Semi-supervised IFA with prior knowledge on the mixing process. An application to a railway device diagnosis.

Etienne Côme, Cherfi Zohra, Latifa Oukhellou, Patrice Aknin

Semi-supervised IFA with prior knowledge on the mixing process
An application to a railway device diagnosis

Etienne Côme[(1)], Zohra Cherfi[(1)], Latifa Oukhellou[(1)(2)], Patrice Aknin[(1)]

[(1)] INRETS-LTN, 2 av du Gal. Malleret Joinville, 94114 Arcueil Cedex, France
[(2)] CERTES-Université Paris XII, 61, av du Gal de Gaulle, 94100 Créteil, France

{come,cherfi,oukhellou,aknin}@inrets.fr

Abstract

Independent Factor Analysis (IFA) is a well known method used to recover independent components from their linear observed mixtures without any knowledge on the mixing process. Such recovery is possible thanks to the hypothesis that the components are mutually independent and non-Gaussians. The IFA model assumes furthermore that each component is distributed according to a mixture of Gaussians. This article investigates the possibility of incorporating prior knowledge on the mixing process and partial knowledge on the cluster belonging of some samples to estimate the IFA model. In this way, other learning contexts can be handled such as semi-supervised or partially supervised learning. Such information is valuable to enhance estimation accuracy and remove indeterminacy commonly encountered in unsupervised IFA such as the permutation of the sources. The proposed method is illustrated by a railway device diagnosis application and results are provided to show its effectiveness for this type of problem.

1. Introduction

A considerable amount of research has been devoted to solving diagnosis problems that aim to assign any measurement signal represented by a feature vector to one of labeled classes. Research in this area has employed either discriminative methods which directly focus on learning class boundaries, or generative approaches that aim to model the underlying distributions of the classes. The choice between these two approaches is closely linked to the classification problem. The generative technique seems to be interesting when a prior knowledge about the measurement process or about the data could be incorporated.

Traditionally, two learning paradigms are possible: supervised learning and unsupervised learning. In the recent years, other paradigms have emerged to mix them as semi-supervised [1] or partially supervised learning [2]. In the former approach, one uses a mix of unlabeled and labeled examples, whereas in the latter, one can define constraints on the possible classes of the examples. A more general framework involving partially or imprecisely labeled samples can also be considered to handle situations where only imperfect knowledge on class labels is available. In this case, possibilist or belief function based labels can be used [3]. The importance for such problems comes from the fact that labeled data are often difficult to obtain, while unlabeled or partially labeled data are easily available.

In this article, we present a generative diagnosis approach which allows to taking advantage from prior knowledge on the dependencies between the latent variables (linked to the system defects) and the observed variables (features extracted from the measurement signal). The generative model involved here assumes that observed variables are generated by a linear mixture of independent and non Gaussians latent variables (sources). Furthermore, each latent variable is assumed to be non Gaussian but generated according to a mixture of Gaussians. In this context, the well known method so called Independent Factor Analysis which is traditionally used in signal processing can be applied to recover the independent components from observed variables [7] [8].

This generative model is often considered within unsupervised learning framework. Both the mixing process and the source densities are only learned from the observed data. The paper investigates the possibility of incorporating partial knowledge on the latent variables to estimate such model. In the general case, this partial information will be encoded by a Dempster-Shafer mass function over the set of clusters describing each source

but it can also be adapted to handle more specific learning problems such as the semi-supervised and the partially supervised cases. In this way, the mixture model of each source density will be supplied by the component origins of a subset of training samples. Such information is valuable to enhance estimation accuracy and remove indeterminacy commonly encountered in unsupervised independent factor analysis such as the permutation of the sources.

The approach will be illustrated applying it to a railway device diagnosis. This component can be considered as a complex system made up of a series of spatially related subsystems: the presence of a defect in one subsystem modifies not only its own signature but also those of subsystems located upstream. The aim of the diagnosis system is to identify the working state of the global component and localize the defective subsystem. This kind of system is present in many other applications such as electrical power distribution systems, gas or water distribution networks, telephone networks, road traffic,…

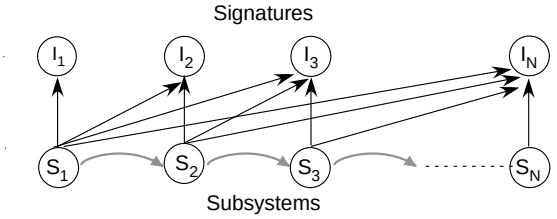

Figure 1. Diagnosis of a complex system made up of a series of spatially related subsystems S_1, S_2,…,S_N

The structure of the problem supplies therefore prior knowledge on the mixing process. As there is no influence between a subsystem S_i state and signatures of subsystems located upstream, some elements of the mixing matrix are null. This information will be introduced in the model estimation.

This article is organized as follows. We will first present Independent Component Analysis (ICA) and Independent Factor Analysis (IFA) as maximum likelihood estimation problems. The problem of learning the IFA model with prior knowledge on labels will then be addressed in Section 3. In Section 4, the method is applied to the diagnosis of a railway track circuit. The impact of using constraints on the mixing matrix and of semi-supervised learning will then be evaluated. The paper ends with a conclusion.

2. Independent Factor Analysis

2.1. Background on ICA

Independent Component analysis aims at recovering independent latent components $\{z_1,\ldots,z_S\}$ from their observed linear mixtures [4] [5]. In its general formulation the relationship between latent and observed variables takes the following form:

$$\mathbf{x} = A.\mathbf{z} + \varepsilon \tag{1}$$

where ε is a Gaussian noise independent from **z** and *A* a mixing matrix of size ($D \times S$). The independence assumption is translated to a factorization of the joint distribution:

$$p(\mathbf{z}) = \prod_{s=1}^{S} p_s(z_s) \tag{2}$$

The problem concerns the estimation of both the mixing matrix *A* and the realizations of the latent variables **z**. This general problem is handled differently depending on the number of the observed and latent variables *D* and *S*. If D > S the system is over-determined and a pre-processing is classically performed (by a Principal Component Analysis (PCA) to transform the observed variables **x** to *S* new variables, which also leads to remove the noise. After this pre-processing the noiseless ICA model is assumed to be:

$$\mathbf{x} = A.\mathbf{z} \tag{3}$$

with *A* a square matrix. The distribution of the observed variables is given by:

$$p(\mathbf{x}) = \frac{1}{\left|det(A)\right|} \prod_{s=1}^{S} p_s\left(\left(A^{-t}.\mathbf{x}\right)_s\right) \tag{4}$$

Considering a set of *N* samples, the relation (3) can be written in matrix form as:

$$\mathbf{Z} = \mathbf{X}.A^{-1} \tag{5}$$

where **X** is the data matrix of size ($N \times S$) and **Z** is the source matrix of size ($N \times S$). The log likelihood for the N observations has the form

$$\begin{aligned} &L\left(A;\{\mathbf{x}\}_1^N\right) = \\ &-N\,log\left(\left|det(A)\right|\right) + \sum_{i=1}^{N}\sum_{s=1}^{S} log\left(p_s\left(\left(A^{-t}.\mathbf{x}_i\right)_s\right)\right) \end{aligned} \tag{6}$$

Once particular source density models have been chosen, the estimation of the mixing matrix *A* can be performed by differentiating the log likelihood with respect to either *A* or the un-mixing matrix $W=A^{-1}$ its inverse:

$$\Delta A = N.A^{-t} - \Phi^t.\boldsymbol{Z} \tag{7}$$

where Φ is the score function of each sample $\Phi = \phi\left(\boldsymbol{X}\,A^{-t}\right)$, $\phi(\mathbf{z}) = \left(-\frac{\delta log\, p_1(z_1)}{\delta z_1}, \cdots, -\frac{\delta log\, p_S(z_S)}{\delta z_S}\right)$, and $\boldsymbol{Z}$ the matrix of estimated latent variables. In order to

take account of known constraints on the mixing process, the log likelihood will be differentiated with respect to *A* rather than to *W*.

The maximum likelihood mixing matrix could be found with an ascended gradient procedure. A better optimization strategy is obtained by using the natural gradient [6] in which the climbing direction is found by multiplying the gradient by a first order Hessian approximation. The updating of the matrix *A* is achieved by the following learning rule

$$A^{(q+1)} = A^{(q)} + \tau\, A^{(q)} \left(\Phi^t . \boldsymbol{Z} - N\, \boldsymbol{I} \right) \tag{8}$$

where τ is the learning rate that must be tuned, $\boldsymbol{I}$ the identity matrix of size (S x S).

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 [4]. An alternative solution investigated by several authors [7] [8], 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.

In IFA model, each latent variable is assumed to be distributed according to a mixture model given by:

$$p_s\left(z_s\right) = \sum_{k=1}^{K_s} \pi_k^s\, N\left(z_s ; \mu_k^s, \sigma_k^S\right) \tag{9}$$

where N refers to the unidimensional Gaussian distribution. The model parameters are the mixing matrix A and the parameters of latent variable distributions. The set of all model parameters is denoted $\theta = (A, \pi^1, \ldots, \pi^S, \theta^1, \ldots, \theta^S, \mu^1, \ldots, \mu^S, \sigma^1, \ldots, \sigma^S)$. π^S is the vector of clusters proportions of source s which sum to one, μ^S and σ^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 independently and identically distributed learning set use the likelihood function which can be obtained by substituting the density function p_s in (6) by its definition given in (9):

$$L\left(A; \{\mathbf{x}\}_1^N\right) = -N \log\left(\left|det\left(A\right)\right|\right) + \sum_{i=1}^{N} \sum_{s=1}^{S} log\left(\sum_{k=1}^{K_s} \pi_k^s\, N\left(A_s^{-t} . x_i, \mu_k^s, \sigma_k^s\right) \right) \tag{10}$$

As outlined previously a gradient climbing algorithm can be set up to optimize such function with respect to A if the source densities are frozen. Similarly, with A kept fixed this likelihood function can be optimized with respect to the parameters of each source model by means of an Expectation Maximization (EM) algorithm. These remarks lead to the definition of an alternating optimization algorithm which takes the form a Generalized Expectation Maximization algorithm (GEM) since the objective function is only increased during the maximization step [9]. It should be noted that two indeterminacies affect the ICA model: the scaling and the permutation of sources.

3. Partially-supervised learning in IFA

The IFA model is often used in unsupervised learning context. The idea that we investigate in this paper is to incorporate partial knowledge on the cluster belonging of some samples in the learning process. In this way, other learning contexts can be handled such as semi-supervised or partially supervised learning. For that purpose, an objective function generalizing the likelihood function needs to be defined and an EM algorithm dedicated to its optimization has to be set up.

3.1. Derivation of a Generalized likelihood criterion

We shall assume a learning set of the form $X^{iu} = \{(x_1, m_1), \ldots, (x_N, m_N)\}$, where $m_i = \left[m_i^1, m_i^2, \cdots, m_i^S \right]$ is a set of basic belief assignments or Dempster-Shafer mass functions [10] encoding our knowledge on the cluster belonging of sample i for source s. The set of all possible cluster for source s will be denoted $\mathcal{Y}^s = \{c_1, \ldots, c_{Ks}\}$. The observed data x_i will be assumed to be generated according to the IFA model defined in Section 2. Depending on the choice of the mass functions, this formulation can therefore be seen as addressing a more general issue which encompasses unsupervised, supervised and semi-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. Furthermore, selecting the simple hypothesis with highest plausibility given the observations is a natural decision strategy in the belief functions framework.

The proposed estimation principle to search the value of θ is based on the maximization of the conditional plausibility given the data [11]:

$$\hat{\theta} = \underset{\Theta}{arg\,max}\; pl^{\Theta}\left(\theta \,|\, \boldsymbol{X}^{iu}\right) \tag{11}$$

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	$m_i^s(c_k)=1$	$pl_{ik}^s=1,\ pl_{ik'}^s=0 \quad \forall k' \neq k$
Partially supervised	$m_i^s(C)=1$	$pl_{ik}^s=1 \quad if\ c_k \in C$ $pl_{ik}^s=0 \quad if\ c_k \notin C$
"soft" supervised	m_i^s ?	$pl_{ik}^s \in [0,1]$

Previous work on mixture model estimation with belief function based labels [11] using such principle can easily be extended to the IFA model. It can be shown that the logarithm of the conditional plausibility of the IFA parameters given the dataset can be expressed as:

$$L\left(A;\{\mathbf{x}\}_1^N\right) = -N\,log\left(\left|det\left(A\right)\right|\right) + \sum_{i=1}^{N}\sum_{s=1}^{S} log\left(\sum_{k=1}^{K_s} pl_{ik}^s\ \pi_k^s\ N\left(A_s^{-t}.x_i,\mu_k^s,\sigma_k^s\right)\right) \quad (12)$$

where pl_{ik}^s are the plausibilities of each cluster k of source s for each sample i. Once the criterion is defined, the remaining work concerns its optimization. The next section details this approach in the particular context of semi-supervised context.

3.2. Semi-supervised learning in IFA model

In semi-supervised learning approach, the IFA model is built from a combination of M labeled and N-M unlabeled samples. Consequently the log-likelihood can be decomposed in two parts corresponding, respectively, to the supervised and unsupervised learning examples. Criterion (12) can then be rewritten as:

$$L\left(A;\{\mathbf{x}\}_1^N\right) = -N\,log\left(\left|det\left(A\right)\right|\right) + \sum_{i=1}^{M}\sum_{s=1}^{S}\sum_{k=1}^{K_s} l_{ik}^s\ log\left(\pi_k^s\ N\left(A_s^{-t}.x_i,\mu_k^s,\sigma_k^s\right)\right) + \sum_{i=M+1}^{N}\sum_{s=1}^{S} log\left(\sum_{k=1}^{K_s}\pi_k^s\ N\left(A_s^{-t}.x_i,\mu_k^s,\sigma_k^s\right)\right) \quad (13)$$

with $l_{ik}^s \in \{0,1\}^{K_S}$ are binary variables encoding the class of sample i, $l_{ik}^s=1$ if sample i comes from cluster c_k, and $l_{ik}^s=0$ otherwise.

The GEM algorithm can easily be adapted to optimize this function. The modification affects only the E step, where the posterior probabilities t_{ik}^s are only computed for unlabeled observations. During the M step, the known labels are used instead of the t_{ik}^s, for the labelled data.

Algorithm GEM, pseudo-code for semi-supervised IFA

Inputs: centered data matrix $\{\mathbf{x_i}\}_{i=1\cdots N}$ and labels $\{l_{ik}^s\}_{i=1\ldots M\ ,k=1\ldots Ks}^{s=1\ldots S}$

Initialize parameters vector

$\theta^{(0)} = \left(A^{(0)},\pi^{1(0)},\ldots,\pi^{S(0)},\mu^{1(0)},\ldots,\mu^{S(0)},\sigma^{1(0)},\ldots,\sigma^{S(0)}\right)$

$q \leftarrow 0$

While increment in log likelihood > precision threshold **do**

$\mathbf{Z} = \mathbf{X}\,A^{-1}$ // Source Update

For all sources $s \in \{1,\ldots,S\}$

For all clusters $k \in \{1,\ldots,K_s\}$ **do** // E step

$t_{ik}^{s(q)} = l_{ik}^s \ ,\ i=1\cdots M$

$t_{ik}^{s(q)} = \dfrac{\pi_k^{s(q)}\ N(\mathbf{Z}_{is};\mu_k^{s(q)},\sigma_k^{s(q)})}{\sum_{k'=1}^{K_s}\pi_{k'}^{s(q)}\ N(\mathbf{Z}_{is};\mu_{k'}^{s(q)},\sigma_{k'}^{s(q)})} \ ,\ i=M+1\cdots N$

End for

End for

For all sources $s \in \{1,\ldots,S\}$

For all clusters $k \in \{1,\ldots,K_s\}$ **do** // M step

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

$$\boldsymbol{\mu}_k^{s(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{s(q)}}\sum_{i=1}^{N} t_{ik}^{s(q)}\mathbf{Z}_{is}$$

$$\sigma_k^{s(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{s(q)}}\sum_{i=1}^{N} t_{ik}^{s(q)}(\mathbf{Z}_{is}-\boldsymbol{\mu}_k^{s(q+1)})^2$$

End for

$\Phi_{.S} = \phi^{(q+1)}\left(\mathbf{Z}_{.S}\right)$ // Score function update

End for

$A^{(q+1)} = A^{(q)} + \tau\ A^{(q)}\left(\Phi^t.\mathbf{Z} - N\,I\right)$ // Mixing matrix update

$q \leftarrow q+1$

End while

Outputs : estimated model parameters

$\theta^{ML} = \left(A,\pi^1,\ldots,\pi^S,\mu^1,\ldots,\mu^S,\sigma^1,\ldots,\sigma^S\right)$

4. Railway application

4.1. Track circuit principle

The application considered in this paper concerns the automatic diagnosis of a railway track/vehicle transmission system. On high speed lines, signalling information is transmitted to the train driver using modulated currents that are injected into the rail and picked up by antennas mounted under the train. This system, called a *track circuit*, provides a great deal of information that is useful for train movement, such as the speed limit on a given track section.

So, the railway track is divided into different sections, each section is equipped by a track circuit consisting of (see Figure 2): 1) a transmitter connected to one of the two section ends that delivers a frequency modulated alternating current 2) the two rails that can be considered as a transmission line 3) at the other end of the track section, a receiver that detects the presence or the absence

of train on the section 4) trimming capacitors connected between the two rails at constant spacing to compensate for the inductive behaviour 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.

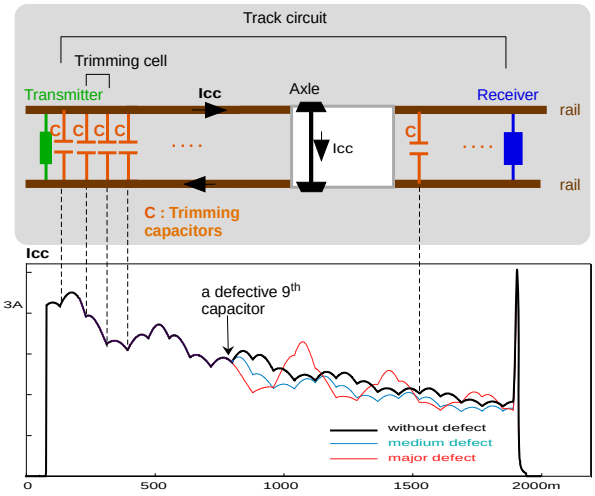

Figure 2. Track circuit representation and inspection signal without defect and with a defective 9th capacitor

4.2. Diagnosis purpose and methodology

The different parts of the system are subject to malfunctions (due to aging, or track maintenance operations) that must be detected as soon as possible in order to maintain the system at the required safety and availability levels. In the 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 defects that affect capacitor internal capacitance. Figure 2 shows examples of signals simulated along a 1500m track circuit: one of them corresponds to an absence of defect, while the others correspond to a defective 9th capacitor. The idea is to consider the track circuit as a global system S, and each trimming capacitor as a subsystem S_i. A defect on one subsystem is represented by a continue value of the capacitance parameter. The proposed diagnosis method must take account of a variable number of subsystems depending on the track circuit length and a spatial relationship between subsystems.

A generative model can be build where the observed variables are the coefficients of the local polynomial approximating the measuring signal located between two subsystems, and the latent variables are the capacitances of each trimming capacitor. In this way, the diagnosis of a shorter system (N'<N) simply uses a sub-model extracted from the global one and the model structure can also take advantages from prior knowledge (downstream-upstream dependencies) as shown in Figure 1 which leads to a block lower triangular mixing matrix.

5. Results and discussion

To access the performances of the method, 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 of each capacitor where 1000 are used in the training phase while the 1500 others are employed for the test phase to estimate the performances. A piecewise approach is then 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 generative model.

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, the S most correlated with the capacitances being kept and the others being considered as noise. We have investigated the influence on the method results of using different amount of labeled samples and imposing constraints on the mixing matrix.

The comparison between the different settings will be quantified by the correlation between the true capacitances and their estimates calculated on the test set. The results presented in Figure 3 show that the source permutation is avoided when the components origins of a sufficient amount of training samples is provided. The non detection rate is also decreased in this case since no capacitor is weakly correlated with its estimate.

Figure 4 shows the mean correlation between estimated latent variables and capacitor capacitances function of the number of labeled samples when the mixing matrix is constrained or not. Fifty random starting points were used for the GEM algorithm and only the best solution according to the likelihood was kept. To avoid the source permutations which occur when not enough labeled samples are supplied, the mean correlation was computed according to the best permutation of the sources.

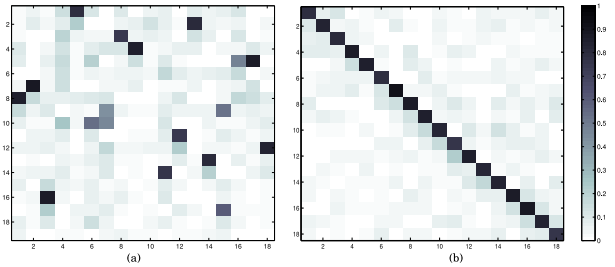

Figure 3. Correlation between estimated latent variables and true capacitor capacitances, computed on a test set for unsupervised IFA (a) and semi-supervised IFA with constraints on the mixing matrix and 40% of labeled samples among 1000 training samples (b).

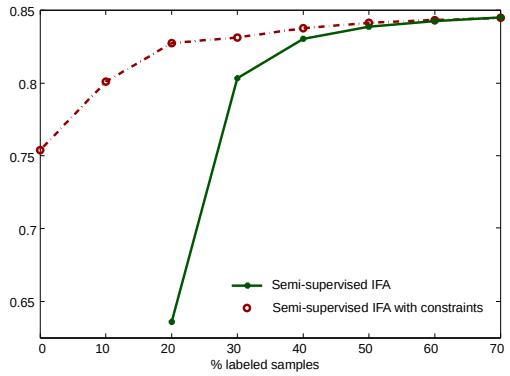

Figure 4. Mean correlation between estimated latent variables and true capacitor capacitances evaluated on a test set, function of the percentage of laleled samples among the 1000 training samples, for IFA with and without constraints on the mixing matrix.

As expected, when the number of labeled data increases, the mean correlation also increases. Furthermore, when prior information on the mixing matrix is provided, the sources are well estimated. The use of labeled data seems less influent when prior information such as constraints on the mixing matrix are available.

6. Conclusion

Here we presented a diagnosis method based on independent factor analysis which aims to recover the latent variables linked to the defects from their linear observed mixtures. In this paper, we investigate the possibility of incorporating knowledge on the cluster belonging of some samples and also on the structure of the mixing matrix to estimate the IFA model. A generalized maximum likelihood criterion was defined and a GEM algorithm dedicated to its optimization was given. The proposed approach was illustrated on a railway device diagnosis application. The results show that our solution is able to take advantage of information on class labels and on the mixing form. The benefits are in terms of source estimation accuracy and no permutation of sources. Further studies will be carried out to apply the approach on real signals in order to take account of imprecise and uncertain labels (label noise).

7. References

[1] O. Chapelle, B. Schölkopf, A. Zien (Eds.), Semi-Supervised Learning, MIT Press, Cambridge, Ma, 2006.

[2] C. Ambroise, G. Govaert, EM algorithm for partially known labels, in: Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS-2000), Springer, Namur, Belgium, 2000, pp. 161–166.

[3] T. Denoeux, L.M. Zouhal, Handling possibilistic labels in pattern classification using evidential reasoning, Fuzzy Sets and Systems 122 (3) (2001) 47–62.

[4] A. Hyvarinen, Independent Component Analysis, Inter-Science, Wiley, 2001.

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

[6] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In Advances in Neural Information Processing Systems, volume 8, pages 757-763. The MIT Press, 1996.

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

[8] E. Moulines, J. Cardoso, and E. Cassiat, Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models, In ICASSP, volume 5, pages 3617-3620, 1997.

[9] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, B 39:1-38, 1977.

[10] A. P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Annals of Mathematical statistics, 38:325-339, 1967.

[11] E. Côme, L. Oukhellou, T. Denoeux, and P. Aknin, learning from partially supervised data using mixture models and belief functions, Pattern recognition, accepted for publication, 2008.