Noname manuscript No.
(will be inserted by the editor)

Fault diagnosis of a railway device using semi-supervised independent factor analysis with mixing constraints

the date of receipt and acceptance should be inserted later

Abstract Independent factor analysis (IFA) defines a generative model for observed data that are assumed to be linear mixtures of some unknown non-Gaussian, mutually independent latent variables (also called sources or independent components). The probability density function of each individual latent variable is modelled by a mixture of Gaussians (MOG). Learning in the context of this model is usually performed within an unsupervised framework in which only unlabelled samples are used. Both the mixing matrix and the parameters of latent variable densities are learned from observed data. This paper investigates the possibility of estimating an IFA model in its noiseless setting when two kinds of prior information are incorporated; namely, constraints on the mixing process and partial knowledge on the cluster membership of some training samples. Semi-supervised or partially-supervised learning frameworks can thus be handled. The investigation of these two kinds of prior information was motivated by a real-world application concerning the fault diagnosis of railway track circuits. Results from both this application and simulated data are provided to demonstrate the capacity of our approach to enhance estimation accuracy and remove the indeterminacy commonly encountered in unsupervised IFA, such as source permutations.

Keywords Independent factor analysis · Semi-supervised learning · Mixing constraints · Mixture models · Maximum likelihood · Fault diagnosis

1 Introduction

Generative latent variable models aim to describe observed variables in terms of a set of unobservable (or latent) variables. Depending on the assumptions made on the latent and observed variable distributions, different kind of models can be distinguished, including principal component analysis (PCA) (17; 26), factor analysis (FA) (5; 14), and independent component analysis (ICA) (18; 6; 8). This paper deals with a particular model of this last family recently proposed by (2; 22) in the context of signal processing known as independent factor analysis (IFA).

The generative model involved in IFA assumes that observed variables are generated by a linear mixture of independent and non-Gaussian latent variables, as in the ICA model. Furthermore, it assumes that each individual latent variable has its own distribution, which is modelled semi-parametrically by a mixture of Gaussians (MOG). The general IFA model includes data that are noisy, or it can be considered in its noiseless setting. In the latter case, the model is similar to standard ICA but it includes a mixture of Gaussians model for the source densities.

Learning in the context of the IFA and ICA models is often considered within an unsupervised or blind framework. The model parameters and latent variables are learned exclusively from the observed data. These models yield reliable results, provided that the independence assumption is satisfied and the postulated mixing model is suited to the physics of the system. Otherwise, they may fail to recover the sources. Several extensions of the basic ICA model have been proposed to improve its performance, such kinds of approach are commonly denoted as semi-blind independent component analysis (4) for a review. The main approaches exploit nonlinear mixtures (16), temporal correlation (3), positivity (8) or sparsity (15; 19; 28). An approach to modelling class-conditional densities based on an IFA model was also recently

Address(es) of author(s) should be given

introduced in (21). The method proposed here fall in the broad class of semi-blind approach and may takes advantage of two kind of prior information

The method proposed here fall in this class of semi-blind approaches and may takes advantage of two kind of prior information through two extensions of the basic noiseless IFA model. The first one concerns the possibility of incorporating independence hypotheses with respect to a subset of 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 (5) and, more specifically, in the domain of structural equation modelling (7). The second extension incorporates additional information on the cluster membership of certain samples to estimate the IFA model. In this way, semi-supervised learning can be addressed. Considering the graphic model of IFA shown in Figure 1, the mixing process prior consists in omitting some connections between some set of observed variables X and some set of latent variables Z. The second prior means that additional information on the discrete latent variables Y is taken into account. Indeed, it can be seen that the IFA model provides two levels of interpretation corresponding to discrete and continuous latent variables. For each one of the S latent variables, the discrete latent variable Y_{si} encodes the cluster from which each sample i is drawn. Therefore, partial labelling over these discrete latent variables may help to improve the performance. Such an approach was very recently also explored in the context of Non-negative matrix factorisation using quite different tools by (27), but it is up to our knowledge the first attempt to use such information in the context of IFA.

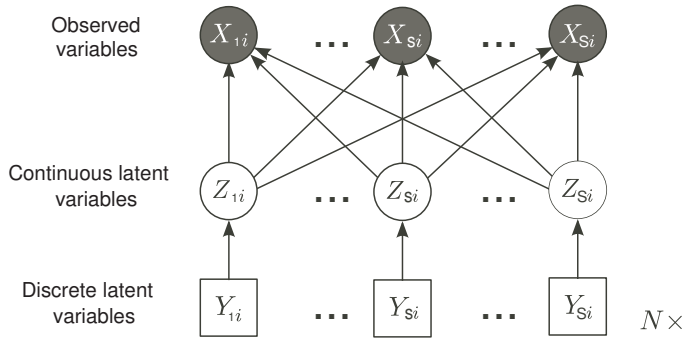

Fig. 1 Graphical model for Independent Factor Analysis.

This article is organised as follows. We first review the noiseless IFA model and show how it can be estimated using maximum likelihood. In Section 3, the problem of learning in the context of the IFA model with prior knowledge on the mixing process is addressed. Section 4 focuses on the problem of incorporating additional information on the cluster membership of some training samples to learn the model. A generalised maximum likelihood criterion is defined, and the algorithm for its optimisation is also detailed. In Sections 5 and 6, the approach presented here is applied to both simulated data and to a diagnosis problem through which the impact of using priors is evaluated. The paper ends with some concluding remarks on the analysis presented.

2 Noiseless IFA

2.1 Background on noiseless IFA

IFA was introduced by (22; 2). It originates from both ordinary factor analysis (FA) in applied statistics (24; 25) and independent component analysis (ICA) in signal processing (18; 6). IFA aims to recover independent latent variables from their observed linear mixtures. The latent variables are assumed to be mutually independent and non-Gaussian. In the noiseless form that is used throughout this paper, the IFA model can be expressed as:

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

where A is a square matrix of size $S \times S$, $\mathbf{x}$ is the random vector where elements $(\mathbf{x}_1, \ldots, \mathbf{x}_S)$ are the mixtures and $\mathbf{z}$ is the random vector where elements $(\mathbf{z}_1, \ldots, \mathbf{z}_S)$ are the latent variables. In this noiseless setting, the un-mixing matrix plays an important role, and we note that $W = A^{-1}$ in this paper. Thanks to

the noiseless setting, a deterministic relationship between the distributions of observed and latent variables can be expressed as:

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

Unlike the ICA model in which the probability density functions of the latent variables are fixed using prior knowledge or according to some indicator that allows switching between sub- and super-Gaussian densities (18), each latent variable density in IFA is modelled as a mixture of normally-distributed components (i.e., mixture of Gaussians MOG) so that a wide class of densities can be approximated (22; 2):

$$f^{\mathcal{Z}_s}(z_s) = \sum_{k=1}^{K_s} \pi_k^s \varphi(z_s; \mu_k^s, \nu_k^s), \tag{3}$$

Note that $\varphi(.; \mu, \nu)$ denotes a normal density function with vector mean μ and variance ν. The Equation (3) means that each latent variable is described as a mixture of K_s Gaussians with mean μ_k^s, variance ν_k^s and mixing proportions π_k^s.

2.2 Noiseless IFA model Learning

The learning problem associated with the IFA model consists in estimating both the mixing matrix A and the MOG parameters from the observed variables alone. Consider an i.i.d random sample of size N. Using the Equation (2) under the latent variable independence hypothesis, the log-likelihood has the form:

$$\mathcal{L}(A; \mathbf{X}) = -N \log(|\det(A)|) + \sum_{i=1}^{N} \sum_{s=1}^{S} \log\left(f^{\mathcal{Z}_s}\left((A^{-1}\mathbf{x}_i)_s\right)\right), \tag{4}$$

By substituting the density distribution by its expression given in (3), the log-likelihood can be rewritten as:

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

Note $\boldsymbol{\psi}$ is the IFA parameter vector $\boldsymbol{\psi} = (A, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^S, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^S, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^S)$, A is the mixing matrix; $\boldsymbol{\pi}^j$ is the vector of mixing proportions of source j that sum to 1; and $\boldsymbol{\mu}^j$ and $\boldsymbol{\nu}^j$ are vectors of size K_j containing the means and the variances of each cluster, respectively.

All IFA parameters can therefore be estimated by maximising the likelihood function. When the latent variable densities are known, the mixing matrix estimation is based on gradient methods that maximize the likelihood. The stochastic gradient of the log-likelihood defined in (5) can be derived as:

$$\begin{aligned}
\Delta A \propto \frac{\partial \mathcal{L}(A; \mathbf{X})}{\partial A} &\propto -N.(A^{-1})^t + \sum_{i=1}^{N} (A^{-1})^t \mathbf{g}\left(A^{-1}.\mathbf{x}_i\right) (A^{-1}\mathbf{x}_i)^t \\
&\propto -(A^{-1})^t + \frac{1}{N} \sum_{i=1}^{N} (A^{-1})^t \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i^t \\
&\propto (A^{-1})^t \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i^t - \mathbf{I}\right),
\end{aligned} \tag{6}$$

where $\mathbf{z}_i = A^{-1}\mathbf{x}_i$ and $\mathbf{g}(\mathbf{z}) = \left[\frac{-\partial \log(f^{\mathcal{Z}_1}(z_1))}{\partial z_1}, \ldots, \frac{-\partial \log(f^{\mathcal{Z}_S}(z_S))}{\partial z_S}\right]^t$.

The update rule of the mixing matrix is thus given by:

$$A^{(q+1)} = A^{(q)} + \tau \Delta A^{(q)}, \tag{7}$$

Note that τ is the gradient step that can be adjusted by means of line search methods such as the backtracking line search method (23, page 37), which decrease the learning rate by τ by a factor $\rho \in [0, 1]$ until sufficient decrease condition are met.

The convergence of this algorithm can be improved by using the natural gradient, which is based on the geometrical structure of the parameter space. This can be obtained by multiplying the left-hand side by the matrix AA^t, which leads to the following mixing matrix update rule (18):

$$\Delta_{nat} A = AA^t \Delta A = A \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i^t - \mathbf{I} \right). \tag{8}$$

with $\mathbf{z}_i^{(q)} = W^{(q)} \mathbf{x}_i$ and I the identity matrix.

The maximum likelihood of all model parameters can be achieved by an alternating optimisation strategy. The gradient algorithm is indeed well suited to optimise the log-likelihood function with respect to the mixing matrix A when the parameters of the source marginal densities are held constant. In contrast, when A is fixed, an EM algorithm can be used to optimise the likelihood function with respect to the parameters of each source. These observations have led to the development of a generalised EM algorithm (GEM) that simultaneously maximises the likelihood function with respect to all model parameters (13; 20).

3 Constraints on the mixing process

This section investigates the possibility of incorporating independence hypotheses concerning relationships between some subset of 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, namely, $X_h \perp\!\!\!\perp Z_g$, which means that X_h is statistically independent from Z_g. This kind of hypothesis constrains the form of the mixing matrix, as shown by the following proposition:

Proposition 1 *In the noiseless ICA model,*

$$X_h \perp\!\!\!\perp Z_g \Leftrightarrow A_{hg} = 0. \tag{9}$$

Proof. Independence can be defined as:

$$X_h \perp\!\!\!\perp Z_g \Leftrightarrow f^{\mathcal{X}_h \times \mathcal{Z}_g}(x_h, z_g) = f^{\mathcal{X}_h}(x_h)\ f^{\mathcal{Z}_g}(z_g). \tag{10}$$

In the case of the noiseless IFA model, the joint probability density function on $\mathcal{X}_h \times \mathcal{Z}_g$ is given by:

$$\begin{aligned} f^{\mathcal{X}_h \times \mathcal{Z}_g}(x_h, z_g) &= \int_{\mathbb{R}^{S-1}} f^{\mathcal{X}_h \times \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_S}(x_h, z_1, \ldots, z_S) \prod_{l=1, l\neq g}^{S} \mathrm{d}z_l & (11) \\ &= \int_{\mathbb{R}^{S-1}} \prod_{s=1}^{S} f^{\mathcal{Z}_s}(z_s) \times \delta(x_h - A_{h.}\mathbf{z}) \prod_{l=1, l\neq g}^{S} \mathrm{d}z_l \\ &= f^{\mathcal{Z}_g}(z_g) \times \left(\int_{\mathbb{R}^{S-1}} \prod_{l=1, l\neq g}^{S} f^{\mathcal{Z}_l}(z_l) \times \delta(x_h - A_{h.}\mathbf{z}) \mathrm{d}z_l \right), & (12) \end{aligned}$$

Note that $A_{h.}$ is the h^{th} row of the mixing matrix A. Using the independence assumption, we identify:

$$X_h \perp\!\!\!\perp Z_g \Leftrightarrow f^{\mathcal{X}_h}(x_h) = \int_{\mathbb{R}^{S-1}} \prod_{l=1, l\neq g}^{S} f^{\mathcal{Z}_l}(z_l) \times \delta(x_h - A_{h.}\mathbf{z}) \mathrm{d}z_l, \tag{13}$$

Note that δ is the Dirac function. The integral must not depend on z_g (the g^{th} row of $\mathbf{z}$), which is possible only if A satisfies $A_{hg} = 0$. □

The log-likelihood must be maximised under the constraint that some of the mixing coefficients are null, and the gradient ascent is only performed with respect to the non-null coefficients. In this case, the initialisation and the update rules of the mixing matrix are given by:

$$\begin{aligned} A^{(0)} &= C \bullet A^{(0)} \\ A^{(q+1)} &= A^{(q)} + \tau\, C \bullet \Delta A^{(q)}, \end{aligned} \tag{14}$$

Note that $\bullet$ denotes the Hadamard product between two matrices, and C is a binary matrix defined by $C_{hk} = 0$ if $Z_k \perp\!\!\!\perp X_h$, and $C_{hk} = 1$, otherwise.

4 Semi-supervised learning in noiseless IFA

The expectation maximization (EM) algorithm provides a general solution to problems involving missing data (13). Here, we are interested in one of its most classical application, which concerns mixture estimation problems.

4.1 The GEM algorithm for semi-supervised IFA learning

In a semi-supervised learning context, the IFA model is built from a combination of M labelled and $N-M$ unlabelled samples (10; 11). The maximum likelihood criterion can be decomposed into two parts respectively corresponding, to the supervised and unsupervised learning examples; the log-likelihood criterion then becomes:

$$\mathcal{L}(A;\mathbf{X}) = -N\log(|\det(A)|)+ \sum_{i=1}^{M}\sum_{s=1}^{S}\sum_{k=1}^{K_s} l_{ik}^s \log\left(\pi_k^s \varphi\left((A^{-1}\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right) + \sum_{i=M+1}^{N}\sum_{s=1}^{S}\log\left(\sum_{k=1}^{K_s}\pi_k^s\varphi\left((A^{-1}\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right). \quad (15)$$

Note that $l_{ik}^s \in \{0,1\}^{K_s}, l_{ik}^s = 1$ if sample i comes from component c_k of sources s and $l_{ik}^s = 0$ otherwise.

A generalised EM algorithm (GEM), (also noted here as Algorithm 1) can be designed to simultaneously maximise the likelihood function with respect to all model parameters. This algorithm is therefore similar to the EM algorithm used to estimate IFA parameters in an unsupervised setting, except for the E-step during which the posterior probabilities t_{ik}^s are only computed for the unlabelled samples. The score function g of each latent variable density is thus given by:

$$\begin{aligned} g_s(z_{is}) &= \begin{cases} \frac{-\partial\log\left(\sum_{k=1}^{K_s} l_{ik}^s \pi_k^s \varphi(z_{is};\mu_k^s,\nu_k^s)\right)}{\partial z_{is}}, \text{ if } i \le M \\ \frac{-\partial\log\left(\sum_{k=1}^{K_s} \pi_k^s \varphi(z_{is};\mu_k^s,\nu_k^s)\right)}{\partial z_{is}}, \text{ if } i > M \end{cases} \\ &= \begin{cases} \sum_{k=1}^{K_s} l_{ik}^s \frac{(z_{is}-\mu_k^s)}{\nu_k^s}, \text{ if } i \le M \\ \sum_{k=1}^{K_s} t_{ik}^s \frac{(z_{is}-\mu_k^s)}{\nu_k^s}, \text{ if } i > M \end{cases} \end{aligned} \quad (16)$$

Note that t_{ik}^s is the posterior probability that the sample i belongs to component k of the latent variable s, given $z_{is} = (W\mathbf{x}_i)_s$ and the labels:

$$t_{ik}^s = \frac{\pi_k^s \varphi(z_{is};\mu_k^s,\nu_k^s)}{\sum_{k'=1}^{K_s}\pi_{k'}^s\varphi(z_{is};\mu_{k'}^s,\nu_{k'}^s)}. \quad (17)$$

The proposed algorithm which deals with mixing constraints and semi-supervised datasets takes as input a centred observation matrix $\mathbf{X}$ (of size $N \times S$), the available labels : $l_{ik}^s, \forall i \in \{1,\dots,M\}, \forall s \in \{1,\dots,S\}$ and $\forall k \in \{1,\dots,K_s\}$ and a constraints matrix (a binary matrix of size $S \times S$) with zeros indicating the desired independence assumptions.

Algorithm 1 Pseudo-code for Semi-supervised IFA with GEM algorithm

*Inputs:*Centred observation matrix $\mathbf{X}$, cluster membership for the M labelled data l_{ik}^s, constraints matrix C

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

while convergence test **do**

`//Latent variable update`

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

`//Update of the latent variable parameters (EM)`

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

`//E-Step`

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

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

end for

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

`//M-step, Update of the parameter vector of each latent variable`

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

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

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

end for

`//Update of the score matrix G (16)`

$\mathbf{G} = \mathbf{g}^{(q+1)}(\mathbf{Z})$

`//Natural gradient (8)`

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

`//Linear search` τ `(gradient step)`

$\tau^* = \text{Linearsearch}(A^{(q)}, \Delta A)$

`//mixing matrix update`

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

`//Latent variable normalization to remove scale indetermination`

for all $s \in \{1, \ldots, S\}$ **do**

$\sigma_s^2 = \sum_{k=1}^{Ks} \pi_k^{s(q+1)} (\nu_k^{s(q+1)} + \mu_k^{s(q+1)^2}) - \left(\sum_{k=1}^{Ks} \pi_k^{s(q+1)} \mu_k^{s(q+1)}\right)^2$

for all $k \in \{1, \ldots, K_s$ **do**

$\mu_k^{s(q+1)} = \mu_k^{s(q+1)} / \sigma_s$

$\nu_k^{s(q+1)} = \nu_k^{s(q+1)} / \sigma_s^2$

$A_{s.}^{(q+1)} = A_{s.}^{(q+1)} / \sigma_s$

end for

end for

$q \leftarrow q + 1$

end while

*Outputs:*Estimated parameters : $\widehat{\boldsymbol{\psi}}^{ml}$, estimated latent variables : $\widehat{\mathbf{Z}}^{ml}$

Remark 1 The convergence of the algorithm can be checked in different ways. The increment in log likelihood can be monitored and a test such as:

$$\frac{\mathcal{L}(\boldsymbol{\psi}^q; \mathbf{X}) - \mathcal{L}(\boldsymbol{\psi}^{(q-1)}; \mathbf{X})}{|\mathcal{L}(\boldsymbol{\psi}^{(q-1)}; \mathbf{X})|} < \epsilon, \tag{18}$$

where ϵ is a precision threshold set to a small value (10^{-6}), can be used to check the convergence. Another convergence test can be based on differences between successive estimates of the parameters.

In the following sections, we study the performance of semi-supervised noiseless IFA with mixing constraints using both simulated and real data sets. We compare our results with the results of classic unsupervised IFA. The experiments were designed to illustrate the capability of the proposed approach to enhance estimation accuracy and remove the indeterminacy commonly encountered in unsupervised IFA, such as indeterminacy related to permutation of sources.

5 Experiments using simulated data

To better understand our approach as compared to the unsupervised IFA model, different experiments were carried out to show the influence of incorporating prior information on estimation accuracy as well

as on the complexity of the optimisation problem. The first set of experiments was designed to show the impact of mixing constraints on estimation accuracy. In the second set of experiments, we investigate the benefits of learning the IFA model when information regarding the component membership of some training samples is introduced. We show that such information can be exploited to enhance estimation accuracy and remove indeterminacy commonly encountered in unsupervised IFA, such as the indeterminacy due to permutation of sources.

Several simulated data sets were built as follows. Six independent latent variables were considered with the densities shown in Figure 2. The mixtures were then generated using the IFA model given in (1), where each coefficient of the 6×6 mixing matrix was randomly generated according to a standard normal distribution. We aimed to recover six latent variables from six observed ones.

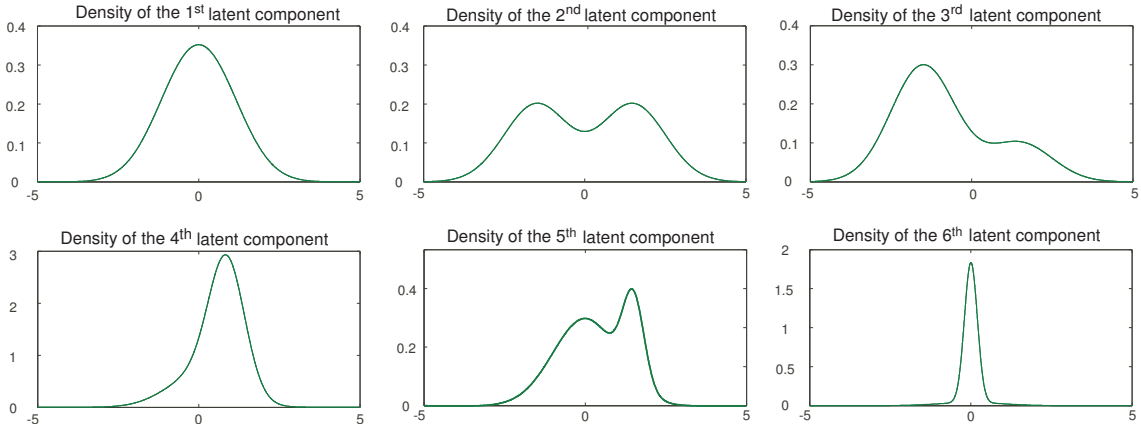

Fig. 2 Simulated sources densities.

5.1 Influence of mixing constraints on estimation accuracy

We now present simulation results that highlight the practical usefulness of introducing mixing constraints when estimating the IFA model. The mixing coefficients have been randomly generated according to a standard normal distribution, and zeros have been randomly introduced in the mixing matrix; each coefficient has a probability equal to 0.3 of being null. The experiment aims to compare the results of the classic noiseless IFA and the constrained IFA in which prior information on the mixing process is incorporated. The comparison has been made on data sets of different size $N \in \{100, \ldots, 700\}$ and assessed using the Amari performance index (9). This index is a classical measure of performance in the context of ICA, and it measures the closeness of the estimated un-mixing matrix to the inverse of the mixing matrix, which is assumed to be known:

$$A_p(Q) = \frac{1}{S}\sum_{i=1}^{S}\left(\left(\sum_{j=1}^{S}\frac{|Q_{ij}|}{max(|Q_{i.}|)} - 1\right) + \left(\sum_{j=1}^{S}\frac{|Q_{ji}|}{max(|Q_{.i}|)} - 1\right)\right), \tag{19}$$

$$Q = A\widehat{W}, \tag{20}$$

Note that A is the exact mixing matrix and $\widehat{W} = \widehat{A}^{-1}$ is the estimated un-mixing matrix. The lower the Amari index is, the more accurate the latent variable estimation is.
The Amari indexes obtained for different sample sizes and for the 2 noiseless IFA models (constrained IFAc and without mixing constraints IFA) are presented in Figure 3. The results were averaged over 30 different training sets, from which 25 random starting points were used for the GEM algorithm. Only the best solution according to likelihood was kept. The results show that constrained IFA (IFAc) outperforms classic IFA regardless of the sample size; the index improvement is indeed approximately 0.5.

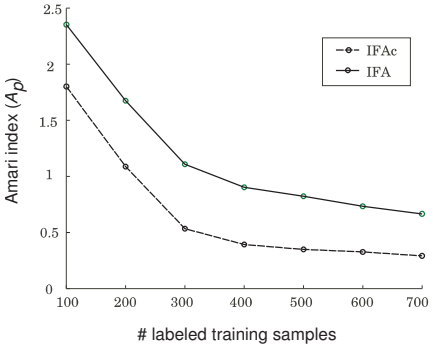

Fig. 3 The Amari performance index function of the training sample size obtained using classic IFA (IFA) and the IFA with prior information on mixing matrix (IFAc).

5.2 Influence of labelling on estimation accuracy

The next experiment aims to illustrate the influence of the number of labelled samples on IFA model performance. At first, three different learning strategies were compared, namely, unsupervised IFA, semi-supervised IFA using 25% of labelled samples for three sources and semi-supervised IFA using 25% of observations labelled over all sources. Figure 4 shows the performance measures obtained for these strategies when 500 simulated samples are used during the learning phase. The performances were quantified using the absolute correlation between the true sources and their estimates calculated based on a test set of $5,000$ samples. It can be seen that permutation is avoided for sources when labelled samples are provided; correlation is close to 1 in the diagonal terms. This result is because, in the semi-supervised setting, the likelihood is not invariant under permutations of the sources, as the numbering of the source is fixed by the known labels. This avoids the a posteriori analysis usually employed to identify the correspondences between data and sources. Visually, one also sees that the results are improved when semi-supervised IFA is employed, as a better contrast in the correlation matrix is achieved.

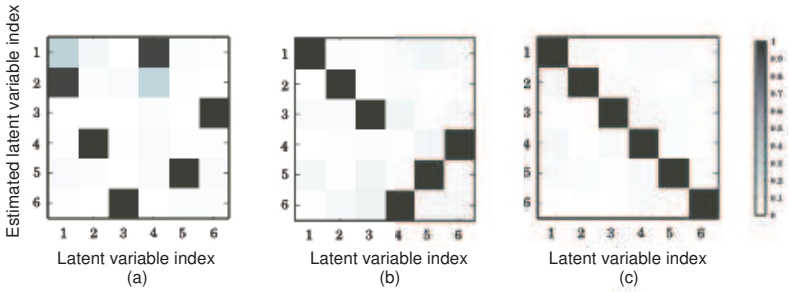

Fig. 4 Absolute correlation between estimated and latent variables, computed on a test set for (a) unsupervised IFA, (b) semi-supervised IFA with 25% labelled samples for three sources (b) and (c) semi-supervised IFA with 25% of samples labelled over all sources (c).

Figure 5 provides the Amari performance index (1) and the Pearson's correlation cœfficient r^2 between the true latent variables and their estimates when the number of labelled samples over all sources varies between 5% and 50%. These results were computed on 30 different learning data sets of 500 samples each. Fifty initialisations were performed for the GEM algorithm, and only the best solution according to likelihood was kept to avoid the problem of local maxima. 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 labelled samples increases (i.e., to greater than 20%), the model behaves better, as both the mean correlation and the Amari performance index are significantly improved. When there is not enough prior information (i.e., less than 20%) provided to the model, the variability of the r^2 indicator seems to be important, which is mainly because the GEM algorithm converges to local maxima corresponding to the permutations of sources.

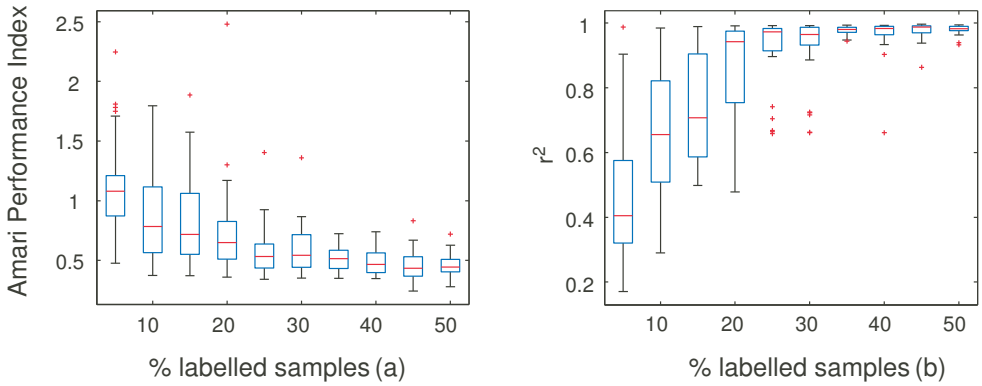

Fig. 5 Influence of the proportion of labelled samples on the estimation of the semi-supervised IFA model: Boxplot of the Amari perfomance index and correlation cœfficient function of the percentage of labelled samples.

5.3 Influence of labelling on the likelihood landscape

The following experiment was designed to show the potential benefit of incorporating labels in terms of simplification of the optimisation problem. Figure 6(a) displays the computation time required for the GEM algorithm convergence function of the amount of labelled samples. These results were computed on 30 different learning data sets of 500 samples each. It can be seen that the computation time (or the number of iterations) decreases when the amount of labelled samples increases. The number of different local maxima found by the GEM algorithm was also counted for 100 random initialisations. This experiment was repeated 10 times with different data sets of 500 samples each, that have been generated by the IFA model with the latent variable densities shown on Figure 2. The labelling concerns all the sources. Figure 6(b) displays the number of local maxima[1] function of the amount of labelled samples used in the training phase. It can be seen that the number of local maxima exponentially decreases when the amount of labelled samples increases. This consideration has high practical interest as the problem of local maxima is very important in the unsupervised learning context.

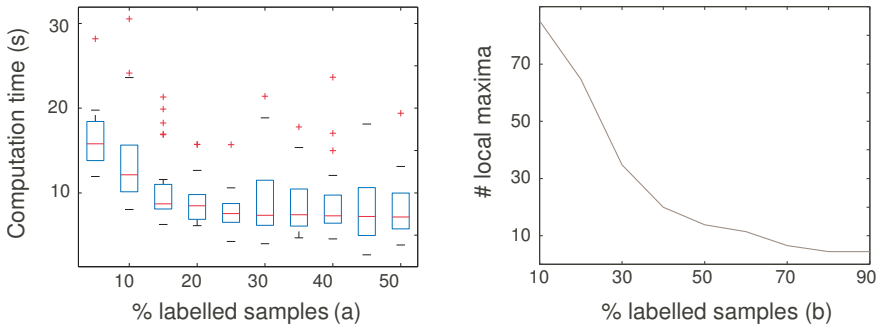

Fig. 6 Influence of the proportion of labelled samples on the likelihood landscape: Boxplot of the computation time function of the percentage of labelled samples and number of local maxima detected over 100 random initializations for the GEM algorithm, as a function of labelled samples.

6 Fault Diagnosis in Railway Track Circuits

The application considered in this paper concerns fault diagnosis in railway track circuits. This device is first described and the problem to be addressed is then explained. Finally, an overview of the proposed diagnosis method is presented.

[1] Two local maxima were supposed to be distinct if the distance between the vectors containing all the parameters was larger than 0.001.

6.1 The track circuit principle

The track circuit is an essential component of the automatic train control system (12). 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 and vehicle transmission system. It uses a specific carrier frequency to transmit coded data to the train regarding, for example, the maximum authorised 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 (see Figure 7):

- 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.

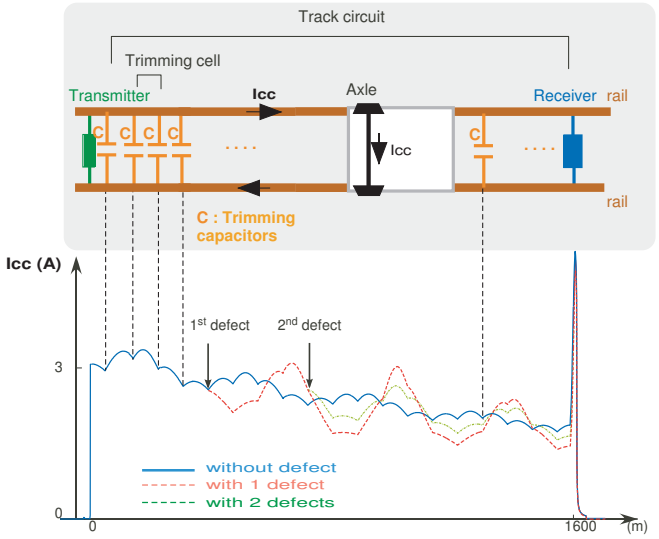

Fig. 7 Diagram of a track circuit and example of a denoised inspection signal.

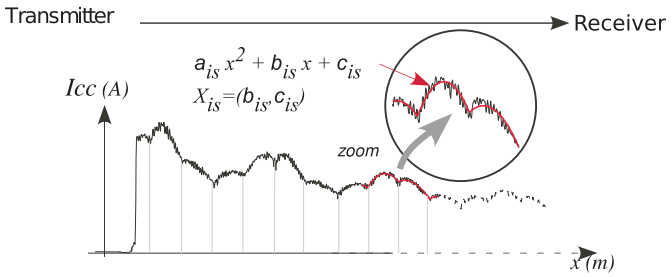

Fig. 8 Polynomial parametrization of inspection signal.

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 the track circuit signal due to shorting by train wheels. The drop in the received signal below a preset threshold indicates that

the section is occupied. In order to make the transmitted information specific to each track section and to minimize the influence of both longitudinal interference and transverse crosstalk, four frequencies are used for adjacent and parallel track circuits. Neighbouring track circuits are also isolated electrically by using tuned circuits (capacitor and inductance) on both the transmitter and the receiver.

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, a malfunction may cause an unfortunate attenuation of the transmitted signal that leads to the stopping of the train. The purpose of diagnosis is to inform maintainers about track circuit failures based on an analysis of a specific current as recorded by an inspection vehicle. This paper will focus on trimming capacitor faults that affect their capacitance. Figure 8 shows an example of the inspection signal when the system is fault-free, while the other figures correspond to a defective 5^{th} capacitor and defective 5^{th} and 9^{th} capacitors. The aim of the diagnosis system is to detect faults in the track circuit and localise the defective capacitor by analysing the measurement signal.

6.2 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 8):

- 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 $ax^2 + bx + c$;
- Each observed arch is influenced by the capacitors located upstream, that is, on the transmitter side.

The proposed method consists in extracting features from the measurement signal and building a generative model as shown in Figure 9 in which each observed variable X_{is} corresponds to cœfficients (b_{is}, c_{is}) of the local polynomial approximating the arch located between two capacitors. Only two coefficients are used because of continuity constraints between each polynomial, as there exists a linear relationship between the third cœfficient and the three cœfficients of the previous polynomial. The continuous latent variable Z_{is} of the generative model is linked the capacitance of the i^{th} capacitor and the discrete latent variable Y_{is} corresponds to the categorisation of the capacitor's state into one of the three states, namely, fault-free, minor defect, and 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 is also introduced in the model estimation by using constraints on the mixing matrix.

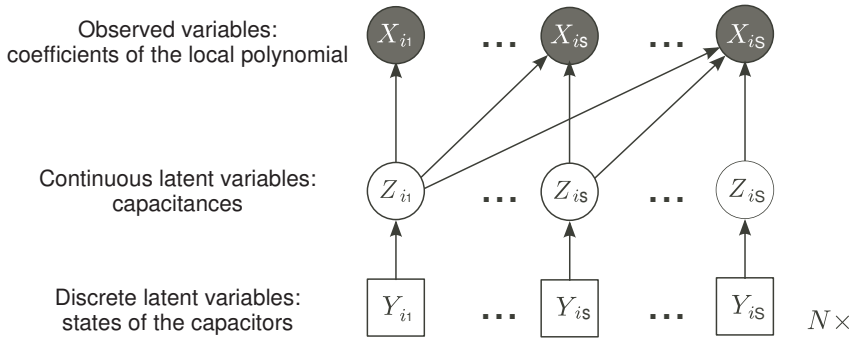

Fig. 9 A generative model for the diagnosis of track circuits represented by a graphical model including both continuous and discrete latent variables.

We can clearly see that this model is closely linked to the IFA model represented 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 source parameters. When the parameters are estimated it's possible to recover the latent components (i.e., capacitances) and the estimated capacitor's state from the observed variables alone. The

capacitances are estimated by computing $A^{-1}\mathbf{x}_i$ and the most probable capacitors state is estimated by setting it to the source component with maximum a posteriori probability.

Moreover, learning the IFA model with mixing constraints and component membership of some training samples may also be considered during the learning phase. As already explained, a piecewise approach is adopted for signal representation; each arch is approximated by a second degree polynomial in which two cœfficients are used as observed variables for each node in the model of Figure 1, resulting in $2 * S$ observed variables. Given an observation matrix, the aim is to recover S latent variables from $2*S$ observed variables with the hope that they will be strongly correlated with the variable of interest, i.e., capacitance. As prior information on the mixing matrix is available, PCA cannot be used for preprocessing because the mixing structure would be lost. Note that $2 * S$ 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 modelled by simple Gaussian distributions.

Remark 2 We note that with standard IFA model, S latent variables can be recovered from $2 \times S$ observed ones. However, in this paper, we consider the noiseless IFA model, which is a straightforward way to incorporate prior information and to recover sources from the data. In this case, the number of latent variables must be equal to the number of observed variables.

6.3 Results and discussion

To assess the performance of this approach, we considered a track circuit of $S = 18$ subsystems (i.e., capacitors) and a database containing $2,500$ noised signals obtained for different values of the capacitance of each capacitor. We note that 500 signals were used in the training phase, while an additional $2,000$ signals were employed for the test phase. The experiments aim to illustrate the influence of both the number of labelled samples and the use of the mixing matrix constraints on the results. The signals considered here are obtained for only one frequency (four frequencies are used for adjacent and parallel track circuits). The experiments given at this frequency will indeed show the number of observations that have to be labelled in order to obtain satisfactory performance. For the other frequencies, the idea is to label afterwards only this amount of observations since this task is cost effective.

The model provides two levels of interpretation corresponding to discrete and continuous latent variables. We first discuss the results for the continuous latent variables and then discuss results for the discrete latent variables.

Three different learning settings are compared in this section :

- supervised IFA without constraints : in this setting only labelled training data are used to learn the IFA model.
- semi-supervised IFA without constraints : in this setting labelled data are completed with the remaining training samples (over the 500 training signals) this part of the training dataset is unlabelled.
- semi-supervised IFA with constraints : in this settings unlabelled training data are used and the constraints on the mixing matrix are also used.

Furthermore, the number of labelled data was varied between 0 and 500. Note that when the number of labelled samples is equal to zero the case of semi-supervised IFA without constraints illustrates the performance of the traditional IFA model (without any prior), which are indeed very poor in this task. Figure 10(a) shows the mean of the absolute value of the correlation between estimated latent variables and capacitances as a function of the number of labelled training samples when the mixing matrix is both constrained and not constrained. When more labelled samples are used, the performances reach a more satisfactory level. For all the settings, 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 labelled samples is small. As expected, when the number of labelled data increases, the mean correlation also increases to reach a maximal value of 0.84 for the constrained IFA model with 250 labelled sampled and for the unconstrained one with 350 labelled samples. When a sufficient amount of labelled 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 unlabelled samples improve the performances of the approach, particularly when their are few labelled data available. In such case, the unlabelled data are of great help in the estimation process. Figure 10(b) shows the boxplot of the

computation time as the proportion of labelled samples increases from 0 to 500. It can be seen that both the mean value and the variability of the computation time decrease when the amount of labelled samples increases. Further improvement of the overall performance level would require a non-linear model.

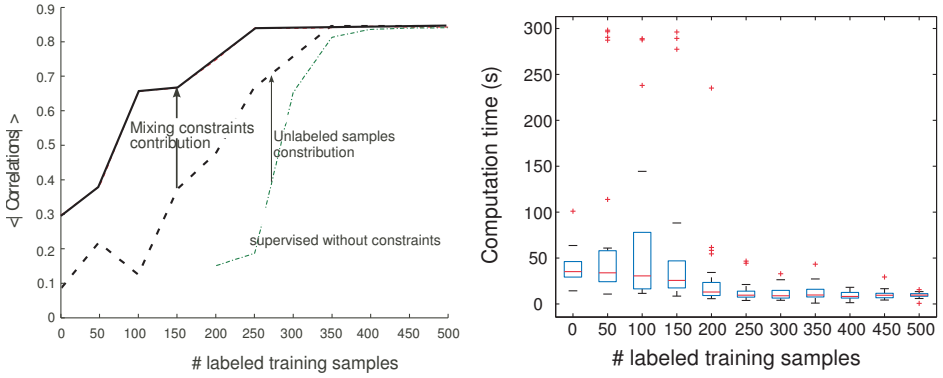

Fig. 10 Influence of prior information : (a) Results of semi-supervised IFA with constraints (–), without constraints (- -) and of supervised IFA without constraints (-. -.) when the number of labelled samples varies between 0 and 500. (b) Boxplot of the computation time function of the number of labelled samples for semi-supervised IFA without constraints.

In the following section, the results are analysed according to the discrete latent variables. With regards to the three classes (i.e., fault-free ω_0, minor defect ω_1, and major defect ω_2), the confusion matrix between the true and the estimated classes for all capacitances observed in the data set is reported in Table 1 when 30% of samples are labelled. The results show that a good classification performance is obtained despite some confusions between neighbourhood classes (i.e., between ω_0 and ω_1 and ω_1 and ω_2). A detection matrix can be also computed by combining the two defect cases (i.e., $\omega_1 \cup \omega_2$), which allows for an evaluation of the usual indicators such as false alarm (FA), correct detection (CD) and false detection (FD) rates.

In order to accurately quantify the affect of the proportion of labelled samples, the evolution of both false alarm and correct detection rates is presented as a function of the number of labelled samples when the mixing matrix is either constrained or not constrained (Figure 11). When the amount of labelled samples provided to the model is small (i.e., less than 100), the false alarm rate is too high to exploit the results. However, when the proportion of labelled samples is sufficiently large to obtain a weak false alarm rate (i.e., 1.5%), the benefit of prior information on the mixing constraint and the cluster membership of some samples is clearly visible. With only 250 labelled samples, 92% correct detection is reached. Further improvement of the overall performance level would require a non-linear model.

	ω_0	ω_1	ω_2
D_0	98.45	11.24	1.73
D_1	1.33	75.25	25.18
D_2	0.22	13.51	73.09

Table 1 Confusion matrix between true classes (ω_0,ω_1,ω_2) and their estimates (D_0,D_1,D_2), computed on the test set of 2000 track circuits when 30% of samples are labelled.

$$CD = nb_{(\omega_1 \cup \omega_2, D_1 \cup D_2)} / nb_{(\omega_1 \cup \omega_2)} = 92.43\% \quad (21)$$

$$FA = nb_{(\omega_0, D_1 \cup D_2)} / nb_{\omega_0} = 1.55\% \quad (22)$$

$$FD = nb_{(\omega_0, D_1 \cup D_2)} / nb_{D_1 \cup D_2} = 24.22\% \quad (23)$$

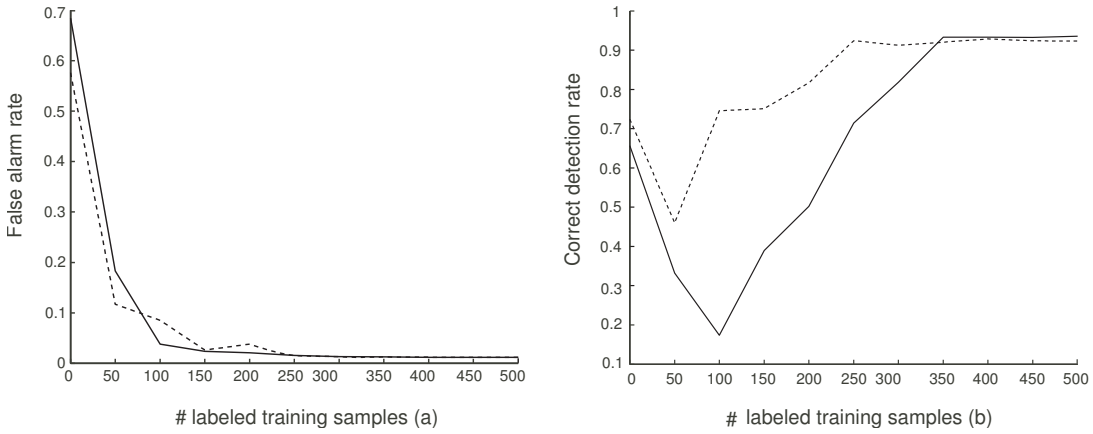

Fig. 11 Evolution of false alarm (a) and correct detection (b) rates when the number of labelled samples varies between 0 and 500 for semi-supervised IFA with (- -) and without constraints (–). The rates have been evaluated on the test set.

7 Conclusions

In this paper, we have proposed a method for learning parameters of the noiseless IFA model that incorporates two kinds of prior information related to the mixing process, on the one hand, and cluster membership of some subset of training samples, on the other hand. The main assumption of this generative model is that the observed data have been generated by a linear mixture of latent variables assumed to be independent and non-Gaussian. Moreover, each individual latent variable has its own distribution, which is modelled semi-parametrically by a mixture of Gaussians (MOG). In this context, a criterion can be defined, and a GEM algorithm dedicated to its optimisation can be subsequently described. The main limitations of the approach concerns the hypothesis underlying the IFA model; if the mixing system is not linear or if the sources are not independents the method is likely to fail. However, the proposed method has been applied to both simulated data as well as real data related to fault diagnosis in railway track circuits with success. The diagnosis system aims to recover the latent variables linked to track circuit defects using features extracted from the inspection signal and IFA hypothesis. A comparison between standard and proposed IFA models has been carried out to show that our approach takes advantage of prior information, thus significantly improving estimation accuracy and removing indeterminacy associated with unsupervised IFA, such as indeterminacy related to the permutation of sources. Further studies should be carried out to take into account imprecise and uncertain cluster memberships, such as those yielded subjectively by human experts. Indeed, soft labels seem to be appropriate when dealing with the imprecise knowledge that experts have regarding the monitoring data of complex systems.

References

1. 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 1996.
2. H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.
3. H. Attias. Independent factor analysis with temporally structured factors. In *Proceedings of the 12th NIPS Conference*, pages 386–392. MIT Press, 2000.
4. M. Babaie-zadeh and C. Jutten. Semi-Blind Approaches for Source Separation and Independent component Analysis. In *Proceedings of the European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 301–312, 2006.
5. D. J. Bartholomew and K. Martin. *Latent variable models and factor analysis*. Arnold, London, 1999. Seconde édition.
6. A. J. Bell and T. J. Sejnowski An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.
7. K. A. Bollen. Structural Equations with Latent Variables. Wiley, 1989.

8. A. Cichocki and and R. Zdunek and A.H. Phan and S. Amari. *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.
9. A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing*. Wiley, 2002.
10. E. Côme and L. Oukhellou and T. Denœux and P. Aknin. Partially-supervised learning in independant factor analysis. ESANN 17th European Symposium on Artificial Neural Networks. Bruges, pages 53–58, 2009.
11. E. Côme and L. Oukhellou and T. Denœux and P. Aknin. Independent Factor Analysis with mixing constraints in a semi-supervised framework. Application to railway device fault diagnosis. ICANN 19th International Conference on Artificial Neural Networks, Limassol, Chypre, pages 416–425,2009.
12. A. Debiolles and L. Oukhellou and 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.
13. A. P. Dempster and N. M. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B:39–138, 1977.
14. B.S. Everitt. An Introduction to Latent Variable Models. Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York, 1984.
15. O. Georgiev, F. Theis and A. Cichocki Blind source separation and sparse component analysis of overcomplete mixtures. In *Proceedings of ICASSP*, pages 493–596, 2004.
16. A. Honkela and H. Valpola and A. Ilin and J. Karhunen. Blind Separation of Nonlinear Mixtures by Variational Bayesian Learning. Digital Signal Processing, 5(17):914–934, 2007.
17. H. Hotelling. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology,24:417–441,498–520, 1933.
18. A. Hyvärinen, J. Karhunen, E. Oja. *Independent Component Analysis*. Wiley, 2001.
19. A. Hyvärinen and R. Karthikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*, 49(1):151–162, 2002.
20. G. J. Mclachlan and T. Krishnan, The EM algorithm and Extension Wiley, 1996.
21. A. Montanari and D.G. Calo and C. Viroli. Independent factor discriminant analysis. Computational Statistics & Data Analysis, 52(6):3246–3254, 2008.
22. 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.
23. J. Nocedal, S.J. Wright. Numerical Optimization. Springer Series in Operations Research, Springer, 1999.
24. C. Spearman. General intelligence, objectively determined and measured. American Journal of psychology, 15:201–293, 1904.
25. L.L Thurstone. Multiple Factor Analysis. University of Chicago Press, 1947.
26. M.E. Tipping and C. Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B, 61:611–622,1997.
27. C. Wang, S. Yan, L. Zhang and H. Zhang ; Non-Negative Semi-Supervised Learning In *Proceedings of AISTATS*,
28. K. Zhang and L. W. Chan. ICA with sparse connections. In *Proceedings of Intelligent Data Engineering and Automated Learning Conference (IDEAL)*, pages 575–582, 2009.