RAILWAY DEVICE DIAGNOSIS USING SPARSE INDEPENDENT COMPONENT ANALYSIS

Zohra L. Cherfi[1], *Etienne Côme*[1], *Latifa Oukhellou*[1,2], *and Patrice Aknin*[1]

[1] INRETS-LTN, Le Descartes 2, 2 rue de la butte verte, 93160 Noisy-le-Grand, France
[2] CERTES-Université Paris XII, 61 av. Gal. de Gaulle, 94100 Créteil, France
phone: + (33) 1 45 92 56 47, fax: + (33) 1 45 92 55 01, email: {cherfi, come, oukhellou, aknin}@inrets.fr

ABSTRACT

This paper presents a study on the potential interest of sparse Independent Component Analysis (ICA) for the diagnosis of a complex railway infrastructure device. This complex system is composed of several spatially related subsystems, i.e. a defective subsystem modifies not only its own inspection data but also those of other subsystems. In this context, the ICA model will be used to extract from inspection data indicators of each subsystem state. We assume here that inspection data are observed variables generated by a linear mixture of independent and nongaussian latent variables linked to the defects. Furthermore, physical knowledge on the inspection system provides prior information on the mixing structure. We will then investigate the ability of sparse ICA to recover this structure and to provide meaningful defect indicators. Also, we will show that introducing sparsity in the mixing process slightly improves the results.

1. INTRODUCTION

The diagnosis of a complex system consists in detecting and identifying defects appearances from inspection measurements. According to whether labeled data are available or not, two learning frameworks are possible : supervised or unsupervised. In many real-world applications, labeled data are often difficult to obtain while unlabeled ones are easily available. In this case, the main task is to develop models which are able to catch the structure of the analyzed system underlying the observations [5].

In this article, we explore the use of a generative approach to achieve the diagnosis of a railway device, namely the track circuit, in an unsupervised context. This component can be considered as a complex system made up of spatially related subsystems, i.e. the presence of a defect in one subsystem modifies its own inspection data but also the inspection data pertaining to other subsystems. The aim of the diagnosis is to identify and localize defects appearance on subsystems on the basis of a specific inspection signal analysis. The model involved here assumes that the observed variables extracted from the signal have been generated from a linear mixing of the latent variables linked to the subsystems defects. Under hypothesis that the latent variables are mutually independent and nongaussian the Independent Component Analysis (ICA) model suits the problem well [7]. We also investigate the possibility of making the unknown mixing matrix as sparse as possible. Indeed, detecting zero entries, could reduce the model complexity and thus provide a more reliable estimation of the parameters. It could also make the structure of the generative model more interpretable which seems to be interesting in our application where the observations are affected by only a smaller subset of the latent variables [8][10].

This paper is organized as follows. In section 2 the operation of the railway device and the purpose of its diagnosis are described. The section 3, gives the ICA background and details the incorporation of penalty functions and constraints on the mixing matrix to produce sparse parameters. The efficiency of using the ICA model in its traditional or sparse form is then evaluated on the railway application.

2. RAILWAY TRACK CIRCUIT

2.1 Track circuit principle

The track circuit is an essential component of the automatic train control system. Its main function is to detect the presence or absence of vehicle traffic on a given 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, such as the authorized speed on a given section. The railway track is divided into different sections (Figure 1). Each section has a specific track circuit consisting of:

- a transmitter connected to one of the two section ends, which supplies a frequency modulated alternating current
- the two rails that can be considered as a transmission line
- a receiver at the other end of the track section
- trimming capacitors connected between the two rails at constant spacing to compensate the inductive behavior of the track. An electrical tuning is then performed to limit the attenuation of the transmitted current and improve the transmission level

A train is detected when its wheels and axles short-circuit the track, which induces the loss of the track circuit signal. The drop of the received signal below a threshold indicates that the section is occupied.

2.2 Diagnosis purpose and Methodology

The different parts of the track circuit are subject to many constraints (mechanic, electric, atmospheric...) that may lead to a defective behavior of the system. In the most extreme cases, significant attenuation of the transmitted signal can occur, which may induce signalling problems (the section can be considered as occupied even if it is not). To avoid such inconvenience and inform maintainers about failures, an inspection car equipped with a sensor in the first axle is used

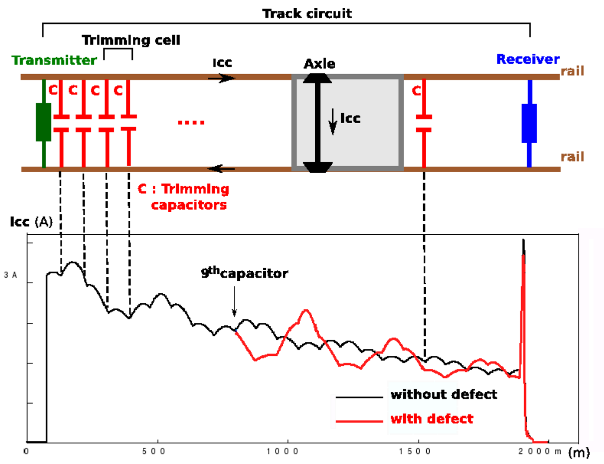

Figure 1: Track circuit representation and inspection signal without defect and with a defective 9^{th} capacitor.

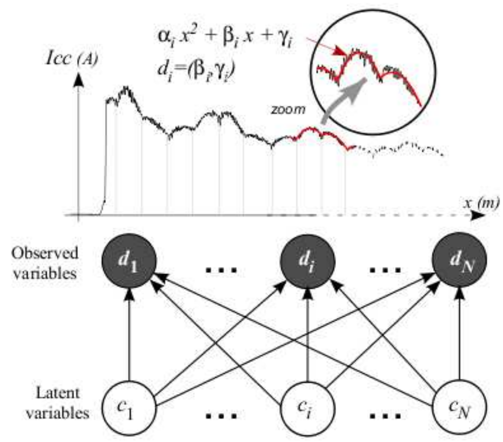

Figure 2: Generative graphical model

to pick up the carrier current level of the short circuit current (Icc, see Figure 1). This signal is recorded at each position of the train while the track is shunted by the inspection train itself. In this paper, we will focus on defects that affect trimming capacitors. Figure 1 shows two examples of denoised inspection signals along a 1500m track circuit when the system is fault-free and when one capacitor (the 9^{th} one) is defective . The diagnosis system aims to detect the operating mode of the track circuit by analyzing the measurement signal (Icc) which is closely linked to electrical trimming capacitors characteristics.

The proposed method is based on the following observations:

- the inspection signal has a specific structure, which is a succession of local arches that can be approximated by quadratic polynomials
- the trimming capacitors have a spatial relationship between them, i.e. the presence of a defect in a trimming capacitor only affects the signal between the defect and the receiver leaving the signal upstream unchanged (see Figure 1)

The idea is to consider the track circuit as a global system Σ made up of a series of N subsystems $S_1,\dots,S_N$ that correspond to the N trimming capacitors. A defect on one subsystem S_i is represented by a continue value of the capacitance parameter c_i. A generative model can be built where the latent variables c_i for $i = 1..N$ are the capacitances of trimming capacitors and the observed variables d_i for $i = 1..N$ are extracted by approximating each arch of the inspection signal (Icc) by a quadratic polynomial $\alpha_i x^2 + \beta_i x + \gamma_i$. Because of the continuity between arches, each observed variable d_i is only described by two coefficients of the local polynomial (β_i, γ_i) and the whole measurement signal is thus described by a total of $2N$ coefficients. This kind of model is useful because its structure can take advantages from prior knowledge on the spatial relationship between subsystems.

Assuming a linear relationship between observed and latent variables (Figure 2), the generative model could be described by:

$$\mathbf{d} = A.\mathbf{c} \tag{1}$$

where $\mathbf{d} = (d_1, d_2, ..., d_N)^t$ is the vector of observed variables, $\mathbf{c} = (c_1, c_2, ..., c_N)^t$ the vector of independent latent variables and A a matrix with a sparse structure that transcribes the spatial dependencies of the model.

Considering the previous problem as a blind source separation problem, the ICA model can be used to estimate the mixing matrix A and thereby to recover the latent components (sources) $c_1, c_2, ..., c_N$ from the observed variables alone. Moreover, ICA with a sparse mixing matrix can also be considered to take account of prior information on the mixing process. In this way, the connections between the observed and recovered variables can be closely linked to the physical behavior of the system.

3. INDEPENDENT COMPONENT ANALYSIS

3.1 Independent Component Analysis principle

The classic version of the ICA model can be expressed as [7]:

$$\mathbf{d} = A.\mathbf{c} \tag{2}$$

where $\mathbf{d} = (d_1, d_2, ..., d_N)^t$ are the observed variables, $\mathbf{c} = (c_1, c_2, ..., c_N)^t$ are the latent variables assumed to be non-gaussian and mutually independent, and A is an unknown mixing matrix. The problem consists of estimating both the mixing matrix and the realizations of the latent variables from the obseved variables alone. For simplicity, we assume here that the number of latent variables is equal to that of the observed ones. The estimation of the model could be done by maximizing the log-likelihood, which thanks to the independance assumption takes the following form [7, p. 204]:

$$\mathscr{L}(A) = -T\log(|\det A|) + \sum_{k=1}^{T}\sum_{i=1}^{N} \log p_i[(A^{-1}\mathbf{d}_k)_i] \tag{3}$$

where p_i is the density of the source i. The estimation of the mixing matrix A could be done thanks to gradient ascent algorithms [2]. The natural gradient is well suited to this problem [1]. In this case, the update rule for A takes the following form:

$$\triangle A \propto A\left(E\{\psi(\mathbf{c})\mathbf{c}^t\} - I\right) \tag{4}$$

where E stands for empirical expectation, I is the identity matrix, $\mathbf{c}$ is the current estimate of the sources $\mathbf{c} = A^{-1}\mathbf{d}$ and $\psi(\mathbf{c}) = (\psi_1(c_1),..,\psi_N(c_N))^t$ with $\psi_i(c_i) = -\frac{p_i^{'}(c_i)}{p_i(c_i)}$.

The densities p_i can be adapted online to extract both supergaussian and subgaussian sources. In our experiments, this was done by using one of the following ψ functions : $\psi_i(c_i) = -2tanh(c_i)$, $\psi_i(c_i) = tanh(c_i) - c_i$, according to an indicator of the source nature : respectively supergaussian or subgaussian [7, p. 205-210].

3.2 ICA with sparse mixing matrix

3.2.1 Constraints on the mixing matrix

In many real-world applications, the structure of the problem supplies a prior knowledge on the mixing process that can be interesting to introduce in the model estimation. The mixing matrix could be constrained to a specific form in order to take account of this kind of prior. In our application, as there is no influence between a trimming capacitor operating mode and inspection data related to subsytems located upstream from it, some elements of the mixing matrix can indeed be constrained to be null. In this case, the estimation problem of the ICA model has to be reformulated to take account of conditional independencies of some sources given some observed variables. It is shown in [3] that making this kind of hypothesis constraints the form of the mixing matrix to be as:

$$\mathbf{d}_i \perp\!\!\!\perp \mathbf{c}_j \Leftrightarrow a_{ij} = 0 \tag{5}$$

where $\mathbf{d}_i \perp\!\!\!\perp \mathbf{c}_j$ means that the latent variable $\mathbf{c}_j$ doesn't influence the observed variable $\mathbf{d}_i$.

The maximization of the log-likelihood under these constraints is done by considering only the gradient of the nonzero coefficients. The initialization and the natural gradient update rule of A become then :

$$A^{(0)} = M_\bullet A^{(0)} \tag{6}$$

$$\triangle A_{Ct} \propto M_\bullet A\left(E\{\psi(\mathbf{c})\mathbf{c}^t\} - I\right) \tag{7}$$

where $_\bullet$ defines the entrywise product and M is a binary matrix defined by:

$$m_{ij} = \begin{cases} 0 & if\ \mathbf{d}_i \perp\!\!\!\perp \mathbf{c}_j \\ 1 & else \end{cases} \tag{8}$$

3.2.2 Penalized mixing matrix

In [8], sparse priors are introduced in the basic model to penalize mixing matrix with a large number of significantly non-zero parameters. Sparsity can also be achieved by applying certain penalty functions to the mixing parameters as in linear regression problems [9][10]. Different functions globally expressed as $P_\lambda(a)$ can be used as penalties, where P_λ defines a sparsity measure on the mixing matrix and λ the degree to which the penalty influences the solution. The following table (Table 1) gives the expression of three penalty functions used in the experiments: the L_1, the L_2 and the Smoothly Clipped Absolute Deviation (*SCAD*) penalties [4].

The sparse mixing matrix is then simply derived from the log-likelihood together with the penalty as the objective function [4] :

$$\mathscr{L}_P = \frac{1}{T}\left(\mathscr{L}(A)\right) - \sum_{i,j=1}^{N} P_\lambda(a_{ij}) \tag{9}$$

Table 1: Expression of the penalty functions P_λ

Penalty	$\mathbf{P}_\lambda(\mathbf{a})$
L_1	$\lambda\|a\|$
L_2	$\lambda\|a\|^2$
$SCAD_{(\alpha>2)}$	$\begin{cases} \lambda\|a\| & \|a\| \le \lambda, \\ \frac{-(a^2 - 2\alpha\lambda\|a\| + \lambda^2)}{[2(\alpha-1)]} & \lambda < \|a\| \le \alpha\lambda, \\ (\alpha+1)\lambda^2/2 & \|a\| > \alpha\lambda. \end{cases}$

The corresponding natural gradient learning rule for A is obtained as:

$$\triangle A_{Pen} \propto A\left(E\{\psi(\mathbf{c})\mathbf{c}^t\} - I - A^t[P_\lambda^{'}(a_{ij})]\right) \tag{10}$$

where $[P_\lambda^{'}(a_{ij})]$ denotes the matrix whose $(i,j)^{th}$ element is $P_\lambda^{'}(a_{ij})$.

4. RESULTS

To assess the performance of the approach, we use a database where the information about sources (latent variables) is available. We considered a track circuit of $N = 19$ subsystems (or capacitors) and a database of 2500 noised signals with different values of the capacitance of each capacitor. 1000 signals were used for the training phase while the 1500 others were reserved for the test phase. At the end, the aim was to recover N latent variables from $2N$ observed ones (2 coefficients per signal arch). To conserve the specific spatial structure of the mixing process, the dimensionality reduction on the observation matrix using a principal component analysis was kept out. We chose to extract $2N$ latent variables and to keep the N ones strongly correlated with the variables of interest, the N others being considered as noise. The right order of the estimated variables is deduced according to the obtained correlations.

The experiments were designed to illustrate the performance of the ICA method to achieve the diagnosis task and also to quantify the influence of introducing sparsity in the mixing process. Basic and sparse ICA models using natural gradient-based algorithm were thus performed on the observation matrix. For the penalized model, the penalties described in Table 1 were used (*SCAD* parameters was chosen as suggested in [4]). The results of the different settings were quantified through the correlation between the true sources (capacitances) and their estimates calculated on the test set, and by checking the mixing matrix structure.

Figure 3 shows the performances of the algorithm using the traditional ICA model and the penalized one with L_2 penalty (for a $\lambda = 4$). Obviously, the correlations between the estimated sources and the true capacitances (after permutation) are stronger in the penalized case. As expected, the estimated mixing matrix shown seems to be block lower triangular, which validates the spatial relationship between the system variables. Visually, one sees that the mixing structure is more legible when applying penalty in the ICA model.

However, a more accurate quantification of the benefits of penalized ICA needs a suitable choice of the regularization parameter λ. Different computations on the test set according to a variable $\lambda = 0...25$ were then achieved for the different penalties. Note that the case $\lambda = 0$ illustrates the performances of the traditional ICA model (without any penalty).

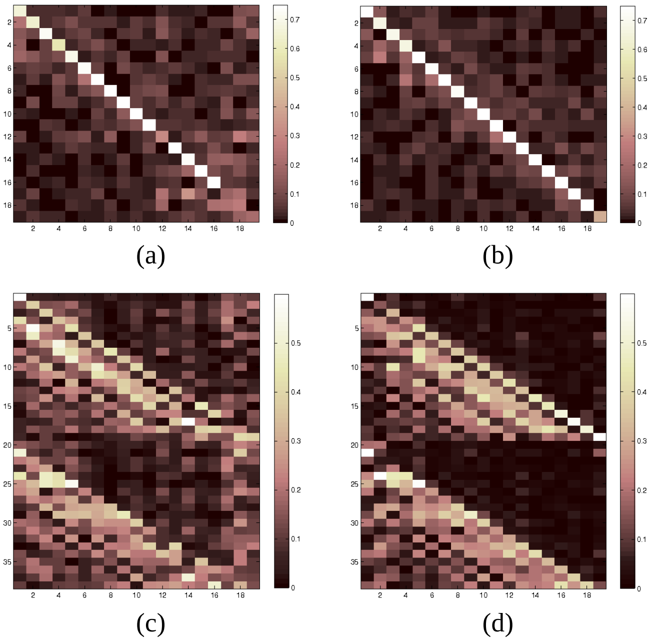

Figure 3: Absolute value of the correlations between estimated latent variables and true capacitances computed on the test set using the traditional ICA model and the penalized one resp. (a)(b), and absolute value of the estimated mixing matrix computed on the test set using the traditional ICA model and the penalized one, (c)(d) respectively.

As shown in Figure 4(a), the values of the mean correlation between the estimated and the true sources can be seen to increase with increasing the regularization parameter λ until a certain value. The decreasing is normal for too large values of λ because the data is then neglected and only the penalized term is taken into account. A suitable value of λ should be chosen equal to 8 for L_1 penalty, 10 for SCAD penalty, while the L_2 penalty seems to be more effecient than the others for a lower λ value ($\lambda = 4$).

We also used a measure of sparsity to evaluate the consistency of the structure obtained on the estimated mixing matrix knowing the structure implied by the prior knowledge. This sparsity measure is proposed in [6] to quantify the energy contained into the components of a vector. We adapted this measure on the mixing matrix to evaluate its sparsity:

$$sparsity(A) = \frac{\sqrt{N \times 2N} - \left(\sum |a_{ij}|\right) / \sqrt{\sum a_{ij}^2}}{\sqrt{N \times 2N} - 1} \quad (11)$$

This measure is equal to one if and only if A contains only a single non-zero component, and takes a value of zero if all components are equal (up to signs). In the ICA with mixing constraints, half of the mixing matrix elements are zero and the mean sparsity measure estimated on 30 runs is equal to 0.5. For the penalized ICA, figure 4(b) illustrates the sparsity measure as function of the regularization parameter λ. In this case, the sparsity increases with the value of λ. A suitable choice of the parameter λ is thus given by the mean correlation between estimated and real sources rather than the sparsity.

We also compared the results of penalized ICA estimation with respect to one element ($a_{1,17}$) of A expected to be at zero thanks to prior information. Figure 4(c) shows the mean estimate of the entry $a_{1,17}$ of A for each λ value over 30 runs of the algorithm. According to all previous observations, we can see that from a certain λ value, the last three methods perform better than traditional ICA and provide a more stable estimate.

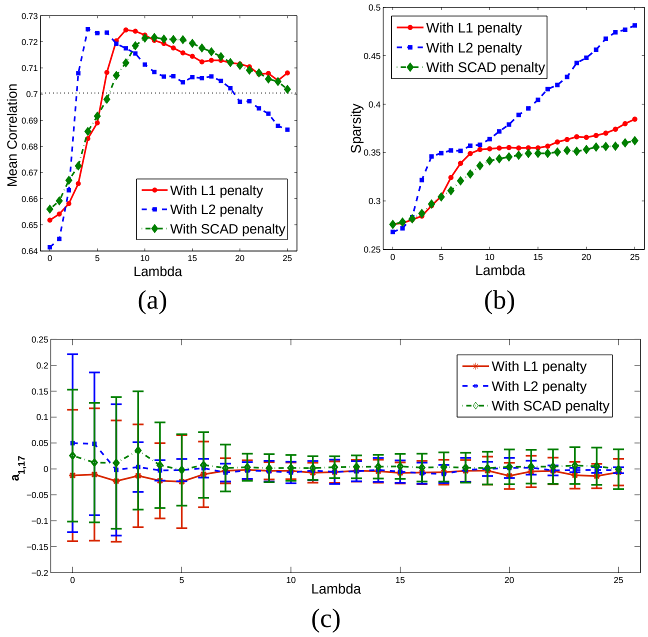

Figure 4: Mean correlation between the true capacitances and the estimated ones over 30 runs at different λ values (a). Mean sparsity measure as function of the regularization parameter λ over 30 runs (b). The estimate of the entry $a_{1,17}$ (the expected value is 0) at different λ values, the error bars denotes the standard deviation of the results over 30 runs (c).

A representation of the penalization impact for an appropriate choice of the penalty function and the λ value is summerized in Table 2. The correlations ($r_{\hat{c}_i,c_i}$) between each real source and its estimate are computed from the traditional ICA model, the ICA with constraints on the mixing matrix and the penalized ICA (L_2 penalty with $\lambda = 4$). For each source, the mean of the absolute value of the correlations over 30 runs is given in the table. Taking into account that only the correlations higher than 0.7 are meaningful, the basic ICA allows us to recover 10 capacitances while, the ICA with constraints allows us to recover 13 capacitances and the penalized ICA 15 capacitances. Clearly, the sparse ICA models improve efficiently the estimation of the sources that are not detected by the traditional ICA model. The model with penalty provides better results than the others which could be explained by the fact that the part of the mixing matrix considered as noise is also penalized and thus it less influences the estimation of the model.

5. CONCLUSION

We have investigated in this paper the possibility of using ICA model for a railway infrastructure component diagnosis. The proposed approach aims at recovering the latent variables linked to the defects from their linear observed mixtures that correspond to the features extracted from the inspection signal. The ICA without and with sparse connection have been tested on a database for which the sources are known. Suitable results were obtained that show the effectiveness of the approach to be used in a diagnosis objective. Introducing sparsity seems to be more efficient to establish the mixing matrix structure and thus the data connections. It also provides a better coherence between the estimated sources and the model characteristics.

Table 2: The 19 mean correlations between the real sources and their estimates computed on the test set over 30 runs, with the traditional ICA, the ICA with mixing constraints and the penalized ICA for a penalty L_2 and $\lambda = 4$.

Sources	**1**	**2**	**3**	**4**	**5**
$<\lvert r_{\hat{c}_i,c_i}\rvert>$	0.63	0.61	0.72	0.59	0.73
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>_{Ct}$	0.74	0.64	0.67	0.61	0.77
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>_{L_2\ (\lambda=4)}$	0.73	0.64	0.68	0.63	0.74
Sources	**6**	**7**	**8**	**9**	**10**
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>$	0.72	0.71	0.78	0.72	0.72
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>_{Ct}$	0.71	0.76	0.71	0.77	0.75
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>_{L_2\ (\lambda=4)}$	0.73	0.71	0.76	0.77	0.74
Sources	**11**	**12**	**13**	**14**	**15**
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>$	0.68	0.68	0.75	0.65	0.75
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>_{Ct}$	0.76	0.77	0.73	0.75	0.80
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>_{L_2\ (\lambda=4)}$	0.77	0.79	0.75	0.79	0.80
Sources	**16**	**17**	**18**	**19**	$<>$
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>$	0.74	0.36	0.39	0.20	**0.64**
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>_{Ct}$	0.74	0.69	0.56	0.32	**0.71**
$<\lvert \overline{r_{\hat{c}_i,c_i}}\rvert>_{L_2\ (\lambda=4)}$	0.85	0.81	0.74	0.31	**0.72**

REFERENCES

[1] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[2] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, pages 757–763. MIT Press, 1996.

[3] 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.

[4] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statist. Assoc.*, 96:1348–1360, 2001.

[5] Z. Ghahramani. Unsupervised learning. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 72–112. Springer, 2004.

[6] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[7] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[8] A. Hyvarinen and R. Kathikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*, 49:151–162, 2001.

[9] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal statistical Society*, 58:267–288, 1996.

[10] K. Zhang and L. Chan. Ica with sparse connections. In *IDEAL*, pages 530–537, 2006.