Independent Component Analysis

PGM Project

Nathan de Lara
École Polytechnique

Florian Tilquin
ENS Cachan

Vincent Vidal
ENS Ulm

December 8, 2015

Contents

Abstract

We decided to work on the Independent Component Analysis problem. We intent to implement, apply and compare several algorithms while being certain to understand the link between the likelihood maximisation and the mutual information.

1 Definition of the problem

1.1 Introduction

The general 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 will write:

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

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 will 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 will consider other contrast functions, invariant by permutation, scaling on coordinates and maximal for independent ones.

1.2 Information Theory

Till the end, for a random variable $X \in \mathbb{R}^n$, we will note $P(X)$ his density and Σ_X his covariance matrix.

In the space of measures, we will consider $\mathcal{G}$ the manifold of gaussian distributions, $\mathcal{P}$ the manifold of "product" distributions and $\mathcal{P}\wedge\mathcal{G}$ the manifold of gaussian "product" distribution. We can remark that these manifolds are exponential families.

The main property 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 will write this projection $P^{\mathcal{E}}$.

We can then 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**:

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

The **non-gaussianity**:

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

The **correlation**:

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

We can show the following relations:

$$\mathrm{I}(Y) + \sum_i \mathrm{G}(Y_i) = \mathrm{G}(Y) + \mathrm{C}(Y). \tag{9}$$

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 the $\widetilde{W}$ as follows:

$$\widetilde{W}_{ij} = f(y_i)g(y_j). \tag{10}$$

2.2 EASI algorithm

For a given contrast function ψ and a cost function J_ψ, we iterate as follows:

$$W_{t+1} = \left(I_n - \lambda_t \nabla J_\psi(y_t)\right) W_t. \tag{11}$$

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

2.3 Jade algorithm

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

$$c\,(x) = \sum_{i,k,l} \left|\mathrm{Cum}\,(x_i, x_i^*, x_k, x_l)\right|^2. \tag{12}$$

2.4 FastICA algorithm

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

$$\widetilde{W}_{t+1} = \mathbb{E}\big[\,X.f(W_t^\intercal X)^\intercal\,\big] - \mathbb{E}\big[\,f''(W_t^\intercal X)\,\big]\,W_t, \tag{13}$$

with W_t the normalise vector of $\widetilde{W}_t$. In our case, we will use $f(x) = \frac{x^4}{4}$. We may use $f(x) = \log\cosh x$ or $f(x) = \exp\left(-\frac{x^2}{2}\right)$ too.

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.