Independent Component Analysis

PGM Project

Nathan de Lara
École polytechnique

Florian Tilquin
ENS Cachan

Vincent Vidal
ENS Ulm

January 11, 2016

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**:

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

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**:

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

Using the Pythagorean theorem and the two decompositions of $\mathrm{K}\big(P \parallel P^{\mathcal{P}\wedge\mathcal{G}}\big)$, through $P^{\mathcal{P}}$ or $P^{\mathcal{G}}$, shown in the Figure 1, we can prove that:

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

Because the non gaussianity is invariant under invertible affine transforms, minimising the mutual independence according to W is equivalent to minimise $\mathrm{C}(Y) - \sum_i \mathrm{G}(Y_i)$. We can then define a set of contrast function, for $\alpha \geq 0$:

$$\phi_\alpha(Y) = \alpha \mathrm{C}(Y) - \sum_i \mathrm{G}(Y_i). \tag{10}$$

For more information see [Car03].

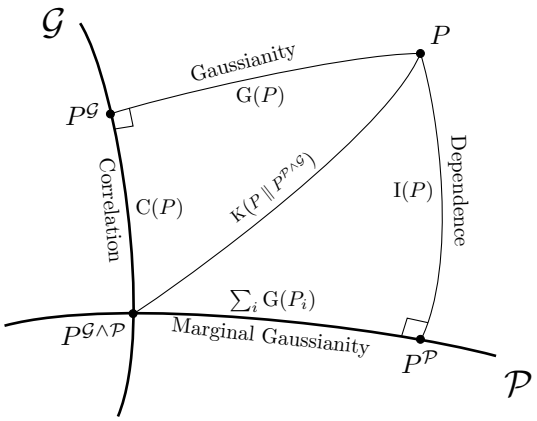

Figure 1: Representation of a distribution P and the different projections on the exponential families $\mathcal{P}$ and $\mathcal{G}$. On the paths between the distributions are the quantities associated to the Kullback-Leibler divergence between those distributions.

1.3 ICA and Maximum Likelihood

As presented in [HO00a], 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 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:

$$c\,(x) = \sum_{i,k,l} \left|\mathrm{Cum}\,(x_i, x_i^*, x_k, x_l)\right|^2. \tag{14}$$

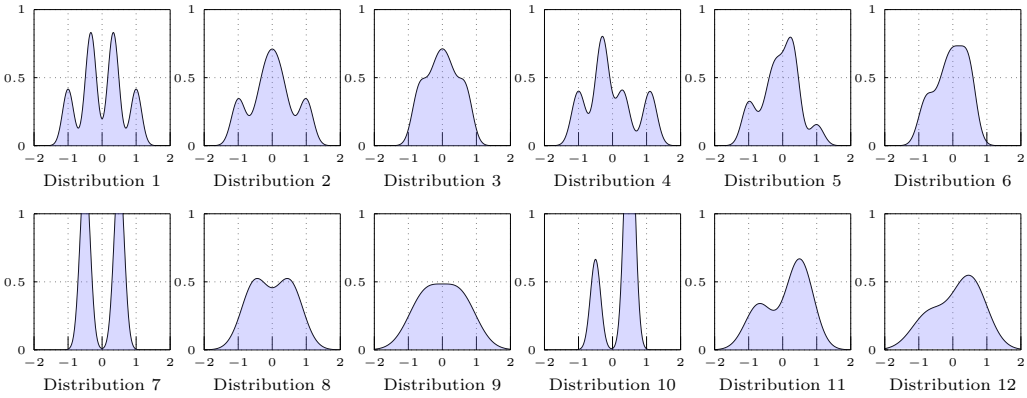

Figure 2: Distributions used to test the algorithms.

2.3 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:

$$\widetilde{W}_{t+1} = \mathbb{E}\big[\, X.f(W_t^\mathsf{T} X)^\mathsf{T} \big] - \mathbb{E}\big[\, f''(W_t^\mathsf{T} X) \,\big]\, W_t, \tag{15}$$

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.

2.4 Kernel ICA algorithm

Given a reproducing kernel Hilbert space $\mathcal{F}$, this algorithm seeks to minimize the Kernel Generalized Variance defined as:

$$\widehat{\delta}_{\mathcal{F}} = -\frac{1}{2}\log\prod_i (1-\rho_i^2) \tag{16}$$

where the ρ_i are the kernel canonical correlations between the observations components, obtained with computations over the observations Gram matrices.

3 Results

Performance measure The "Amari divergence", equation 17, gives a criterion of proximity between two matrices, to evaluate the performance of an algorithm. If U and V are two n-by-n matrices, the Amari error is defined by:

$$d(U,V) = \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}|a_{ij}|}{\max_j |a_{ij}|} - 1\right) + \frac{1}{2n}\sum_{j=1}^{n}\left(\frac{\sum_{i=1}^{n}|a_{ij}|}{\max_i |a_{ij}|} - 1\right) \tag{17}$$

with $a_{ij} = (UV^{-1})_{ij}$. This function, which is not an actual distance, has the advantage to be invariant by scaling factors and permutations of the matrices components.

	JADE	HJ	FastICA	Kernel ICA
1	5.67	6.37	6.51	**3.35**
2	9.47	8.15	11.54	**5.11**
3	6.76	**6.74**	8.39	10.55
4	6.21	6.82	7.66	**3.32**
5	5.99	6.42	7.30	**2.93**
6	9.32	**9.25**	12.18	9.37
7	2.90	**2.28**	3.68	2.74
8	8.52	**8.38**	10.31	12.53
9	17.52	**11.65**	21.95	28.21
10	13.94	10.30	17.71	**2.95**
11	9.57	8.98	12.31	**6.57**
12	19.19	**12.33**	22.93	13.76

m	N	JADE	HJ	FastICA	Kernel ICA
2	250	8.35	7.56	10.45	**6.06**
	1000	3.66	3.39	4.42	**2.38**
4	1000	11.88	58.52	12.20	**9.72**
	4000	5.33	83.07	5.80	**3.86**
8	2000	19.75	X	19.40	**19.15**
	4000	13.06	X	13.29	**9.71**
16	4000	31.81	X	**28.92**	X
	8000	**20.56**	X	27.89	X

Figure 3: **Left:** Average Amari divergence re-scaled by 100 obtained with the listed algorithms for random mix $m = 2$ sources of size $N = 250$ sampled with twelve different distributions. **Right:** Same measure for m sources of size N whose distributions are randomly selected among the twelve. The best results are in bold font. An X is put when a standard desktop computer could not compute the result.

Figure 4: Application of JADE algorithm to images separation. The first line presents the original sources, the second one the mix and the last one the estimations.

References

[ACY+96] Shun-ichi Amari, Andrzej Cichocki, Howard Hua Yang, et al. A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, pages 757–763, 1996.

[BJ03] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48, 2003.

[BS95] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[Car89] Jean-Francois Cardoso. Source separation using higher order moments. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 2109–2112. IEEE, 1989.

[Car97] Jean-Francois Cardoso. Infomax and maximum likelihood for blind source separation. 1997.

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

[Com94] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[HO00a] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[HO00b] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[JH91] Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991.

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