Independent Component Analysis

Group6 : Armandine, Samuael, Verlon

African Master's in Machine Intelligence (AMMI)

Supervised by Pr. MOUSTAPHA CISSE

April 12, 2023

Overview

AIMS African Institute for Mathematical Sciences SENEGAL

Overview

- Two people speaking simultaneously in a room.
- Speeches are recorded by two microphones in separate locations.

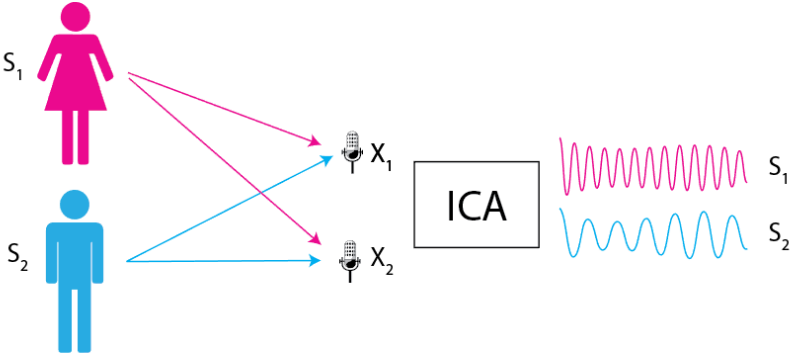

Image ref

Overview

Problem formalization

- d-**speakers** are speaking simultaneously at a party.

Problem formalization

- d-**speakers** are speaking simultaneously at a party.
- d differents **mircophones** place in the same room records an overlapping combination of the d speakers.

Problem formalization

- *d*-**speakers** are speaking simultaneously at a party.
- *d* differents **mircophones** place in the same room records an overlapping combination of the d speakers.
- Using *d*-**microphones**, can you separate out the *d*- **speakers** speech signals?

Problem formalization

- *d*-**speakers** are speaking simultaneously at a party.
- *d* differents **mircophones** place in the same room records an overlapping combination of the d speakers.
- Using *d*-**microphones**, can you separate out the *d*- **speakers** speech signals?
- To solve this problem, we use Independent Component Analysis (ICA) which is a statistical method for separating a multivariate signal into independent components.

Problem formalization

Let us consider some data $S^{(i)} \in \mathbb{R}^d$ that is generated via d independent sources at time i as :

Problem formalization

Let us consider some data $S^{(i)} \in \mathbb{R}^d$ that is generated via d independent sources at time i as :

$$X^{(i)} = AS^{(i)} \quad , X^{(i)} \in \mathbb{R}^n \tag{1.1}$$

Where :

- A is an unknown square matrix called **mixing matrix**

Problem formalization

Let us consider some data $S^{(i)} \in \mathbb{R}^d$ that is generated via d independent sources at time i as :

$$X^{(i)} = AS^{(i)} \quad , X^{(i)} \in \mathbb{R}^n \tag{1.1}$$

Where :

- A is an unknown square matrix called **mixing matrix**
- $X^{(i)}$ is the signal of each microphone at time i

Problem formalization

Let us consider some data $S^{(i)} \in \mathbb{R}^d$ that is generated via d independent sources at time i as :

$$X^{(i)} = AS^{(i)} \quad , X^{(i)} \in \mathbb{R}^n \tag{1.1}$$

Where :

- A is an unknown square matrix called **mixing matrix**
- $X^{(i)}$ is the signal of each microphone at time i
- $S^{(i)}$ is the signal of each speaker records at time i

Problem formalization

Let us consider some data $S^{(i)} \in \mathbb{R}^d$ that is generated via d independent sources at time i as :

$$X^{(i)} = AS^{(i)} \quad , X^{(i)} \in \mathbb{R}^n \tag{1.1}$$

Where :

- A is an unknown square matrix called **mixing matrix**
- $X^{(i)}$ is the signal of each microphone at time i
- $S^{(i)}$ is the signal of each speaker records at time i
- $S_j^{(i)}$ is the signal from speaker j at time i

Problem formalization

Let us consider some data $S^{(i)} \in \mathbb{R}^d$ that is generated via d independent sources at time i as :

$$X^{(i)} = AS^{(i)} \quad , X^{(i)} \in \mathbb{R}^n \tag{1.1}$$

Where :

- A is an unknown square matrix called **mixing matrix**
- $X^{(i)}$ is the signal of each microphone at time i
- $S^{(i)}$ is the signal of each speaker records at time i
- $S_j^{(i)}$ is the signal from speaker j at time i
- $X_j^{(i)}$ is the signal of one microphone j records at time i

Let $W = A^{-1}$ be the **unmixing matrix**. Our goal is to find W. for notation convenience, also let w_i^T denote the $i-th$ row of W, so that

$$W = \begin{bmatrix} -w_1^T- \\ -w_2^T- \\ \vdots \\ -w_d^T- \end{bmatrix}$$

Thus, $w_i \in \mathbb{R}^d$, and the $j-th$ source can be revovered as $s_j^{(i)} = w_j^T x^{(i)}$

Overview

- The number of sources is equal to the number of recorders.

- The number of sources is equal to the number of recorders.
- The Sources are statistically independent.

- The number of sources is equal to the number of recorders.
- The Sources are statistically independent.
- The independent component have a non gaussian distribution.

- The number of sources is equal to the number of recorders.
- The Sources are statistically independent.
- The independent component have a non gaussian distribution.
- The mixing matrix A is inversible.

ICA Ambiguities

- The pertubation of the original sources doesn't matter for most applications.

ICA Ambiguities

- The pertubation of the original sources doesn't matter for most applications.

- If the w_i found by the algorithm is scalled by any non zero real number, the corresponding recovered source S_i will be scalled by the same factor. This usually doesn't matter.

ICA Ambiguities

- The pertubation of the original sources doesn't matter for most applications.

- If the w_i found by the algorithm is scalled by any non zero real number, the corresponding recovered source S_i will be scalled by the same factor. This usually doesn't matter.

- Since ICA separtes sources by maximizing their non-Gaussianity, perfect Gaussian sources can't be separated.

Densities and linear transformations

Let's consider a random variable S draw according to some density $Ps(S)$.
Let x defined according to $x = AS = W^{-1}S, \quad (S = Wx$ with $W = A^{-1})$.
So,

$$P_x(x) = P_s(S) = P_s(Wx) \quad \text{"}\textit{is} \quad \textit{not} \quad \textit{correct}\text{"} \tag{3.1}$$

Densities and linear transformations

Let's consider a random variable S draw according to some density $Ps(S)$.

Let x defined according to $x = AS = W^{-1}S, \quad (S = Wx$ with $W = A^{-1})$.

So,

$$P_x(x) = P_s(S) = P_s(Wx) \quad \text{"}\textit{is not correct}\text{"} \tag{3.1}$$

and in particular the right formula is :

$$P_x(x) = P_s(Wx).|W| \tag{3.2}$$

with $|W|$ the determinant of matrix W.

Densities and linear transformations

Choice of $P_s(S_j)$:
We can choose density or as equivalent, we can choose a CDF. For the historical reason and for convenience we need to choose some function that increase from 0 to 1. And we can't choose a Gaussian. For convenience we choose the CDF to be a sigmoïd function.

ICA Algorithm

Two principles of ICA estimations :

• **Minimization of the mutual information (ICA gradient ascend and Fast ICA)** and

• **Maximumum Likelihood Estimation (MLE)**

MLE

We suppose that the distribution of each source S_j is given by a density P_s and that the joint distribution of the sources S is given by

$$P(S) = \prod_{j=1}^{d} P_s(S_j)$$

$$P_x(x) = \left(\prod_{j=1}^{d} P_s(W_j^T x) \right) |W|$$

$$L(W) = \log \left(\prod_{i=1}^{n} \left(\prod_{j=1}^{d} P_s(W_j^T x^{(i)}) \right) |W| \right)$$

$$L(W) = \sum_{i=1}^{n} \left(\left(\sum_{j=1}^{d} \log(P_s(W_j^T x^{(i)}) \right) + \log(|W|) \right) \qquad (4.1)$$

Let's consider

$$g(S_j) = \frac{1}{1+\exp(-S_j)} \quad \text{with} \quad S_j = W_j^T x$$

So

$$P_s(S_j) = g'(S_j) = g(S_j)(1 - g(S_j))$$

Then,

$$L(W) = \sum_{i=1}^{n} \left(\left(\sum_{j=1}^{d} \log(g'(S_j)) \right) + \log(|W|) \right) \qquad (4.2)$$

ICA Algorithm

Let us compute the gradient of $L(W)$ given in (4.2) w.r.t W So,

$$\nabla_W L(W) = \sum_{i=1}^{n}\left(\left(\sum_{j=1}^{d}\nabla_W \log(g'(S_j))\right) + \nabla_W \log(|W|)\right)$$

- $$\nabla_W \log(|W|) = \frac{\nabla_w |W|}{|W|} = (W^{-1})^T = (W^T)^{-1},$$

because

$$\nabla_W |W| = |W|(W^{-1})^T$$

- Set $f(g'(S_j)) = \log(g'(S_j))$. So,

$$\nabla_W f = \frac{\partial f}{\partial g'(S_j)}\frac{\partial g'(S_j)}{\partial S_j}\frac{\partial S_j}{\partial W} \qquad \text{(Chain rule's)}$$

ICA Algorithm

$$\frac{\partial S_j}{\partial w} = x, \frac{\partial g'(S_j)}{\partial S_j} = g'(S_j)(1 - 2g(S_j)) \quad \text{and} \quad \frac{\partial f}{\partial g'(S_j)} = \frac{1}{g'(S_j)}$$

Then,

$$\nabla_W f = (1 - 2g(S_j))$$

Therefore,

$$\nabla_W L(W) = \sum_{i=1}^{n} \left((1 - 2g(w_j^T x^{(i)}) x^{(i)T} + (W^T)^{-1} \right) \text{ where } S_j^{(i)} = w_j^T x^{(} \tag{4.3}$$

ICA Algorithm

Then, for the training example $x^{(i)}$, the update rule using the stochastic gradient ascent is :

$$W := W + \alpha \left(\begin{bmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_d^T x^{(i)}) \end{bmatrix} x^{(i)T} + (W^T)^{-1} \right)$$

Where α is the learning rate.

We then compute $S^{(i)} = Wx^{(i)}$ to recover the orginal sources. After the algorithm converges.

References

Andrew Ng, Tengyu.
CS229 Lecture Notes, 2022.

Hyvarinen, Aapo.
"Independent component analysis, a tutorial.
http://www. cs. helsinki. fi/u/ahyvarin/papers/NN00new, pdf (1999).

Andrew Ng.
Lecture 15| Machine Learning
https://www.youtube.com/watch?v=QGd06MTRMHslist=PLFyHXClT3SSje86s3ukLqTP8_3eS_xc-j

Acknowledgements