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

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

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 recovered from only their linear observed mixtures. Both the mixing process and the source 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 in its noiseless setting when two kinds of prior information 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 fault diagnosis of a railway device. Results from this application are provided to demonstrate the ability of our approach to enhance estimation accuracy and remove indeterminacy commonly encountered in unsupervised IFA such as source 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 nongaussian latent variables. Furthermore, when the IFA model is considered, each latent variable has its own distribution, modeled semi-parametrically by a mixture of Gaussians (MOG) and the number of mixtures can differ from the number of sources. The IFA model introduced by [5] can indeed handle both square noiseless mixing and the general case where the data are noisy. These models yield reliable results provided 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. The main approaches exploit temporal correlation [6], positivity [7, 3, 11] or sparsity [8, 9].

In this paper, we propose two extensions of the basic noiseless IFA model. The first one concerns the possibility of incorporating independence hypotheses between some latent and observed variables. Such hypotheses can be derived from physical knowledge available on the mixing process. This kind of approach has not been applied within the framework of IFA, but it has been widely considered in factor analysis [10] and, more specifically, in the structural equation modeling domain [12]. The second extension incorporates additional information on cluster membership of some samples to estimate the IFA model. In this way, the semi-supervised learning framework can be handled. Considering the graphical model of IFA shown in Figure 1, the mixing process prior consists in omitting some connections between observed (X) and latent (Z) variables. The second prior means that additional information on the discrete latent variables (Y) is taken into account.

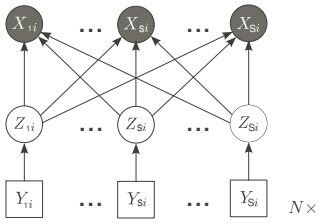

Fig. 1. Graphical model for Independent Factor Analysis.

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 samples will be addressed. In Section 5, the approach will be applied to diagnosis problem for which the impact of using priors will be evaluated. The paper ends with a conclusion.

2 Background on Independent Factor Analysis

ICA and IFA aim at recovering independent latent components from their observed linear mixtures. In its noiseless formulation (used throughout this paper), the ICA model can be expressed as $\mathbf{x} = A\,\mathbf{z}$, where A is 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 probability density functions of the sources can be fixed using prior knowledge, or according to some indicator that allows switching between sub and super Gaussian densities [1]. An alternative solution,

referred to as IFA, consists in modeling 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. This model is close to ICA with a Mixture of Gaussians model for the sources. The problem consists in 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 parameter 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. Maximum likelihood of the model parameters can be achieved by an alternating optimization strategy. The gradient algorithm [14] 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 have led to the development of a Generalized EM algorithm (GEM) able to simultaneously maximize the likelihood function with respect to all the model parameters [18].

3 Constraints on the Mixing Process

This section investigates the possibility of incorporating independence hypotheses concerning relationships between some latent and observed variables in the ICA model. Such hypotheses are often deduced from physical knowledge of the mixing process. The hypothesis that we consider in this section has the following form: $X_h \perp\!\!\!\perp Z_g$, which means that X_h is statistically independent from Z_g. Making this kind of hypothesis constraints the form of the mixing matrix as shown by the following proposition, which has been proven in [16]:

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

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

The log-likelihood has to be maximized under the constraint that some of the mixing coefficients are null, and gradient ascent is only performed with respect to the non-null coefficients. In this case, the initialization and the update rule of the mixing matrix are given by $A^{(0)} = C \bullet A^{(0)}$, $A^{(q+1)} = A^{(q)} + \tau\, C \bullet \Delta A^{(q)}$ where $\bullet$ denotes the Hadamard product between two matrices 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 (in its noiseless setting) in a partially-supervised learning context where partial knowledge of 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 will 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 [15, 16] 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 source s. Depending on the choice of the mass functions, this formulation can therefore be seen as addressing a more general framework that encompasses unsupervised, supervised and partially-supervised learning paradigms as mentioned in Table 1. The concept of likelihood function has strong relations with that of possibility and, more generally, plausibility, as already noted by several authors [15]. 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 parameter value with maximal conditional plausibility given the data: $\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})$.

Table 1. Different learning paradigms and soft labels.

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

Parameter estimation in a mixture model with belief function-based labels was already addressed in [16]. In this context, a likelihood criterion taking into account *soft* labels has been defined and an EM algorithm dedicated to its optimization has been presented. In this article, we propose an extension of such study to the IFA model in which partial knowledge of some cluster memberships is incorporated. The following proposition, proved in [16], gives the expression of the generalized likelihood criterion for the IFA model.

Proposition 2. *If the labels are assumed to be mutually independent 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:*

$$\log\left(pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})\right) = -N\log(|\det(A)|) + \sum_{i=1}^{N}\sum_{s=1}^{S}\log\left(\sum_{k=1}^{K_s} pl_{ik}^s \pi_k^s \varphi\left((A^{-1}\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right) + cst. \quad (4)$$

where pl_{ik}^s is the plausibility that the sample i belong to cluster k of the latent variable s, (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 $pl_{ik}^s = l_{ik}^s \in \{0,1\}^{K_s}$, $l_{ik}^s = 1$ if sample i comes from cluster c_k of sources s and $l_{ik}^s = 0$ otherwise. 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 (4) can be rewritten as:

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

A Generalized EM algorithm (GEM), (Algorithm 1) can be designed to simultaneously maximize the likelihood function with respect to all the model parameters. This algorithm is similar to the 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. The updating of the mixing matrix also takes into account the mixing constraints and depends not only of the latent variables, but also of the labels.

5 Fault Diagnosis in Railway Track Circuits

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 [17]. Its main function is to detect the presence or absence of vehicle traffic within a specific section of railway track. 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;

Algorithm 1: Pseudo-code for noiseless 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**

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

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

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

$q \leftarrow q + 1$

– 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 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

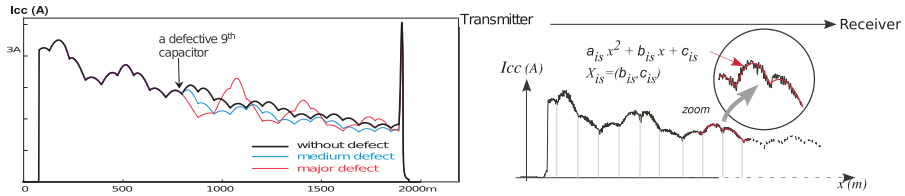

Fig. 2. Examples of inspection signals.

occupied. The different parts of the system are subject to malfunctions 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 capacitor faults that affect their capacitance. Figure 2 shows an example of the inspection signal when the system is fault-free while the others correspond to a defective 9^{th} capacitor. The aim of the diagnosis system is to detect faults in 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 continuous 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 $ax^2 + bx + c$. Second, each observed arch is influenced by the capacitors located upstream (on the transmitter side). The proposed method consists in extracting features from the measurement signal, and building 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 of continuity constraints between each polynomial, as their exists a linear relationship between the third coefficient and the three coefficients of the previous polynomial. The continuous 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 states: fault-free, minor defect, major defect. As there is no influence between a trimming capacitor state and the inspection signal located upstream from it, some connections between latent and observed variables are omitted. This information will be also introduced in the model estimation using constraints on the mixing matrix. We can clearly see that this model is closely linked to the IFA model represented in Figure 1. Considering the diagnosis task

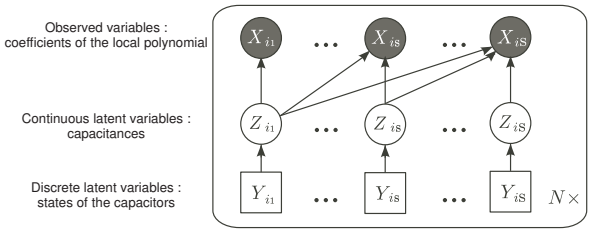

Fig. 3. Generative model for the diagnosis of track circuits represented by a graphical model including both continuous and discrete latent variables.

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 (capacitances) from the observed variables alone. As already explained, a piecewise approach is adopted for the signal representation: each arch is 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 results in $2\times S$ observed variables. Given an observation matrix, the aim is to recover S latent variables from $2 \times S$ observed ones with the hope that they will be strongly correlated with the variable of interest, i.e., capacitances. As prior information on the mixing matrix is available, PCA cannot be used for preprocessing because the mixing structure would be lost. $2 \times S$ latent variables are therefore extracted: S latent variable densities corresponding to capacitances are assumed to be mixtures of three Gaussian components, one for each state of the capacitors while the S other variables are assumed to be noise variables and are thus modeled by simple Gaussian distributions. It can be noticed that with standard IFA model, S latent variables can be recovered from $2 \times S$ observed ones. But in this paper, we consider the noiseless IFA model which seems to be straightforward to incorporate prior information and to recover the sources from the data.

6 Results and Discussion

To assess the performances of the approach, we considered a track circuit of $S = 18$ subsystems (capacitors) and a database containing 2500 noised signals obtained for different values of the capacitance of each capacitor. 500 were used in the training phase, while the 2000 others were employed for the test phase. The experiments aim at illustrating the influence of both the number of labeled samples and the use of the mixing matrix constraints on the results. The model provides two levels of interpretation corresponding to discrete and continuous latent variables, but we only discuss in this paper the results for the continuous latent variables. Figure 4 shows the mean of the absolute value of the correlation between estimated latent variables and capacitances as a function of the number of labeled training samples, when the mixing matrix is constrained or

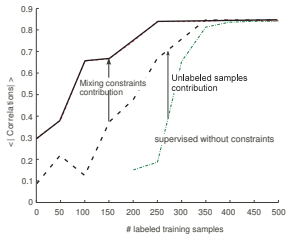

Fig. 4. Results of IFA with (–), without constraints (- -) when the number of labeled samples varies between 0 and 500 and supervised IFA without constraints (-. -.).

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 source permutation. When more labeled samples are used, the permutations of the sources are avoided and the performances reach a more satisfactory level. Twenty random starting points were used for the GEM algorithm and only the best solution according to the likelihood was kept. This figure clearly highlights the benefit of using constraints when the amount of labeled samples is small. 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 improve the performances of the approach, particularly when the size of the labeled learning data is small. Further improvement of the overall performance level would require a non-linear model.

7 Conclusion

In this paper, we have proposed a method for learning parameters of the IFA model while incorporating two kinds of prior information related to the mixing process on the one hand, and the cluster membership of some training samples on the other hand. In this context, a criterion was defined and a GEM algorithm dedicated to its optimization was described. The proposed method has been applied to fault diagnosis in railway track circuits. The diagnosis system aims at recovering 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 has been carried out to show that our approach is able to take advantage of prior information, thus significantly improving estimation accuracy and removing indeterminacy of the unsupervised IFA such as permutation of sources. Further studies will be carried out to incorporate nonlinearity and also to take into account imprecise and uncertain cluster memberships such as supplied by human experts.

References

1. A. Hyvärinen, J. Karhunen, E. Oja. *Independent 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 NIPS Conference*, 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 satistics, 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. E. Côme Apprentissage de modèles génératifs pour le diagnostic de systèmes complexes avec labellisation douce et contraintes spatiales PhD thesis, Université de Technologie de Compiègne, 2009.
14. 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.
15. G. Shafer. A mathematical theory of evidence. *Princeton University Press*, 1976.
16. 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.
17. A. Debiolles, L. Oukhellou, T. Denoeux and P. Aknin. Output coding of spatially dependent subclassifiers in evidential framework. Application to the diagnosis of railway track-vehicle transmission system.. In *Proceedings of FUSION 2006*,Florence, Italy, July 2006.
18. G. J. Mclachlan and T. Krishnan, The EM algorithm and Extension Wiley, 1996.
19. A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing*. Wiley, 2002.