

# Fundamentals of Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Independent Vector Analysis (IVA)

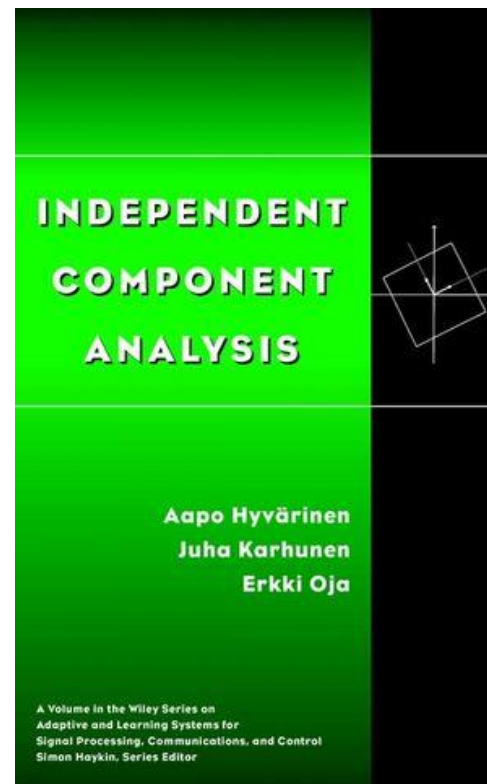
***Dr Mohsen Naqvi***

*Lecturer in Signal and Information Processing,  
School of Electrical and Electronic Engineering,  
Newcastle University*

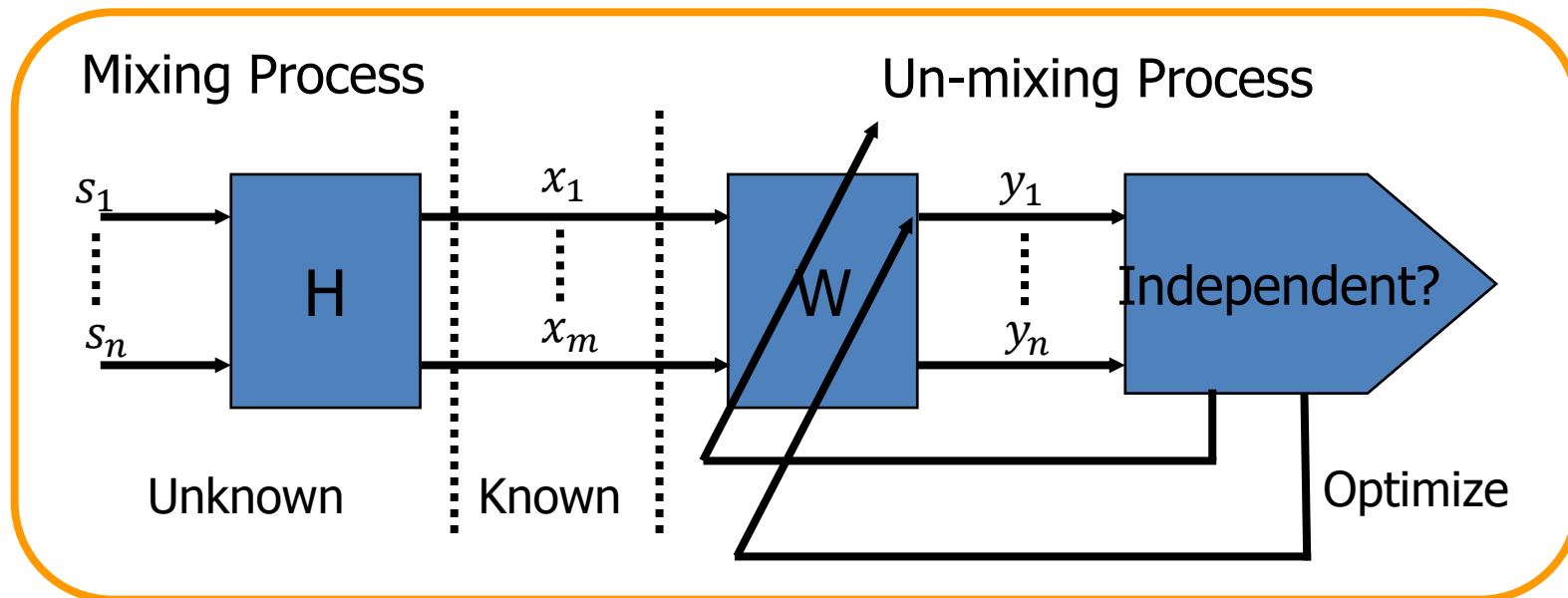
***Mohsen.Naqvi@ncl.ac.uk***

## Acknowledgement and Recommended text:

Hyvarinen et al., Independent  
Component Analysis, Wiley,  
2001

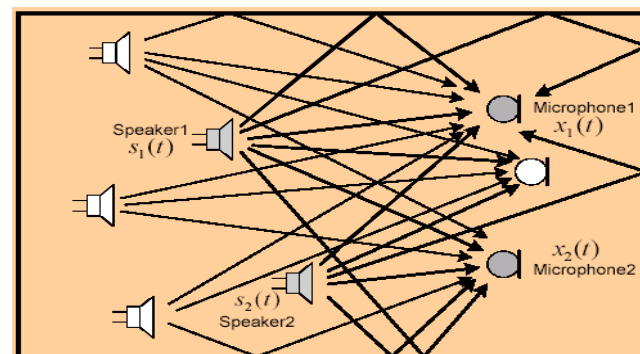


# Source Separation and Component Analysis



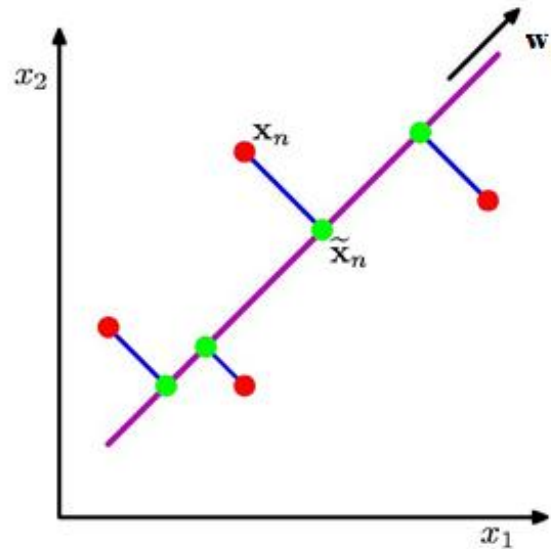
Mixing Model:  $x = Hs$

Un-mixing Model:  $y = Wx = WHs = PDs$



# Principal Component Analysis (PCA)

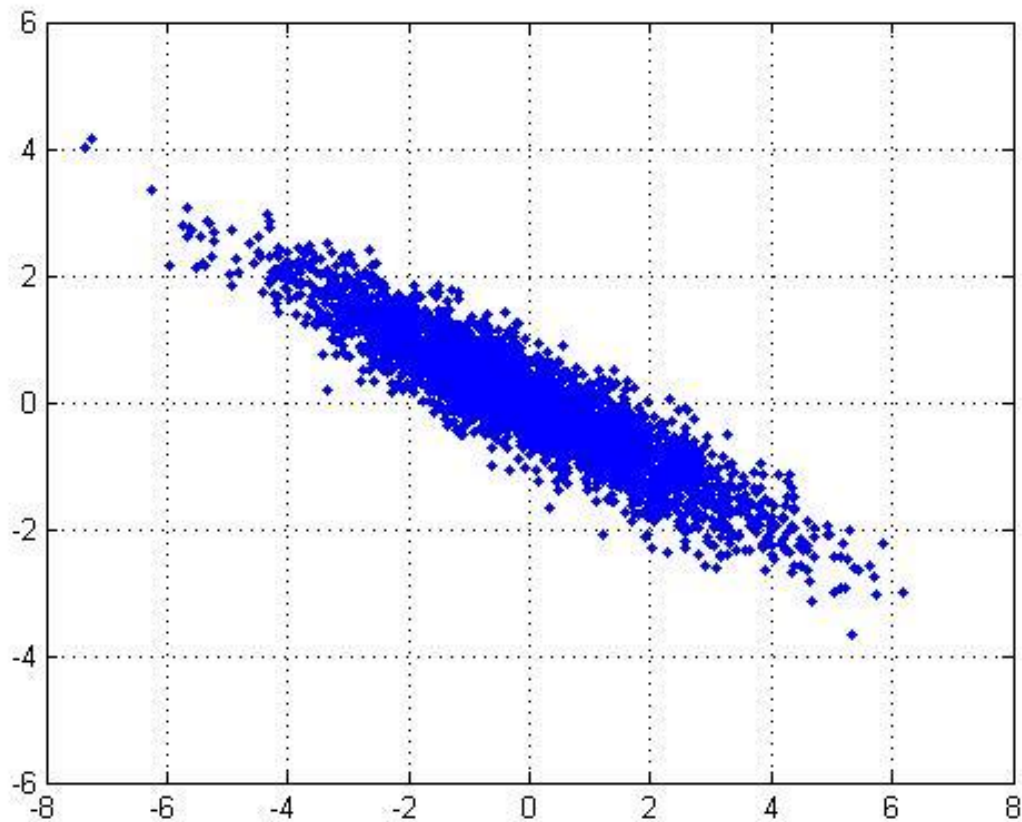
Orthogonal projection of data onto lower-dimension linear space:



- i. maximizes variance of projected data
- ii. minimizes mean squared distance between data point and projections

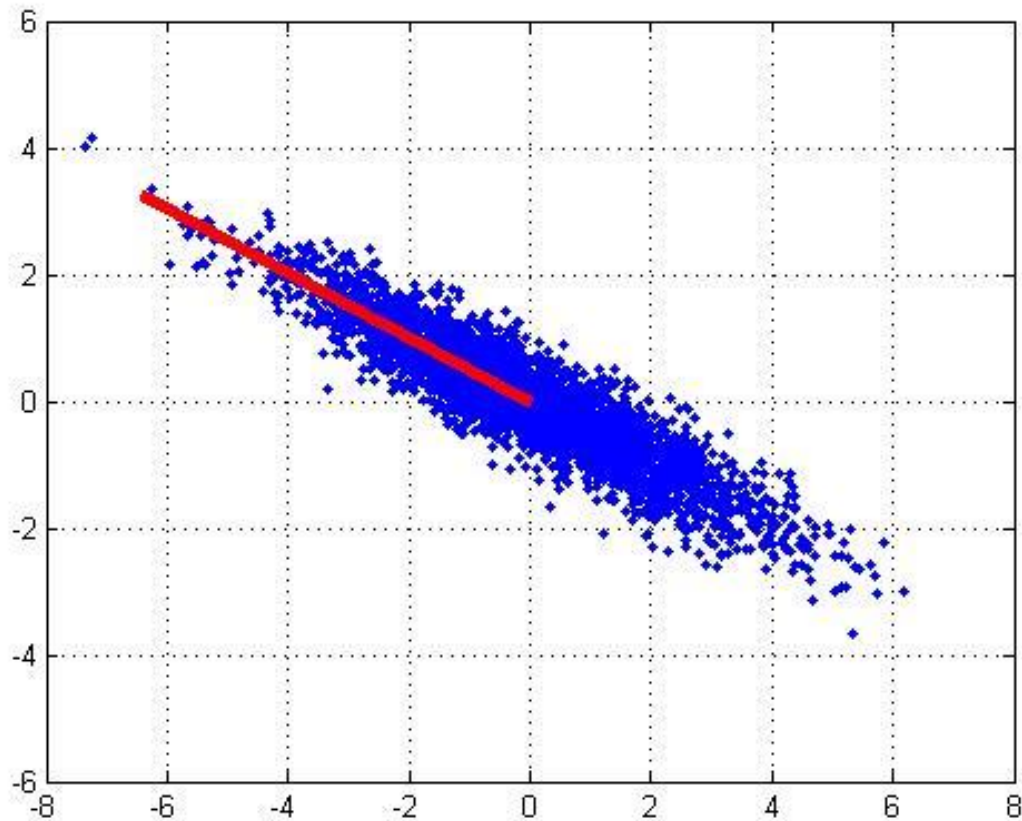
# Principal Component Analysis (PCA)

Example: 2-D Gaussian Data



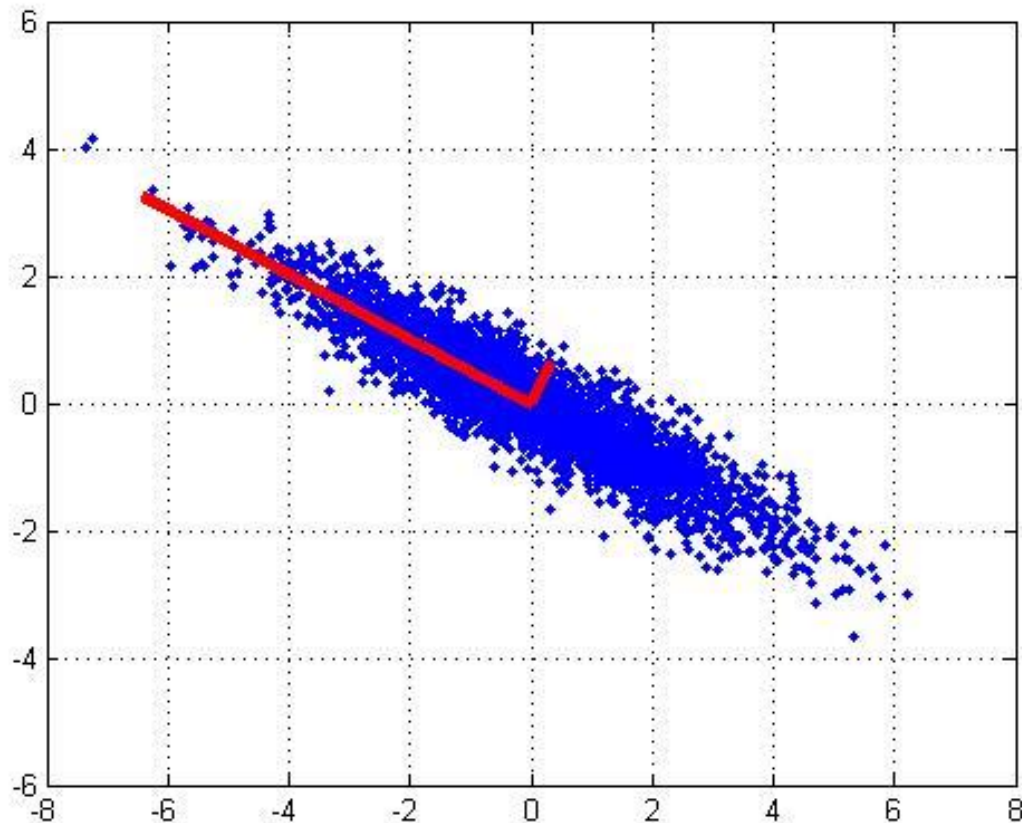
# Principal Component Analysis (PCA)

Example: 1<sup>st</sup> PC in the direction of the largest variance



# Principal Component Analysis (PCA)

Example: 2<sup>nd</sup> PC in the direction of the second largest variance and each PC is orthogonal to other



# Principal Component Analysis (PCA)

- In PCA the redundancy is measured by correlation between data elements
- Using only the correlations as in PCA has the advantage that the analysis can be based on second-order statistics (SOS)

In the PCA, first the data is centered by subtracting the mean

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



# Principal Component Analysis (PCA)

The data  $x$  is linearly transformed

$$y_1 = \sum_{j=1}^n w_{j1} x_j = w_1^T x$$

subject to  $\|w_1\| = 1$

Mean of the first transformed factor

$$E\{y_1\} = E\{w_1^T x\} = w_1^T E\{x\} = 0$$

Variance of the first transformed factor

$$E\{y_1^2\} = E\{(w_1^T x)^2\} = E\{(w_1^T x)(x^T w_1)\}$$

# Principal Component Analysis (PCA)

$$\begin{aligned} E\{y_1^2\} &= E\{(w_1^T x)(x^T w_1)\} = w_1^T E\{xx^T\} w_1 \\ &= w_1^T R_{XX} w_1 \end{aligned} \quad (1)$$

The above correlation matrix is symmetric

$$a^T R_{XX} b = b^T R_{XX} a$$

Correlation matrix  $R_{XX}$  is not in our control and depends on the data  $x$ .

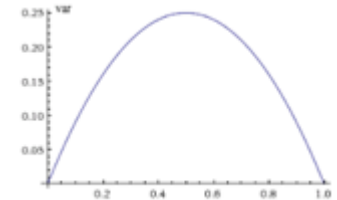
**We can control  $w$ ?**

# Principal Component Analysis (PCA)

## (PCA by Variance Maximization)

At maximum variance a small change in  $w$  will not affect the variance

$$\begin{aligned}
 E\{y_1^2\} &= (w_1 + \Delta w)^T R_{XX} (w_1 + \Delta w) \\
 &= w_1^T R_{XX} w_1 + 2\Delta w^T R_{XX} w_1 + \Delta w^T R_{XX} \Delta w
 \end{aligned}$$



where  $\Delta w$  is very small quantity and therefore

$$E\{y_1^2\} = w_1^T R_{XX} w_1 + 2\Delta w^T R_{XX} w_1 \quad (2)$$

By using (1) and (2) we can write

$$\Delta w^T R_{XX} w_1 = 0 \quad (3)$$

# Principal Component Analysis (PCA)

As we know that

$$\|w_1 + \Delta w\| = 1$$

therefore

$$(w_1 + \Delta w)^T (w_1 + \Delta w) = 1$$

$$w_1^T w_1 + 2\Delta w^T w_1 + \Delta w^T \Delta w = 1$$

where  $\Delta w$  is very small quantity and therefore

$$1 + 2\Delta w^T w_1 + 0 = 1$$

$$\Delta w^T w_1 = 0 \tag{4}$$

The above result shows that  $w_1$  and  $\Delta w$  are orthogonal to each other.

## Principal Component Analysis (PCA)

By careful comparison of (3) and (4) and by considering  $R_{XX}$

$$\Delta w^T R_{XX} w_1 - \Delta w^T \lambda_1 w_1 = 0$$

$$\Delta w^T [R_{XX} w_1 - \lambda_1 w_1] = 0$$

since  $\Delta w^T \neq 0$

$$\therefore [R_{XX} w_1 - \lambda_1 w_1] = 0$$

$$\rightarrow R_{XX} w_i = \lambda_i w_i \quad i = 1, 2, \dots, n \quad (5)$$

where

$\lambda_1, \lambda_2, \dots, \dots, \lambda_n$  are eigenvalues of  $R_{XX}$

and

$w_1, w_2, \dots, \dots, w_n$  are eigenvectors of  $R_{XX}$

## Principal Component Analysis (PCA)

$$R_{XX} w_i = \lambda_i w_i, \quad i=1, 2, \dots, n$$

$$\lambda_1 = \lambda_{max} > \lambda_2 > \dots > \lambda_n$$

and

$$E = [w_1, w_2, \dots, w_n]$$

$$\therefore R_{XX}E = E\Lambda \quad (6)$$

where

$$\Lambda = \text{dig} [\lambda_1, \lambda_2, \dots, \lambda_n]$$

We know that E is orthogonal matrix

$$E^T E = I$$

and we can update (7)

$$E^T R_{XX}E = E^T E \Lambda = I\Lambda = \Lambda \quad (7)$$

## Principal Component Analysis (PCA)

In expanded form (7) is

$$w_i^T R_{XX} w_j = \begin{cases} \lambda_i & i = j \\ 0 & i \neq j \end{cases}$$

$R_{XX}$  is symmetric and positive definite matrix and can be represented as [Hyvarinen et al.]

$$R_{XX} = \sum_{i=1}^n \lambda_i w_i^T w_j$$

We know

$$w_i^T w_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

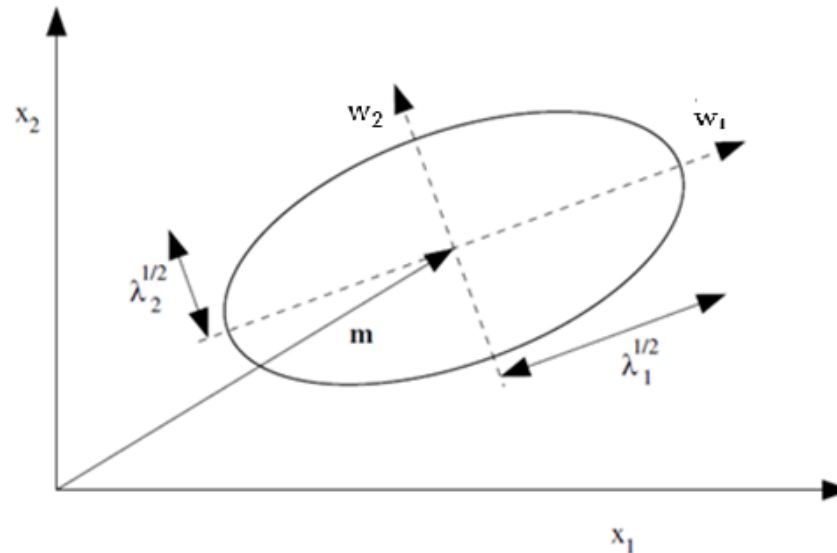
Finally, (1) will become

$$2, \dots, n \quad (8) \quad E\{y_i^2\} = w_i^T R_{XX} w_i = \lambda_i, \quad i=1,$$

Can we define the principal components now?

# Principal Component Analysis (PCA)

In a multivariate Gaussian probability density, the **principal components are in the directions of eigenvectors  $w_i$**  and the respective variances are the eigenvalues  $\lambda_i$ .



There are  $n$  possible solutions for  $w$  and

$$y_i = w_i^T x = x^T w_i \quad i = 1, 2, \dots, n$$



# Principal Component Analysis (PCA)

## Whitening

Data whitening is another form of PCA

The objective of whitening is to transform the observed vector  $z = Vx$ , which is uncorrelated and with variance equal to identity matrix

$$E\{zz^T\} = I$$

The unmixing matrix,  $W$ , can be decomposed into two components:

$$W = UV$$

Where  $U$  is rotation matrix and  $V$  is the whitening matrix and

$$V = \Lambda^{-1/2} E^T$$

# Principal Component Analysis (PCA)

## Whitening

$$z = Vx$$

$$\begin{aligned} E\{zz^T\} &= VE\{xx^T\}V^T = \Lambda^{-1/2} E^T E\{xx^T\}E \Lambda^{-1/2} \\ &= \Lambda^{-1/2} E^T R_{xx}E \Lambda^{-1/2} = I \end{aligned}$$

The covariance is unit matrix, therefore  $z$  is whitened

Whitening matrix,  $V$ , is not unique. It can be pre-multiplied by an orthogonal matrix to obtain another version of  $V$

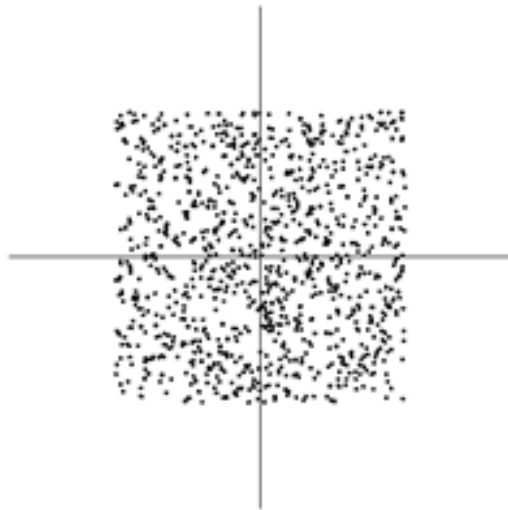
### Limitations:

PCA only deals with second order statistics and provides only decorrelation. Uncorrelated components are not independent

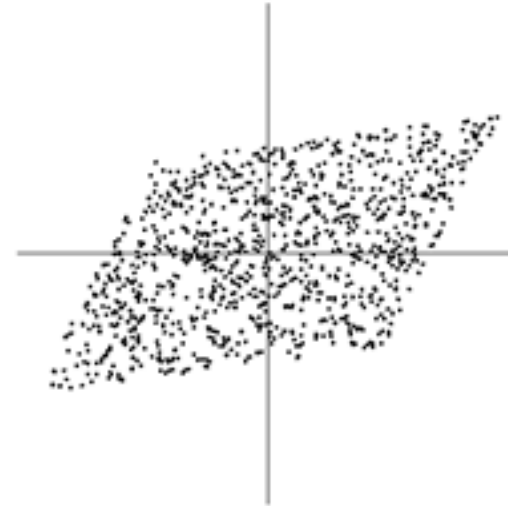
# Principal Component Analysis (PCA)

## Limitations

Two components with uniform distributions and their mixture



The joint distribution of ICs  $s_1$  (horizontal axis) and  $s_2$  (vertical axis) with uniform distributions.

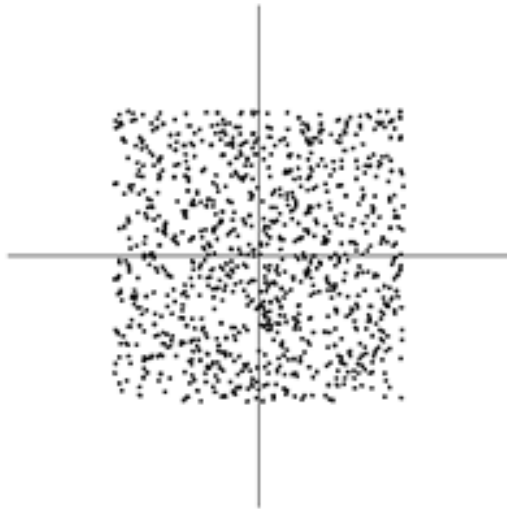


The joint distribution of observed mixtures  $x_1$  (horizontal axis) and  $x_2$  (vertical axis).

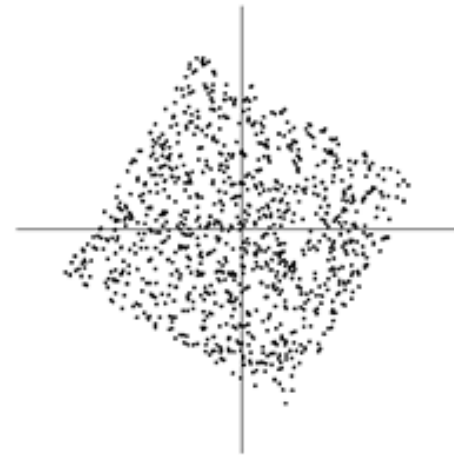
# Principal Component Analysis (PCA)

## Limitations

Two components with uniform distributions and after PCA



The joint distribution of ICs  $s_1$  (horizontal axis) and  $s_2$  (vertical axis) with uniform distributions.



The joint distribution of the whitened mixtures of uniformly distributed ICs

PCA does not find original coordinates

➤ Factor rotation problem

# Principal Component Analysis (PCA)

## Limitations: uncorrelation and independence

If two zero mean random variables  $s_1$  and  $s_2$  are uncorrelated then their covariance is zero:

$$\text{cov}(s_1, s_2) = E\{s_1 s_2\} = 0$$

If  $s_1$  and  $s_2$  are independent, then for any two functions,  $g_1$  and  $g_2$  we have

$$E\{g_1(s_1)g_2(s_2)\} = E\{g_1(s_1)\}E\{g_2(s_2)\}$$

If random variables are independent then they are also uncorrelated

If  $s_1$  and  $s_2$  are discrete valued and follow such a distribution that the pair are with probability 1/4 equal to any of the following values: (0,1), (0,-1), (1,0) and (-1,0). Then

$$E\{s_1^2 s_2^2\} = 0 \neq \frac{1}{4} = E\{s_1^2\}E\{s_2^2\}$$

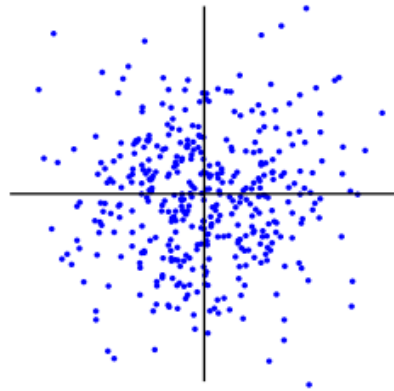
Therefore, uncorrelated components are not independent

# Principal Component Analysis (PCA)

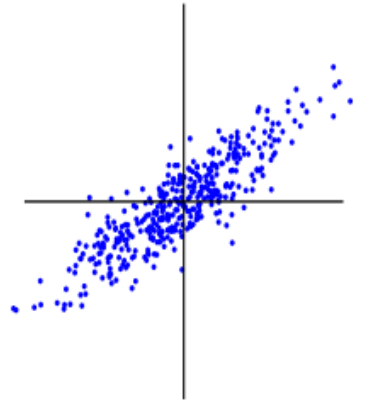
## Limitations: uncorrelation and independence

However, uncorrelated components of Gaussian distribution are independent

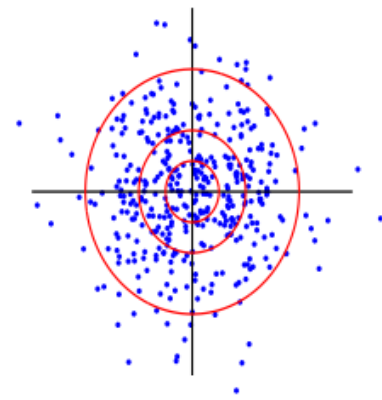
Original components,



observed mixtures,



PCA

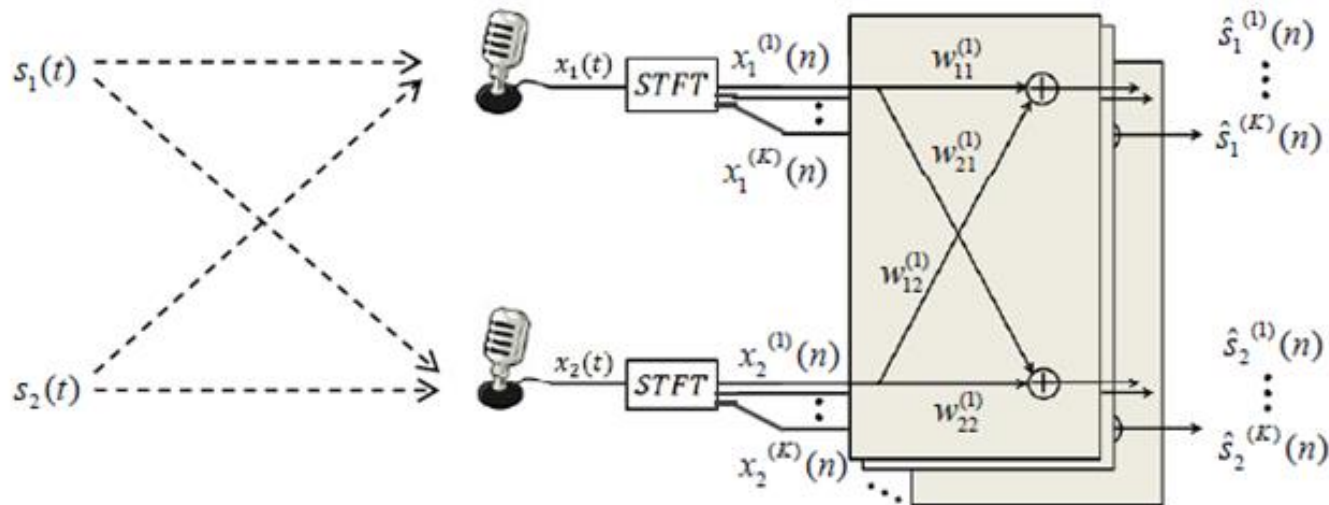


Distribution after PCA is same as distribution before mixing

# Independent Component Analysis (ICA)

Independent component analysis (ICA) is a **statistical and computational technique** for **revealing hidden factors** that underlie sets of random variables, measurements, or signals.

ICA separates the sources at each frequency bin



# Independent Component Analysis (ICA)

The fundamental restrictions in ICA are:

- i. The sources are assumed to be **statistically independent** of each other. Mathematically, independence implies that the **joint probability density function  $p(s)$  of the sources can be factorized**

$$p(s) = \prod_{i=1}^n p_i(s_i)$$

where  $p_i(s_i)$  is the marginal distribution of the  $i$ -th source

- ii. **All but one of the sources must have non-Gaussian distributions**



# Independent Component Analysis (ICA)

## Why?

The joint pdf of two Gaussian ICs,  $s_1$  and  $s_2$  is:

$$p(s_1, s_2) = \frac{1}{2\pi} \exp\left(-\frac{s_1^2 + s_2^2}{2}\right) = \frac{1}{2\pi} \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right)$$

If the mixing matrix  $\mathbf{H}$  is orthogonal, the joint pdf of the mixtures  $x_1$  and  $x_2$  is:

$$p(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{\|\mathbf{H}^T \mathbf{s}\|^2}{2}\right) |\det \mathbf{H}^T|$$

$\mathbf{H}$  is orthogonal, therefore  $\|\mathbf{H}^T \mathbf{s}\|^2 = \|\mathbf{s}\|^2$  and  $|\det \mathbf{H}^T| = 1$

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

Orthogonal mixing matrix does not change the pdf and original and mixing distributions are identical. Therefore, there is no way to infer the mixing matrix from the mixtures

## Independent Component Analysis (ICA)

- iii. In general, the mixing matrix  $H$  is square ( $n = m$ ) and invertible
- iv. Methods to realize ICA are more sensitive to data length than the methods based on SOS

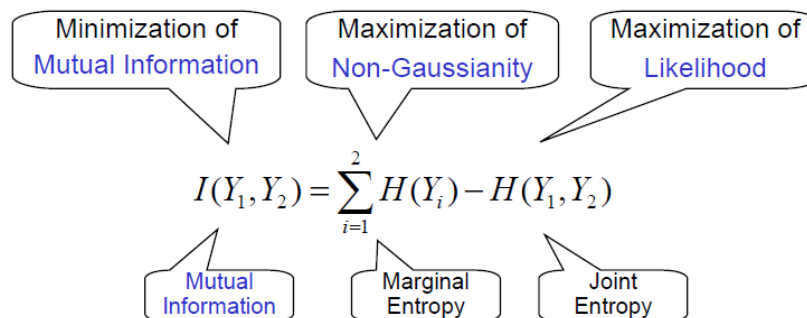
Mainly, ICA relies on two steps:

A **statistical criterion**, e.g. nongaussianity measure, expressed in terms of a **cost/contrast function**  $J(g(y))$

An **optimization technique** to carry out the minimization or maximization of the cost function  $J(g(y))$

# Independent Component Analysis (ICA)

1. Minimization of **Mutual Information**  
(Minimization of **Kullback-Leibler Divergence**)
2. Maximization of **Likelihood**
3. Maximization of **Nongaussianity**



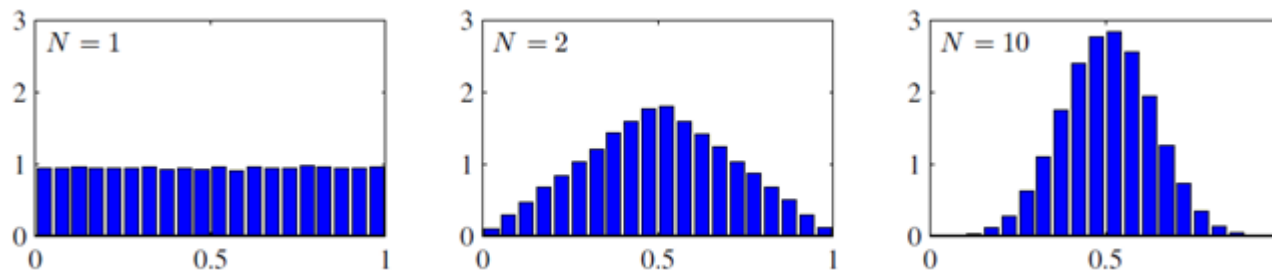
$H(\cdot)$ : Entropy

➤ All solutions are **identical**

# Independent Component Analysis (ICA)

## Nongaussainity and Independence

**Central limit theorem:** subject to certain mild conditions, the sum of a set of random variables has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases



Histogram plots of the mean of  $N$  uniformly distributed numbers for various values of  $N$ . As  $N$  increases, the joint distribution tends towards a Gaussian.

So roughly, any mixture of components will be more Gaussian than the components themselves

# Independent Component Analysis (ICA)

## Nongaussainity and Independence

**Central limit theorem:** The distribution of Independent random variables tends towards a Gaussian distribution, under certain conditions

If  $z_i$  is independent and identically distributed random variable then

$$x_n = \sum_{i=1}^n z_i$$

Mean and variance of  $x_n$  grow without bound when  $n \rightarrow \infty$ , and  $y_n = (x_n - \mu_{x_n}) / \delta_{x_n}^2$

It has been shown that the distribution of  $y_n$  converges to a Gaussian distribution with zero mean and unit variance when  $n \rightarrow \infty$ . This is known as the central limit theorem

# Independent Component Analysis (ICA)

## Nongaussainity and Independence

In BSS, if  $s_j, j = 1, \dots, n$  are unknown source signals and mixed with coefficients  $h_{ij}, i = 1, \dots, m$  then

$$x_i = \sum_{j=1}^n h_{ij} s_j$$

The distribution of the mixture  $x_i$  is usually near to Gaussian when the number of sources  $s_j$  is fairly small

We know that

$$y_i = \sum_{j=1}^m w_{ji} x_j = w_i^T \mathbf{x}$$

How could the central limit theorem be used to determine  $w_i$  so that it would equal to one of the rows of the inverse of mixing matrix  $H$  ( $y \approx H^{-1}x$ ) ?

# Independent Component Analysis (ICA)

## Nongaussainity and Independence

In practice, we **couldn't determine** such a  $w_i$  **exactly**, because the **problem is blind** i.e.  $H$  is unknown

We can find **an estimator** that gives good **approximation** of  $w_i$ .

By **varying** the coefficients in  $w_i$ , and **monitoring** the change in distribution of  $y_i = w_i^T x$

Hence, maximizing the non- Gaussainity of  $w_i^T x$  provides one of the Independent Components

**How we can maximize the nongaussainity ?**

# Independent Component Analysis (ICA)

## Nongaussainity by Kurtosis

For ICA estimation based on nongaussainity, we require a qualitative measures of nongaussainity of a random variable

**Kurtosis** is the name given to the **forth-order cumulants** of a random variable e.g.  $y$ , the kurtosis is defined as:

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

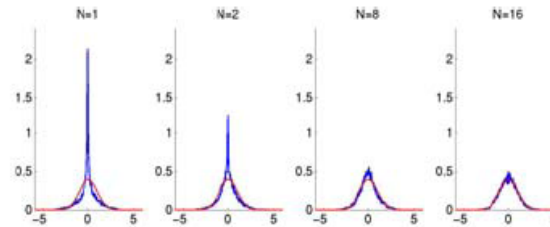
If the data is whitened, e.g. zero mean and unit variance, then  $\text{kurt}(y) = E\{y^4\} - 3$  is normalized version of the fourth moment  $E\{y^4\}$



# Independent Component Analysis (ICA)

## Nongaussainity by Kurtosis

For Gaussian variable  $y$ , the fourth moment is equal to  $3(E\{y^2\})^2$ . Therefore, **kurtosis is zero for Gaussian random variables**



N	1	2	8	16		Gaussian
Kurtosis	2.1	1.8	0.70	0.39		0

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

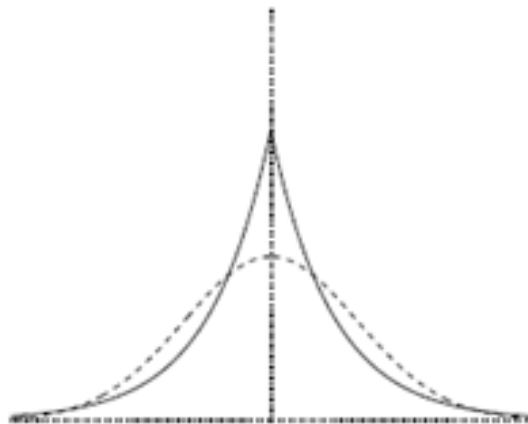
Typically nongaussainity is measured by the absolute value of kurtosis

# Independent Component Analysis (ICA)

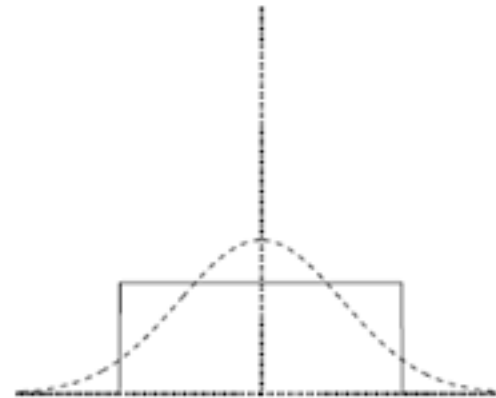
## Nongaussian Distributions

Random variable that have a **positive kurtosis** are called **supergaussian**.  
A typical example is Laplacian distribution

Random variable that have a **negative kurtosis** are called **subgaussian**.  
A typical example is the uniform distribution



Laplacian and Gaussian distributions.



Uniform and Gaussian distributions.

# Independent Component Analysis (ICA)

## Limitations of Kurtosis

Kurtosis is not a robust measure of nongaussainity

Kurtosis can be very sensitive to outliers. For example, in 1000 samples of a random variable of zero mean and unit variance, if one value is equal to 10, then

$$\text{kurt}(y) = \frac{10^4}{1000} - 3 = 7$$

A single value can make kurtosis large and the value of kurtosis may depends only on few observations in the tail of the distribution

**We have a better measure of nongaussainity than kurtosis?**

# Independent Component Analysis (ICA)

## Nongaussainity by Negentropy

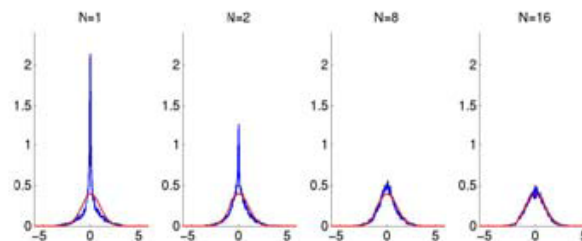
A fundamental result of information theory is that a Gaussian variable has the largest entropy among all variables of equal variance.

A robust but computationally complicated measure of nongaussainity is negentropy

# Independent Component Analysis (ICA)

## Nongaussainity by Negentropy

A measure that is **zero for Gaussian variables** and always **nonnegative** can be obtained from differential entropy, and called **negentropy**



# sources $N$	1	2	8	16	Gaussian
Entropy $H$	1.19	1.33	1.39	1.40	1.41
Negentropy $N$	0.225	0.087	0.025	0.012	0

$$H(y) = \sum_{i=1}^n p_i \log \frac{1}{p_i}$$

$$N(y) = H(x_{\text{gauss}}) - H(y)$$

# Independent Component Analysis (ICA)

## Nongaussainity by Negentropy

A classical method of approximating **negentropy** based on **higher-order cumulants** for zero mean random variable  $y$  is:

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

For **symmetric distributions**, the **first term** on right hand side of the above equation is zero.

**Therefore, a more sophisticated approximation of negentropy is required**

# Independent Component Analysis (ICA)

## Nongaussainity by Negentropy

The method that uses nonquadratic function  $G$ , and provides a simple way of approximating the negentropy is [Hyvarinen et al]:

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

where  $v$  is a standard random Gaussian variable

By choosing  $G$  wisely, we can obtain approximation of negentropy that is better than the one provided by kurtosis

# Independent Component Analysis (ICA)

## Nongaussainity by Negentropy

By choosing a  $G$  that does not grow too fast, we can obtain more robust estimator.

The following choices of  $G$  have proved that these are very useful

$$G_1(y) = \frac{1}{a_1} \log \cosh a_1 y, \quad G_2(s) = \exp\left(-\frac{y^2}{2}\right)$$

where  $1 < a_1 < 2$  is a constant and mostly equal to one

The derivatives of above contrast functions are used in ICA algorithms

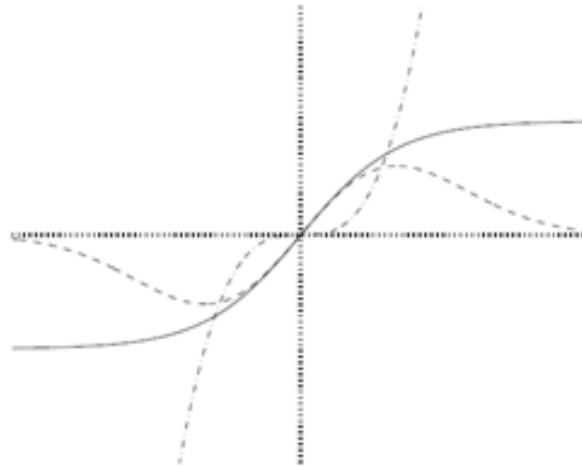
$$\begin{aligned} g_1(y) &= \tanh(a_1 y), \\ g_2(y) &= y \exp\left(-\frac{y^2}{2}\right), \\ g_3(y) &= y^3 \end{aligned}$$



# Independent Component Analysis (ICA)

## Nongaussainity by Negentropy

Different nonlinearities are required, which depends on the distribution of ICs



The robust nonlinearities  $g_1$  (solid line),  $g_2$  (dashed line) and  $g_3$  (dash-dotted line).

Can we optimize the local maxima for nongaussainity of a linear combination  $y = \sum_i w_i x_i$  under the constraint that the variance of  $y$  is constant?

# Independent Component Analysis (ICA)

## ICA Estimation Principle 1

To find the unmixing matrix  $W$ , so that the components  $y_i$  and  $y_j$ , where  $i \neq j$ , are uncorrelated, and the their nonlinear functions  $g(y_i)$  and  $h(y_j)$ , are also uncorrelated.

The above nonlinear decorrelation is the basic ICA method

# Independent Component Analysis (ICA)

## ICA Learning Rules

In gradient descent, we minimize a cost function  $J(W)$  from its initial point by computing its gradient at that point, and then move in the direction of steepest descent

The update rule with the gradient taken at the point  $W_k$  is:

$$W_{k+1} = W_k - \mu \frac{\partial J(W)}{\partial W}$$

According to Amari et al., the largest increase in  $J(W+\partial W)$  can be obtained in the direction of natural gradient

$$\frac{\partial J(W)}{\partial W_{nat}} = \frac{\partial J(W)}{\partial W} W^T W$$

# Independent Component Analysis (ICA)

## ICA Learning Rules

Therefore, the update equation of natural gradient descent is:

$$W_{k+1} = W_k - \mu \frac{\partial J(W)}{\partial W} W_k^T W_k$$

The final update equation of the natural gradient algorithm is:

$$W_{k+1} = W_k + \mu [I - y^T g(y)] W_k$$

where  $\mu$  is called the step size or learning rate

For stability and convergence of the algorithm  $0 < \mu \ll \infty$ . If we select a very small value of  $\mu$  then it will take more time to reach local minima or maxima

# Independent Component Analysis (ICA)

## ICA Learning Rule

Initialize the unmixing matrix

$$W_0 = \text{randn}(n, n)$$

For  $n = 2$

$$Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad g(Y) = \begin{bmatrix} g(y_1) \\ g(y_2) \end{bmatrix}$$

The target is:

$$[I - yg(y)^T] = 0$$

$$\Rightarrow E\{I\} = E\{yg(y)^T\} = R_{yy}$$

How we can diagonalize  $R_{yy}$ ?

# Independent Component Analysis (ICA)

## ICA Learning Rule

If

$$\Delta W = W_{k+1} - W_k$$

and

$$R_{yy} = \begin{bmatrix} y_1 g(y_1) & y_1 g(y_2) \\ y_2 g(y_1) & y_1 g(y_2) \end{bmatrix},$$

Then

$$\Delta W = \mu \begin{bmatrix} c_1 - y_1 g(y_1) & y_1 g(y_2) \\ y_2 g(y_1) & c_2 - y_1 g(y_2) \end{bmatrix} W \Rightarrow 0$$

Update  $W$  so that  $y_1$  and  $y_2$  become mutually independent

# Independent Component Analysis (ICA)

## ICA Estimation Principle 2

To find the local maxima of nongaussainity of a linear combination  $y = \sum_i w_i x_i$  under the condition that the variance of  $y$  is constant. Each local maxima will provide one independent component

ICA by maximization of nongaussainity is one of the most widely used techniques

# Independent Component Analysis (ICA)

## ICA Learning Rule

We can derive a simple gradient algorithm to maximize negentropy

First we whiten the data  $z = Vx$ , where  $V$  is the whitening matrix

We already know that the function to measure negentropy is:

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

We know that  $y = w^T z$  and take the derivative of  $J(y)$  with respect to  $w$

$$\frac{\partial J(y)}{\partial w} \propto 2[E\{G(w^T z)\} -$$



# Independent Component Analysis (ICA)

## ICA Learning Rule

$$\frac{\partial J(y)}{\partial \mathbf{w}} \propto 2[E\{G(\mathbf{w}^T \mathbf{z})\} - E\{G(v)\}] \frac{\partial}{\partial \mathbf{w}} [E\{G(\mathbf{w}^T \mathbf{z})\} - E\{G(v)\}]$$

$$\propto 2[E\{G(\mathbf{w}^T \mathbf{z})\} - E\{G(v)\}] E\{g(\mathbf{w}^T \mathbf{z}) \frac{\partial}{\partial \mathbf{w}} (\mathbf{w}^T \mathbf{z})\}$$

$$\propto 2[E\{G(\mathbf{w}^T \mathbf{z})\} - E\{G(v)\}] E\{\mathbf{z}g(\mathbf{w}^T \mathbf{z})\}$$

If  $\gamma = 2[E\{G(\mathbf{w}^T \mathbf{z})\} - E\{G(v)\}]$  then

$$\frac{\partial J(y)}{\partial \mathbf{w}} \propto \gamma E\{\mathbf{z}g(\mathbf{w}^T \mathbf{z})\}$$

Hence, the update equations of a gradient algorithm are:

$$\Delta \mathbf{w} \propto \gamma E\{\mathbf{z}g(\mathbf{w}^T \mathbf{z})\} \quad \text{and } \mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

# Independent Component Analysis (ICA)

## Implementation Summary

- i. Center the data  $\mathbf{x} = \mathbf{x} - E\{\mathbf{x}\}$
- ii. Whiten the data  $\mathbf{z} = \mathbf{V}\mathbf{x}$
- iii. Select a nonlinearity  $g$
- iv. Randomly initialize  $\mathbf{w}$ , with  $\|\mathbf{w}\| = 1$ , and  $\gamma$
- v. Update  $\Delta\mathbf{w} \propto \gamma E\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\}$
- vi. Normalize  $\mathbf{w} \leftarrow \mathbf{w}/\|\mathbf{w}\|$
- vii. Update, if required,  $\Delta\gamma \propto [E\{G(\mathbf{w}^T\mathbf{z})\} - E\{G(v)\}] - \gamma$

Repeat from step-v, if not converged.

# Independent Component Analysis (ICA)

## Fast Fixed-point Algorithm (FastICA)

FastICA is based on a fixed point iteration scheme to find a maximum of the nongaussianity of IC  $y$

The above gradient method suggests the following fixed point iteration

$$w = E\{zg(w^T z)\}$$

the coefficient  $\gamma$  is eliminated by the normalization step

We can modify the iteration by multiplying  $w$  with some constant  $\alpha$  and add on both sides of the above equation

$$(1+\alpha)w = E\{zg(w^T z)\} + \alpha w$$

# Independent Component Analysis (ICA)

## Fast Fixed-point Algorithm (FastICA)

$$(1+\alpha)w = E\{zg(w^T z)\} + \alpha w$$

Due to the subsequent normalization of  $w$  to unit norm, the above equation gives a fixed point iteration. And  $\alpha$  plays an important role in fast convergence

The optima of  $E\{G(w^T z)\}$  under the constraint  $E\{G(w^T z)^2\} = \|w\|^2 = 1$  are obtained at points when the gradient of the Lagrangian is zero [Hyvarinen et al.]

$$\frac{\partial J(y)}{\partial w} = E\{zg(w^T z)\} + \beta w$$

We can take derivative of the above equation to find the local maxima

$$\frac{\partial^2 J(y)}{\partial w^2} = E\{zz^T g'(w^T z)\} + \beta I$$

# Independent Component Analysis (ICA)

## Fast Fixed-point Algorithm (FastICA)

Data is whitened therefore

$$\frac{\partial^2 J(\mathbf{y})}{\partial \mathbf{w}^2} = E\{g'(\mathbf{w}^T \mathbf{z})\} \mathbf{I} + \beta \mathbf{I}$$

According to Newton's method

$$\mathbf{w} \leftarrow \mathbf{w} - \left[ \frac{\partial^2 J(\mathbf{y})}{\partial \mathbf{w}^2} \right]^{-1} \frac{\partial J(\mathbf{y})}{\partial \mathbf{w}}$$

Therefore Newton iteration is:

$$\mathbf{w} \leftarrow \mathbf{w} - [E\{g'(\mathbf{w}^T \mathbf{z})\} + \beta]^{-1} (E\{\mathbf{z}g(\mathbf{w}^T \mathbf{z})\} + \beta \mathbf{w})$$

# Independent Component Analysis (ICA)

## Fast Fixed-point Algorithm (FastICA)

$$w \leftarrow w - [E\{g'(w^T z)\} + \beta]^{-1} (E\{zg(w^T z)\} + \beta w)$$

Multiply both sides of above equation with  $E\{g'(w^T z)\} + \beta$  and by simplifying, we obtain basic fixed-point iteration in FastICA

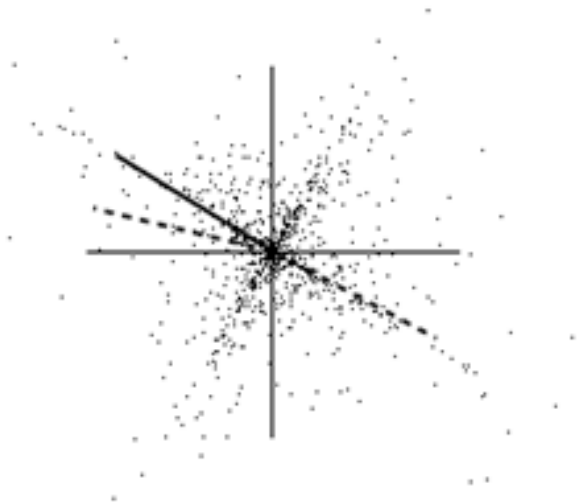
$$w \leftarrow E\{zg(w^T z)\} - E\{g'(w^T z)\}$$

and

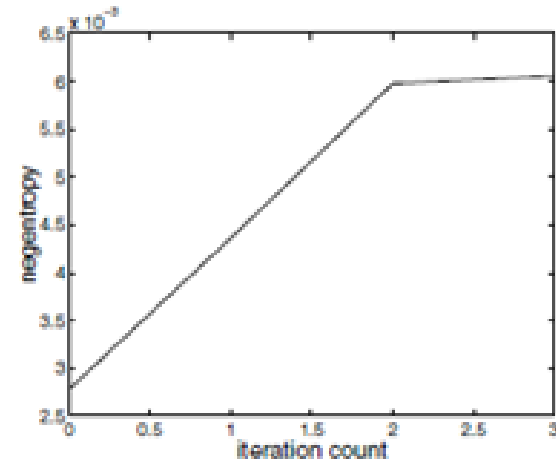
$$w \leftarrow \frac{w}{\|w\|}$$

# Independent Component Analysis (ICA)

## Fast Fixed-point Algorithm (FastICA)



Results with FastICA using negentropy. Dotted and solid lines show  $w$  after first and second iterations respectively (not at actual scale).



The convergence of FastICA using negentropy, for supergusasain ICs (not at actual scale)

# Independent Component Analysis (ICA)

## Implementation Summary of FastICA Algorithm

- i. Center the data  $\mathbf{x} = \mathbf{x} - E\{\mathbf{x}\}$
- ii. Whiten the data  $\mathbf{z} = \mathbf{V}\mathbf{x}$
- iii. Select a nonlinearity  $g$
- iv. Randomly initialize  $\mathbf{w}$ , with constraint  $\|\mathbf{w}\| = 1$
- v. Update  $\mathbf{w} \leftarrow E\{\mathbf{z}g(\mathbf{w}^T \mathbf{z})\} - E\{g'(\mathbf{w}^T \mathbf{z})\}$
- vi. Normalize  $\mathbf{w} \leftarrow \mathbf{w} / \|\mathbf{w}\|$

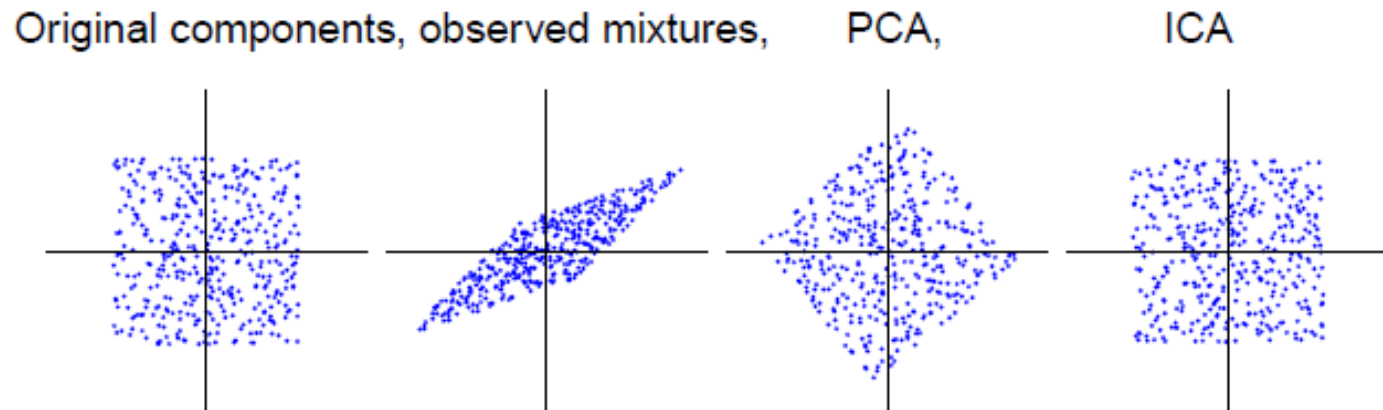
Repeat from step-v till converge



# Independent Component Analysis (ICA)

## Illustration of PCA vs. ICA

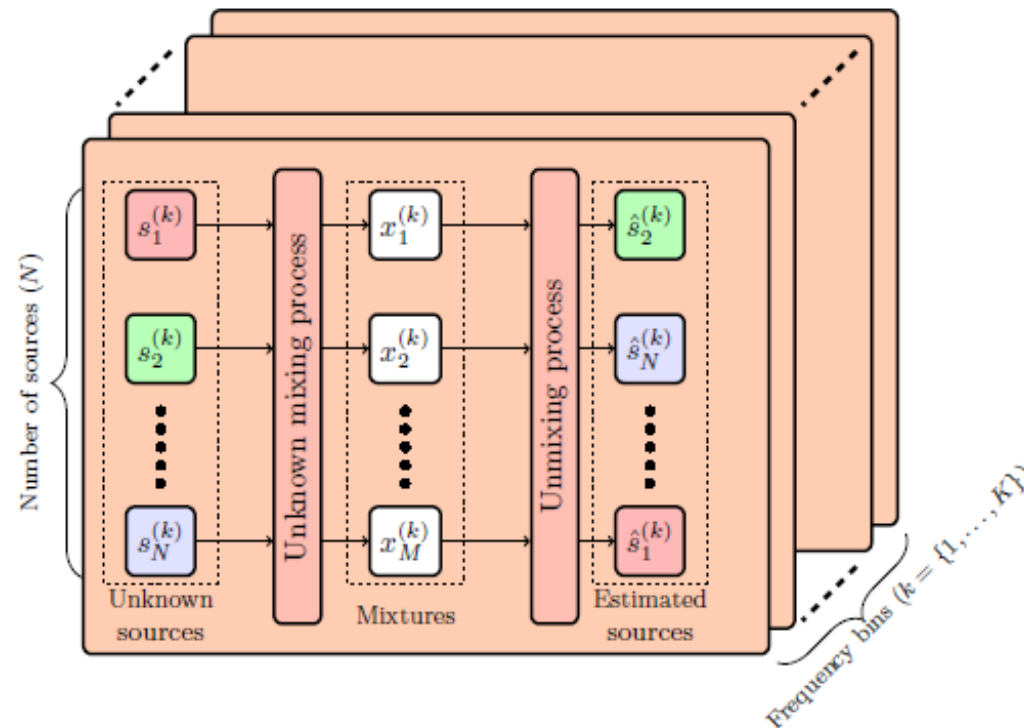
Two components with uniform distributions:



# Independent Component Analysis (ICA)

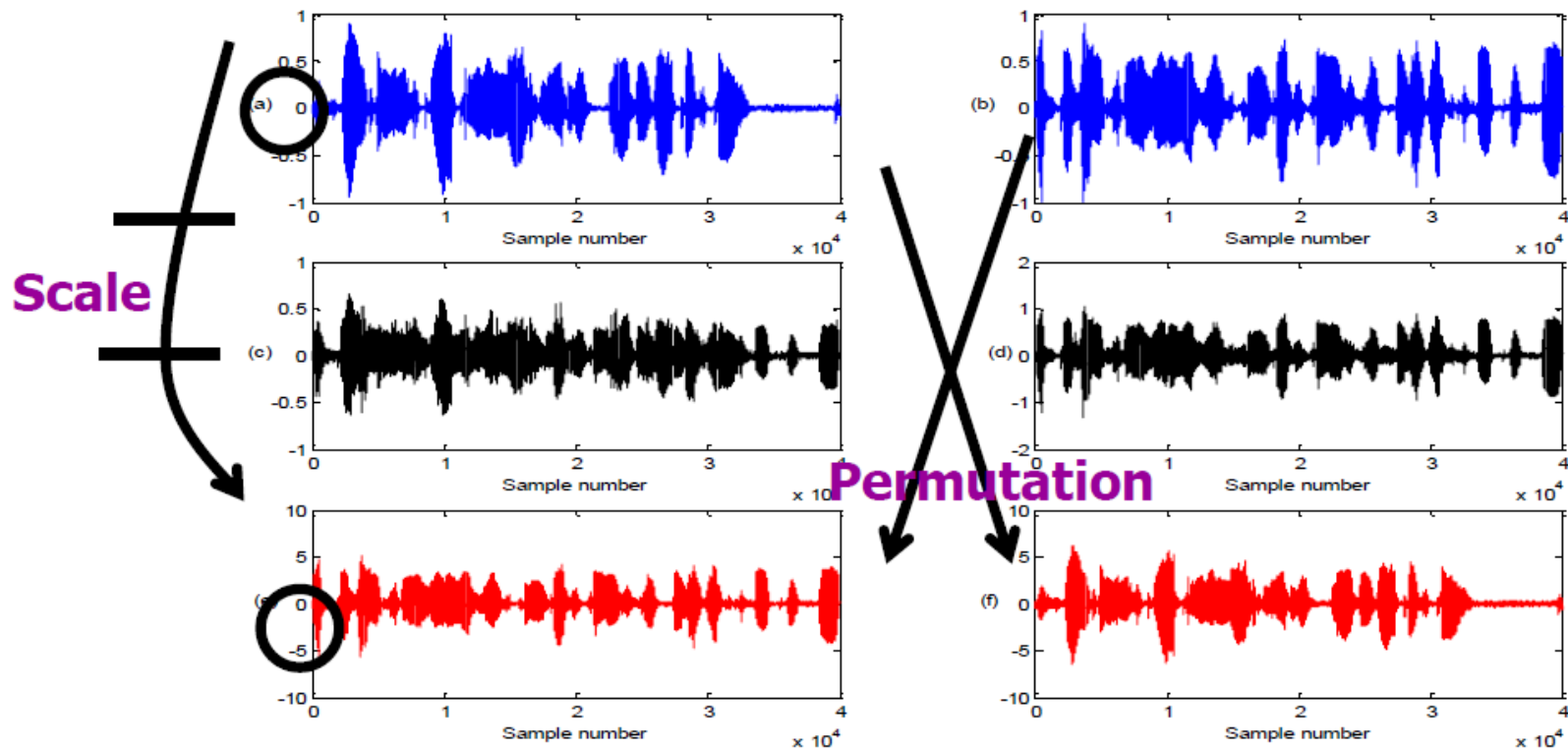
## Limitations of ICA

Permutation and scaling ambiguities



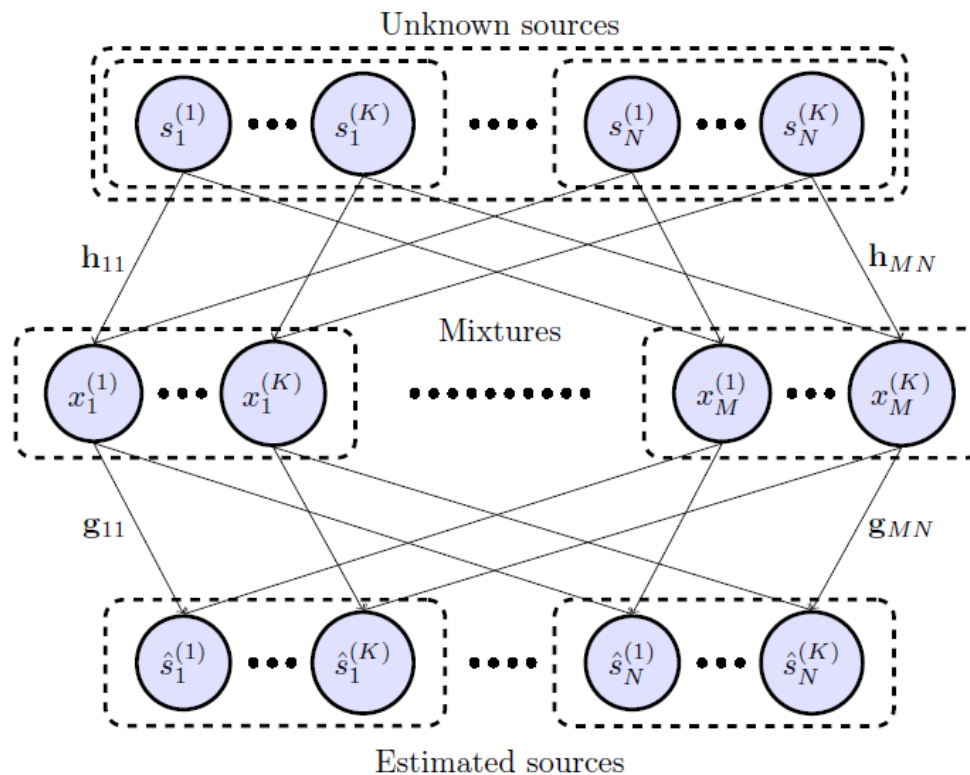
# Independent Component Analysis (ICA)

## Limitations of ICA



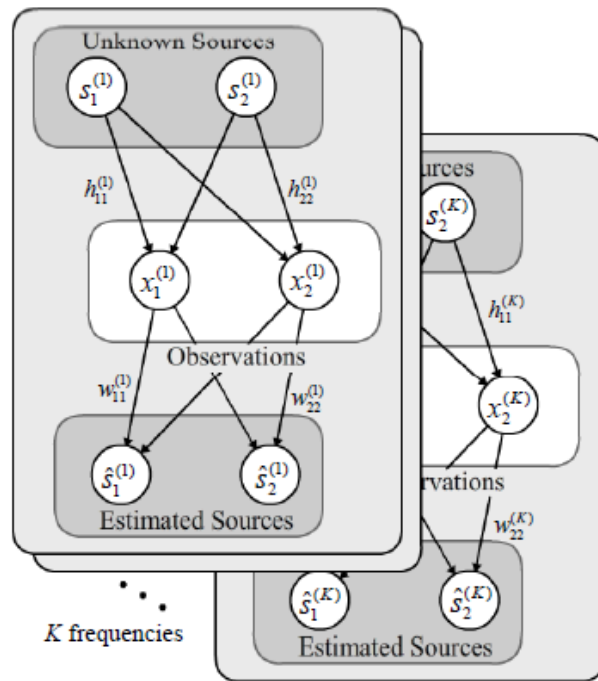
# Independent Vector Analysis (IVA)

IVA models the statistical independence between sources. And maintains the dependency between frequency bins of each source

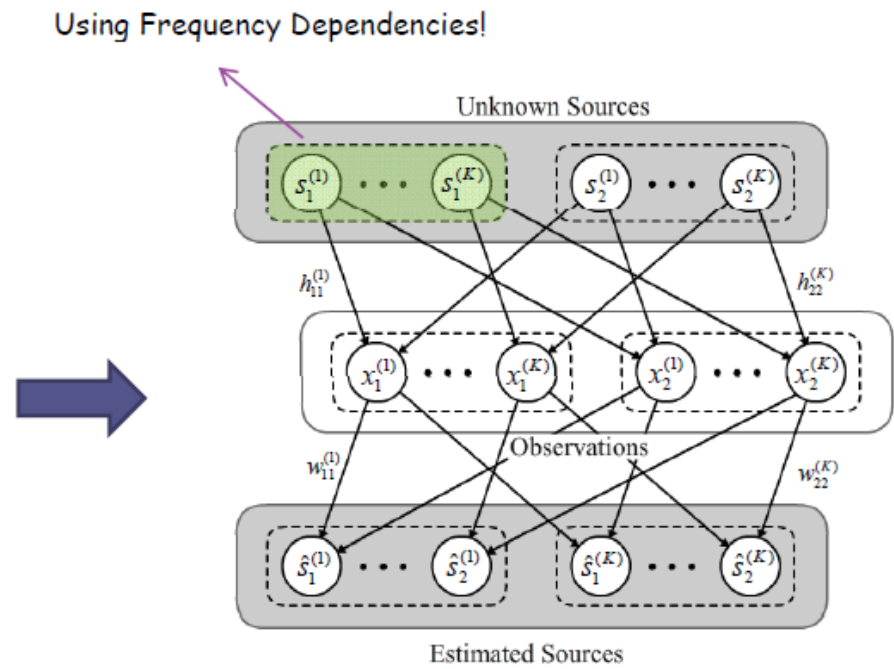


# Independent Vector Analysis (IVA)

## IVA vs ICA



Conventional ICA



Independent Vector Analysis

# Independent Vector Analysis (IVA)

- Modified Multivariate Cost Function

$$C = KL \left( p(\hat{s}_1, \hat{s}_2) \parallel \prod_{i=1}^2 p(\hat{s}_i) \right)$$

Components → Vectors!

$$\hat{\mathbf{s}}_i = (\hat{s}_i^{(1)}, \dots, \hat{s}_i^{(K)})^T$$

Conventional ICA Score Function

$$\begin{aligned} \varphi(\hat{s}_i^{(k)}) &= -\frac{\partial \log p(\hat{s}_i^{(k)})}{\partial \hat{s}_i^{(k)}} \\ &= \frac{\hat{s}_i^{(k)}}{|\hat{s}_i^{(k)}|} = \exp(j \cdot \arg(\hat{s}_i^{(k)})). \end{aligned}$$



- Using a Gradient Descent Algorithm:

$$w_{ij}^{(k) \text{ new}} = w_{ij}^{(k) \text{ old}} + \eta \Delta w_{ij}^{(k)},$$

$$\Delta w_{ij}^{(k)} = -\frac{\partial C}{\partial w_{ij}^{(k)}} = w_{ij}^{-T(k)} - E\{\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)})\} x_j^{\star(k)}$$

- Using Natural Gradient Learning:

$$\Delta w_{ij}^{(k)} = \sum_{l=1}^2 (I_{il} - E\{\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) \hat{s}_i^{\star(k)}\}) w_{ij}^{(k)}$$

IVA Score Function (Multivariate)

$$\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) = \frac{\partial \sqrt{\sum_{k=1}^K |\hat{s}_i^{(k)}|^2}}{\partial \hat{s}_i^{(k)}} = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^K |\hat{s}_i^{(k)}|^2}}$$

**Multivariate function exploiting frequency dependency!**

## Independent Vector Analysis (IVA)

The cost function is derived as

$$\begin{aligned}
 J_{IVA} &= \mathcal{KL}(p(\hat{\mathbf{s}}_1 \dots \hat{\mathbf{s}}_N) \parallel \prod_{i=1}^N q(\hat{\mathbf{s}}_i)) \\
 &= \int p(\hat{\mathbf{s}}_1 \dots \hat{\mathbf{s}}_N) \log \frac{p(\hat{\mathbf{s}}_1 \dots \hat{\mathbf{s}}_N)}{\prod_{i=1}^N q(\hat{\mathbf{s}}_i)} d\hat{\mathbf{s}}_1 \dots \hat{\mathbf{s}}_N \\
 &= \int p(\mathbf{x}_1 \dots \mathbf{x}_M) \log p(\mathbf{x}_1 \dots \mathbf{x}_M) d\mathbf{x}_1 \dots \mathbf{x}_M \\
 &\quad - \sum_{k=1}^K \log |\det G^{(k)}| - \sum_{i=1}^N \int p(\hat{\mathbf{s}}_i) \log q(\hat{\mathbf{s}}_i) d\hat{\mathbf{s}}_i \\
 &= \text{const.} - \sum_{k=1}^K \log |\det G^{(k)}| - \sum_{i=1}^N E[\log q(\hat{\mathbf{s}}_i)],
 \end{aligned}$$

## Independent Vector Analysis (IVA)

Partial derivative of the cost function is employed to find gradient

$$\Delta g_{ij}^{(k)} = -\frac{\partial J_{IVA}}{\partial g_{ij}^{(k)}} = g_{ij}^{(k)-H} - E[\varphi^{(k)}(\hat{s}_i^{(1)}, \dots, \hat{s}_i^{(K)})]x_j^{(k)*}$$

where  $\left[ (G^{(k)-1})^H \right]_{ii} \equiv g_{ij}^{(k)-H}$ .

The natural gradient update becomes

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^N (\delta_{il} - E[\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)})\hat{s}_i^{(k)*}])g_{lj}^{(k)},$$

where  $\delta_{il}$  is the Kronecker delta, i.e. when  $i = l$ ,

$\delta_{il} = 1$ , and zero otherwise.



# Independent Vector Analysis (IVA)

Multivariate score function

$$\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) = -\frac{\partial \log q(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)})}{\partial \hat{s}_i^{(k)}}$$

A source prior applied in the original IVA

$$q(\mathbf{s}_i) \propto \exp\left(-((\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i))^{\frac{1}{2}}\right),$$

For zero mean and unit variance data, the non-linear score function becomes

$$\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^K |\hat{s}_i^{(k)}|^2}}$$

and plays important role in maintaining the dependencies between the frequency bins of each source.

## Conclusions

- PCA can provide uncorrelated data and is limited to second order statistics.
- ICA is nongaussian alternative to PCA.
- ICA finds a linear decomposition by maximizing nongaussianity of the components.
- Application of ICA for convolutive source separation is constraint by the scaling and permutation ambiguities.
- IVA can mitigate the permutation problem during the learning process.
- Robustness of IVA is proportional to the strength of source priors.
- Above methods are applicable to determined and over-determined cases.

## References

- Aapo Hyvärinen, “Survey on Independent Component Analysis”, Neural Computing Surveys, Vol. 2, pp. 94-128, 1999.
- A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. Neural Computation, 9(7):1483-1492, 1997.
- J.-F. Cardoso and Antoine Souloumiac. “Blind beamforming for non Gaussian signals”, In IEE Proceedings-F, 140(6):362-370, December 1993
- A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, E. Moulines. “A blind source separation technique based on second order statistics”, IEEE Trans. on S.P., Vol. 45, no 2, pp 434-44, Feb. 1997.
- A. Mansour and M. Kawamoto, “ICA Papers Classified According to their Applications and Performances”, IEICE Trans. Fundamentals, Vol. E86-A, No. 3, March 2003, pp. 620-633.
- Buchner, H. Aichner, R. and Kellermann, W. (2004). Blind source separation for convolutive mixtures: A unified treatment. In Huang, Y. and Benesty, J., editors, Audio Signal Processing for Next-Generation Multimedia Communication Systems, pages 255–293. Kluwer Academic Publishers.
- Parra, L. and Spence, C. (2000). Convolutive blind separation of non-stationary sources. IEEE Trans. Speech Audio Processing, 8(3):320–327.

## References

- J. Bell, and T. J. Sejnowski, “An information-maximization to blind separation and blind deconvolution”, *Neural Comput.*, vol. 7, pp. 1129-1159, 1995
- S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing 8*, pages 757-763. MIT Press, Cambridge, MA, 1996.
- Te-Won Lee, *Independent component analysis: theory and applications*, Kluwer, 1998 .
- Araki, S. Makino, S. Blin, A. Mukai, and H. Sawada, (2004). Underdetermined blind separation for speech in real environments with sparseness and ICA. In *Proc. ICASSP 2004*, volume III, pages 881–884.
- M. I. Mandel, S. Bressler, B. Shinn-Cunningham, and D. P. W. Ellis, “Evaluating source separation algorithms with reverberant speech,” *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 7, pp. 1872–1883, 2010.
- Yi Hu; Loizou, P.C., "Evaluation of Objective Quality Measures for Speech Enhancement," *Audio, Speech, and Language Processing*, *IEEE Transactions on* , vol.16, no.1, pp.229,238, Jan. 2008.

Thank you!

*Acknowledgement: Various sources for visual material*

