An extension of the ICA model using latent variables

Selwa Rafi, Marc Castella, Wojciech Pieczynski

AN EXTENSION OF THE ICA MODEL USING LATENT VARIABLES

Selwa Rafi, Marc Castella, and Wojciech Pieczynski

Institut Telecom; Telecom SudParis
Département CITI; UMR-CNRS 5157
9 rue Charles Fourier, 91011 Evry Cedex, France

ABSTRACT

The Independent Component Analysis (ICA) model is extended to the case where the components are not necessarily independent: depending on the value a hidden latent process at the same time, the unknown components of the linear mixture are assumed either mutually independent or dependent. We propose for this model a separation method which combines: (i) a classical ICA separation performed using the set of samples whose components are conditionally independent, and (ii) a method for estimation of the latent process. The latter task is performed by Iterative Conditional Estimation (ICE). It is an estimation technique in the case of incomplete data, which is particularly appealing because it requires only weak conditions. Finally, simulations validate our method and show that the separation quality is improved for sources generated according to our model.

Index Terms— Independent Component Analysis (ICA), blind source separation, Iterative Conditional Estimation (ICE)

1. INTRODUCTION

For the last decades, Blind Source Separation (BSS) has been an active research problem: this popularity comes from the wide panel of potential applications such as audio processing, telecommunications, biology,... In the case of a linear multi-input/multi-output instantaneous system, BSS corresponds to Independent Component Analysis (ICA), which is now a well recognized concept [6]. Contrary to other frameworks where techniques take advantage of a strong information on the diversity, for instance through the knowledge of the array manifold in antenna array processing, the core assumption in ICA is much milder and reduces to the statistical mutual independence between the inputs. However, the latter assumption is not mandatory in BSS. For instance, in the case of static mixtures, sources can be separated if they are only decorrelated, provided that their nonstationarity or their color can be exploited. Other properties such as the fact that sources belong to a finite alphabet can alternatively be utilized [5, 11] and do not require statistical independence.

We investigate a particular model which combines an ICA model with a probabilistic model on the sources, making them either dependent or independent at different time instants. Our method exploits the "independent part" of the source components. Although it is possible to refine our model by introducing a temporal dependence, it assumes neither nonstationarity nor color of the sources. To our knowledge, only few references have tackled this issue in such a context [2, 4, 8], although the interest in dependent sources has been witnessed by some works in applied domains [1, 9]. Finally we would like to outline the difference between our work and [1]: the latter assumes a conditional independence of the sources, whereas, depending on a hidden process, we assume either conditional independence or dependence.

In the whole paper, $\mathcal{N}(\mathbf{0}, \mathbf{I})$ stands for a zero mean Gaussian law, with identity covariance matrix. $\mathcal{L}(\lambda)$ denotes the scalar Laplace distribution with parameter λ. We will use $\sim$ to indicate the distribution followed by a random variable and $\mathbf{r} \,|\, \mathbf{X}; \theta$ will refer to the law of $\mathbf{r}$ conditionally on $\mathbf{X}$ under parameter values θ.

2. EXTENDED ICA MODEL

2.1. Linear mixture

We consider a set of T samples of vector *observations*. At each time instant $t \in \{1, \ldots, T\}$ the observed vector is denoted by $\mathbf{x}(t) \triangleq (x_1(t), \ldots, x_N(t))^{\mathrm{T}}$. We assume that these observations result from a linear mixture of N unknown and unobserved *source* signals. In other words, there exists a matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ and a vector valued process $\mathbf{s}(t) \triangleq (s_1(t), \ldots, s_N(t))^{\mathrm{T}}$ such that:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \qquad \forall t \in \{1, \ldots, T\}. \tag{1}$$

Let $\mathbf{X} \triangleq (\mathbf{x}(1), \ldots, \mathbf{x}(T))$ be the $N \times T$ matrix with all observations and $\mathbf{S} \triangleq (\mathbf{s}(1), \ldots, \mathbf{s}(T))$ be the $Q \times T$ matrix with all sources. The matrix $\mathbf{A}$ is unknown and the objective consists in recovering $\mathbf{S}$ from $\mathbf{X}$ only: this is the so-called *blind source separation* problem. We will assume here that $\mathbf{A}$ is invertible and the problem thus resumes to the estimation of $\mathbf{A}$ or its inverse $\mathbf{B}$. A solution has been developed for long and is known as ICA [6]. It generally requires two assumptions: the source components should be non Gaussian –except possibly one of them– and they should be statistically mutually independent. With these assumptions, it is known that one

can estimate a matrix $\mathbf{B} \in \mathbb{R}^{N\times N}$ such that $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$ restores the sources up to some ambiguities, namely ordering and scaling factor. In the following, $\mathbf{B}$ denotes an inverse of $\mathbf{A}$ up to these ambiguities.

2.2. Latent variables

In this work, we extend ICA methods and relax in a certain way the independence assumption. The basic idea consists in introducing a hidden process $r(t)$ such that, depending on the particular value of $r(t)$ at instant t, the independence assumption is relaxed at time t. Let $\mathbf{r} \triangleq (r(1), \ldots, r(T))$. We assume more precisely:

A1. Conditionally on $\mathbf{r}$, the components $\mathbf{s}(1), \ldots, \mathbf{s}(T)$ of $\mathbf{S}$ at different times are independent.

A2. The hidden process $r(t)$ has values in $\{0, 1\}$ and, conditionally on $r(t)$:

(i) when $r(t) = 0$, the components of $\mathbf{s}(t)$ are mutually independent and non Gaussian, except possibly one of them;

(ii) when $r(t) = 1$, the components of $\mathbf{s}(t)$ are dependent.

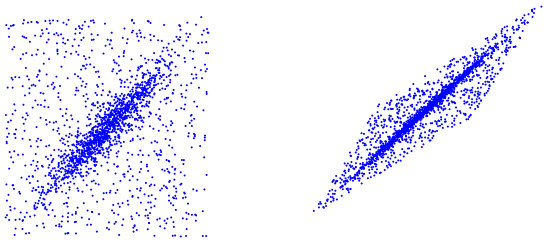
In a BSS context, one can see that, if $\mathbf{r}$ were known, one could easily apply ICA techniques by discarding the time instants where the sources are dependent. To be more precise, let $\mathcal{I}_0 \triangleq \{t \in \{1, \ldots, T\} \mid r(t) = 0\}$ be the set of time instants where the components of $\mathbf{s}(t)$ are independent. Then the subset $\mathbf{X}_0 \triangleq (\mathbf{x}(t))_{t\in\mathcal{I}_0}$ of the whole set $\mathbf{X}$ of the observations satisfies the assumptions usually required by ICA techniques. The main idea in our article consists in performing alternatively and iteratively an estimation of $\mathbf{B}$ (corresponding to $\mathbf{A}^{-1}$) and of the hidden data $\mathbf{r}$.

2.3. Typical sources separated by our method

The sources that we consider satisfy Assumptions A1 and A2. To illusrate Assumption A2, let us give an example with $N = 2$ sources which will be considered in simulations. Let

$$\mathbf{u} \triangleq \frac{1}{\sqrt{2}} \begin{pmatrix} s_1(t) + s_2(t) \\ s_1(t) - s_2(t) \end{pmatrix}. \tag{2}$$

When $r(t) = 1$, that is, when the sources are dependent, we will consider in A2-(ii) that $\mathbb{P}(\mathbf{s}(t) \,|\, r(t) = 1)$ is such that $\mathbf{u} \sim \mathcal{L}(\lambda) \times \mathcal{N}(0, 1)$ with $\lambda = 2$. The components of $\mathbf{u}$ are hence independent and follow respectively a Laplace and Gaussian law. It is not required to specify further A2-(i), but for illustration, we will consider that $\mathbb{P}(\mathbf{s}(t) \,|\, r(t) = 0)$ is such that the components of $\mathbf{s}(t)$ are independent and uniformly distributed. Such a distribution density is illustrated by simulated values in Figure 1(a). A density of some mixed observations is also shown in Figure 1(b). Considering $\mathbf{X}_0$ only amounts to removing the cloud set of dependent points on the distributions of Figure 1.



(a) (b)

Fig. 1. (a) Density plot of the sources $\mathbf{s}$: independent components with probability $\mathbb{P}(r(t) = 0) = p = 0.7$. (b) Observations $\mathbf{x} = \mathbf{A}\mathbf{s}$ with: $\mathbf{A} = \left(\begin{smallmatrix} 0.3 & 0.8 \\ 0.1 & 0.1 \end{smallmatrix}\right)$

3. PARAMETER ESTIMATION FROM INCOMPLETE DATA

3.1. Context

Let us denote by θ the set of parameters to be estimated from the data: here, θ consists of the matrix $\mathbf{B}$ and of the parameters which give the distribution of $\mathbf{r}$. Let us call $(\mathbf{r}, \mathbf{X})$ the set of *complete* data, whereas $\mathbf{X}$ alone is the set of *incomplete* data: since $\mathbf{r}$ is a hidden process, the model described in Section 2 corresponds to the situation where only incomplete data is available for estimation of the searched parameters θ. Note that the adjective *blind* is used to emphasize that $\mathbf{S}$ is unavailable, whereas *incomplete* emphasizes that $\mathbf{r}$ is unavailable.

3.2. Iterative conditional estimation

Iterative conditional estimation (ICE) is an iterative estimation method that applies in the context of incomplete data and that has been proposed in the problem of image segmentation [10, 12]. Starting from an initial guess of the parameters, the method consists in finding iteratively a sequence of estimates of θ, where each estimate is based on the previous one. More precisely, if $\hat{\theta}^{[0]}$ is the first guess, the sequence of ICE estimates is defined by:

$$\hat{\theta}^{[q]} = \mathbb{E}\{\hat{\theta}(\mathbf{r}, \mathbf{X}) \mid \mathbf{X}; \hat{\theta}^{[q-1]}\} \tag{3}$$

where $\mathbb{E}\{. \mid \mathbf{X}; \hat{\theta}^{[q-1]}\}$ denotes the expectation conditionally on $\mathbf{X}$ and with parameter values $\hat{\theta}^{[q-1]}$. In practice, the previous conditional expectation can be replaced by a sample mean, that is (3) can be replaced by:

$$\hat{\theta}^{[p]} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}(\mathbf{r}^{(k)}, \mathbf{X}) \tag{4}$$

where $K \in \mathbb{N}^*$ is fixed and each $\mathbf{r}^{(k)}$ is drawn according to the a posteriori law $\mathbf{r}|\mathbf{X}; \hat{\theta}^{[q-1]}$. The prerequisites in order to apply ICE are thus the following:

- there exist an estimator from complete data $\hat{\theta}(\mathbf{r}, \mathbf{X})$,
- one is able either to calculate $\mathbb{E}\{. \mid \mathbf{X}; \hat{\theta}^{[q-1]}\}$ or to draw random variables according to $\mathbf{r}|\mathbf{X}; \hat{\theta}^{[q-1]}$.

These two conditions are very weak, which is the reason of our interest in ICE. In fact concerning the former one, there would be no hope to perform incomplete data estimation if no complete data estimator exists, whereas the second requirement consists only in being able to simulate random values according to the a posteriori law.

4. ASSUMED DISTRIBUTION FOR $(\mathbf{r}, \mathbf{S})$

In this Section, we describe the assumptions on which our method relies. These assumptions on the probabilistic model describing the unknown data $(\mathbf{r}, \mathbf{S})$ do not exactly correspond to the source model given in section 2.3, although they are used in the ICE method.

As specified above, the observed data is given by a linear transformation of $\mathbf{S}$ according to Equation (1). Combining a statistical model on $(\mathbf{r}, \mathbf{S})$ with an ICA model on $(\mathbf{S}, \mathbf{X})$ yields a genuine model. In a more classical context, we would indeed have:

- either the ICA model of Equation (1) with an independence assumptions on $\mathbf{s}(t)$, which is much simpler than assumption A2.
- or a probabilistic model written on $(\mathbf{r}, \mathbf{X})$ and no particular relation such as (1).

For simplicity, we assume in the following $N = 2$, which also corresponds to our simulation settings. However, the method is theoretically valid for any value of N.

4.1. Specification of assumption A2

In the estimation algorithm, we assume, similarly to A1, that the probability law $\mathbb{P}(\mathbf{r}, \mathbf{S})$ factorizes:

$$\mathbb{P}(\mathbf{r}, \mathbf{S}) = \mathbb{P}(\mathbf{r}) \prod_{t=1}^{T} \mathbb{P}(\mathbf{s}(t) \,|\, r(t))$$

where $\mathbb{P}(\mathbf{s}(t) \,|\, r(t) = 1)$ is such that the components $u_i, i = 1, 2$ of the vector $\mathbf{u}$ in (2) both follow $u_i \sim \frac{1}{2}\mathcal{L}(\lambda) + \frac{1}{2}\mathcal{N}(0, 1)$ with $\lambda = 2$. In the other words, when $r(t) = 1$, each component of $\mathbf{u}$ is modelized as a mixture of the probability densities of a Gaussian and a Laplace law. This distribution is a symmetrical analog to the one in Section 2.3: although it is not the real distribution of the sources, such a symmetry is necessary in our method because ICA methods generally leave permutation ambiguities. Finally, when $r(t) = 0$, we consider $\mathbb{P}(\mathbf{s}(t) \,|\, r(t) = 0) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Again, because of A2-(i), this does not correspond to the true distribution but to the model used by the ICE method.

4.2. Hidden process r model

In the simplest version of our method, no assumption is made on $\mathbf{r}$. This amounts to modeling $\mathbf{r}$ as an i.i.d. Bernoulli process, that is $\mathbb{P}(\mathbf{r}) = \prod_{t=1}^{T} \mathbb{P}(r(t))$ with $\mathbb{P}(r(t) = 0) = p$ and $\mathbb{P}(r(t) = 1) = 1 - p$. The parameters p can be estimated by the ICE method.

An extension consists in modeling $\mathbf{r}$ as a stationary Markov chain, in which case the temporal dependence of $\mathbf{r}$ is taken into account. In this case, the probability distribution of $\mathbf{r}$ is given by: $\mathbb{P}(\mathbf{r}) = \mathbb{P}(r(1)) \prod_{t=2}^{t=T} \mathbb{P}(r(t) \,|\, r(t-1))$ where $\mathbb{P}(r(t) \,|\, r(t-1))$ is given by a transition matrix independent of t. The main advantage of considering a Markov model is that $\mathbb{P}(\mathbf{r} \,|\, \mathbf{X}; \theta)$ can be calculated by an efficient forward-backward algorithm [7], making ICE method applicable [10].

4.3. Data parameters

In our simulations, the process $\mathbf{r}$ has been generated either i.i.d. (Section 5.2.1) with $\mathbb{P}(r(t) = 0) = p$, $\mathbb{P}(r(t) = 1) = 1 - p$ or as a Markov chain (Section 5.2.2). The sources are generated following the description in Section 2.3 and mixed according to (1). The mixing matrix $\mathbf{A}$ is drawn randomly.

5. SIMULATIONS

5.1. Summary of the algorithm

Our proposed method combines ICA and ICE. The complete data estimator $\hat{\theta}(\mathbf{r}, \mathbf{X})$ in Section 3.2 is provided by one of the existing ICA separation methods. We denote it by `ICA` in the summary of our algorithm which follows:

Initialize the parameters $\hat{\theta}^{[0]} = (\hat{\mathbf{B}}^{[0]}, \hat{p}^{[0]})$.
For $q = 1, 2, \ldots, q_{\max}$, repeat:

- calculate $\mathbb{P}(\mathbf{r} \,|\, \mathbf{X}; \hat{\theta}^{[q]})$ and draw $\hat{\mathbf{r}}^{[q]}$ according to this distribution ,
- set: $\hat{\mathcal{I}}_0^{[q]} = \{t \mid \hat{r}^{[q]}(t) = 0\}$ and $\hat{\mathbf{X}}_0^{[q]} = (x(t))_{t \in \hat{\mathcal{I}}_0^{[q]}}$
- $\hat{\mathbf{B}^{[q+1]}} = \texttt{ICA}(\hat{\mathbf{X}}_0^{[q]})$
- update the parameters of the process $\mathbf{r}$.

In the simplest case where $\mathbf{r}$ is i.i.d., the last step above consists in updating the estimated value of p according to $\hat{p}^{[q+1]} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}(r(t) = 0 \,|\, \mathbf{x}(t), \hat{\theta}^{[q]})$. Finally, note that the above method estimates p according to (3), whereas $\mathbf{B}$ is estimated according to (4) with $K = 1$.

5.2. Results

We compare now the separation quality using only a classical ICA method or the same ICA method combined with ICE as proposed. In the simulations, the chosen ICA algorithm is the JADE algorithm for real data [3]. The number of ICE iterations has been fixed empirically to $q_{\max} = 20$. We show

here the values of the mean square error (MSE) on the two retrieved sources. All the given results have been obtained by an average over a set of 1000 Monte-Carlo realizations.

5.2.1. i.i.d. latent variable

We have first tried our method with $\mathbf{r}$ i.i.d. Figure 2 shows the MSE for different values of p and for different methods: JADE applied on $\mathbf{X}$, our method combining JADE and ICE (JADE + ICE) and JADE applied on $\mathbf{X}_0$ (JADE + $\mathbf{r}$). The latter case is an ideal situation which would only apply when complete data is available. We clearly see that our method succeeded in separating sources for values of p greater than 0.4. The quality improvement of our method in comparison to JADE is clearly observed for p between 0.4 and 0.9 approximately. We also studied the influence of the number of samples. The results are presented in Table 5.2.1, where we have considered the value $p = 0.5$. We can see from Table 5.2.1 that our method is advantageous for all samples sizes.

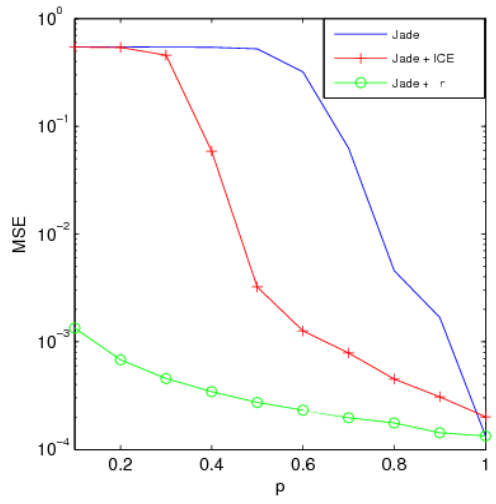


Fig. 2. Average MSE on the sources depending on p and for $T = 5000$ samples.

T	JADE	JADE + $\mathbf{r}$	JADE + ICE
1000	$3.7\ 10^{-1}$	$1.4\ 10^{-3}$	$3.8\ 10^{-2}$
2000	$4.4\ 10^{-1}$	$0.6\ 10^{-3}$	$1.4\ 10^{-2}$
5000	$5.1\ 10^{-1}$	$0.2\ 10^{-3}$	$2.9\ 10^{-2}$
10000	$5.4\ 10^{-1}$	$0.1\ 10^{-3}$	$1.3\ 10^{-2}$

Table 1. Average MSE for different sample sizes and $p = 0.5$.

5.2.2. Markov latent variable

We have tested our method in the case where the process $\mathbf{r}$ of the simulated data is a stationary Markov chain with transition matrix $\left(\begin{smallmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{smallmatrix}\right)$. We have compared the MSE values obtained with our method "JADE + ICE" based on a Markov model for $\mathbf{r}$ and "JADE + ICE" based on an i.i.d. model for $\mathbf{r}$. The results in Table 5.2.2 show that taking into account the Markov dependence of $\mathbf{r}$ significantly improves the quality result.

T	Markov model	i.i.d. model
1000	$0.9\ 10^{-1}$	$1.7\ 10^{-1}$
5000	$2.1\ 10^{-3}$	$3.3\ 10^{-3}$

Table 2. MSE for the i.i.d. model and the Markov model.

6. REFERENCES

[1] S. Akaho. Conditionally independent component analysis for supervised feature extraction. *Neurocomputing*, 49:139–150, December 2002.

[2] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. ICASSP '98. Seattle*, 1998.

[3] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. *IEE-Proceeding-F*, 140(6):362–370, December 1993.

[4] M. Castella and P. Comon. Blind separation of instantaneous mixtures of dependent sources. In *Proc. of ICA'07*, volume 4666 of *LNCS*, pages 9–16, London, UK, September 2007.

[5] P. Comon. Contrasts, independent component analysis, and blind deconvolution. *Int. Journal Adapt. Control Sig. Proc.*, 18(3):225–243, Apr. 2004.

[6] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[7] P. A. Devijver. Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, 3:369–373, 1985.

[8] A. Hyvärinen and S. Shimizu. A quasi-stochastic gradient algorithm for variance-dependent component analysis. In *Proc. International Conference on Artificial Neural Networks (ICANN2006)*, pages 211–220, Athens, Greece, 2006.

[9] F. Kohl, G. Wübbeler, D. Kolossa, C. Elster, M. Bär, and R. Orglmeister. Non-independent BSS: A model for evoked MEG signals with controllable dependencies. In *Proc. of ICA'09*, volume 5441 of *LNCS*, pages 443–450, Paraty-RJ, Brazil, March 2009.

[10] P. Lanchantin, J. Lapuyade-Lahorgue, and W. Pieczynski. Unsupervised segmentation of randomly switching data hidden with non-gaussian correlated noise. *Signal Processing*, 91:163–175, February 2011.

[11] T.-H. Li. Finite-alphabet information and multivariate blind deconvolution and identification of linear systems. *IEEE-IT*, 49(1):330–337, January 2003.

[12] W. Pieczynski. Statistical image segmentation. *Machine GRAPHICS & VISION*, 1(1/2):262–268, 1992.