AIMS | African Institute for Mathematical Sciences SENEGAL

AFRICAN MASTER'S IN MACHINE INTELLIGENCE (AMMI)

Foundations of ML/DL

Independent Component Analysis

Authors:

Armandine Sorel Kouyim Meli askmeli@aimsammi.org
Samuael Adnew sadnew@aimsammi.org
Verlon Roel Mbingui vrmbingui@aimsammi.org

Group 6

2022-2023

Contents

Introduction

Imagine that you are in a room where two people are speaking simultaneously. You have two microphones, which are located in different distance. The microphones give you two recorded time signals. How to separate out the signal emitted by each speaker? This problem is known as " *A cocktail-party problem*" [1–5]. We can solve this problem by using Independent Component Analysis (ICA) method [3, 5]. ICA is a generative model and was first introduced in the 80s by J. Herault, C. Jutten and B. Ans, and the authors proposed an iterative real-time algorithm [5] .Independent Component Analysis (ICA) is a statistical and computational technique used in machine learning to separate a multivariate signal into its independent non-Gaussian components. ICA has been used in many applications such as : speech separation (speech enhancement and recognition), image processing (image denoising, filling in missing data), biomedical signal processing etc. There exist two popular methods for ICA estimation : Minimization of Mutual Information and Maximum Likelihood Estimation [5]. In this report we will focus on the estimation method based on the Maximum Likelihood Estimation(MLE).

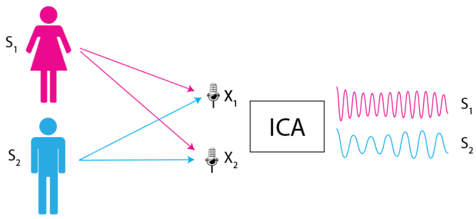


Figure 1: cocktail-party problem (Image ref)

We can express this as :

$$\begin{cases} x_1(t) = a_{11}s_1 + a_{12}s_2 \\ x_2(t) = a_{21}s_1 + a_{22}s_2 \end{cases} . \tag{1}$$

1 Problem formalization

In this section, we want to generalize the cocktail-party problem as mention in the introduction.

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

$$X^{(i)} = AS^{(i)} \quad , X^{(i)} \in \mathbb{R}^n. \tag{2}$$

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 recovered by $s_j^{(i)} = w_j^T x^{(i)}$.

2 Independent Component Analysis (ICA)

In this part, we want to talk about some ICA assumptions and Ambiguities

2.1 Assumptions

To perform the ICA, We require some assumptions such that:

(i) The components $S^{(i)}$ are statistically independent.

(ii) The independent component must have a non-gaussian distributions.

(iii) We also assume the unknown mixing matrix is square and invertible.

2.2 Ambiguities

(i) We cannot determine the order of the independent components,

(ii We cannot determine the variance of the independents components .

3 ICA Methods

Estimation method are used for estimating parameters of statistical models given a set of observations. In ICA, this method is used for estimating the unmixing matrix W which provides the best fit for extracted signals. The estimation is based on three principles methods:

(a) Minimizing Mutual information.

(b) Maximizing a non-gaussianity.

(c) Maximizing the likelihood estimation.

3.1 Densities and linear transformations

Let us 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,

$$P_x(x) = P_s(Wx).|W| \tag{3}$$

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

In ICA, We can choose probability density function(PDF) or as equivalent, we can choose a Cumulative density function (CDF). For the historical reason and for convenience we need to choose some function that increase from 0 to 1. And we can not choose a Gaussian. For convenience we choose the CDF to be a sigmoïd function.

3.2 MLE estimation method

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

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

Then

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

Given n observations of x, the log-likelihood of W which is denoted by W is given by

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

therefore

$$L(W) = \sum_{i=1}^{n}\left(\left(\sum_{j=1}^{d} \log(P_s(w_j^T x^{(i)})\right) + \log(|W|)\right) \tag{4}$$

3.3 Update the unmixing matrix by using Stochastic Gradient Ascent(SGA)

In this part, we want to update the unmixing matrix. For that, we want first to compute the gradient of $L(W)$ define by (4).

Let us 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,

$$L(W) = \sum_{i=1}^{n}\left(\left(\sum_{j=1}^{d} \log(g'(S_j))\right) + \log(|W|)\right) \tag{5}$$

By applying the chain rules and by using $\nabla_W |W| = |W|(W^{-1})^T$ in (5), we get :

$$\nabla_W L(W) = \sum_{i=1}^{n} \left((1 - 2g(w_j^T x^{(i)}) x^{(i)T} + (W^T)^{-1} \right). \tag{6}$$

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

$$W := W + \alpha \left(\begin{bmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_d^T x^{(i)}) \end{bmatrix} x^{(i)T} + (W^T)^{-1} \right) \quad \text{where } \alpha \text{ is the learning rate.} \tag{7}$$

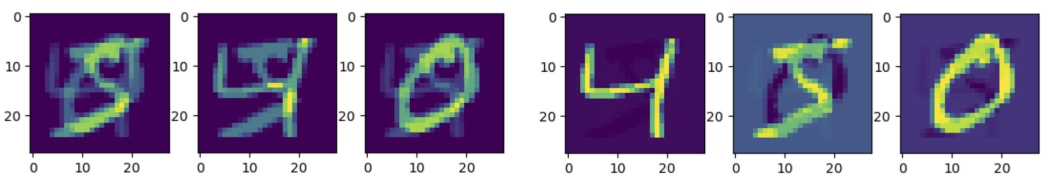
We then compute $S^{(i)} = Wx^{(i)}$ to recover the orginal sources after the algorithm converges.

Algorithm 1 : ICA algorithms with MLE and SGA

Require: $W,\ X,\ D,\ num - epochs,\ learning\ rate\ (lr)$
  $X \leftarrow X - mean(X)$ // **Data centering**
  change in likelihood $\leftarrow 0$
  W $\leftarrow$ random matrix of shape (D,D) // **Initialization of W**
  epoch count $= 0$
  repeat
    change in likelihood $\leftarrow \nabla_W$(change in likelihood)
    $W \leftarrow W$ + lr$\times$(change in likelihood) // **Update W**
    epoch count $\leftarrow$ epoch count +1
  until (epoch count <= num-epochs and lr$\times$ (change in likelihood)<lr)
  $S \leftarrow WX$ //**Find S**
  return S

4 Implementation

You can find here the python code used to implement the previous algorithm.



(a) mixed image (b) unmixed image

These are the result of separating three different mixed image observation (a) into three independent image source (b).

Conclusion

In ICA, observed random data are linearly transformed into independent components through maximizing the non-Gaussianity, maximizing the likelihood, or minimizing mutual information between independent components. Compare to PCA which is a dimensionality reduction method. Although, we can not derive the exact sign, magnitude, and variance of unmixing matrix, and the order of independent components in the output, we can separate mixed signal, and can be used for feature extraction, etc. However ICA has also some limitations : if the sources are non Gaussian, mixed non linearly, ICA may not be effective. It can also suffer on convergence problem that means it may not always able to find a solution. This can be case for complex data sets with many sources.

References

[1] Aapo Hyvarinen, Juha Karhunen, and Erkki Oja. Independent component analysis. *Studies in informatics and control*, 11(2):205–207, 2002.

[2] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[3] Andrew Ng. *CS229 Lecture notes*. 2000.

[4] James V Stone. *Independent component analysis*. A Bradford Book, 2004.

[5] Alaa Tharwat. Independent component analysis: An introduction. *Applied Computing and Informatics*, 17(2):222–249, 2021.