Tutorial on Blind Source Separation and Independent Component Analysis

Lucas Parra

Adaptive Image & Signal Processing Group

Sarnoff Corporation

February 09, 2002

Linear Mixtures

... problem statement ...

$$\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{S}$$

Q: Given $\boldsymbol{X}$ can you tell what $\boldsymbol{A}$ and $\boldsymbol{S}$ is?

A: Yes! Use prior information on $\boldsymbol{A}$ and $\boldsymbol{S}$.

Mixing of Independent Sources

... basic physics often leads to linear mixing ...

Think of $\boldsymbol{S}$ as sources $s_i(t)$, $\boldsymbol{X}$ as sensor readings $x_j(t)$, and $\boldsymbol{A}$ as a physical mixing process with coefficients a_{ij}.

samples / sensors: $\boldsymbol{X}$ = Components / sensors: $\boldsymbol{A}$ * samples / components: $\boldsymbol{S}$

$$\boldsymbol{X} = \boldsymbol{A} * \boldsymbol{S}$$

Examples are:

Acoustic array	microphone	= room response	* sound amplitude
Spectroscopy	spectra	= concentration	* emission spectra
Hyperspectral	image	= abundance	* reflection spectra
EEG	elect. potential	= elect. coupling	* electrical potential
MEG	magn. field	= magn. coupling	* electrical current

In source separation prior knowledge is
statistical independence
of sources $\boldsymbol{S}$.

Separation Based on Independence

... non−Gaussianity, non−stationarity, non−whiteness ...

Statistical independence implies for all $i \neq j, t, \tau, n, m$:

$$E[s_i^n(t)\, s_j^m(t+\tau)] = E[s_i^n(t)]\, E[s_j^m(t+\tau)]$$

For M sources and N sensors each t, τ, n, m gives $M(M-1)/2$ conditions for the NM unknowns in $\boldsymbol{A}$.

Sufficient conditions if we use multiple:

use	sources assumed	resulting algorithm
n, m	non−Gaussian	ICA
t	non−stationary	multiple decorrelation
τ	non−white	multiple decorrelation

Multiple Decorrelation

... solution given by generalized eigenvalues ...

Second order independence implies diagonal cross−correlation for the sources $\Lambda_{\boldsymbol{s}}(\tau) = E[\boldsymbol{s}(t)\, \boldsymbol{s}^T(t+\tau)]$.

The measured cross−correlation $\boldsymbol{R}_{\boldsymbol{x}}(\tau) = E[\boldsymbol{x}(t)\, \boldsymbol{x}^T(t+\tau)]$ is then

$$\boldsymbol{R}_{\boldsymbol{x}}(\tau) = \boldsymbol{A}\, \Lambda_s(\tau)\, \boldsymbol{A}^{\boldsymbol{T}}$$

Combing these equations for two time delays $\tau = \tau_1, \tau_2$ leads to a generalized eigenvalue problem for $\boldsymbol{A}$,

$$\boldsymbol{R}_{\boldsymbol{x}}(\tau_1)\, \boldsymbol{R}_{\boldsymbol{x}}(\tau_2)^{-1} \boldsymbol{A} = \boldsymbol{A}\, \Lambda_s(\tau_1)\, \Lambda_s(\tau_2)^{-1}$$

Quickie BSS

... source separation in two lines ...

```
[W,D] = eig(X*X',R);   % compute unmixing matrix W
S = W'*X;              % compute sources S
```

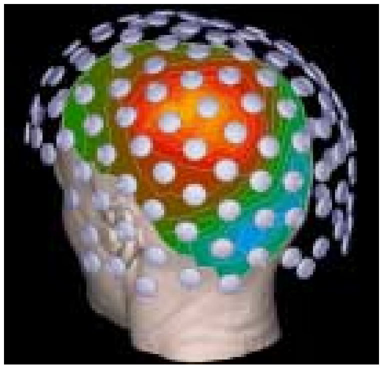
`X` is *NT* matrix containing *T* samples of *N* sensor readings presumably generated by `X=A*S`, with `A=inv(W')`.
`[V,D]=eig(A,B)` is the generalized eigenvalue procedure such that `A*V=B*V*D`, i.e. `V` jointly diagonalizes `A` and `B`.

<u>use</u>	<u>sources assumed</u>
`R` = Cross−correlation at some delay τ_2	non−white
`R` = Covariance at different time t_2	non−stationary
`R` = Cumulant of some higher order *m*	non−Gaussian

More robust if one diagonalizes more than two matrices. Details and references at `quickiebss.html`

Source Separation in MEG

... prior knowledge: sources decorrelated and non−white ...



$$X = A * S$$

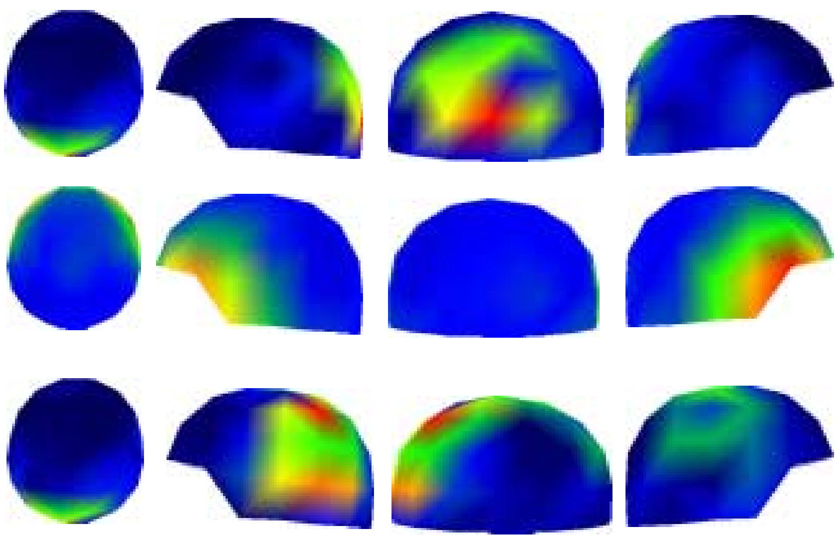
X	A	S
Magnetic fields measured in SQUID sensors	Magnetic coupling or attenuation due to geometry	Effective current flows in neuronal populations

i^{th} column in $\boldsymbol{A}$

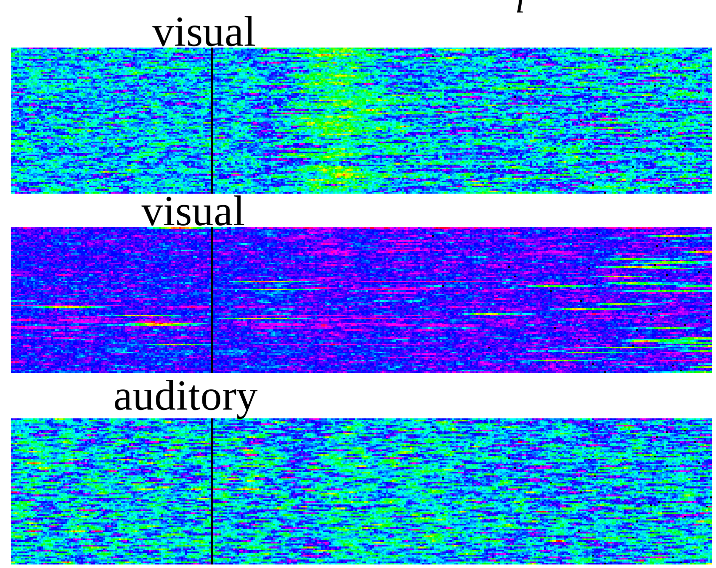
$i=6$

$i=20$

$i=50$



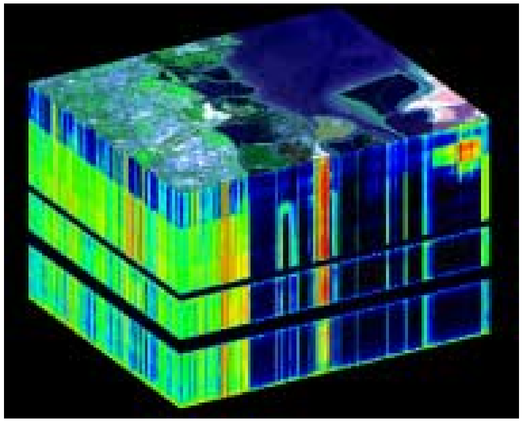
stimulus locked $s_i(t)$



Data and results provided by Akaysha Tang and Barak Pearlmutter from UNM. To appear in in *Neural Computations*, 2002

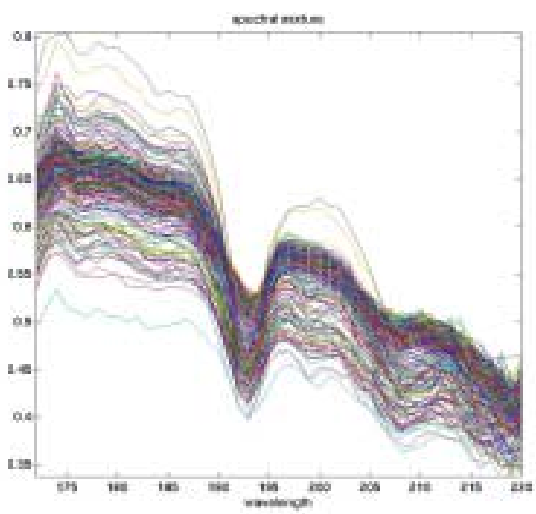
Source Separation in Hyperspectral Imaging

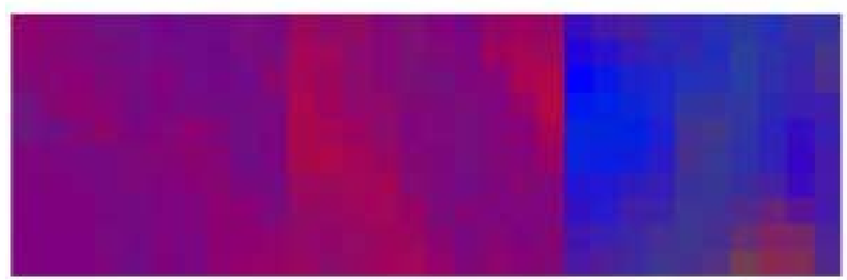
... prior knowledge: innovation process independent ...

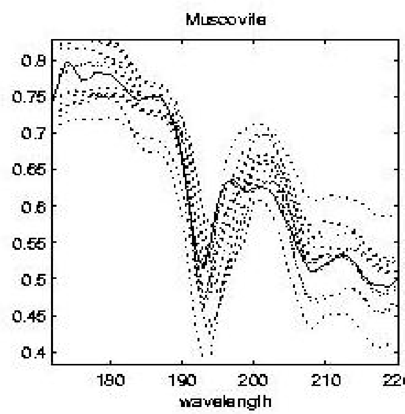


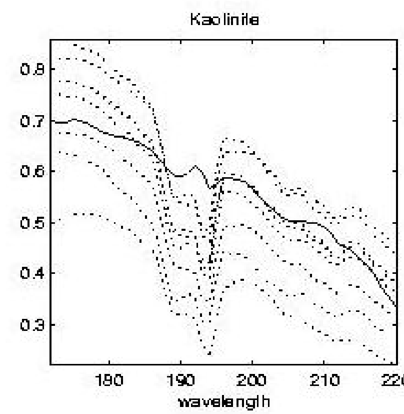
$$X = A * S$$

X	=	A	*	S
Images at multiple wavelengths		Material abundance in each pixel		Reflectance of each material for different wavelengths



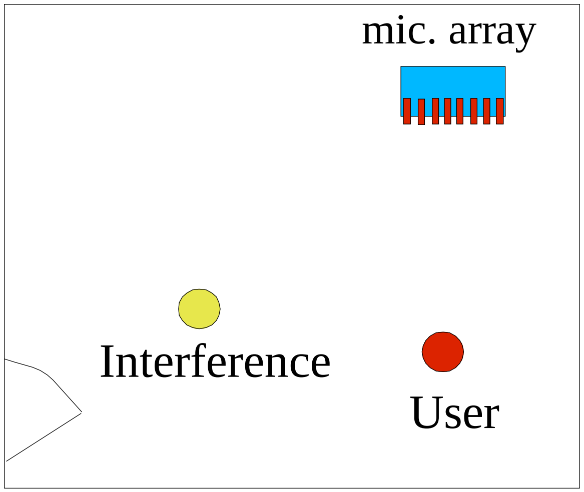






Source Separation in Acoustics

... prior knowledge: source decorrelated and non−stationary...

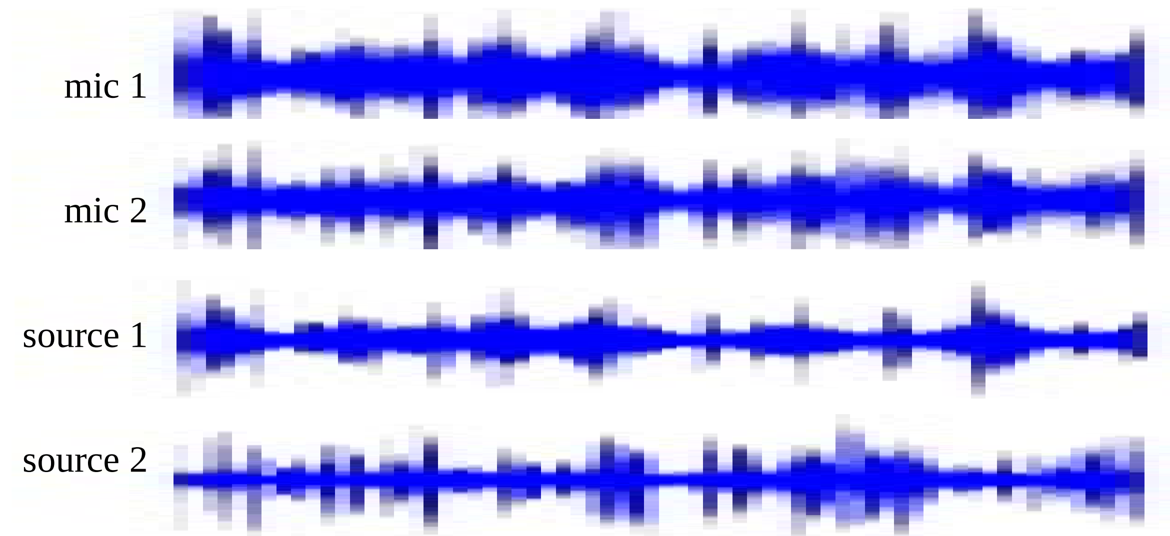


$\boldsymbol{X}(\omega) \quad = \quad \boldsymbol{A}(\omega) \quad * \quad \boldsymbol{S}(\omega)$

Microphone recordings for each frequency bin ω

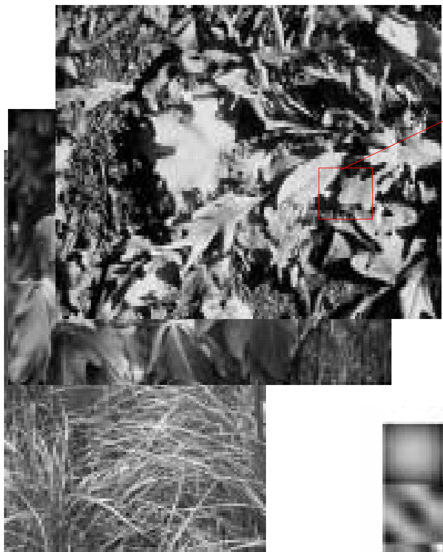
Convolutive room response now a matrix of FIR filters

Acoustic sources for each frequency



Independent Linear Basis of Images

... useful for denoising and compression ...

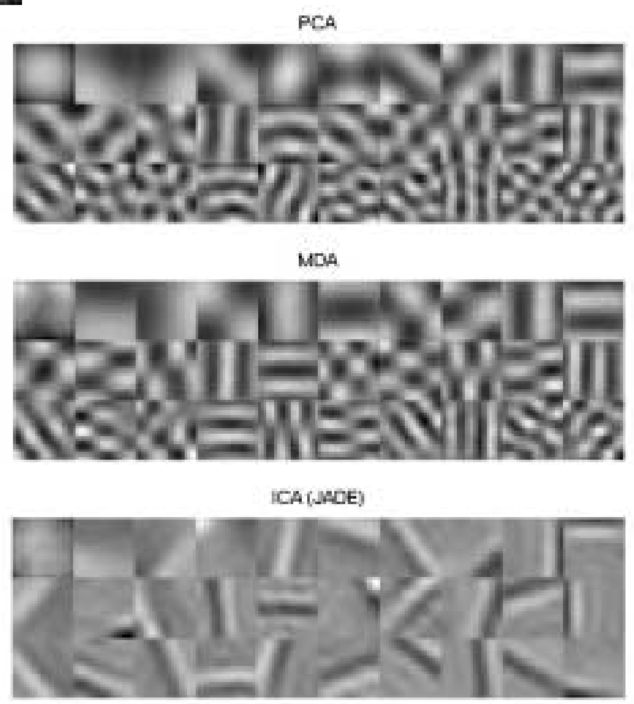


$$\boldsymbol{X} \quad = \quad \boldsymbol{A} \quad * \quad \boldsymbol{S}$$

Image intensities for image patches

Inverse of linear transform, similar to "mother" function in wavelets

Independent linear decomposition coefficients



Resulting bases produces components that are:

- sparse −useful for denoising.
- Non−redundant − useful for encoding.

Similarity to receptive fields supports Barlow’s minimum redundancy argument for visual processing.

Independent Linear Basis of Speech

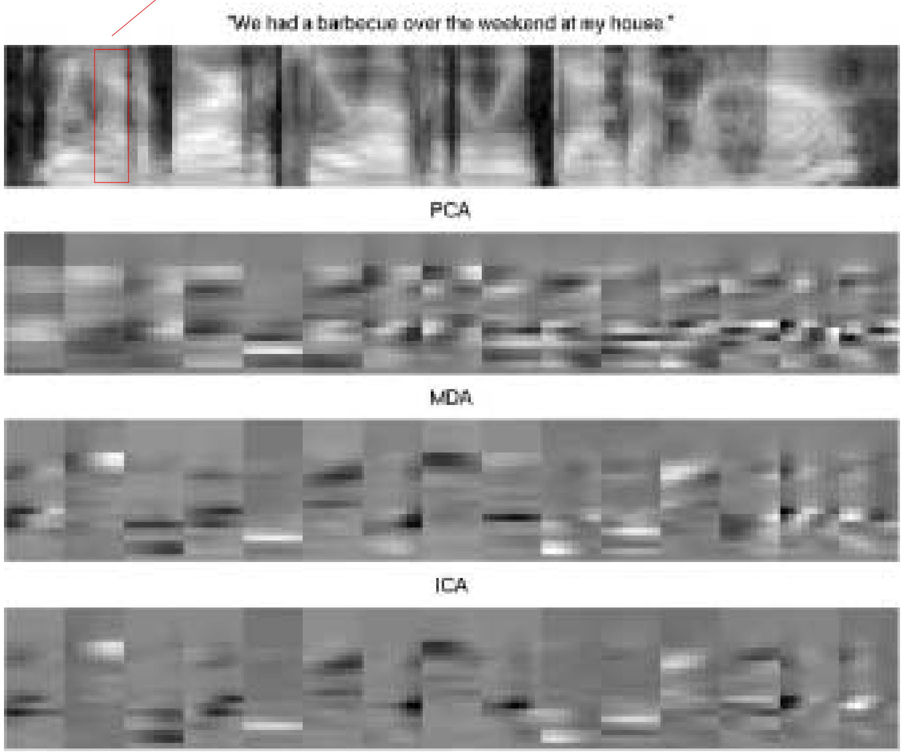
... Speech is spanned by non−stationary independent features ...

$$X \quad = \quad A \quad * \quad S$$

Recorded power in spetro−temporal window

Spectro−temporal basis set or "features" that constitute speech

Independent contributions of speech "features".



Finding: Non−stationary assumption (MDA) and higher order independence (ICA) give the same components.

Conclusion: Speech can be understood as a linear superposition of non−stationary independent components. Linearity is justified by acoustics. However, non−stationary, independent "features" are a property of speech.

ICA as Density Estimation

... Maximum Likelihood ...

Statistical independence implies that:

$$p(\boldsymbol{s}) = p(s_1)\, p(s_2) \ldots p(s_M)$$

Likelihood of observations as function of model, ***s* = *W x***:

$$p_x(\boldsymbol{x}|\boldsymbol{W}) = \left|\frac{\partial \boldsymbol{s}}{\partial \boldsymbol{x}}\right| p_s(\boldsymbol{s}) = |\boldsymbol{W}|\, p_s(\boldsymbol{W}\,\boldsymbol{x}) = |\boldsymbol{W}| \prod_{i=1}^{M} p_s(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\, \boldsymbol{x})$$

Minimize log−likelihood with stochastic gradient ascent gives for the k^{th} i.i.d. sample $\boldsymbol{x}(k)$:

$$\Delta \boldsymbol{W} = \boldsymbol{W}^{-T} + \boldsymbol{u}(k)\, \boldsymbol{x}^T(k) \qquad\qquad \boldsymbol{u}(k) = \nabla_{\boldsymbol{s}} \ln p_s(s(k))$$

With positive definite projection $\boldsymbol{W}^T\boldsymbol{W}$ obtain popular "natural gradient" ICA algorithm:

$$\Delta \boldsymbol{W} = [\boldsymbol{I} + \boldsymbol{u}(k)\, \boldsymbol{s}^T(k)]\, \boldsymbol{W}$$

ICA and Information Theory

... ICA = PCA for Gaussian sources and orthogonal transform...

ICA becomes Principal Component Analysis (PCA) if:

(1) Sources are Gaussian: $p(\boldsymbol{s}) \propto \exp(-\boldsymbol{s}^T \Lambda^{-1} \boldsymbol{s})$

(2) Transformation orthogonal: $\boldsymbol{W}^{-1} = \boldsymbol{W}^T$

With (1) we have, $\boldsymbol{u}(k) = \nabla_{\boldsymbol{s}} \ln p_s(s(k)) = -\Lambda^{-1} \boldsymbol{s}(k)$

Insert into log likelihood gradient, $\boldsymbol{W}^{-T} + \boldsymbol{u}(k) \boldsymbol{x}^T(k)$, sum over all samples and set equal zero:

$$\boldsymbol{W}^T = \frac{\Lambda^{-1}}{K} \sum_{k=1}^{K} \boldsymbol{s}(k) \boldsymbol{x}^T(k) = \frac{\Lambda^{-1}}{K} \sum_{k=1}^{K} \boldsymbol{W}\, \boldsymbol{x}(k) \boldsymbol{x}(k) = \Lambda^{-1} \boldsymbol{W}\, \boldsymbol{R}_{\boldsymbol{x}}$$

Using (2) we obtain, i.e. PCA: $\boldsymbol{R}_{\boldsymbol{x}} = \boldsymbol{W}\, \Lambda\, \boldsymbol{W}^T$

ICA and Information Theory

... Independence and Minimal Mutual Information ...

Mutual Information is defined as the KLD between the joint and the product of the individual variables:

$$KLD[\,p(\boldsymbol{s}),\prod_i p(s_i)] = \int d\,\boldsymbol{s}\; p(\boldsymbol{s}) \ln\left(\frac{p(\boldsymbol{s})}{\prod_i p(s_i)}\right)$$

$$= MI[\,p(\boldsymbol{s})] = \sum_i H[\,p(s_i)] - H[\,p(\boldsymbol{s})]$$

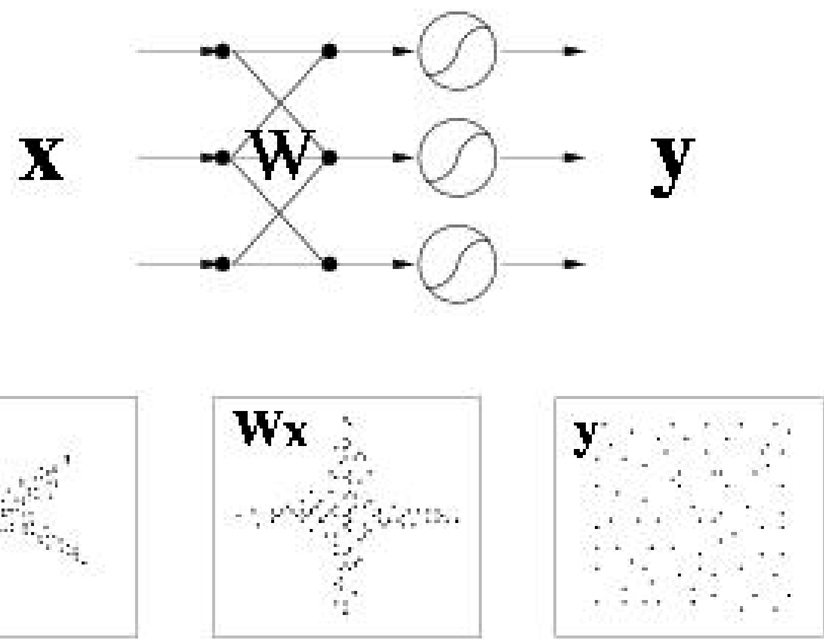
Hence, minimizing the Mutual Information is equivalent with fitting parameters to make distribution independent, e.g ICA for linear transform $\boldsymbol{s} = \boldsymbol{W}\,\boldsymbol{x}$.

Also, if we keep, |W|=1, we find that, $H[p(\boldsymbol{s})] = H[p(\boldsymbol{x})] = const$. And we get ICA by *minimizing* entropy of individual variables.

ICA and Information Theory

... Independence and Maximum Entropy ...

Interestingly, also *maximizing* entropy, after non−linear transform give independent components.



Conclusion

... ...

Source separation gives physically meaningful sources whenever the physical mixing process is linear and there exist independent sources, such as in acoustics, encephalography, and spectroscopy.

ICA may also be useful as linear basis set for density modeling, compression, de−noising even if there are no such independent sources.

Independent basis sets can explain some statistical properties of natural signals and suggest optimal processing.

Blind source separation can exploit non−Gaussianity, non−stationarity, or non−whiteness of independent signals.

Cross−moment methods are often more robust. However, density fitting approaches result sometimes in simpler algorithms.