Independent Component Analysis

PGM Project

Nathan de Lara
École polytechnique

Florian Tilquin
ENS Cachan

Vincent Vidal
ENS Ulm

December 28, 2015

Contents

Abstract

This paper is dedicated to the study of Independent Component Analysis. We intent to implement, apply and compare several algorithms while being presenting some theoretical aspects such as the link between the likelihood maximisation and the mutual information.

1 Problem statement

1.1 Introduction

The general Independent Component Analysis problem can be formalised this way: Suppose we have some random variables $x \in \mathbb{R}^p$ which correspond to a mix of some primitive sources $s \in \mathbb{R}^n$. The aim is to extract from x every source s_i. To do so, we will suppose here that:
- the sources are independents.
- the mix is linear and instantaneous
- at most one source has a Gaussian distribution.

We define:

$$x = As \quad \text{and} \quad y = Wx, \tag{1}$$

where A is the mixing matrix, W the separation matrix and y the estimation of the sources. The goal is then to find a matrix W that maximise a certain measure of independence of y.

As a measure of independence, we consider, for theoretical purpose, the mutual information:

$$\mathrm{I}(Y) = \int_{\mathbb{R}^p} P(Y) \log \frac{P(Y)}{\Pi_i P_i(Y_i)} \mathrm{d}Y. \tag{2}$$

However, as it is too hard to compute, we consider other contrast functions, invariant by permutation, scaling on coordinates and maximal for independent ones.

1.2 Information Theory

Let $X \in \mathbb{R}^n$ be a random variable, we note $P(X)$ his density and Σ_X his covariance matrix.

In the space of measures, let $\mathcal{G}$ be the manifold of Gaussian distributions, $\mathcal{P}$ the manifold of "product" distributions and $\mathcal{P} \wedge \mathcal{G}$ the manifold of Gaussian "product" distributions. Note that these manifolds are exponential families.

The main advantage of this geometric point of view is that the Kullback-Leibler divergence allows the notion of projection on exponential families. The projection of P on the family $\mathcal{E}$ is defined as the vector of $\mathcal{E}$ that minimise the divergence to P. We write this projection $P^{\mathcal{E}}$.

Then, we define:
The **Kullback–Leibler divergence** distribution from Q to P:

$$\mathrm{K}(P \parallel Q) = \int_{\mathbb{R}^n} P(x) \log \frac{P(x)}{Q(x)} \mathrm{d}x. \tag{3}$$

The **entropy**:

$$\mathrm{H}(P) = -\int_{\mathbb{R}^n} P(x) \log P(x) \mathrm{d}x. \tag{4}$$

The **mutual information**:

$$\mathrm{I}(Y) = \mathrm{K}\Big(P(Y) \parallel \Pi_i P_i(Y_i)\Big) = \mathrm{K}\Big(P(Y) \parallel P(Y)^{\mathcal{P}}\Big). \tag{5}$$

The **negentropy**:

$$\mathrm{G}(Y) = \mathrm{H}\Big(\mathcal{N}\left(\mathbb{E}[\,Y\,]\,,\,\Sigma_Y\right)\Big) - \mathrm{H}(Y) = \mathrm{H}\big(P(Y)^{\mathcal{G}}\big) - \mathrm{H}(P(Y)). \tag{6}$$

The **non-gaussianity**:

$$\mathrm{G}(Y) = \mathrm{K}\Big(Y \parallel \mathcal{N}\left(\mathbb{E}[\,Y\,]\,,\,\Sigma_Y\right)\Big) = \mathrm{K}\Big(P(Y) \parallel P(Y)^{\mathcal{G}}\Big). \tag{7}$$

The **correlation**:

$$\begin{aligned} \mathrm{C}(Y) &= \mathrm{K}\Big(\mathcal{N}\left(\mathbb{E}[\,Y\,]\,,\,\Sigma_Y\right) \parallel \mathcal{N}\left(\mathbb{E}[\,Y\,]\,,\,\mathrm{Diag}\,\Sigma_Y\right)\Big) \\ &= \mathrm{K}\Big(P(Y)^{\mathcal{G}} \parallel P(Y)^{\mathcal{P}\wedge\mathcal{G}}\Big) \\ &= \frac{1}{2} \log \frac{\det\left(\mathrm{Diag}(\Sigma_Y)\right)}{\det\left(\Sigma_Y\right)}. \end{aligned} \tag{8}$$

Using the fact that:

$$\mathrm{K}\Big(P(Y) \parallel P(Y)^{\mathcal{P}}\Big) + \sum_i \mathrm{K}\Big(P(Y_i) \parallel P(Y_i)^{\mathcal{G}}\Big) = \mathrm{K}\Big(P(Y) \parallel P(Y)^{\mathcal{G}}\Big) + \mathrm{K}\Big(P(Y)^{\mathcal{G}} \parallel P(Y)^{\mathcal{P}\wedge\mathcal{G}}\Big) \tag{9}$$

We have:

$$\mathrm{I}(Y) + \sum_i \mathrm{G}(Y_i) = \mathrm{G}(Y) + \mathrm{C}(Y). \tag{10}$$

For more information see [Car03].

1.3 ICA and Maximum Likelihood

As presented in [HO00], it is possible to consider ICA as a maximum likelihood problem linked to the infomax principle. With the previously introduced notations, the log-likelihood is defined as:

$$L = \sum_t \sum_i \log f_i(w_i^T x(t)) + T \log(|det(W)|) \tag{11}$$

Where f_i is the density function of s_i. The expectation of this likelihood is:

$$\mathbb{E}[L] = \sum_i \mathbb{E}[\log f_i(w_i^T x(t))] + \log(|det(W)|) \tag{12}$$

In the case where f_i is the actual distribution of $w_i^T x(t)$, the first term becomes $-\sum_i H(w_i^T x(t))$ which is one of the independence measures listed in 1.2.

2 Algorithms for ICA

2.1 Hérault and Jutten (HJ) algorithm

This method is based on the neural network principle. We write $W = \left(I_n + \widetilde{W}\right)^{-1}$ and for a pair of given functions $(f,\ g)$, we adapt $\widetilde{W}$ as follows:

$$\widetilde{W}_{ij} = f(y_i)g(y_j). \tag{13}$$

2.2 EASI algorithm

For a given contrast function ψ and a cost function J_ψ, the algorithm iterates as follows:

$$W_{t+1} = (I_n - \lambda_t \nabla J_\psi(y_t))\, W_t. \tag{14}$$

The EASI algorithm corresponds to the functions $\psi(y) = \sum |y_i|^4$ and $J_\psi(y) = \mathbb{E}[\,\psi(y)\,]$.

2.3 Jade algorithm

Several methods are based on the cumulants. The goal here is to annul all the cross cumulants of order 4. Thus, the idea is to diagonalize the cumulant tensor which is equivalent to minimise the following contrast function:

$$c\,(x) = \sum_{i,k,l} \left|\mathrm{Cum}\,(x_i, x_i^*, x_k, x_l)\right|^2. \tag{15}$$

2.4 FastICA algorithm

The FastICA algorithm is based on the information theory. The goal here is to maximise the marginal non-gaussianity on the whitened data, relying on a non linear quadratic function f with the following rule:

$$\widetilde{W}_{t+1} = \mathbb{E}[\,X.f(W_t^\mathsf{T} X)^\mathsf{T}\,] - \mathbb{E}[\,f''(W_t^\mathsf{T} X)\,]\, W_t, \tag{16}$$

with W_t the normalise vector of $\widetilde{W}_t$. In our experiments, we used $f(x) = \frac{x^4}{4}$. But it is possible to use $f(x) = \log\cosh x$ or $f(x) = \exp\left(-\frac{x^2}{2}\right)$ as well.

3 Results

First, we decided to apply the main algorithm to simple simulated data. For three sinusoidal signals of arbitrary frequencies and phases, we obtain the results in figure **??**, which are good for all methods. We only have to remark that FastICA doesn't retrieve the sign of the signal, which is irrelevant.

References

[Car03] Jean-François Cardoso. Dependence, correlation and gaussianity in independent component analysis. *The Journal of Machine Learning Research*, 4:1177–1203, 2003.

[HO00] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[LeB] Hervé LeBorgne. Analyse en composantes indépendantes. chapter 3.