Independent Component Analysis
Probabilistic Graphical Model

Nour Bouayed, Yvann Le Fay, Zineb Bentires

November 2022

Abstract

Independent Component Analysis (ICA) is a special case of blind source separation (BSS) methodology wherein the objective is to recover an unknown mixing-process f from noisy mixed-observations of the process $\mathbf{x} = f(\mathbf{s}) + \boldsymbol{\varepsilon}$ for some statistically independent sources $\mathbf{s}$ and some noise $\boldsymbol{\varepsilon}$. It typically proceed by maximizing proxies of the statistical independence of the estimated components such as non-Gaussianity of the estimates. In this paper, we review existing frameworks which have been introduced over the past two decades, with a focus on the linear mixing process case, $\mathbf{x} = A\mathbf{s} + \boldsymbol{\varepsilon}$. We further exhibits the existing links between ICA and identifiability theory, allowing for exact inference of both, the sources and the latent mixing process using more recent techniques such as identifiable variational autoencoders. Finally, we experimentally validate the ICA algorithms on toy models, showcasing reasonable results on real-word applications Y: pas clair. introduire le modèle génératif..

1 Introduction

Assume there are m conversations held by several people in a room, and n microphones recording the background sound. In that case, the records can be seen as a complex and maybe noisy transformation, let it be f, which is dependent on the room's shape, size and acoustics as well as the orientation of the microphones with respect to the people. The blind source separation problem consists in performing statistical inference to recover the original conversations from the sound records.Y: ref où pour la première fois ça a été mentionné

Formally, let us assume we observe n signals $\mathbf{x}(t) = [x_1(t), \dots, x_n(t)]^\top \in \mathbb{R}^n$ which we assume to be a transformation of some latent sources $\mathbf{s}(t) = [s_1(t), \dots, s_m(t)]^\top \in \mathbb{R}^m$ through an unknown but invertible mixing process $f : \mathbb{R}^n \to \mathbb{R}^m$ plus an additional independent noise $\boldsymbol{\varepsilon}$, that is,

$$\mathbf{x} = f(\mathbf{s}) + \boldsymbol{\varepsilon}. \tag{1}$$

In the context of ICA, we further make the assumption the s_i's are statistically independent and non-Gaussian variables. Given some observed data

$\mathbf{X} = [\mathbf{x}(t_1), \dots, \mathbf{x}(t_T)] \in \mathbb{R}^{n \times T}$, we are interested in estimating the mixing process such that the estimated sources are as independent as possible using different proxies, such as maximizing non-Gaussianity measures or minimizing the mutual information. Once the transformation estimated, the sources are obtained by applying the inverse transformation to the observations.

While this statistical framework is a reasonable starting point to perform estimation of the sources, it has some drawbacks. 1) It is built upon assumptions which are difficult to test in real-world scenarios: while the Non-Gaussianity assumption is reasonable Y: ref+ecrire une petite phrase sur la tête des distributions des signaux observés, the independency assumption is difficult to test for. Moreover, the law of $\boldsymbol{\varepsilon}$ has to be specified to perform estimation. 2) It suffers from the identifiability issue. Indeed, without further knowledge on f, we are only able to recover the sources up to some transformations that keep invariant (1).

The remainder of this report is organized as follows:

1. In section 2, we focus on the linear case where the observed signals are assumed to be a linear mixture of the sources. In particular, we review the mutual information approach and the maximum likelihood approach, upon which, the FastICA algorithm **??** is built.

2. In section 3, we relax the linear hypothesis and show inference is still possible if some additional auxiliary variables are observed, using more recent techniques such sa variational autoencoders and identifiability theory.

3. In section 4, we provide numerical evidence for the statistical performances of, first, the FastICA algorithm and its extension, second, the VAE approach.

2 Linear ICA

The ICA approach to BSS was first proposed in Hyvärinen and Oja [2000], in which the observation process is assumed to a noiseless linear invertible transformation of the sources, in that case (1) becomes

$$\mathbf{x} = A\mathbf{s}, \tag{2}$$

where A is assumed to admit a pseudo-inverse W. Hyvärinen and Oja [2000] propose to estimate W such that the recovered sources has minimal mutual information. Indeed, we know the mutual information of $\hat{\mathbf{s}} = \hat{W}\mathbf{x}$ will be zero if and only if $\hat{\mathbf{s}}$ has statistically independent components, and otherwise, non-negativeY: ref. Thus, the ICA can be seen as a minimization program of the mutual information of $\hat{W}\mathbf{x}$ over the invertible linear transformations $\hat{W} \in \mathbb{R}^{m \times n}$. To adress the scale-indeterminacy issue, we restrict ourselves to ...Y: ecrire le set de W sur lesquels on veut optimiser. Y: write the mutual information in the discrete case +the vector is Wx

(3)

However, the mutual information is not tractable without further estimating the density functions of the s_i's from X, which makes the overall approach difficult to implement. Instead, Hyvärinen and Oja [2000] remark that using Y: write link between MI and negentropy

(4)

for which, empirically performant approximations have been developed Y: ref. Let G be...

Y: MI then Negentropy, then approximation of the negentropy using G then fastICA algorithm

2.1 The MLE approach

Y: specify the log density on each source. Write the log-likehood of the model. MI = MLE in that case. Motivation for chosing a different G for each source. Discriminating using sub ou supergaussianity. Y:

Y: less subsection, dont use subsubsection

2.2 The MLE approach

Let us denote by p_i the density associated with the signal s_i, we can write the loglikelihood of the model l as :

$$l(W|x_1, ...x_n) = \sum_{t=1}^{T} \sum_{i=1}^{n} log(p_i(w_i^T \mathbf{x}(t)) + Tlog|det(W)| \tag{5}$$

$$\mathbf{y} = W\mathbf{x} \tag{6}$$

This can be written also, omitting the terms independent of W, and denoting $G_i(s) = -\log(p_i(x))$ as:

$$l(W) = \sum_{i=1}^{n} \mathrm{E}(G_i(s_i)) + Tlog|det(W)| \tag{7}$$

Supposing we have access to the those densities, and denoting by $g_i = \frac{\delta p_i(u)}{\delta u}$ then using the hypotesis of uncorrelated and unit variance data, we can compute the following step by using the orthornomal constraint on W and the Newton natural gradient descent:

$$\Delta W = [I + \mathbb{E}(g(W\mathbf{x})(W\mathbf{x})^T]W \tag{8}$$

We don't have the p_i and so far we have $G_i = -\log(p_i)$, we do now that is not a gaussian distribution, empirically it has been shown **?** that it is sufficient to know if the distribution is sub or super gaussian, 2 popular choice for g are (the constant are non relevant):

- for the subgaussian : $g(z) = tanh(z) - z$
- for the supergaussian : $g(z) = -tanh(z) - z$

And then this learning rule can be simplified as:

$$\Delta W = [I - sg(k_4)tanh(W\mathbf{x})(W\mathbf{x})^T - (W\mathbf{x})(W\mathbf{x})^T]W \tag{9}$$

with $sg(k_4)$ is diagonal matrix such as $sg(k_4)(w_i^T\mathbf{x}) = \mathbb{E}((w_i^T\mathrm{x})^4) - 3$, the normalized kurtosis (unit variance data).

FastICA Writing this as fixed point algorithm instead of a gradient descend we have:

$$\Delta w_i = -\frac{\mathbb{E}(\mathbf{x}g_i(w_i^T\mathbf{x}) - \beta w_i}{\mathbb{E}(g_i'(w_i^T\mathbf{x}) - \beta} \tag{10}$$

where $\beta = \mathbb{E}(w_i^T\mathbf{x}g_i(w_i^T\mathbf{x})$
Writing this in the matrix form :

$$\Delta W = W[\mathbb{E}(W\mathbf{x}g(\mathbf{x^T}W)) - \mathrm{diag}(\beta_i)]D \tag{11}$$

where $D = \mathrm{diag}(\beta_i - \mathbb{E}(g'(W\mathbf{x}))$, then we apply an orthogonalization to W.

3 Non-linear case, VAE

4 Experiences

The JAX code to reproduce the experiments listed below can be found at the following address: https://github.com/hallelujahylefay/independent_component_analysis.

5 Individual contributions

The implementation and experimental evaluation is equally due to all three members, in particular, FastICA algorithm is due to YLF, the MLE approach is due to ZB (or NB). The VAE implementation is due to ... The ex The Abstract and Motivation sections were written by YLF, ...the linear ICA section were written by both NB and ZB with inputs from YLF, after which all authors reviewed the manuscript.

6 Notes

Supplementary Material

7 Linear ICA

7.1 ICA as minimization of mutual information

Z: Comment utiliser les commandes th et proof?

Theorem 7.1 (mon theoreme)**.** *Proof.* ma preuve □

$$I(X;Y) = D_{KL}((X,Y)|X \otimes Y)$$

$$D_{KL}((X,Y)|X \otimes Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_{x,y} p(x,y) \log \left(\frac{1}{p(x)} + \frac{1}{p(y)} - \frac{1}{p(x,y)} \right)$$

$$= \sum_{x,y} p(x,y) \log \left(\frac{1}{p(x)} \right) + \sum_{x,y} p(x,y) \log \left(\frac{1}{p(y)} \right) - \sum_{x,y} p(x,y) \log \left(\frac{1}{p(x,y)} \right)$$

$$= H(X) + H(Y) - H((X,Y)) = I(X;Y)$$

References

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265. 2, 3