Noiseless Independent Factor Analysis with mixing constraints in a semi-supervised framework. Application to railway device fault diagnosis.

Etienne Côme, Latifa Oukhellou, Patrice Aknin, Thierry Denoeux

Independent Factor Analysis with mixing constraints in a semi-supervised framework. Application to railway device fault diagnosis.

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

1- INRETS-LTN, 2 Av Malleret Joinville, 94114 Arcueil- France,
2- Université Paris 12- CERTES, 61 av du Gal de Gaulle, 94100 Créteil- France
3- Heudiasyc, UTC - UMR CNRS 6599, B.P 20529, 60205 Compiègne - France

Abstract. In Independent Factor Analysis (IFA), latent components (or sources) are only recovered from their linear observed mixtures. Both the mixing process and the sources densities (that are assumed to be generated according to mixtures of gaussians) are learned from observed data. This paper investigates the possibility of estimating the IFA model when two prior knowledge are incorporated : constraints on the mixing process and partial knowledge on the cluster membership of some examples. Semi-supervised or partially supervised learning frameworks can thus be handled. These two proposals have been initially motivated by a real-world application that concerns a fault diagnosis of a railway device. Results on this application are provided to demonstrate its ability to enhance estimation accuracy and remove indeterminacy commonly encountered in unsupervised IFA such as the sources permutations.

Key words: Independent Factor Analysis, mixing constraints, semi-supervised learning, diagnosis, railway device

1 Introduction

The generative model involved in Independent Component Analysis (ICA) assumes that observed variables are generated by a linear mixture of independent and non Gaussian latent variables (or sources). Furthermore, when the IFA model is considered, each latent variable has its own distribution, modeled semi-parametrically by a mixture of Gaussians (MOG). These models (ICA or IFA) lead to reliable results if only the independence assumption is satisfied and the postulated mixing model suited to the physics of the system. Otherwise, they fail to recover the sources. Several extensions of the basic ICA model have been proposed to improve its performance. They take account of prior information that could concern either the mixing process, the latent variables or both of them. The main approaches exploit priors like temporal correlation [6], positivity [7, 3, 11] or sparsity [8, 9].

In this paper, we propose two extensions of the basic IFA model. The first one concerns the possibility of incorporating independence hypotheses between

some latent and observed variables, hypotheses that can be derived from physical knowledge available on the mixing process. This kind of approach, not yet applied within the framework of IFA, has been widely considered in the Factorial Analysis [10, pages 43-44, 175-176] and more specifically in the structural equation modeling domain [12]. The second proposition consists to incorporate additional information on the cluster membership of some samples to estimate the IFA model. In this way, semi-supervised learning framework is handled. Considering the graphical model of the IFA as shown in Figure 1, the prior knowledge on the mixing process consists to omit some connections between observed and latent variables while the second prior means that additional information on the value of the discrete (Y) latent variables encoding our knowledge on the cluster membership of some samples is tacking into account.

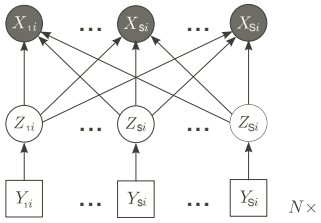

Fig. 1. Graphic model for the Independent Factor Analysis.

This article is organized as follows. We will first present IFA model estimation by maximum likelihood in a noiseless setting. In section 3 and 4, the problem of learning the IFA model with prior knowledge on the mixing process and on the cluster membership of some examples will then be addressed. In Section 5, the approach will be illustrated applying it to a railway device diagnosis on which the impact of using priors will be evaluated. The paper ends with a conclusion.

2 Background on Independent Factor Analysis

ICA and IFA aims at recovering independent latent components from their observed linear mixtures. In its noiseless formulation (the formulation used throughout of this paper), the model can be expressed as $\mathbf{x} = A\,\mathbf{z}$ 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. Thanks to the noiseless setting a deterministic relationship between the distributions of observed and latent variables can be expressed as: $f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(A)|} f^{\mathcal{Z}}(A^{-1}\,\mathbf{x})$. 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 (MOG) so that a wide class of densities can be approximated [4, 5] :

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

with $\varphi(.; \mu, \nu)$ the density of a gaussian random variable of mean μ and variance ν. The problem consists of estimating both the mixing matrix, and the MOG parameters from the observed variables alone. Considering an iid random sample of size N, the log-likelihood has the form:

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}) = -N \log(|\det(A)|) + \sum_{i=1}^{N} \sum_{s=1}^{S} \log \left(\sum_{k=1}^{K_s} \pi_k^s \varphi\left((A^{-1}\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right) \right). \tag{2}$$

where $\boldsymbol{\psi}$ is the IFA parameters vector $\boldsymbol{\psi} = (A, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^S, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^S, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^S)$, with A the mixing matrix, $\boldsymbol{\pi}^s$ the vector of cluster proportions of source s which sum to 1, $\boldsymbol{\mu}^s$ and $\boldsymbol{\nu}^s$ the vectors of size K_s containing the means and the variances of each cluster. The estimation of the IFA model parameters by the maximum likelihood can be achieved by an alternating optimization strategy. The gradient algorithm [13] is indeed well suited to optimize the log-likelihood function with respect to the mixing matrix A when the parameters of the source marginal densities are frozen. Conversely, with A 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 Constraints on the mixing process

This section investigates the possibility of incorporating independence hypotheses between some latent and observed variables in the ICA model, hypotheses often supplied by the physical knowledge of the mixing process. The hypothesis that we consider in this section have the following form: $X_h \perp\!\!\!\perp Z_g$. Making this kind of hypothesis constraints the form of the mixing matrix as it is shown by the following proposition :

Proposition 1. *In the noiseless ICA model, we have :*

$$X_h \perp\!\!\!\perp Z_g \Leftrightarrow A_{hg} = 0. \tag{3}$$

Proof. The independence can be defined as $X_h \perp\!\!\!\perp Z_g \Leftrightarrow f^{\mathcal{X}_h \times \mathcal{Z}_g}(x_h, z_g) = f^{\mathcal{X}_h}(x_h) \times f^{\mathcal{Z}_g}(z_g)$. In the case of noiseless ICA model, the joint probability

density function on $\mathcal{X}_h \times \mathcal{Z}_g$ is given by :

$$f^{\mathcal{X}_h \times \mathcal{Z}_g}(x_h, z_g) = \int_{\mathbb{R}^{S-1}} f^{\mathcal{X}_h \times \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_S}(x_h, z_1, \ldots, z_S) \prod_{l=1, l\neq g}^{S} \mathrm{d}z_l \tag{4}$$

$$= \int_{\mathbb{R}^{S-1}} \prod_{s=1}^{S} f^{\mathcal{Z}_s}(z_s) \times \delta(x_h - A_{h.}\mathbf{z}) \prod_{l=1, l\neq g}^{S} \mathrm{d}z_l$$

$$= f^{\mathcal{Z}_g}(z_g) \times \left(\int_{\mathbb{R}^{S-1}} \prod_{l=1, l\neq g}^{S} f^{\mathcal{Z}_l}(z_l) \times \delta(x_h - A_{h.}\mathbf{z}) \mathrm{d}z_l \right). \tag{5}$$

Using (**??**), we identify $X_h \perp\!\!\!\perp Z_g \Leftrightarrow f^{\mathcal{X}_h}(x_h) = \int_{\mathbb{R}^{S-1}} \prod_{l=1, l\neq g}^{S} f^{\mathcal{Z}_l}(z_l) \times \delta(x_h - A_{h.}\mathbf{z}) \mathrm{d}z_l$, where $A_{h.}$ is the h^{th} raw of the mixing matrix A and δ the Dirac function. The integral must not depend on z_g (the g^{th} raw of $\mathbf{z}$), which is possible only if A satisfies $A_{hg} = 0$. □

The estimation problem of the ICA model has to be reformulated to take account of conditional independencies of some sources given some observed variables. Indeed, the log-likelihood has to be maximized under the constraint that some of the mixing coefficients are nulls. The gradient ascent is only achieved respectively to the non-nulls coefficients. In this case, the initialization and the update rule of the mixing matrix are given by:

$$\begin{aligned} A^{(0)} &= C \bullet A^{(0)} \\ A^{(q+1)} &= A^{(q)} + \tau\, C \bullet \Delta A^{(q)}, \end{aligned} \tag{6}$$

where $\bullet$ denotes the Hadamard product between two matrices (element-by-element product) and C a binary matrix of which the elements are $C_{hk} = 0$ if $Z_k \perp\!\!\!\perp X_h$, $C_{hk} = 1$ otherwise.

4 Semi-supervised learning in IFA

The IFA model is often considered within an unsupervised learning framework. This section considers the learning of this model within partially-supervised learning context where partial knowledge on the cluster membership of some samples is available. For that purpose, a generalized likelihood function has to be defined and an EM algorithm dedicated to its optimization has to be set up. In the general case, we shall assume a learning set of the form: $\mathbf{X}^{iu} = \{(\mathbf{x}_1, m_1^{\mathcal{Y}_1}, \ldots, m_1^{\mathcal{Y}_S}), \ldots, (\mathbf{x}_N, m_N^{\mathcal{Y}_1}, \ldots, m_N^{\mathcal{Y}_S})\}$, where $m_i^{\mathcal{Y}_1}, \ldots, m_i^{\mathcal{Y}_S}$ is a set of basic belief assignments or Dempster-Shafer mass functions [14, 15] 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. Depending on the choice of the mass functions, this formulation can therefore be seen as addressing a more generale framework which encompasses unsupervised, supervised and partially-supervised learning paradigms as mentioned in Table 1.

	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	$m_i^s(C) = 1$	$pl_{ik}^s = 1 \; if \; c_k \in C, \; pl_{ik}^s = 0 \; if \; c_k \notin C$

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 [14]. 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 [15]. 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 class labels of a subset of samples is incorporated.

Proposition 2. *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:*

$$\log\left(pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})\right) = -N\log(|\det(A)|) + \sum_{i=1}^{N}\sum_{s=1}^{S}\log\left(\sum_{k=1}^{K_s} pl_{ik}^s \pi_k^s \varphi\left((A^{-1}\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right) + cst. \quad (7)$$

where pl_{ik}^s *is the plausibility that the sample* i *belong to 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* $\boldsymbol{\psi}$*.*

In a semi-supervised learning context, the IFA model is built from a combination of M labeled and $N - M$ unlabeled samples. For labeled samples, the plausibilities used as labels are crisp and we have $pl_{ik}^s = l_{ik}^s \in \{0,1\}^{K_s}$ binary variables encoding the cluster membership of labeled sample i, $l_{ik}^s = 1$ if sample i comes from cluster c_k of sources s and $l_{ik}^s = 0$ otherwise. Whereas for unlabeled samples $pl_{ik}^s = 1$ for all clusters k and sources s. Consequently, the criterion can be decomposed into two parts corresponding respectively, to the supervised and

unsupervised learning examples and criterion 6 can be rewritten as:

$$\mathcal{L}(A;\mathbf{X}) = -N\log(|\det(A)|) + \sum_{i=1}^{M}\sum_{s=1}^{S}\sum_{k=1}^{K_s} l_{ik}^s \log\left(\pi_k^s \varphi\left((A^{-1}\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right) + \sum_{i=M+1}^{N}\sum_{s=1}^{S}\log\left(\sum_{k=1}^{K_s}\pi_k^s\varphi\left((A^{-1}\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right). \quad (8)$$

A Generalized EM algorithm (GEM), Algorithm 1 able to simultaneously maximize the likelihood function with respect to all the model parameters can be used. This algorithm is similar to EM algorithm used to estimate IFA parameter in an unsupervised setting, except for the E step, where the posterior probabilities t_{ik}^s are only computed for the unlabeled samples and the updating of the mixing matrix which takes account of the mixing constraints and depends not only of the latent variables but also of the labels.

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

Input: Centered observation matrix $\mathbf{X}$, cluster belonging for the M labeled data l_{ik}^s, constraints matrix encoding independence hypothesis C.

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

while *Convergence test* **do**

$\mathbf{Z} = \mathbf{X}\left(A^{(q)^{-1}}\right)^t$ *# Source update*

forall $s \in \{1,\ldots,S\}$ *and* $k \in \{1,\ldots,K_s\}$ **do**

$t_{ik}^{s(q)} = l_{ik}^s, \quad \forall i \in \{1,\ldots,M\}$

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

forall $s \in \{1,\ldots,S\}$ *and* $k \in \{1,\ldots,K_s\}$ **do**

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

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

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

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

 # Natural gradient

$\Delta A = \left(A^{(q)^{-1}}\right)^t \left(\frac{1}{N}\sum_{i=1}^N \mathbf{g}\left(\mathbf{z}_i^{(q)}\right)\mathbf{z}_i^{(q)^t} - \mathbf{I}\right)$

$\tau^* = \text{Linearsearch}(A^{(q)}, C \bullet \Delta A)$ *# Linear Search for* τ

$A^{(q+1)} = A^{(q)} + \tau^*.C \bullet \Delta A$ *# mixing matrix Update*

 # source normalization to remove scale indetermination

$q \leftarrow q+1$

5 Fault diagnosis in railway track circuit

The application considered in this paper concerns fault diagnosis in railway track circuits. This device will first be described and the problem addressed will be exposed. An overview of the proposed diagnosis method will be presented.

5.1 Track circuit principle

The track circuit is an essential component of the automatic train control system. Its main function is to detect the presence or absence of vehicle traffic within a specific section of railway track. The signalling system uses the occupation of track section to protect trains from coming into conflict. On French high speed lines, the track circuit is also a fundamental component of the track/vehicle transmission system. It uses a specific carrier frequency to transmit coded data to the train, for example the maximum authorized speed on a given section on the basis of safety constraints. The railway track is divided into different sections. Each one of them has a specific track circuit consisting of the following components:
- A transmitter connected to one of the two section ends, which delivers a frequency modulated alternating current
- The two rails that can be considered as a transmission line;
- At the other end of the track section, a receiver that essentially consists of a trap circuit used to avoid the transmission of information to the neighboring section;
-Trimming capacitors connected between the two rails at constant spacing to compensate for the inductive behavior of the track. Electrical tuning is then performed to limit the attenuation of the transmitted current and improve the transmission level. The number of compensation points depends on the carrier frequency and the length of the track section.
The rails themselves are part of the track circuit, and a train is detected when

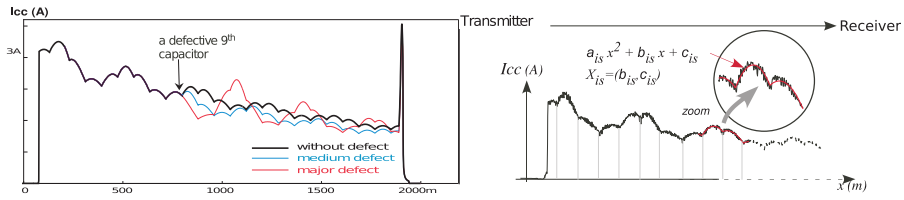

Fig. 2. Examples of inspection signals.

its wheels and axles short-circuit the track. The presence of a train in a given section induces the loss of track circuit signal due to shorting by train wheels. The drop of the received signal below a preset threshold indicates that the section is occupied. The different parts of the system are subject to malfunctions

(due to aging, ...) that must be detected as soon as possible in order to maintain the system at the required safety and availability levels. In the most extreme case, this causes an unfortunate attenuation of the transmitted signal that leads to the stop of the train. The purpose of diagnosis is to inform maintainers about track circuit failures on the basis of the analysis of a specific current, recorded by an inspection vehicle. This paper will focus on trimming capacitors faults that affect capacitor capacitance. Figure 2 shows an example of the inspection signal : one of them corresponds to an absence of fault, while the others correspond to a defective 9^{th} capacitor. The aim of the diagnosis system is to detect the operating mode of the track circuit and localize the defective capacitor by analyzing the measurement signal.

5.2 Overview of the diagnosis method

The track circuit can be considered as a large-scale system made up of a series of spatially related subsystems that correspond to the trimming capacitors. A defect on one subsystem is represented by a continue value of the capacitance parameter. The proposed method is based on the following two observations (see Figure 2). First, the inspection signal has a specific structure, which is a succession of so many arches as capacitors; an arch can be approximated by a quadratic polynomial $ax^2 + bx + c$, next each observed arch is influenced by the capacitors located upstream from it. The proposed method consists in extracting features from the measurement signal, and build a generative model as shown in figure 3, where each observed variable X_{is} corresponds to the coefficients (b_{is}, c_{is}) of the local polynomial approximating the arch located between two subsystems. Only two coefficients are used because continuity constraints between each polynomials are used, therefore their exist a linear relationship between the third coefficient and the three coefficients of the previous polynomial. The continue latent variable Z_{is} is the capacitance of the i^{th} capacitor and the discrete latent variable Y_{is} corresponds to the membership of the capacitor state to one of the three operating modes (fault-free, weak defect, major defect). As there is non influence between a trimming capacitor state and the inspection signal located upstream from it, some connections between latent and observed variables are omitted as it can be seen in figure 3. This information will be also introduced in the model estimation by the means of constraints on the mixing matrix. One can clearly see that this model is closely linked to the IFA model given in figure 1. Considering the diagnosis task as a blind source separation problem, the IFA model can be used to estimate the mixing matrix A and thereby to recover the latent components (capacitor capacitances) from the observed variables alone. Moreover, learning the IFA model with mixing constraints can also be considered to take account of prior information on the mixing process. The cluster membership of some training samples (represented by the discrete latent variable) will also be incorporated during the learning phase of the IFA model. As already explained a piecewise approach is adopted for the signal representation: each arch was approximated by a second degree polynomial of which two coefficients are used as observed variables for each node in the model of Figure 1 which lead

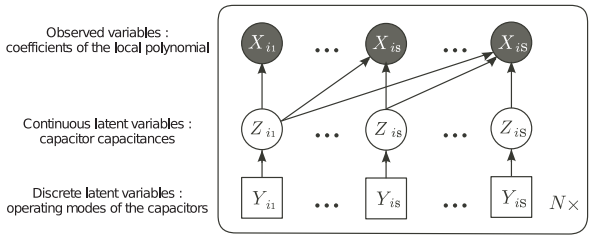

Fig. 3. Generative model for the diagnosis of track circuits represented by a graphical model including both continuous and discrete latent variables.

to $2 * S$ observed variables. Given an observation matrix, the aim is to recover S latent variables from $2 * S$ observed ones with the hope that they will be strongly correlated with the variables of interest that are the capacitor capacitances. As prior information on the mixing matrix is available, PCA cannot be used as a preprocessing because the mixing structure will be lost. $2 * S$ latent variables are therefore extracted, S latent variable densities corresponding to capacitors capacitance are assumed to be mixtures of 3 Gaussian components that correspond to the three operating modes of the capacitors while the S other variables are assumed to be noise variables and are thus modeled by simple gaussian random distributions.

6 Results and discussion

To access the performances of the approach, we considered a track circuit of $S = 18$ subsystems (capacitors) and built a database containing noised signals obtained for different values of the capacitance of each capacitor. 2500 signals are thus obtained where 500 are used in the training phase while the 2000 others are employed for the test phase. The experiments aim to illustrate the influence of both the number of labeled samples and the use of the mixing matrix constraints on the method results. The model supplies two levels of interpretation, discret and continuous latent variable but we only supply in this papern the results for the continuous latent variables. The performances were quantified through the mean absolute correlation between the true sources and their estimates calculated on a test set of 2000 samples. Figure 4 shows the mean of the absolute value of the correlation between estimated latent variables and capacitor capacitances function of the number of labeled training samples when the mixing matrix is constrained or not. Note that the case of unlabeled samples without constraints illustrates the performances of the traditional IFA model (without any prior), which are very poor as our criterion is sensitive to sources permutation. When more labeled samples are used the permutations of the sources are avoided and the performances reach a more interesting level. Twenty random starting points were used for the GEM algorithm and only the best solution according to the

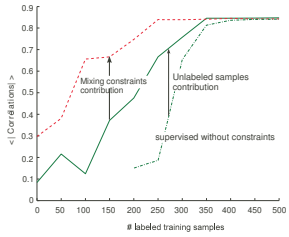

Fig. 4. Results of IFA with (- - red), without constraints (– green) when the number of labeled training samples varies between 0 and 500 and supervised IFA without constraints (-. -. green). The mean correlation between the estimated sources and the capacitor capacitances is estimated on a test set of 2000 samples. Twenty random starting points were used for the GEM algorithm and only the best solution according to the likelihood was kept.

likelihood was kept. This figure clearly highlights the benefit to use constraints when the amount of labeled samples is weak. As expected, when the number of labeled data increases, the mean correlation also increases to reach a maximal value of 0.84 for the constrained IFA model with 250 labeled sampled and for the unconstrained one with 350 labeled samples. When a sufficient amount of labeled samples is provided to the model (> 350), the prior on the mixing process does not significantly improve the performances. It can also be noticed that unlabeled samples allows improving the approach performance particularly when the size of the labeled learning data is weak. An improvement of the global performance (0.84) would require a non-linear model.

7 Conclusion

In this paper we have proposed a learning of the IFA model by incorporating two prior knowledge. The first concerns the mixing process whereas the second uses the cluster membership of some training samples. In this context, a criterion was defined and a GEM algorithm dedicated to its optimization was given. The proposed method have been applied to fault diagnosis in railway track circuits. The diagnosis system aims to recover the latent variables linked to the defects from their linear observed mixtures (features extracted from the inspection signal. A comparison between standard and proposed IFA models have been carried out to 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. Further studies will be carried out to incorporate nonlinearity and also to take account of imprecise and uncertain cluster memberships when they are supplied by human expert.

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. C. Jutten and P. Comon, editors. *Séparation de source 2, au-delà de l'aveugle et application.* Hermès, 2007.
4. 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.
5. H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.
6. H. Attias. Independent factor analysis with temporally structured factors. In *Proceedings of the 12th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 386–392. MIT Press, 2000.
7. S. Moussaoui, H. Hauksdóttir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, D. Douté and J. Benediktosson, *On the decomposition of Mars Hyperspectral data by ICA and Bayesian positive source separation. Neurocomputing for Vision Research; Advances in Blind Signal Processing*, 71:2194–2208, 2008.
8. A. Hyvärinen and R. Karthikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*, 49(1):151–162, 2002.
9. K. Zhang and L. W. Chan. ICA with sparse connections. In *Proceedings of Intelligent Data Engineering and Automated Learning Conference (IDEAL)*, pages 530–537. Springer, 2006.
10. Bartholomew, D. J. and K. Martin. Latent variable models and factor analysis. Kendall's library of satisticsyear, Arnold, London, Second edition, 1999
11. T. Bakir, A. Peter, R. Riley, and J. Hackett. Non-negative maximum likelihood ICA for blind source separation of images and signals with application to hyperspectral image subpixel demixing. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3237–3240, 2006.
12. K. A. Bollen. Structural Equations with Latent Variables. Wiley, 1989.
13. 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.
14. G. Shafer. A mathematical theory of evidence. *Princeton University Press*, 1976.
15. 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.
16. A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing.* Wiley, 2002.