Independent component analysis applied to biophysical time series and EEG
Example: Speech Separation
Cortex

Crane

Synchronie locale

Peau

Relative Independence

Synchronie locale

Electrodes

Cortex
Infomax ICA

\[ Y = [A; B] \]

Linear Combination
\[ X = YW \]

ICA
\[ \tilde{Y} = W^{-1}\tilde{X} \]
Independent component analysis

Mixture of Brain source activity

Cocktail Party
ICA is a method to recover a version, of the original sources by multiplying the data by a unmixing matrix.
ICA activity $U = WX$

**Data X**

\[
\begin{bmatrix}
3 & 2 & 5 & 4 & 3 & 2 \\
0 & -2 & -5 & -1 & 1 & -1 & \ldots \\
-1 & 2 & 0 & 1 & 0 & -3
\end{bmatrix}
\]

**Channel 1**

**Channel 2**

**Channel 3**

**Weight matrix W**

\[
\begin{bmatrix}
5 & 3 & -2 \\
1 & 2 & 4 \\
0 & -1 & 3
\end{bmatrix}
\]

**ICA activity U**

\[
\begin{bmatrix}
3\ast5 + 0\ast3 - 1\ast(-2) & 2\ast5 + (-2)\ast3 + 2\ast(-2) \\
3\ast1 + 0\ast2 - 1\ast4 & 2\ast1 + (-2)\ast2 + 2\ast4 & \ldots \\
5\ast1 - 5\ast2 + 0\ast4 & 5\ast1 - 5\ast2 + 0\ast4
\end{bmatrix}
\]

**Comp. 1**

**Comp. 2**

**Comp. 3**
Data $\mathbf{X} = \mathbf{W}^{-1} \mathbf{U}$

ICA activity $\mathbf{U}$

Inverse weight matrix $\mathbf{W}^{-1}$

Data $\mathbf{X}$

$$\begin{bmatrix}
3 & 2 & 5 & 4 & 3 & 2 \\
0 & -2 & -5 & -1 & 1 & -1 \\
-1 & 2 & 0 & 1 & 0 & -3
\end{bmatrix} \quad \leftarrow \text{Comp. 1}
$$

$$\begin{bmatrix}
3 \cdot 5 + 0 \cdot 3 - 1 \cdot (-2) & 2 \cdot 5 + (-2) \cdot 3 + 2 \cdot (-2) \\
3 \cdot 1 + 0 \cdot 2 - 1 \cdot 4 & 2 \cdot 1 + (-2) \cdot 2 + 2 \cdot 4 \\
5 \cdot 1 - 5 \cdot 2 + 0 \cdot 4 & 5 \cdot 1 - 5 \cdot 2 + 0 \cdot 4
\end{bmatrix} \quad \leftarrow \text{Chan 1}
$$

$$\begin{bmatrix}
1 & 2 & 4 \\
0 & -1 & 3
\end{bmatrix}$$

$$\begin{bmatrix}
3 \cdot 5 + 0 \cdot 3 - 1 \cdot (-2) & 2 \cdot 5 + (-2) \cdot 3 + 2 \cdot (-2) \\
3 \cdot 1 + 0 \cdot 2 - 1 \cdot 4 & 2 \cdot 1 + (-2) \cdot 2 + 2 \cdot 4 \\
5 \cdot 1 - 5 \cdot 2 + 0 \cdot 4 & 5 \cdot 1 - 5 \cdot 2 + 0 \cdot 4
\end{bmatrix} \quad \leftarrow \text{Chan 2}
$$

$$\begin{bmatrix}
1 & 2 & 4 \\
0 & -1 & 3
\end{bmatrix}$$

$$\begin{bmatrix}
3 \cdot 5 + 0 \cdot 3 - 1 \cdot (-2) & 2 \cdot 5 + (-2) \cdot 3 + 2 \cdot (-2) \\
3 \cdot 1 + 0 \cdot 2 - 1 \cdot 4 & 2 \cdot 1 + (-2) \cdot 2 + 2 \cdot 4 \\
5 \cdot 1 - 5 \cdot 2 + 0 \cdot 4 & 5 \cdot 1 - 5 \cdot 2 + 0 \cdot 4
\end{bmatrix} \quad \leftarrow \text{Chan 3}
$$
Historical Remarks

- Herault & Jutten ("Space or time adaptive signal processing by neural network models“, Neural Nets for Computing Meeting, Snowbird, Utah, 1986): Seminal paper, neural network
- Bell & Sejnowski (1995): Information Maximization
- Amari et al. (1996): Natural Gradient Learning
- Cardoso (1996): JADE

- Applications of ICA to biomedical signals
  - EEG/ERP analysis (Makeig, Bell, Jung & Sejnowski, 1996).
  - fMRI analysis (McKeown et al. 1998)
ICA Theory – Cost Functions

Family of BSS algorithms
- Information theory (Infomax)
- Bayesian probability theory (Maximum likelihood estimation)
- Negentropy maximization
- Nonlinear PCA
- Statistical signal processing (cumulant maximization, JADE)

A unifying Information-theoretic framework for ICA
- Pearlmutter & Parra showed that InfoMax, ML estimation are equivalent.
- Lee et al. (1999) showed negentropy has the equivalent property to InfoMax.
- Girolami & Fyfe showed nonlinear PCA can be viewed from information-theoretic principle.
ICA is a method to recover a version, of the original sources by multiplying the data by a unmixing matrix, $U = WX$, While PCA simply decorrelates the outputs (using an orthogonal matrix $W$), ICA attempts to make the outputs statistically independent, while placing no constraints on the matrix $W$. 
ICA and PCA

Principal component analysis

Independent component analysis
Central limit theorem

Scalp channels = linear mixture of A and B
(more gaussian)

Brain source A

Brain source B

Scalp channel 1

Scalp channel 2
ICA Training Process

Central limit theorem

• Remove the mean
  \[ x = x - \langle x \rangle \]

• ‘Sphere’ the data by diagonalizing its covariance matrix,
  \[ x = \langle xx^T \rangle^{-1/2} (x - \langle x \rangle). \]

• Update \( W \) according to
  \[ \Delta W \propto \frac{\partial H(y)}{\partial W} W^T W. \]
Entropy

\[ H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_b p(x). \]

**Dice: 1/6**

\[ H = 6 \left( -\frac{1}{6} \log_2 \left( \frac{1}{6} \right) \right) = 2.58 \]

**Fake dice (make a 6 half of the time): entropy 2.16 (base 2)**

\[ H = 5 \left( -\frac{1}{10} \log_2 \left( \frac{1}{10} \right) \right) - \frac{1}{2} \log_2 \left( \frac{1}{2} \right) = 2.16 \]
Entropy

\[ H(X) = - \sum_{x \in X} p(x) \log_b p(x). \]

Joint entropy

\[ H(X, Y) = - \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log_b p(x, y). \]

Mutual Information

\[ H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2). \]

Shannon in his landmark 1948 paper "A Mathematical Theory of Communication."

From http://planetmath.org/encyclopedia/ShannonsTheoremEntropy.html
## Contingency table for stress and emotionality

<table>
<thead>
<tr>
<th>EMOT=</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>63</td>
<td>64</td>
<td>3</td>
<td>1</td>
<td></td>
<td>142</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>87</td>
<td>84</td>
<td>35</td>
<td>13</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

From http://tecfa.unige.ch/~lemay/thesis/THX-Doctorat/node149.html
## Contingency frequencies for stress and emotionality

<table>
<thead>
<tr>
<th></th>
<th>STRE</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>EMOT=1</td>
<td>0.07</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.24</td>
<td>0.25</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td></td>
</tr>
</tbody>
</table>

Joint entropy 3.46; exercise: compute mutual information

\[
H(X, Y) = - \sum_{(x,y) \in X \times Y} p(x, y) \log_b p(x, y).
\]
ICA learning rule

How to make the outputs statistical independent?

Minimize their redundancy or mutual information.

Consider the joint entropy of two components,

$$H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2).$$

Maximizing $H(y_1, y_2) \implies$ minimizing $I(y_1, y_2)$.

The learning rule:

$$\Delta W \propto \frac{\partial H(y)}{\partial W} W^T W$$

=0 if the two variables are independent

Entropy extremum

Natural gradient (Amari)
Kurtosis, Super- and Sub-Gaussian

Kurtosis: a measure of how peaked or flat of a probability distribution is.

\[ \text{kurt}(X) = \frac{E[(X - \mu)^4]}{\sigma^4} \]

Gaussian Dist. Kurtosis = 0
Super-Gaussian: kurtosis > 0
Sub-Gaussian: kurtosis < 0
Moments

Moments

\[ \mu_x(n) = E\{x^n\} \]

Central moments

\[ m_x(n) = E\{(x - m_x)^n\} \]

Cumulants

\[
\begin{align*}
c_1 &= m_1 = \mu & \text{mean} \\
c_2 &= m_2 = \sigma^2 & \text{variance} \\
c_3 &= m_3 & \text{skewness} \\
c_4 &= m_4 - 3m_2^2 & \text{kurtosis}
\end{align*}
\]
Sub-gaussian

Super-gaussian

Sphering

ICA
InfoMax (Bell & Sejnowski, 1995)

To make the \( u_i \) independent, we need to operate on non-linear transformed output variables, \( y = g(u) \), such as

\[
y = \frac{1}{1 + e^{-u}}, \quad u = Wx + w_0
\]

The non-linear function provides all the higher-order statistics necessary to establish independence.
Independent components of EEG/ERP
ICA/EEG Assumptions

- Mixing is linear at electrodes
- Propagation delays are negligible
- Component time courses are independent
- Number of components less than the number of channels.
Independent Component Categories

- Artifacts
- Stimulus-locked activity
- Response-locked activity
- Non-phase locked activity
- Event-modulated oscillatory activity
Characteristics of Independent Component of the EEG

- Concurrent Activity
- Maximally Temporally Independent
- Overlapping Maps and Spectra
- Dipolar Scalp Maps
- Functionally Independent
- Between-Subject Regularity
Largest 30 Independent Components (single subject)
ICA Decomposition into Independent Components

EEG Scalp Data

Independent Components

unmixing (W)

activations (u=WX)

scalp maps (W⁻¹)
Selective Projection onto Scalp Channels

IC1

IC2

IC4

Artificial-corrected EEG

mixing $W^{-1}$

$1\text{ sec}$

$x_0 = W^{-1}u_0$

VEOG

F3

Cz

Pz
\[ X = W^{-1} U \]

**Data**

ICA activity \( U \)

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & \ldots \\
0 & -2 & -5 & -1 & 1 & -1 & \ldots \\
0 & 0 & 0 & 0 & 0 & 0 & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\end{bmatrix}
\]

\( \rightarrow \) Comp. 1

\( \rightarrow \) Comp. 2

\( \rightarrow \) Comp. 3

**Inverse weight matrix \( W^{-1} \)**

Data \( X \)

**Chan 1**

\[
\begin{bmatrix}
3\times 5 + 0\times 3 - 1\times (-2) & 2\times 5 + (-2)\times 3 + 2\times (-2) & \ldots \\
3\times 1 + 0\times 2 - 1\times 4 & 2\times 1 + (-2)\times 2 + 2\times 4 & \ldots \\
5\times 1 - 5\times 2 + 0\times 4 & 5\times 1 - 5\times 2 + 0\times 4 & \ldots \\
\ldots & \ldots & \ldots \\
\end{bmatrix}
\]

\( \rightarrow \) Chan 1

\( \rightarrow \) Chan 2

\( \rightarrow \) Chan 3

**Inverse weight matrix \( W^{-1} \)**
ICA-based Artifact Removal
Artifact removal using ICA
Two Neck Muscle Processes
Some Independent EMG Components
Sample EEG Decomposition
Two Lateral Alpha Processes
Two Central Alpha Processes
Localization

ICA component scalp maps

Time (ms)

Electrodes

Components

Localization
Frontal midline
Localization of activity
Subject MRI

MNI model (Loreta)

Subject scanned electrode positions

Subject components scalp map

Normalization of subject MRI to MNI brain

EEGLAB

SPM2

EEGLAB

LORETA

EEGLAB

EEGLAB & FILEDTRIP

BRAINVISA

Coregistration

Head mesh
Inverse weight matrix $W^{-1}$

Data $X$ (EEG/MEG time series)

Temporal ICA

ICA activity $U$

$X = W^{-1}U$

Comp. 1
Comp. 2
Comp. 3

Chan 1
Chan 2
Chan 3

Inverse weight matrix $W^{-1}$

Data $X$ (EEG/MEG voxel activities)

Spatial ICA

ICA activity $U$

$X = W^{-1}U$

Comp. 1
Comp. 2
Comp. 3

Time 1
Time 2
Time 3
ICA Applied to fMRI Data

- Task-related
- Arousal
- Measured Signal
- Physiologic Pulsations
- Machine Noise

McKeown et al., Human Brain Map., 1998
Independent fMRI Components

- Consistently task-related
- Transiently task-related
- Abrupt head movement
- Quasi-periodic
- Slowly-varying
- Slow head movement

Activated
Suppressed
Rejection: Raw Data vs. ICA

Artificial artifacts

1 - Transient high-frequency events (20-60 Hz) modeling temporal muscle artifacts

2 - Low-frequency events (1-3 Hz) modeling eye movement artifacts

3 - Signal discontinuities from electrical artifacts

4 - High noise EEG artifacts

5 - Linear trends from electrical artifacts

EEG activity (31 channels x 100 trials of artifact-free EEG)
Rejection: Raw Data vs. ICA

1 - Transient high frequency event (muscle)

2 - Low frequency event (eye movements)

3 - Discontinuity

4 - High noise

5 - Linear trend
Rejection methods

- Detection of peaks of activity (thresholding)
- Detection of linear trends ($R^2$)
- Detection of improbable events (Joint Probability)
- Detection of peaky distributions of activity (Kurtosis)
- Detection of frequency peaks (frequency thresholding)

Known artifacts

\{ Vary artifacts' amplitude \}

\begin{align*}
\text{Optimize the free parameter for each method} \\
\text{Performance of each method}
\end{align*}
Rejection: Raw Data vs. ICA

Mixture

Original data

Infomax ICA