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 ...

\[ X = A \ S \]

Q: Given \( X \) can you tell what \( A \) and \( S \) is?

A: Yes! Use prior information on \( A \) and \( S \).
Mixing of Independent Sources

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

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

$$X = A \ast S$$

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 $S$. 
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 \( t, \tau, n, m \) gives \( M(M-1)/2 \) conditions for the \( NM \) unknowns in \( A \).

Sufficient conditions if we use multiple:

<table>
<thead>
<tr>
<th>Use</th>
<th>Sources assumed</th>
<th>Resulting algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n, m )</td>
<td>non–Gaussian</td>
<td>ICA</td>
</tr>
<tr>
<td>( t )</td>
<td>non–stationary</td>
<td>multiple decorrelation</td>
</tr>
<tr>
<td>( \tau )</td>
<td>non–white</td>
<td>multiple decorrelation</td>
</tr>
</tbody>
</table>
Second order independence implies diagonal cross–correlation for the sources $\Lambda_s(\tau) = E[s(t) s^T(t+\tau)]$.

The measured cross–correlation $R_x(\tau) = E[x(t) x^T(t+\tau)]$ is then

$$R_x(\tau) = A \Lambda_s(\tau) A^T$$

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

$$R_x(\tau_1) R_x(\tau_2)^{-1} A = A \Lambda_s(\tau_1) \Lambda_s(\tau_2)^{-1}$$
Quickie BSS
... source separation in two lines ... 

\[ [W, D] = \text{eig}(X'X', R); \quad \% \text{compute unmixing matrix } W \]
\[ S = W'X; \quad \% \text{compute sources } S \]

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

Use sources assumed non-white
\( R = \text{Cross-correlation at some delay } \tau_2 \)
\( R = \text{Covariance at different time } t_2 \)
\( R = \text{Cumulant of some higher order } m \)

More robust if one diagonalizes more than two matrices. Details and references at \texttt{quickiebss.html}
Source Separation in MEG

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

\[ X = A * S \]

- **X**: Magnetic fields measured in SQID sensors
- **A**: Magnetic coupling or attenuation due to geometry
- **S**: Effective current flows in neuronal populations

*i*th column in **A**

- \( i=6 \)
- \( i=20 \)
- \( i=50 \)

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

\[ X = A \ast S \]

- Images at multiple wavelengths
- Material abundance in each pixel
- Reflectance of each material for different wavelengths

abundance of Kaolinite (red), Muscovite (blue)
Source Separation in Acoustics

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

\[ X(\omega) = A(\omega) \ast S(\omega) \]

Microphone recordings for each frequency bin \( \omega \)

Convolutive room response now a matrix of FIR filters

Acoustic sources for each frequency

*Figure: Diagram showing microphone array, interference, and user with waveforms for microphone 1, microphone 2, source 1, and source 2.*
Independent Linear Basis of Images

... useful for denoising and compression ...

\[ X = A \star S \]

- 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* ...

\[
X = A * S
\]

Recorded power in spectro–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:

\[ p(s) = p(s_1) p(s_2) ... p(s_M) \]

Likelihood of observations as function of model, \( s = W x \):

\[ p_x(x|W) = \left| \frac{\partial s}{\partial x} \right| p_s(s) = |W| p_s(W x) = |W| \prod_{i=1}^{M} p_s(w_i^T x) \]

Minimize log–likelihood with stochastic gradient ascent gives for the \( k \)th i.i.d. sample \( x(k) \):

\[ \Delta W = W^{-T} + u(k) x^T(k) \quad \quad u(k) = \nabla_s \ln p_s(s(k)) \]

With positive definite projection \( W^T W \) obtain popular "natural gradient" ICA algorithm:

\[ \Delta W = [I + u(k) s^T(k)] W \]
ICA and Information Theory

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

ICA becomes Principal Component Analysis (PCA) if:

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

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

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

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

\[
W^T = \frac{\Lambda^{-1}}{K} \sum_{k=1}^{K} s(k) x^T(k) = \frac{\Lambda^{-1}}{K} \sum_{k=1}^{K} W x(k) x(k) = \Lambda^{-1} W R_x
\]

Using (2) we obtain, i.e. PCA: \( R_x = W \Lambda W^T \)
Mutual Information is defined as the KLD between the joint and the product of the individual variables:

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

\[
= MI[p(s)] = \sum_i H[p(s_i)] - H[p(s)]
\]

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

Also, if we keep, \(|W|=1\), we find that, \( H[p(s)] = H[p(x)] = \text{const.} \) 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.
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.