced Statistical Machine Learning/Pattern Recognition

Deterministic Component Analysis

- Goal (Lecture): To present two more component analysis algorithms, Independent Component Analysis (ICA).

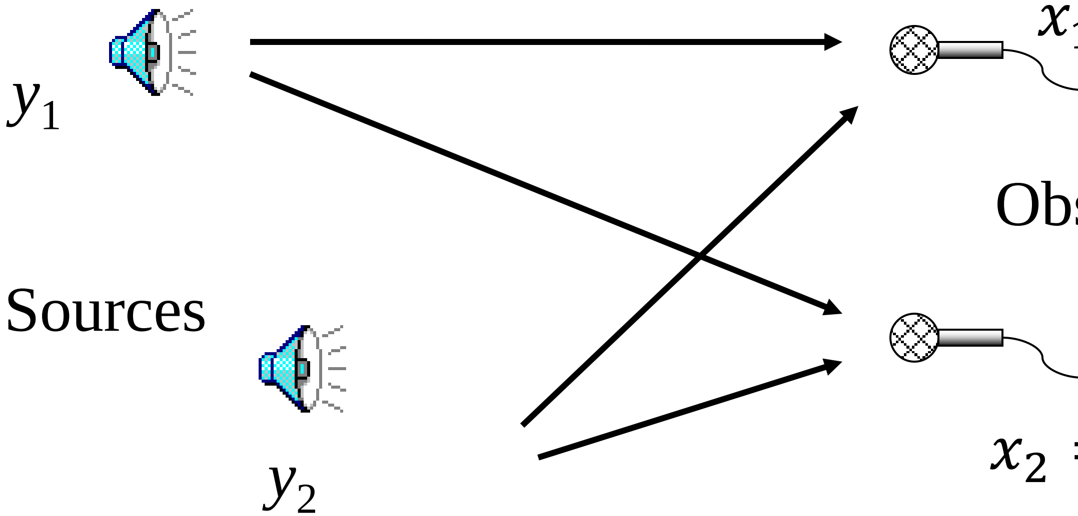
- Goal (Tutorials): To provide the students the necessary mathematical tools for deeply understanding the CA techniques.

Materials

- Pattern Recognition & Machine Learning by C. Bishop Chapter 12
- **ICA:** Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." *Neural networks* 13.4 (2000): 411-430.
- **ICA:** Hyvarinen, Aapo. "Fast and robust fixed-point algorithms for independent component analysis." *Neural Networks, IEEE Transactions on* 10.3 (1999): 626-634.

‘Cocktail party’

Mixing matrix $\mathbf{W}$



$$x_1 = w_{11}y_1 + w_{12}y_2$$

Observations

$$x_2 = w_{21}y_1 + w_{22}y_2$$

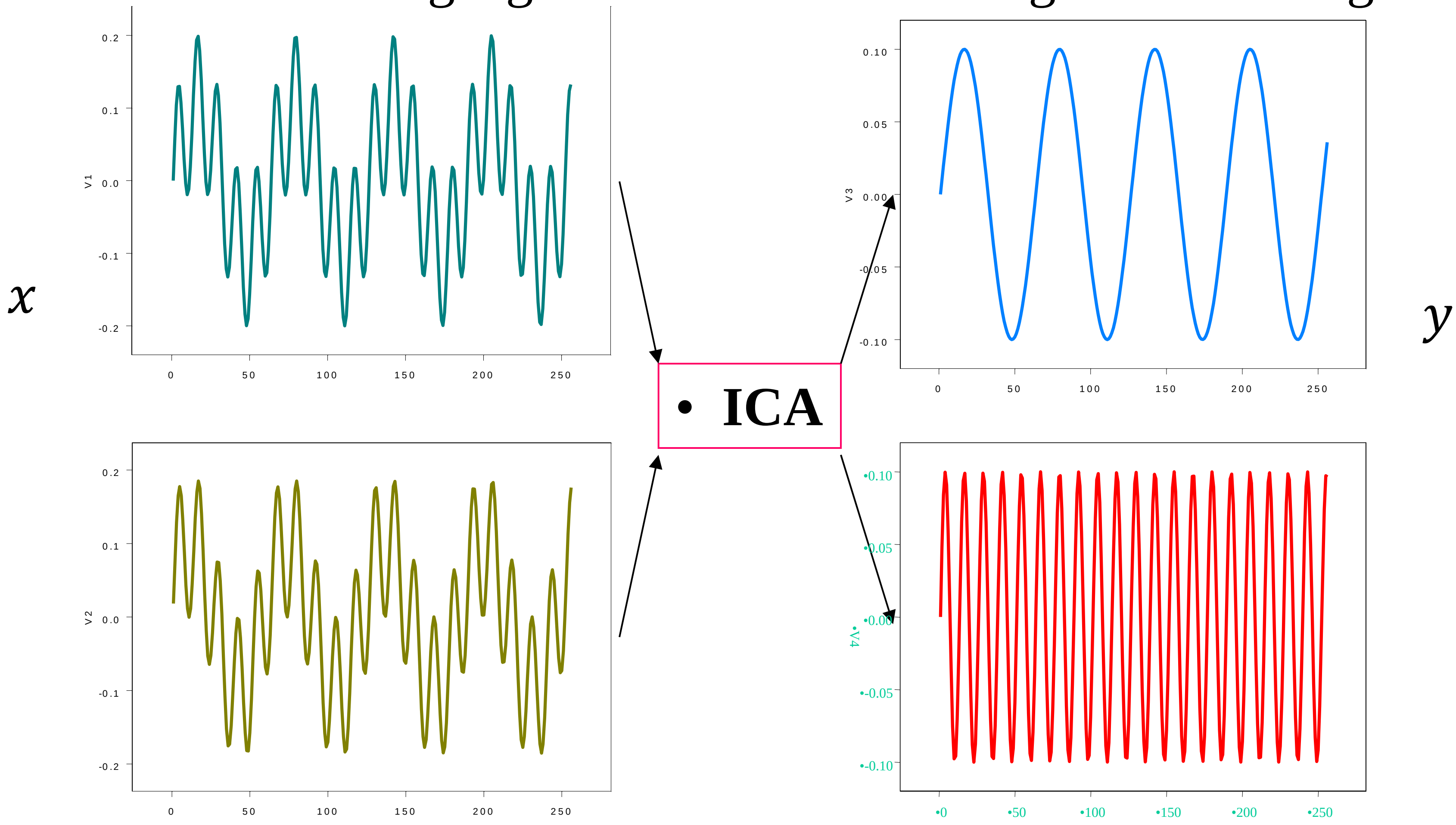
$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{y}$$

d sources (latent space), N observations

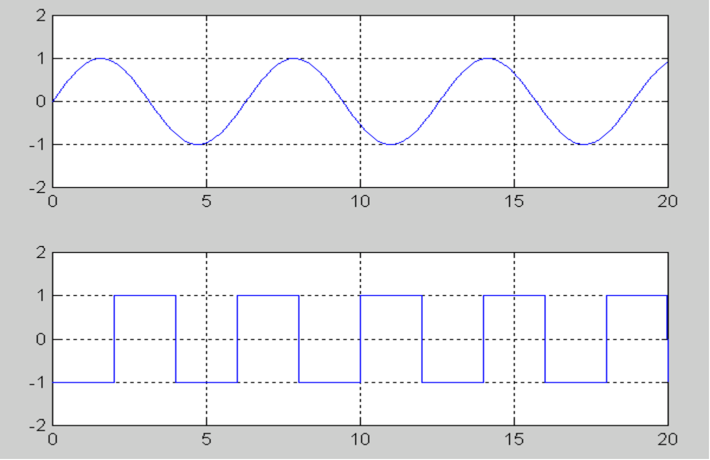
‘Cocktail party’

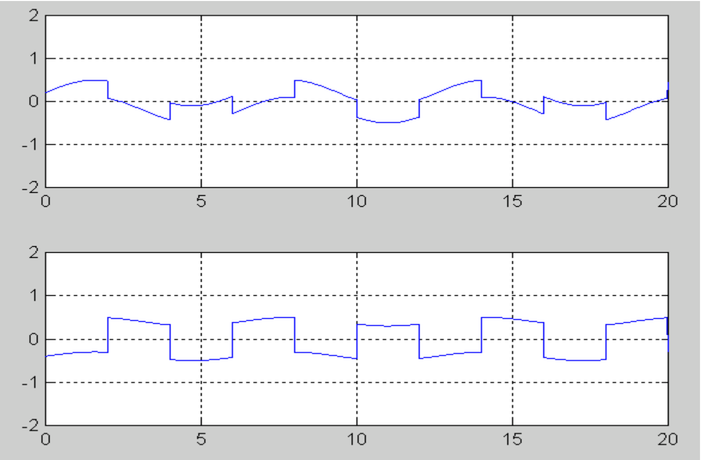
•Observing signals

•Original source signal



‘Cocktail party’





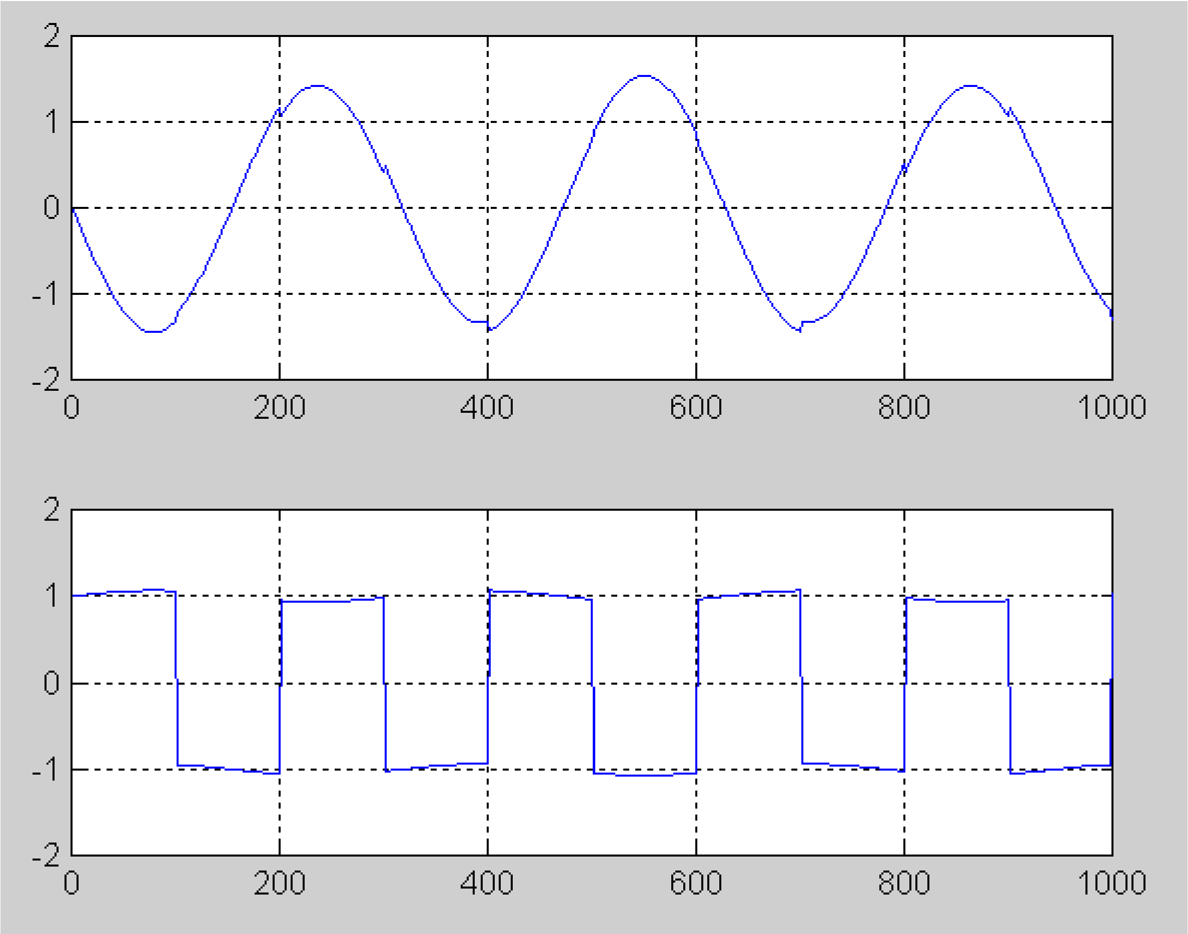
- Two Independent Sources
- Mixture at two Mics

$$x_1(t) = w_{11}\, y_1(t) + w_{12}\, y_2(t)$$

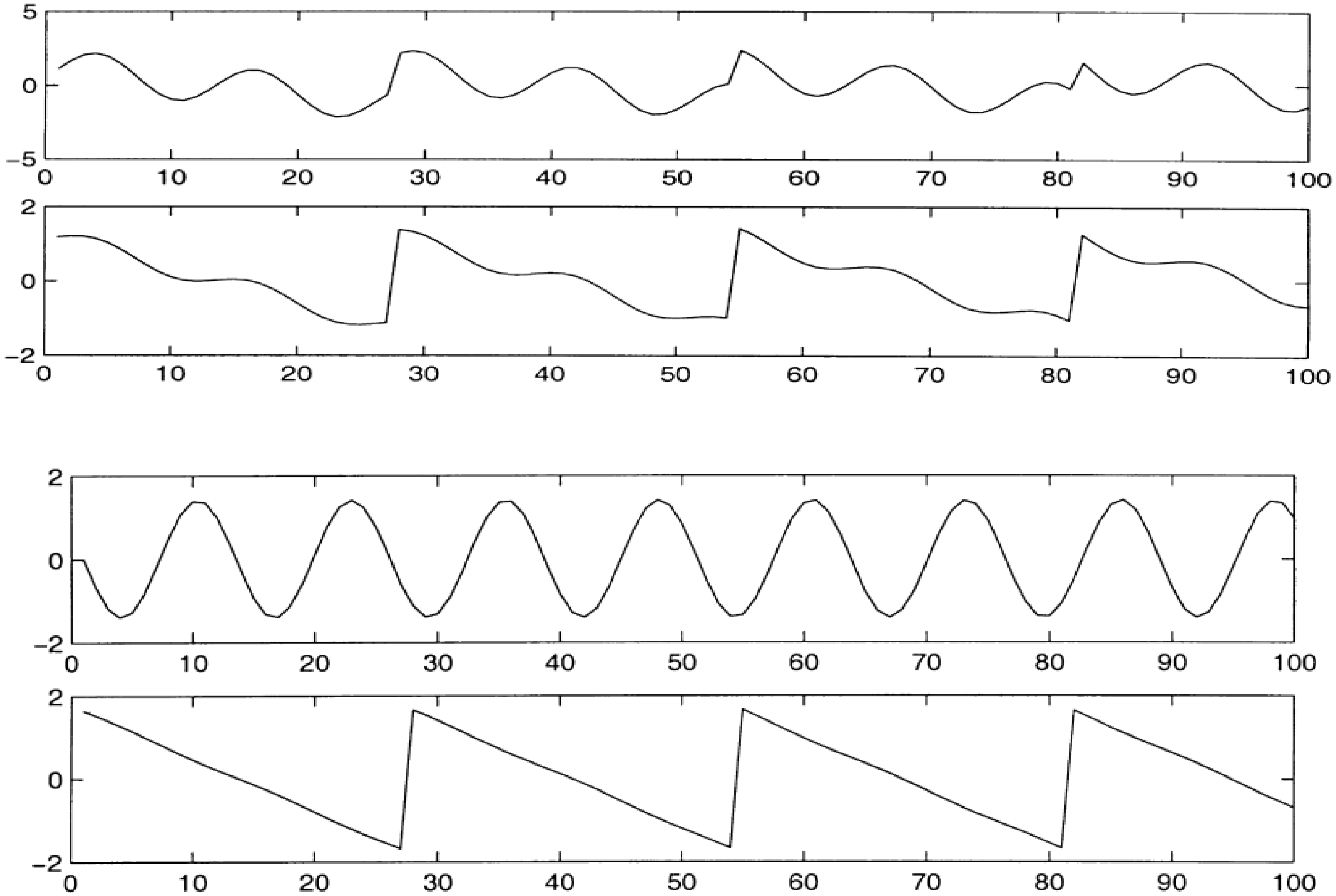
$$x_2(t) = w_{21}\, y_1(t) + w_{22}\, y_2(t)$$

‘Cocktail party’

- Get the Independent Signals out of the Mixture



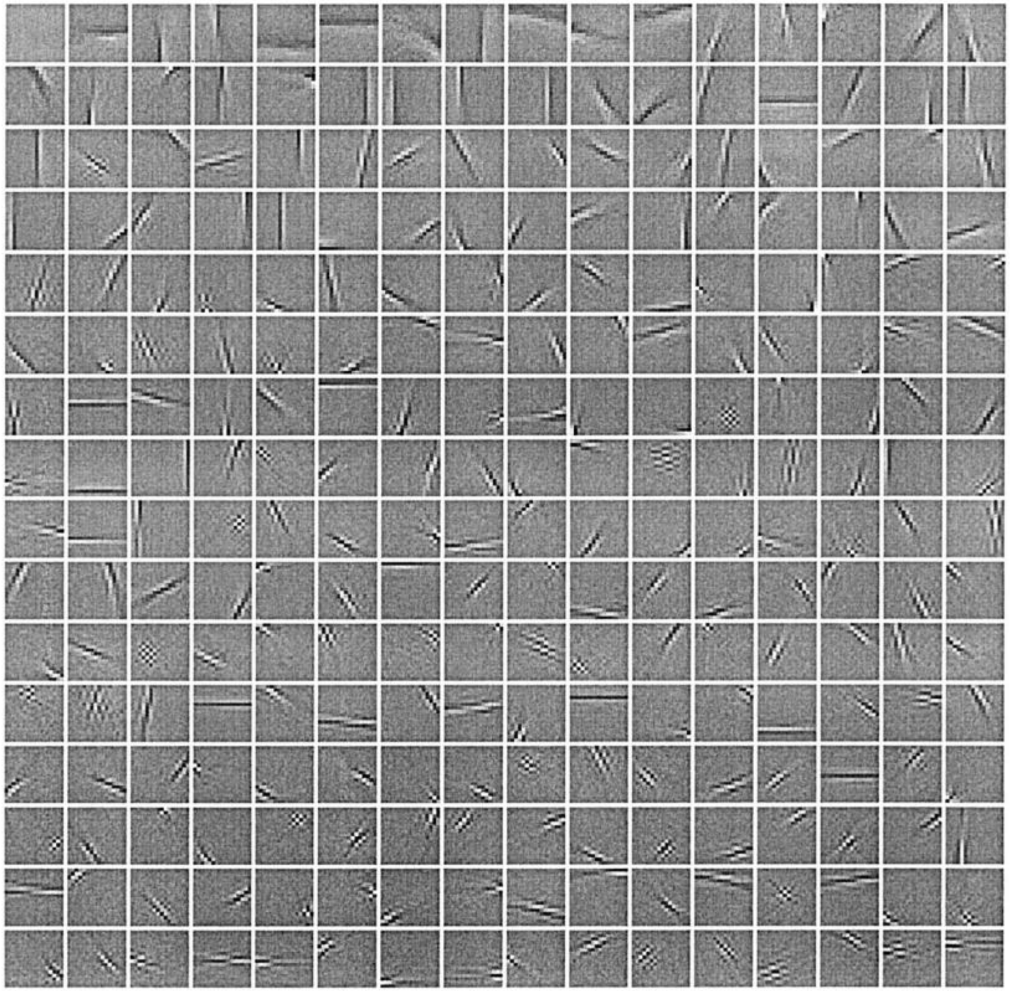
‘Cocktail party’



‘Cocktail party’

Independent Components

of natural images



Formulating the problem

$$x_{i1} = \sum_{k=1}^{d} w_{ik} y_{k1}$$

$$\vdots$$

$$x_{iN} = \sum_{k=1}^{d} w_{ik} y_{kN}$$

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{Y}$$

Find the mixing matrix $\boldsymbol{W}$

and the independent components **Y**

If **W** is a square and invertible then $\boldsymbol{Y} = \boldsymbol{W}^{-1}\boldsymbol{X}$

Formulating the problem

Sources of ambiguity:

1. We cannot determine the variances (energies) of the independent components.

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{Y}$$

$$\boldsymbol{L} = \begin{bmatrix} l_1 & 0 & 0 \\ 0 & l_2 & 0 \\ 0 & 0 & l_3 \end{bmatrix} \qquad \boldsymbol{L}^{-1} = \begin{bmatrix} \frac{1}{l_1} & 0 & 0 \\ 0 & \frac{1}{l_2} & 0 \\ 0 & 0 & \frac{1}{l_3} \end{bmatrix}$$

It can be partially resolved by

$$E[\boldsymbol{y}]=\mathbf{1}$$

Though sign ambiguity cannot be resolved

Formulating the problem

2. We cannot determine the order of the independent components.

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{Y}$$

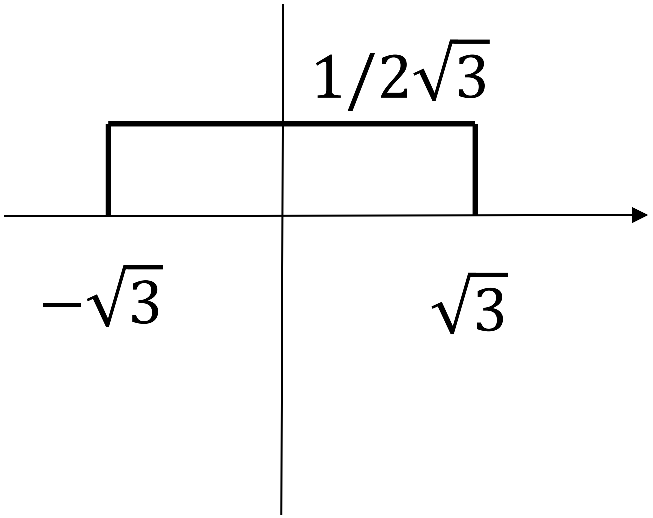
$\boldsymbol{P}$ can be a permutation matrix

$\boldsymbol{P}\boldsymbol{Y}$ are the independent components but in another order

$\boldsymbol{W}\boldsymbol{P}^{-1}$ is the new mixing matrix

A first example

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



$$y_1 \sim p(y) \quad y_2 \sim p(y)$$

$$E[y_1] = 0 \quad E[y_2] = 0$$

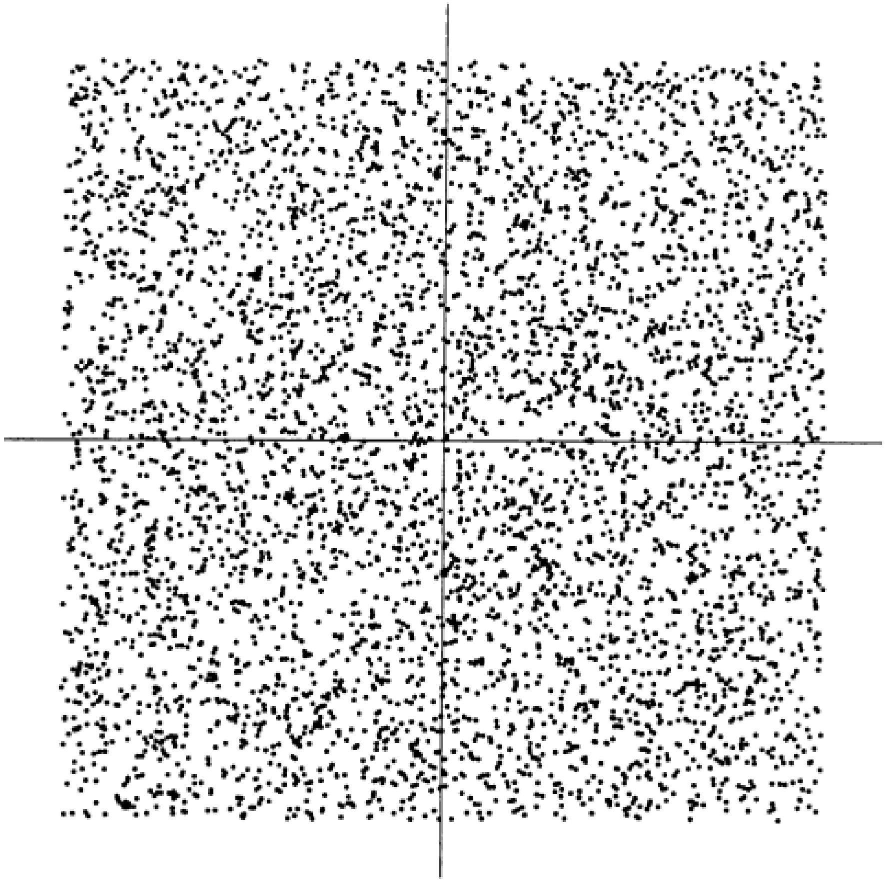
$$\sigma_1{}^2 = E[y_1{}^2] = 1$$

$$\sigma_2{}^2 = E[y_2{}^2] = 1$$

A first example

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

$$y_1 \sim p(y) \quad y_2 \sim p(y)$$

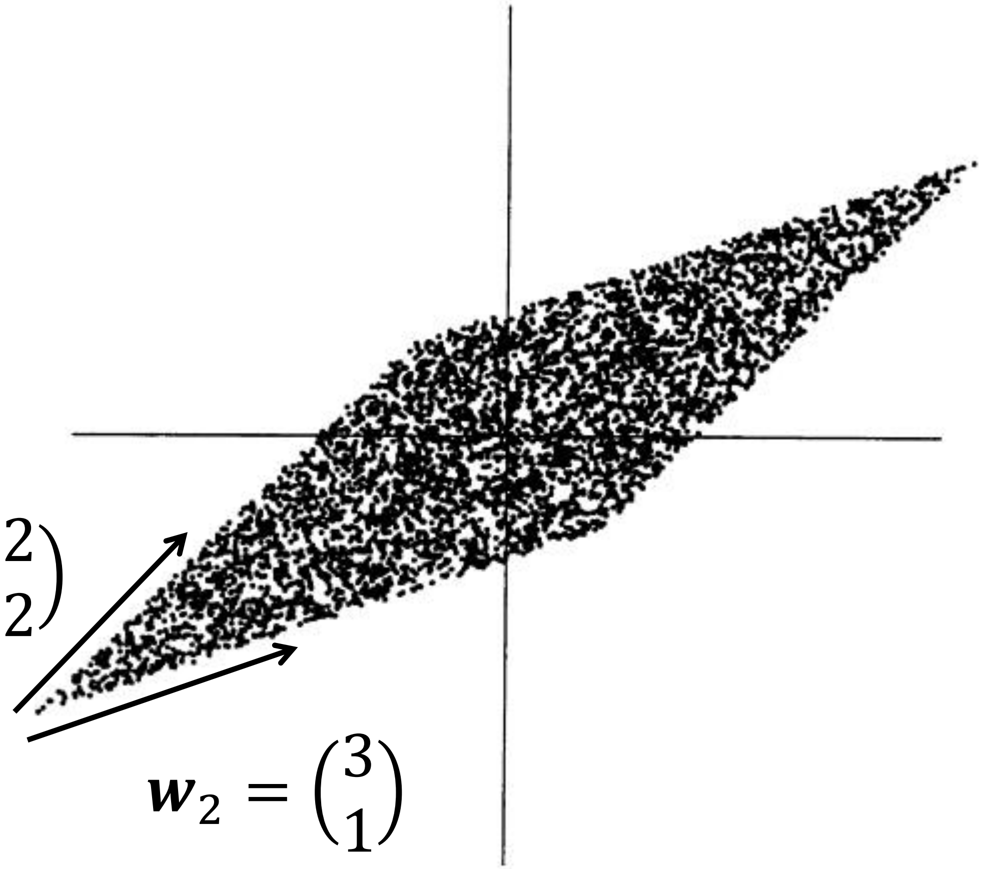


A first example

$$x_1 = 2y_1 + 3y_2$$
$$x_2 = 2y_1 + 1y_2$$

$$\boldsymbol{W} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$

$$\boldsymbol{w}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$



$$\boldsymbol{W}^{-1} = \begin{bmatrix} -0.25 & 0.5 \\ 0.75 & -0.5 \end{bmatrix}$$

$$\boldsymbol{w}_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

Definition of independence

y_1 and y_2 are independent iff $p(y_1, y_2) = p(y_1)p(y_2)$

Alternative definition

$$E(h_1(y_1), h_2(y_2)) = E(h_1(y_1))E(h_2(y_2))$$

$$E(h_1(y_1), h_2(y_2)) = \int_{-\infty}^{+\infty} h_1(y_1)h_2(y_2)p(y_1, y_2)dy_1 dy_2$$

$$= \int_{-\infty}^{+\infty} h_1(y_1)\, p(y_1)dy_1 \int_{-\infty}^{+\infty} h_2(y_2)\, p(y_2)dy_2$$

$$= E(h_1(y_1))E(h_2(y_2))$$

Uncorrelated variables are only partly independent

$h_1(y) = y \quad h_2(y) = y$

$E(y_1, y_2) = E(y_1)E(y_2)$

To show that

$(y_1, y_2) = (0,1)$

$(y_1, y_2) = (0,-1)$

$(y_1, y_2) = (1,0)$

$(y_1, y_2) = (-1,0)$

$E(y_1, y_2) = E(y_1)E(y_2)$

$E({y_1}^2 {y_2}^2) = 0$

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

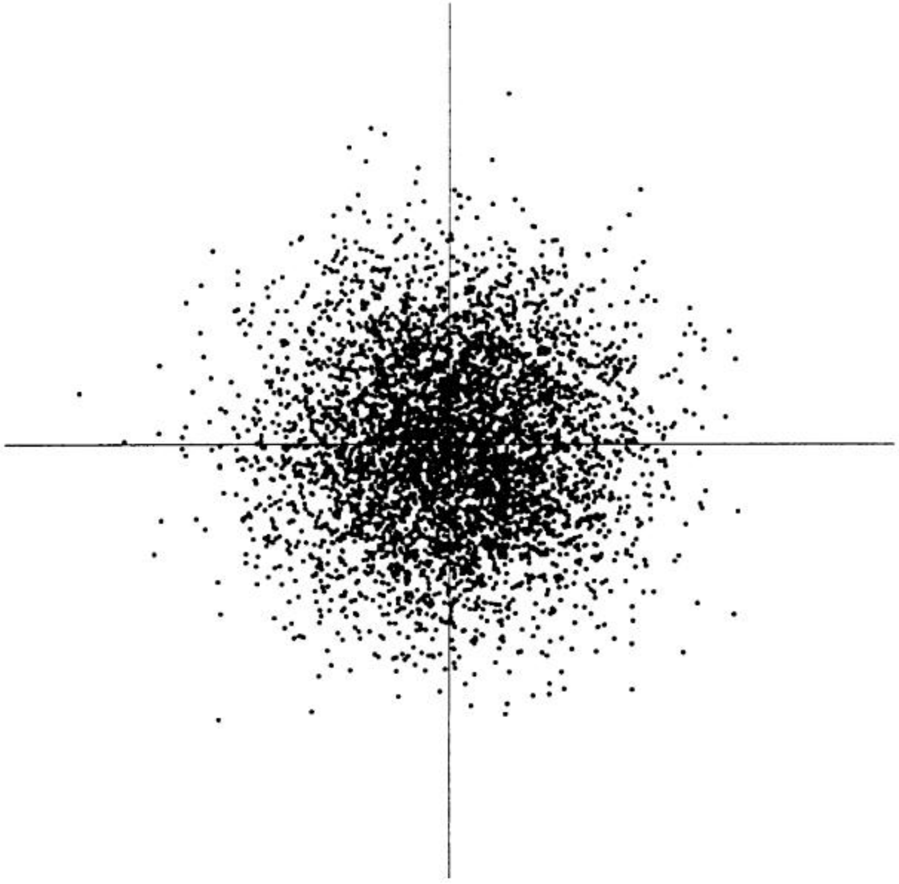
Non Gaussianity

$$p(y_1) = \frac{1}{\sqrt{2\pi}} e^{-{y_1}^2}$$

$$p(y_2) = \frac{1}{\sqrt{2\pi}} e^{-{y_2}^2}$$

W orthogonal it gives

$$p(y_1, y_2) = \frac{1}{\sqrt{2\pi}} e^{-({y_2}^2 + {y_1}^2)}$$



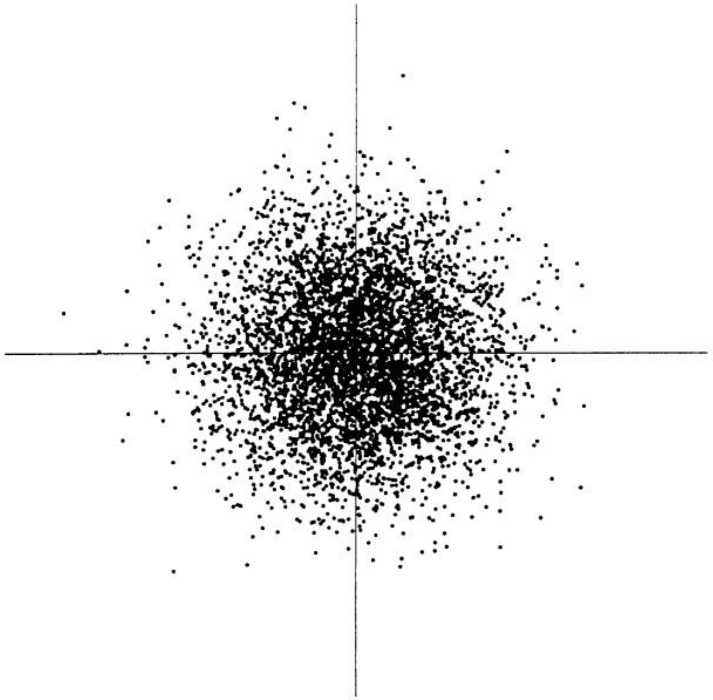
Non Gaussianity

The joint density of unit variance y_1 & y_2 is symmetric. So it doesn‘t contain any information about the directions of the cols of the mixing matrix $\boldsymbol{W}$.

So $\boldsymbol{W}$ can‘t be estimated.

We need non-gaussianity for the independent components(IC)

If only one IC is gaussian, the estimation is still possible.



Non Gaussianity

- Key element is non-gaussianity

$$\boldsymbol{A} = \boldsymbol{W}^{-1} \qquad \qquad \tilde{\boldsymbol{y}} = \boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{X}$$

- If $\boldsymbol{a}$ was one of the rows of the inverse of $\boldsymbol{W}$, this linear combination $\tilde{\boldsymbol{y}}$ would actually equal to one of the independent components.

- How could we use the Central Limit Theorem to determine $\boldsymbol{W}$ so that it would equal to one of the rows of the inverse of $\boldsymbol{A}$?

Non Gaussianity

- Let us make a change of variables

defining $\boldsymbol{z} = \boldsymbol{W}^T\boldsymbol{a} \Rightarrow$

$$\tilde{\boldsymbol{y}} = \boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{a}^T\boldsymbol{W}\boldsymbol{Y} = \boldsymbol{z}^T\boldsymbol{Y}$$

$\tilde{\boldsymbol{y}}$ is thus a linear combination of $\boldsymbol{Y}$, with weights given by $\boldsymbol{z}$

Even the sum of two independent random variables is more Gaussian than the original variables

Non Gaussianity

- Maximize the non-Gaussianity of $\boldsymbol{a}^T\boldsymbol{X}$. This means that $\boldsymbol{a}^T\boldsymbol{X}$ ˆ equals to one of the independent components!

- Maximizing the non-Gaussianity of $\boldsymbol{a}^T\boldsymbol{X}$ thus gives us one

of the independent components.

- In fact, non-Gaussianity in the *n*-dimensional space of vectors $\boldsymbol{a}_i$ has 2*n* local maxima, two for each independent component, corresponding to $\boldsymbol{y}_i$ and $-\boldsymbol{y}_i$.

- To find several independent components, we exploit the fact that different independent components are uncorrelated

Measures of Non Gaussianity

- Assuming a random variable y such that

$$E[y] = 0 \qquad \sigma^2 = E[y^2] = 1$$

- The classical measure of non-Gaussianity is kurtosis or the fourth-order cumulant

$$kurt(y) = E[y^4] - 3(E[y^2])^2 \Rightarrow kurt(y) = E[y^4] - 3$$

kurtosis is zero for a Gaussian random variable. For most (but not quite all) non-Gaussian random variables, kurtosis is non-zero.

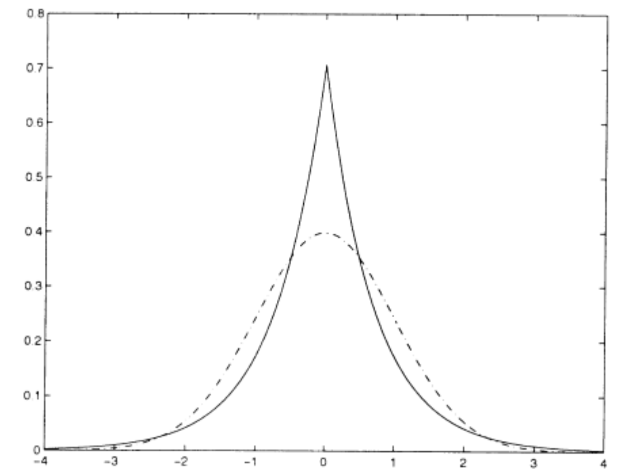
Measures of Non Gaussianity - Kurtosis

- $kurt(y) < 0$ are called sub-Gaussian,
- $kurt(y) > 0$ are called super-Gaussian..

Super-Gaussian random variables have typically a "spiky" pdf with heavy tails, i.e.

Sub-Gaussian random variables, on the other hand, have typically a "flat" pdf,

$$p(y) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|y|}$$



Measures of Non Gaussianity - Kurtosis

- Kurtosis, or rather its absolute value, has been widely used as a measure of non-Gaussianity in ICA and related fields.

- The main reason is its simplicity, both computational and theoretical.

- Computationally, kurtosis can be estimated simply by using the fourth moment of the sample data.

Measures of Non Gaussianity - Kurtosis

- Kurtosis has problems when its value is estimated from a measured sample.
  1. It is very sensitive to outliers.
  2. Its value depends on only a few observations in the tails of the distribution, which may be erroneous or irrelevant observations.

- Thus, other measures of non-Gaussianity might be better than kurtosis in some situations, i.e. negentropy that more or less combine the good properties of both measures.

Measures of Non Gaussianity - Negentropy

- A very important measure of non-Gaussianity is given by negentropy. Negentropy is based on the information-theoretic quantity of (differential) entropy.

- Entropy is the basic concept of information theory (measure of "randomness" of a variable)

- Entropy H is defined for discrete/continuous random variable y as

$$H(y) = -\sum p(y_i) \log p(y_i) \qquad H(y) = -\int p(y) \log p(y) dy$$

Measures of Non Gaussianity - Negentropy

- A fundamental result of information theory is that: *a Gaussian variable has the largest entropy among all random variables of equal variance*.

- Gaussian distribution is the "most random" or the least structured of all distributions.

- How entropy could be used as a measure of non-Gaussianity?

- Negentropy J is defined as follows:

$$\mathrm{J}(y) = H(y_{Gauss}) - H(y)$$

where is a y_{Gauss} Gaussian random variable of the same covariance matrix as y.

Measures of Non Gaussianity - Negentropy

- Negentropy is always non-negative, and is zero if and only if **y** has a Gaussian distribution

- 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

- Estimating negentropy using the definition would require an estimate (possibly non-parametric) of the pdf.

Measures of Non Gaussianity - Negentropy

- The classical method of approximating negentropy is using higher-order moments, for example as follows (zero mean and unit variance)

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

- However, the validity of such approximations may be rather limited.
- In particular, these approximations suffer from the non-robustness encountered with kurtosis

Measures of Non Gaussianity - Negentropy

- Assuming a zero mean and unit variance $\boldsymbol{y}$, a more useful approximation is the following

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

where G is practically any non-quadratic function (higher order than 2), c is an irrelevant constant, and $\boldsymbol{v}$ is a Gaussian variable of zero mean and unit variance (i.e., standardized).

- If we set $\boldsymbol{y} = \boldsymbol{a}^T\boldsymbol{X}$ then negentropy is reformulated as

$$\mathrm{J}(\boldsymbol{a}) = [E(G(\boldsymbol{a}^T\boldsymbol{X})) - E(G(\boldsymbol{v})]^2$$

Measures of Non Gaussianity - Negentropy

Examples of function G

$$G_1(\mathrm{y}) = \frac{1}{4}y^4 \qquad g_1(\mathrm{y}) = y^3$$

$$G_2(\mathrm{y}) = -\frac{1}{c_1}e^{-\frac{c_1}{2}y^2} \qquad g_2(\mathrm{y}) = ye^{-\frac{c_1}{2}y^2}$$

$$G_3(\mathrm{y}) = \frac{1}{c_2}\log\cosh c_2 y \qquad g_3(\mathrm{y}) = \tanh c_2 y$$

$$1 \le c_2 \le 2,\ c_1 \approx 1$$

Measures of Non Gaussianity - Negentropy

$$\boldsymbol{A} = \mathrm{argmax}_{\mathbf{A}}\, \mathrm{J}(\boldsymbol{A}) = \sum_{k=1}^{d} J(\boldsymbol{a}_k)$$
$$\text{s.t. } \boldsymbol{A}^T \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{A} = \boldsymbol{I}$$

Assuming whitened data, i.e. $\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}} = \boldsymbol{I}$

$$\boldsymbol{A} = \mathrm{argmax}_{\boldsymbol{A}}\, \mathrm{J}(\boldsymbol{A}) = \sum_{k=1}^{d} J(\boldsymbol{a}_k)$$
$$\text{s.t. } \boldsymbol{A}^{\boldsymbol{T}} \boldsymbol{A} = \boldsymbol{I}$$

Measures of Non Gaussianity - Negentropy

Let that we want to find one $\boldsymbol{a}$

$$\boldsymbol{a} = \text{argmax}_{\boldsymbol{a}}\, J(\boldsymbol{a}) \quad \text{s.t. } \boldsymbol{a}^T\boldsymbol{a} = 1$$

Lagrangian $\quad \mathrm{L}(\boldsymbol{a},\lambda) = J(\boldsymbol{a}) - \lambda(\boldsymbol{\alpha}^{\mathbf{T}}\boldsymbol{\alpha} - 1)$

$$\frac{\partial \mathrm{L}(\boldsymbol{a},\lambda)}{\partial \boldsymbol{\alpha}} = E[\boldsymbol{x}_i g(\boldsymbol{a}^T\boldsymbol{x}_i)] - \lambda\boldsymbol{a} \Rightarrow \lambda = E[\boldsymbol{a}^T\boldsymbol{x}_i g(\boldsymbol{a}^T\boldsymbol{x}_i)]$$

$$\frac{\partial^2 \mathrm{L}}{\partial \boldsymbol{a}_i \partial \boldsymbol{a}_j} = E[\boldsymbol{x}_i\boldsymbol{x}_i{}^T g'(\boldsymbol{a}^T\boldsymbol{x}_i)] - \lambda\mathbf{I}$$

Measures of Non Gaussianity - Negentropy

Assuming the approximation

$$E[\boldsymbol{x}_i \boldsymbol{x}_i{}^T g'(\boldsymbol{a}^T \boldsymbol{x}_i)] \approx E[\boldsymbol{x}_i \boldsymbol{x}_i{}^T] E[g'(\boldsymbol{a}^T \boldsymbol{x}_i)] = E[g'(\boldsymbol{a}^T \boldsymbol{x}_i)]\mathrm{I}$$

We get the Newton updates

$$\boldsymbol{\alpha}_+{}^{(t)} = \boldsymbol{a}^{(t-1)} - \frac{E[\boldsymbol{x}_i g(\boldsymbol{a}^T \boldsymbol{x}_i)] - \lambda \boldsymbol{a}}{E[g'(\boldsymbol{a}^T \boldsymbol{x}_i)] - \lambda}$$

$$\boldsymbol{a}^{(t)} = \frac{\boldsymbol{\alpha}_+{}^{(t)}}{||\boldsymbol{\alpha}_+{}^{(t)}||}$$

Also setting that $\lambda = E[\boldsymbol{a}^T \boldsymbol{x}_i g(\boldsymbol{a}^T \boldsymbol{x}_i)]$

We get the fix point updates

$$\boldsymbol{\alpha}_+{}^{(t)} = E[\boldsymbol{x}_i g(\boldsymbol{a}^T \boldsymbol{x}_i)] - E[g'(\boldsymbol{a}^T \boldsymbol{x}_i)]\boldsymbol{a}^{(t-1)}$$

Measures of Non Gaussianity - Negentropy

•A simple way of achieving decorrelation is a deflation scheme based on a Gram–Schmidt-like decorrelation.

•We estimate the independent components one by one, i.e. to estimate *d* independent components, or *d* vectors $\boldsymbol{a}_1, \boldsymbol{a}_2,\ldots\ \boldsymbol{a}_d$ we run the one-unit fixed point algorithm for $\boldsymbol{a}_{i+1}$; and after every iteration step we subtract from $\boldsymbol{a}_{i+1}$ the "projections" of the previously estimated *i* vectors, and then renormalize as:

$$\boldsymbol{a}_{i+1} = \boldsymbol{a}_{i+1} - \sum_{j=1}^{i} \boldsymbol{a}_{i+1}{}^{T} \boldsymbol{a}_{\boldsymbol{j}} \boldsymbol{a}_{\boldsymbol{j}}$$

$$\boldsymbol{a}_{i+1} = \frac{\boldsymbol{a}_{i+1}}{||\boldsymbol{a}_{i+1}||}$$

Preprocessing

- Centering

$$\boldsymbol{x}_i = \boldsymbol{x}_i - \boldsymbol{\mu} \quad \text{where} \quad \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$$

- Sphering

$$\widetilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of the possitive eigenvalues and $\boldsymbol{U}$ is the matrix with the corresponding eigenvectors

An overview

PCA: Maximize the global variance

LDA: Minimize the class variance while maximizing the mean variance

LPP: Minimize the local variance

ICA: Maximize independence by maximizing non-Gaussianity

All are deterministic!!