Partially-supervised learning in Independent Factor Analysis

Etienne Côme[1], Latifa Oukhellou[1,2], Patrice Aknin[1] and Thierry Denœux[3]

1- INRETS - LTN
2 av Malleret Joinville, 94114 Arcueil - France

2- Université Paris 12, Val de Marne - CERTES
61 av du Gal de Gaulle, 94100 Créteil- France

3- Heudiasyc, Université Technologique de Compiègne - UMR CNRS 6599
B.P 20529, 60205 Compiègne - France

Abstract. Independent Factor Analysis (IFA) is used to recover latent components (or sources) from their linear observed mixtures within an unsupervised learning framework. Both the mixing process and the source densities are learned from the observed data. The sources are assumed to be mutually independent and distributed according to a mixture of Gaussians. This paper investigates the possibility of incorporating partial knowledge on the cluster belonging of some samples to estimate the IFA model. Semi-supervised and partially supervised learning cases can thus be handled. Experimental results demonstrate the ability of this approach to enhance estimation accuracy and remove indeterminacy commonly encountered in unsupervised IFA such as the permutation of the sources.

1 Introduction

Independent Component Analysis (ICA) defines a generative model for the observed data which are assumed to be linear mixtures of some mutually independent latent variables (also called sources) [1, 2]. Furthermore when the Independent Factor Analysis (IFA) model is considered, each latent component is assumed to be generated according to a mixture of Gaussians [3, 4]. The learning of the IFA model is often performed within an unsupervised framework. Both the mixing process and the latent variables are learned from the observed mixtures alone. In this paper, we investigate the possibility of incorporating partial knowledge on the latent variables to estimate the IFA model. In the general case, this information will be encoded by a Demspter-Shafer mass function over the set of clusters describing each source density but it can also be adapted to handle more specific learning frameworks such as the semi-supervised or the partially supervised cases. In this way, the mixture model of each source density will be provided by the component origins of a subset of training samples .

The paper is organized as follows. We will first present some background on the ICA and IFA models. In Section 3, the problem of learning the IFA model with prior knowledge on cluster memberships will be addressed. Some simulation results will then be presented in Section 4 illustrating the impact of using priors. Conclusions are summarized in Section 5.

2 Independent Factor Analysis (IFA)

2.1 Background on Independent Component Analysis (ICA)

ICA aims at recovering independent latent components from their observed linear mixtures. In its noiseless formulation, the ICA model can be expressed as:

$$\mathbf{x} = A\,\mathbf{z}, \tag{1}$$

with A a square matrix of size $S \times S$, $\mathbf{x}$ the random vector whose elements $(\mathbf{x}_1, \ldots, \mathbf{x}_S)$ are the mixtures and $\mathbf{z}$ the random vector whose elements $(\mathbf{z}_1, \ldots, \mathbf{z}_S)$ are the latent components.

Assuming the unknown mixing matrix to be non singular, the estimation of the un-mixing matrix rather than the mixing one is more appropriate since it allows us to recover the latent variables from the observed ones by simply computing $\mathbf{z} = W\,\mathbf{x}$, where $W = A^{-1}$. Furthermore, thanks to (1) a determinist relationship between the distributions of observed and latent variables can be expressed as:

$$f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(A)|} f^{\mathcal{Z}}(A^{-1}\,\mathbf{x}), \tag{2}$$

The problem consists of estimating both the un-mixing matrix and the realizations of the latent variables from the observed variables alone. Considering random sample of size N, the log-likelihood has the form:

$$\mathcal{L}(W; \mathbf{X}) = \sum_{i=1}^{N} \sum_{s=1}^{S} \log\left(f^{\mathcal{Z}_s}\left((W\mathbf{x}_i)_s\right)\right) + N\log(\det(W)). \tag{3}$$

2.2 Independent Factor Analysis Principle

The ICA model requires the choice of the probability density functions of the sources. They can be fixed by using prior knowledge, or according to some indicator which allows switching between sub and super gaussian densities [1]. An alternative solution investigated by several authors, so called Independent Factor Analysis(IFA), consists to model each source density as a mixture of Gaussians so that a wide class of densities can be approximated [3, 4].

The noiseless IFA model assumes therefore that each marginal density is distributed according to a mixture model given by:

$$f^{\mathcal{Z}_s}(z_s) = \sum_{k=1}^{K_s} \pi_k^s \varphi(z_s; \mu_k^s, \nu_k^s), \tag{4}$$

The vector of parameters is $\boldsymbol{\psi} = (W, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^S, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^S, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^S)$, where W is the un-mixing matrix, $\boldsymbol{\pi}^s$ the vector of cluster proportions of source s which sum to 1, $\boldsymbol{\mu}^s$ and $\boldsymbol{\nu}^s$ are the vectors of size K_s containing the means and the variances of each cluster. Traditional methods for learning these parameters

from an i.i.d learning set use the likelihood function, which can be obtained by substituting the density function in (3) by its definition given in (4):

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}) = N \log(|\det(W)|) + \sum_{i=1}^{N} \sum_{s=1}^{S} \log \left(\sum_{k=1}^{K_s} \pi_k^s \varphi \left((W\mathbf{x}_i)_s, \mu_k^s, \nu_k^s \right) \right). \quad (5)$$

The estimation of the IFA model parameters by the maximum likelihood can be achieved by an alternating optimization strategy. The natural gradient algorithm [5] is indeed well suited to optimize the log-likelihood function with respect to the un-mixing matrix W when the parameters of the source marginal densities are frozen. Conversely, with W kept fixed, an EM algorithm can be used to optimize the likelihood function with respect to the parameters of each source. These remarks naturally lead to develop a Generalized EM algorithm (GEM) able to simultaneously maximize the likelihood function with respect to all the model parameters.

3   Partially-supervised learning in IFA

The IFA model is often considered within an unsupervised learning framework. The idea that we investigate here is to incorporate partial knowledge on the cluster membership of some samples in the learning process. For that purpose, an objective function generalizing the likelihood function has to be defined and an EM algorithm dedicated to its optimization has to be set up.

3.1   Generalized likelihood function

The learning set is $\mathbf{X}^{iu} = \{(\mathbf{x}_1, m_1^{\mathcal{Y}_1}, \ldots, m_1^{\mathcal{Y}_S}), \ldots, (\mathbf{x}_N, m_N^{\mathcal{Y}_1}, \ldots, m_N^{\mathcal{Y}_S})\}$, where $m_i^{\mathcal{Y}_1}, \ldots, m_i^{\mathcal{Y}_S}$ is a set of basic belief assignments or Dempster-Shafer mass functions [6, 7] encoding our knowledge on the cluster membership of sample i for each one of the S sources, $\mathcal{Y}_s = \{c_1, \ldots, c_{K_s}\}$ is the set of all possible clusters for a source s. This means that additional information on the value of the cluster memberships represented by the latent variables $Y_{i1}, \ldots, Y_{iS}$ will be provided to the IFA model shown in Figure 1.
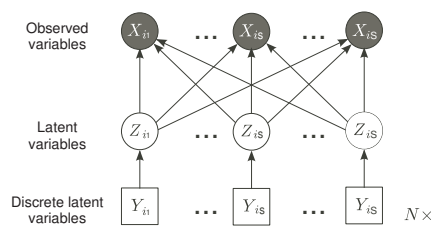

Fig. 1: Graphic model for the Independent Factor Analysis.

The mass functions (referred to hereafter as *soft* labels) may represent different kinds of knowledge, from precise to imprecise and from certain to uncer-

tain. Thus, depending on their choice, this formulation can be seen as addressing a general framework that encompasses unsupervised, supervised, partially-supervised and soft supervised paradigms as mentioned in Table 1. The plausibility pl_{ik}^s that the state of variable Y_{si} for source s and object i is c_k can be computed from mass function $m_i^{\mathcal{Y}_s}$ by using the relationship $pl_{ik}^s = \sum_{C \cap c_k} m_i^{\mathcal{Y}_s}(C)$.

	Mass function	*plausibility*
Unsupervised	$m_i^s(\mathcal{Y}_s) = 1,$	$pl_{ik}^s = 1, \forall k$
Supervised	$m_i^s(c_k) = 1$	$pl_{ik}^s = 1, pl_{ik'}^s = 0, \forall k' \neq k$
Partially supervised	$m_i^s(C) = 1$	$pl_{ik}^s = 1 \ if \ c_k \in C, \ pl_{ik}^s = 0 \ if \ c_k \notin C$
Soft supervised	$m_i^s(C)$?	$pl_{ik}^s \in [0,1]$

Table 1: Different learning paradigms and soft labels.

The concept of likelihood function has strong relations with that of possibility and, more generally, plausibility, as already noted by several authors [6]. Furthermore, selecting the simple hypothesis with highest plausibility given the observations $\mathbf{X}^{iu}$ is a natural decision strategy in the belief function framework. We thus propose as an estimation principle to search for the value of parameter with maximal conditional plausibility given the data: $\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})$.

A previous work on mixture model estimation with belief function based labels has already been addressed in [7]. In this context, a likelihood criterion taking account of *soft* labels has been defined and an EM algorithm dedicated to its optimization has been detailed. In this article, we propose an extension of such study to the IFA model in which partial knowledge on cluster memberships of a subset of samples is incorporated.

Proposition 1 *If the labels are assumed to be independent mutually and independent from the samples* $\mathbf{X}$ *that are i.i.d. generated according to the the generative IFA model setting, then the logarithm of the conditional plausibility of the model parameters vector* $\boldsymbol{\psi}$ *given the learning set* $\mathbf{X}^{iu}$ *is given by:*

$$\log\left(pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})\right) = N \log(|\det(W)|) +$$

$$\sum_{i=1}^{N} \sum_{s=1}^{S} \log\left(\sum_{k=1}^{K_s} pl_{ik}^s \pi_k^s \varphi\left((W\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right) + cst. \quad (6)$$

where pl_{ik}^s *is the plausibility that the sample* i *provides from the cluster* k *of the latent variable* s, *these plausibilities have to be computed from the soft labels* $m_i^{\mathcal{Y}_s}$, *and cst is a constant independent of* $\boldsymbol{\psi}$.

3.2 GEM Algorithm for Partially-Supervised Learning

Once the criterion is defined, the remaining work concerns its optimization which can be performed by a GEM algorithm. This algorithm is therefore the classical EM algorithm, except for the E step, where the posterior probabilities t_{ik}^s are weighted by the plausibilities and the updating of the un-mixing matrix which depends not only of the latent variables but also of the labels.

Algorithm 1: Pseudo-code for IFA with prior knowledge on labels.

Input: Centered observation matrix $\mathbf{X}$, Plausibilities pl_{ik}^s

 # Random initialization of parameters vector $\boldsymbol{\psi}^{(0)}$, $\quad q = 0$

while *Convergence test* **do**

$\quad \mathbf{Z} = \mathbf{X}.W^{(q)^t}$ *# Source update*

$\quad$ **forall** $s \in \{1, \ldots, S\}$ *and* $k \in \{1, \ldots, K_s\}$ *# (E-step)* **do**

$$t_{ik}^{s(q)} = \frac{pl_{ik}^s \pi_k^{s(q)} \varphi(z_{is}; \mu_k^{s(q)}, \nu_k^{s(q)})}{\sum_{k'=1}^{K_s} pl_{ik'}^s \pi_{k'}^{s(q)} \varphi(z_{is}; \mu_{k'}^{s(q)}, \nu_{k'}^{s(q)})}, \quad \forall i \in \{1, \ldots, N\}$$

$\quad$ **forall** $s \in \{1, \ldots, S\}$ *and* $k \in \{1, \ldots, K_s\}$ *(# M-step)* **do**

$$\pi_k^{s(q+1)} = \frac{1}{N} \sum_{i=1}^N t_{ik}^{s(q)}$$

$$\mu_k^{s(q+1)} = \frac{1}{\sum_{i=1}^N t_{ik}^{s(q)}} \sum_{i=1}^N t_{ik}^{s(q)} z_{is}$$

$$\nu_k^{s(q+1)} = \frac{1}{\sum_{i=1}^N t_{ik}^{s(q)}} \sum_{i=1}^N t_{ik}^{s(q)} (z_{is} - \mu_k^{s(q+1)})^2$$

$\quad \mathbf{G} = \mathbf{g}^{(q+1)}(\mathbf{Z})$ *# Update of G*, $g_s(z_{is}) = \sum_{k=1}^{K_s} t_{ik}^{s(q+1)} \frac{(z_{is} - \mu_k^{s(q+1)})}{\nu_k^{s(q+1)}}$,

$\quad \tau^* = linearsearch(W^{(q)}, \Delta W)$

$\quad W^{(q+1)} = W^{(q)} + \tau^*.(\mathbf{I} - \frac{1}{N} \mathbf{G}^t.\mathbf{Z}).W^{(q)^t}$ *# Unmixing matrix Update*

$\quad$ *# Source normalization $q \leftarrow q + 1$*

Output: Model parameters : $\widehat{\boldsymbol{\psi}}^{ml}$, estimated latent variables : $\widehat{\mathbf{Z}}^{ml}$

---

4 Simulations

Several simulated data sets were built as follows. Six latent variables were considered whose densities are shown in Figure 2. The observations are then generated by the IFA model given in (1) where each coefficient of the mixing matrix (6×6) was randomly generated according to a normalized normal distribution. The experiment aims to illustrate the influence of the number of labeled samples on the performance. For this, Figure 3 gives the Amari performance index [8] and the Pearson's correlation coefficient r^2 when the number of labeled samples over all the sources varied between 5% and 50%. These results were computed on 30 different learning data sets of 500 samples each. 50 initializations were performed for the GEM algorithm and only the best solution according to the likelihood was kept to avoid the problem of local minima. From these figures, we can see the benefits of incorporating prior knowledge on labels in the estimation of the IFA model. As expected, when the number of labeled samples increases (> 20%), the model behaves better since both the mean correlation and the Amari performance index are significantly improved.
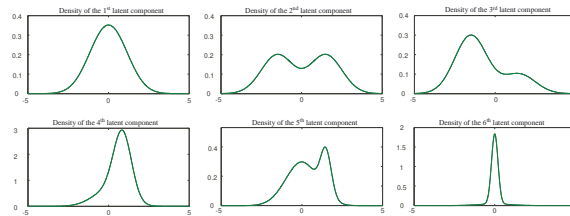
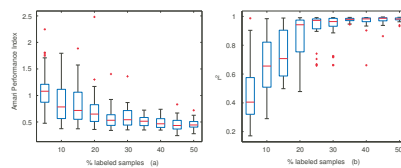Fig. 2: Simulated sources densities.

Fig. 3: Influence of the amount of labeled samples on the estimation of the semi-supervised IFA model: Boxplot of Amari perfomance index and correlation coefficient function of the percentage of labeled samples.

5   Conclusion

In this paper we have proposed a partially supervised learning for the IFA model that offers an interesting way to incorporate partial knowledge on cluster memberships to estimate such model. A generalized maximum algorithm criterion was defined and a GEM algorithm dedicated to its optimization was given. The proposed method have been applied to artificial data. The results show that our approach is able to take advantage of prior information to significantly improve estimation accuracy and to remove indeterminacy of the unsupervised IFA such as permutation of sources.

References

[1] A. Hyvärinen. *Independant Component Analysis*. Wiley, 2001.

[2] A. J. Bell and T. J. Sejnowski An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] E. moulines, J. Cardoso, E. Cassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3617–3620, 1997.

[4] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[5] S. Amari and A. Cichocki and H. H. Yang. A New Learning Algorithm for Blind Signal Separation. In *Proceedings of the 8th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 756–763. MIT Press 1995.

[6] G. Shafer. A mathematical theory of evidence. *Princeton University Press*, 1976.

[7] E. Côme, L. Oukhellou, T. Denœux and P. Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern recognition*, 42:334–348, 2009.

[8] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing*. Wiley, 2002.