Analyse en facteurs indépendants pour le diagnostic d'un composant de l'infrastructure ferroviaire dans un cadre semi-supervisé

Rapport de stage M2 spécialité IAD parcours ADRO – 7 septembre 2009

Nicolas Cheifetz

Encadrante INRETS *:*
Latifa Oukhellou

Encadrant LIP6*:*
Patrick Gallinari

Table des matières

1 Introduction

Ce stage s'est déroulé au Laboratoire des Technologies Nouvelles (LTN) à l'Institut National de Recherche sur les Transports et leur Sécurité (INRETS) sur le site de Marne-la-Vallée. Je tiens à remercier M^r Patrice Aknin pour m'avoir accueilli au sein de l'équipe Diagnostic du LTN. Je remercie également M^lle Latifa Oukhellou pour m'avoir encadré durant ce stage et m'avoir donné le temps de la réflexion. Le stage est la suite directe de la thèse de E. Côme soutenue en janvier (Côme, 2009 [9]). L'étude s'inscrit dans le domaine de l'apprentissage statistique/automatique même si l'un des modèles évoqués (Analyse en Composantes Indépendantes) provient de la communauté du Traitement du signal.

Problématique et état de l'art

L'objectif de ce stage est d'étendre les capacités de diagnostic d'un modèle à base de reconnaissance des formes, développé dans la thèse [9]. Il s'agit d'une approche *générative* dont le but est de caractériser la densité de probabilité sur les variables observées (Jebara, 2004 [16]). On cherche à modéliser les relations entre les descripteurs extraits des signaux réels et des variables latentes (cachées), tout en intégrant un a priori sur leurs relations de dépendance/indépendance. On ajoute un second a priori en fournissant les labels de certaines données d'apprentissage ; on parle alors de cadre *semi-supervisé* (Chapelle, 2006 [7]). Cette approche s'avère particulièrement pertinente pour le diagnostic (Bishop, 2006 [3]) d'un système réparti, constitué de plusieurs sous-systèmes. En effet, il sera possible de détecter et localiser le sous-système défectueux, et d'estimer la gravité du défaut. Le modèle choisi est l'*Analyse en Facteurs Indépendants* [IFA] (Attias, 1999 [1]). Ce modèle généralise l'*Analyse en Composantes Principales*[1] (Pearson, 1901), l'*Analyse Factorielle* (Spearman, 1904) et l'*Analyse en Composantes Indépendantes* (Hérault, 1985).

Dans l'IFA, les variables observées sont supposées être issues d'une combinaison linéaire de variables latentes dont chaque densité s'exprime par un *mélange de gaussiennes* (Pearson, 1894) (McLachlan et al, 2000 [19]). L'idée est que les individus d'une même classe partagent la même loi de probabilité. Notre objectif sera de caractériser chacune de ces lois. Ce problème d'estimation va être résolu par un algorithme GEM (Dempster & al., 1977 [14]). Ce type d'algorithme ne produit pas systématiquement l'estimation optimale. On sélectionne alors la meilleure solution au sens d'un seul critère : la *vraisemblance* (Fisher, 1922). La vraisemblance agrège toutes les informations contenues dans les données observées.

[1] L'ACP est la première méthode (linéaire) proposée pour la réduction de la dimension des données observées ; l'objectif de cette méthode est de déterminer les directions orthogonales de l'espace de projection en maximisant la variance expliquée et en minimisant l'erreur de reconstruction (Saporta et al., 2005 [22]).

Les travaux effectués durant ce stage sont appliqués au diagnostic d'un élément de l'infrastructure ferroviaire : le circuit de voie. Il s'agit d'un système complexe indispensable à la sécurité des lignes à grande vitesse sur le réseau ferroviaire français. La thèse [9] porte sur le diagnostic de circuits de voie avec un nombre fixe de sous-systèmes. Ce stage propose un modèle génératif tiré de l'IFA basé sur l'apprentissage de circuits de voie de différentes tailles.

Tout d'abord, nous présenterons les modèles de mélange et l'algorithme EM dans la section 2, où nous reprendrons en partie les notations de (Bilmes, 1997[2]). Puis nous introduirons le cadre de l'Analyse en variables latentes indépendantes (3.1) dans lequel s'inscrivent ICA (3.2) et IFA (3.3). On présente une extension de l'IFA dans un cadre semi-supervisé (3.3.2). Enfin, nous appliquerons deux versions de l'IFA, y compris la nouvelle extension, au diagnostic des circuits de voie dans un cadre semi-supervisé (section 4).

Remarque 1.1 :

> On considère que les données observées sont ponctuelles. Et, les modèles probabilistes étudiés sont représentés par des modèles graphiques (voir Annexe A.1).
> Les images sont des liens ; cliquez dessus, vous pourrez les visionner en taille réelle sur un navigateur.

2 Algorithme EM et modèles de mélange

L'objet de cette section est d'étudier comment déterminer des groupes homogènes parmi un ensemble d'observations. Il s'agit alors d'estimer les caractéristiques définissant chaque groupe d'observations (ou classe).

2.1 Modèles de mélange et notations

Soit $\mathbb{X}$, un ensemble composé de N observations :
$\mathbb{X} = \{x_1, \ldots, x_i, \ldots, x_N\}$, où $x_i \in \mathbb{R}^p$ (p fixé).
Une hypothèse forte est supposée : les observations sont indépendantes et identiquement distribuées (i.i.d.).

Soit $\mathbb{C}l$, une partition sur les observations, de taille K :
$\mathbb{C}l = \{C_1, \ldots, C_k, \ldots, C_K\}$. En général, $C_k \in \{1, \ldots, K\}$ ou $C_k \in \{0,1\}^K$. Chaque observation x_i est associée à une classe y_i ; on note Y, l'ensemble des variables aléatoires des classes suivant chaque observation :
$Y = \{Y_1, \ldots, Y_i, \ldots, Y_N\}$, où Y_i est la v.a. de y_i.
Enfin, on note l'ensemble des classes des N observations :
$\mathbb{Y} = \{y_1, \ldots, y_i, \ldots, y_N\}$,où $y_i \in \mathbb{C}l$.

Chaque observation x_i est la réalisation d'une variable aléatoire X_i dont la *densité mélange* est :

$$f(x_i) = \sum_{k=1}^{K} \pi_k \cdot f_k(x_i) \tag{2.1}$$

où f_k est la *densité marginale* de la k^e classe du mélange

et π_k est sa proportion a priori :

$$\sum_{k=1}^{K} \pi_k = 1 \quad , \text{ avec } \quad \pi_k = p\left(Y_{\cdot} = C_k\right) \tag{2.2}$$

Un modèle de mélange est une somme de densités simples ayant pour but de modéliser des distributions complexes.

Les densités marginales $(f_k)_{1\leq k\leq K}$ sont des densités inconnues paramétrées par $(\theta_k)_{1\leq k\leq K}$. On suppose que toutes les observations sont distribuées selon la densité mélange (combinaison linéaire des densités marginales). La densité mélange s'écrit :

$$f(x_i;\Phi) = \sum_{k=1}^{K} \pi_k \cdot f_k(x_i;\theta_k), \quad \forall i \in \{1,\dots,N\} \tag{2.3}$$

où $\Phi = (\pi_1,\dots,\pi_K,\theta_1,\dots,\theta_K)$ représente le paramètre global du modèle.

Notre objectif est alors d'estimer ce paramètre global en connaissant les observations $\mathbb{X}$.

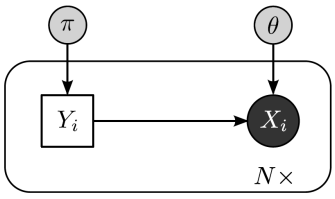


Figure 1: *Modèle graphique de génération des données d'un modèle de mélange.*

Modèles de mélange gaussien

Les densités marginales sont ici des densités gaussiennes. Plus précisément, chaque densité f_k sera paramétrée par la moyenne μ_k et la matrice de variance-covariance Σ_k de la classe k. On pose alors : $\theta_k = (\mu_k, \Sigma_k)$.

$$\forall x \in \mathbb{R}^p, \qquad f_k(x;\theta_k) \;=\; \mathcal{N}(x;\theta_k) \;= \\ \frac{1}{(2\pi)^{p/2}\cdot|\Sigma_k|^{1/2}} \cdot \exp\left(-\tfrac{1}{2}(x-\mu_k)^T \cdot {\Sigma_k}^{-1} \cdot (x-\mu_k)\right) \tag{2.4}$$

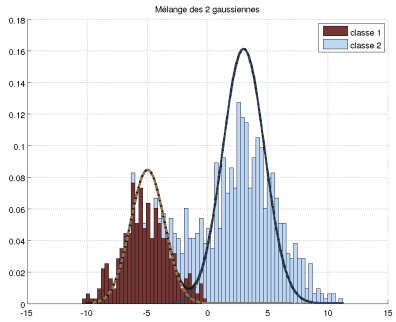


Figure 2: *Exemple d'un mélange de deux gaussiennes.*

Les observations sont toujours supposées i.i.d. et distribuées selon la densité mélange (2.3). Le paramètre global à estimer est de la forme :

$$\Phi = (\pi_1,\dots,\pi_K,\mu_1,\dots,\mu_K,\Sigma_1,\dots,\Sigma_K) \tag{2.5}$$

Le calcul de chaque matrice de covariance Σ_k implique l'estimation des $\frac{p(p+1)}{2}$ coefficients de sa partie supérieure/inférieure (puisque Σ_k est semi-définie positive).

Le nombre important de paramètres à estimer pourrait dégrader la qualité du paramètre global. On utilise alors la parcimonie pour diminuer le nombre total de coefficients à estimer.

Remarque 2.1 : Mélanges gaussiens parcimonieux

Il a été proposé une réduction du nombre de paramètres à estimer, basée sur la décomposition de la matrice de variance-covariance (Celeux et al, 1995 [6]) :

$$\Sigma_k = \lambda_k \cdot D_k \cdot A_k \cdot D_k^T \tag{2.6}$$

où
- $\lambda_k = |\Sigma_k|^{1/p} \in \mathbb{R}$ est le *volume* de la k^e classe
- D_k désigne l'*orientation* de la k^e classe
- A_k indique la *forme* de la k^e classe

D_k est la matrice orthogonale des vecteurs propres de Σ_k, et, $\lambda_k \cdot A_k$ est la matrice diagonale des valeurs propres de Σ_k en ordre décroissant.

G. Celeux et G. Govaert ont proposé 14 modèles parcimonieux en 1995 [6]. En particulier, la forme d'une classe peut être elliptique, sphérique, etc...
Nous allons étudier des données de dimension 1, ce qui ne justifie pas l'emploi de la parcimonie (voir Annexe A.2).

Estimation des paramètres

Pour estimer les paramètres du modèle de mélange, nous cherchons à maximiser la vraisemblance au paramètre global Φ (vraisemblance des données observées) ; elle correspond à la probabilité des observations $\mathbb{X}$ conditionnellement à Φ :

$$L(\Phi;\mathbb{X}) = P(\mathbb{X};\Phi) = \prod_{i=1}^{N} f(x_i;\Phi) \tag{2.7}$$

En général, on calcule plus facilement la Log-vraisemblance (logarithme népérien de la fonction de vraisemblance), qui s'écrit de la forme :

$$\begin{aligned} \mathcal{L}(\Phi;\mathbb{X}) &= \log L(\Phi;\mathbb{X}) \\ &= \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \cdot f_k(x_i;\theta_k)\right) \end{aligned} \tag{2.8}$$

⚠ L'objectif revient à estimer le meilleur paramètre Φ^* au sens de la Log-vraisemblance :

$$\Phi^* = \arg\max_{\Phi} \mathcal{L}(\Phi;\mathbb{X}) \tag{2.9}$$

Si la matrice hessienne de la Log-vraisemblance est définie positive, le problème est convexe. Autrement dit, les maxima locaux sont les maxima globaux. Il nous suffit donc de chercher les valeurs qui annulent $\frac{\partial}{\partial\Phi}\mathcal{L}(\Phi;\mathbb{X})$.
En général, le problème n'est pas convexe et il est difficile de calculer les racines de la différentielle. On utilise alors un algorithme itératif adapté aux modèles de mélange : l'algorithme EM.

2.2 Algorithme EM

Historiquement, l'algorithme EM (Estimation-Maximisation) est une méthode générale formalisée par A. Dempster et al. en 1977 [14] mais d'autres algorithmes similaires avaient été proposés auparavant pour des problèmes particuliers (McLachlan et al, 1997 [18]).

Formulation générale

L'algorithme EM est un algorithme itératif pour estimer le maximum de vraisemblance lorsque l'ensemble des données est partagé entre un ensemble de données observées et de données non observées.

On segmente l'ensemble des "données complètes" $\mathbb{Z}$ en deux sous-ensembles :

$$\mathbb{Z} = \{\mathbb{X}, \mathbb{Y}\}, \text{ où } \begin{cases} \mathbb{X} & : \text{ ensemble des données observées} \\ \mathbb{Y} & : \text{ ensemble des données non observées} \end{cases}$$

On note Z, la variable aléatoire de $\mathbb{Z}$.
La Log-vraisemblance des données complètes est :

$$\begin{aligned} \mathcal{L}(\Phi; Z) &= \log p(X, Y; \Phi) \\ &= \log p(Z|X; \Phi) + \log p(X; \Phi) \end{aligned} \tag{2.10}$$

Explication calculatoire[2]

La Log-vraisemblance des données observées est alors :

$$\begin{aligned} \mathcal{L}(\Phi; X) &= \log p(X; \Phi) \\ &= \log p(Z; \Phi) - \log p(Z|X; \Phi) \end{aligned} \tag{2.11}$$

On cherche toujours à estimer le meilleur paramètre Φ^* tel que : $\Phi^* = \arg\max_{\Phi} \mathcal{L}(\Phi; \mathbb{X})$.

L'algorithme EM résout itérativement le problème d'estimation du paramètre global Φ^* en maximisant l'espérance (sur la variable aléatoire des données complètes Z) de la Log-vraisemblance des données observées $\mathcal{L}(\Phi; X)$, conditionnellement aux données observées $\mathbb{X}$ et au paramètre global de l'itération courante $\Phi^{(q)}$:

$$\begin{aligned} &\mathbb{E}_Z\left(\mathcal{L}(\Phi; X)|X = \mathbb{X}, \Phi^{(q)}\right) \\ &= \underbrace{\mathbb{E}_Z\left(\log p(Z; \Phi)|\mathbb{X}, \Phi^{(q)}\right)}_{Q(\Phi, \Phi^{(q)})} - \underbrace{\mathbb{E}_Z\left(\log p(Z|X; \Phi)|\mathbb{X}, \Phi^{(q)}\right)}_{H(\Phi, \Phi^{(q)})} \end{aligned} \tag{2.12}$$

On remarque que : $\mathbb{E}_Z\left(\mathcal{L}(\Phi; X)|X = \mathbb{X}, \Phi^{(q)}\right) = \mathcal{L}(\Phi; \mathbb{X})$.
On maximise uniquement la fonction Q, car H se dégrade naturellement au cours des itérations (voir Annexe A.3).
Autrement dit :

$$\begin{aligned} \Phi^{(q+1)} &= \arg\max_{\Phi}\{\mathcal{L}(\Phi; \mathbb{X})\} \\ &= \arg\max_{\Phi}\left\{Q(\Phi, \Phi^{(q)})\right\} \end{aligned} \tag{2.13}$$

[2] $p(Z|X) = p(X, Y|X) = p(Y|X, \cancel{X}) \cdot \underbrace{p(X|X)}_{1} = p(Y|X)$

Chaque itération de l'algorithme se décompose en deux étapes : Estimation puis Maximisation de la fonction Q ; l'initialisation s'effectue de manière non déterministe.

Algorithme 1 : Pseudo-code EM dans sa formulation générale

Entrées : données observées centrées-réduites $\mathbb{X}$
 # Initialisation
$q = 0$
$\Phi^{(q)} = \left(\pi_1{}^{(q)}, \ldots, \pi_K{}^{(q)}, \theta_1{}^{(q)}, \ldots, \theta_K{}^{(q)}\right)$
Tant que *non convergence* **faire**
  *# **Etape E** (Estimation) : Calcul de la fonction Q*
  $Q(\Phi, \Phi^{(q)}) = \mathbb{E}_Z\left(\log p(Z; \Phi)|\mathbb{X}, \Phi^{(q)}\right)$
  *# **Etape M** (Maximisation) : Calcul du paramètre $\Phi^{(q+1)}$*
  $\Phi^{(q+1)} = \arg\max_{\Phi} Q(\Phi, \Phi^{(q)})$
  $q \leftarrow q + 1$
Sorties : estimation du paramètre global Φ^*

Propriétés

L'algorithme EM implique souvent des calculs analytiques simples (pas de hessiennes à calculer). Cette caractéristique est un atout car elle facilite l'implémentation. De plus, la vraisemblance des données observées augmente à chaque itération (voir Annexe A.3).

Cependant, il a été démontré que l'algorithme EM converge vers un maximum local, et non global. Une démonstration est notamment proposée par A. Dempster et al. (1977) [14]. La qualité de ce maximum local et la rapidité de convergence dépendent de l'initialisation du paramètre Φ et de la quantité d'information disponible. C'est pourquoi, il faut effectuer plusieurs lancements de l'algorithme et sauver l'estimation de Φ^* pour laquelle la vraisemblance est maximale.

De nombreux tests de convergence peuvent être implémentés (McLachlan et al, 1997 [18]). On peut citer la distance de Kulback-Leibler entre deux paramètres globaux, ou encore la stabilité des distributions a posteriori.
⚠ Notre critère pour tester la convergence des algorithmes étudiés sera toujours la stabilité de vraisemblance :

$$\left|\frac{\mathcal{L}(\Phi^{(q+1)}; \mathbb{X}) - \mathcal{L}(\Phi^{(q)}; \mathbb{X})}{\mathcal{L}(\Phi^{(q)}; \mathbb{X})}\right| < \epsilon \tag{2.14}$$

où ϵ représente le seuil de stabilisation, fixé au départ (usuellement $\epsilon \in [10^{-4}, 10^{-8}]$). Plus d'informations se trouvent dans la thèse de A. Samé [21].

Application aux modèles de mélange

On introduit la nouvelle variable Y_{ik} :

$$Y_{ik} = \mathbb{1}_{|Y_i = C_k} = \begin{cases} 1 & , \text{ si } Y_i = C_k \\ 0 & , \text{ sinon} \end{cases} \tag{2.15}$$

Remarque 2.2 : Explications calculatoires

Comme : $p(y_i) = p(Y_i = y_i) = \prod_{k=1}^{K} [p(Y_i = C_k)]^{Y_{ik}} \quad , \forall i$,

la Log-vraisemblance des données complètes est :

$$\begin{aligned}\mathcal{L}(\Phi; Z) &= \log\left(p(X|Y;\Phi)\cdot p(Y;\Phi)\right)\\ &= \sum_{i=1}^{N}\log\left[\prod_{k=1}^{K}\left(\underbrace{p(X_i|Y_i=C_k;\Phi)}_{f_k(X_i;\theta_k)}\cdot\underbrace{p(Y_i=C_k;\Phi)}_{\pi_k}\right)^{Y_{ik}}\right]\\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} Y_{ik}\cdot\log\left(\pi_k\cdot f_k(X_i;\theta_k)\right)\end{aligned}$$

On calcule alors la fonction Q :

$$\begin{aligned}Q(\Phi,\Phi^{(q)}) &= \mathbb{E}_{X,Y}\left(\mathcal{L}(\Phi;Z)|\mathbb{X},\Phi^{(q)}\right)\\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}_Y\left(Y_{ik}\cdot\log\left(\pi_k\cdot f_k(x_i;\theta_k)\right)|\mathbb{X},\Phi^{(q)}\right)\\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}\underbrace{\mathbb{E}_Y\left(Y_{ik}|\mathbb{X},\Phi^{(q)}\right)}_{{t_{ik}}^{(q)}}\cdot\log\left(\pi_k\cdot f_k(x_i;\theta_k)\right)\end{aligned}$$

Les variables ${t_{ik}}^{(q)}$ représentent les distributions a posteriori des classes à l'itération (q) :

$$\begin{aligned}{t_{ik}}^{(q)} &= p\left(Y_i = C_k|\mathbb{X},\Phi^{(q)}\right)\\ &= \frac{p\left(X_i = x_i, Y_i = C_k;\Phi^{(q)}\right)}{p\left(X_i = x_i;\Phi^{(q)}\right)}\\ &= \frac{{\pi_k}^{(q)}\cdot f_k\left(x_i;{\theta_k}^{(q)}\right)}{\sum_{l=1}^{K}{\pi_l}^{(q)}\cdot f_l\left(x_i;{\theta_l}^{(q)}\right)}\end{aligned} \tag{2.16}$$

Algorithme 2 : Pseudo-code EM appliqué aux modèles de mélange gaussien

Entrées : données observées centrées-réduites $\mathbb{X}$

 # Initialisation

$q = 0$

$\Phi^{(q)} = \left({\pi_1}^{(q)},\dots,{\pi_K}^{(q)},{\mu_1}^{(q)},\dots,{\mu_K}^{(q)},{\Sigma_1}^{(q)},\dots,{\Sigma_K}^{(q)}\right)$

Tant que *non convergence* **faire**

*# **Etape E** (Estimation des lois a posteriori ${t_{ik}}^{(q)}$))*

pour chaque $(i,k) \in \{1,\dots,N\}\times\{1,\dots,K\}$ **faire**

$${t_{ik}}^{(q)} = \frac{{\pi_k}^{(q)}\cdot f_k\left(x_i;{\mu_k}^{(q)},{\Sigma_k}^{(q)}\right)}{\sum_{l=1}^{K}{\pi_l}^{(q)}\cdot f_l\left(x_i;{\mu_k}^{(q)},{\Sigma_k}^{(q)}\right)}$$

*# **Etape M** (Maximisation)* : Calcul de $\Phi^{(q+1)}$

pour chaque $k \in \{1,\dots,K\}$ **faire**

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

$${\mu_k}^{(q+1)} = \sum_{i=1}^{N}{t_{ik}}^{(q)}\cdot x_i \;/\; \sum_{i=1}^{N}{t_{ik}}^{(q)}$$

$$\Sigma_k^{(q+1)} = \frac{\sum_{i=1}^{N}{t_{ik}}^{(q)}\cdot\left(x_i-{\mu_k}^{(q+1)}\right)^T\cdot\left(x_i-{\mu_k}^{(q+1)}\right)}{\sum_{i=1}^{N}{t_{ik}}^{(q)}}$$

$q \leftarrow q+1$

Sorties : estimation du paramètre global Φ^*

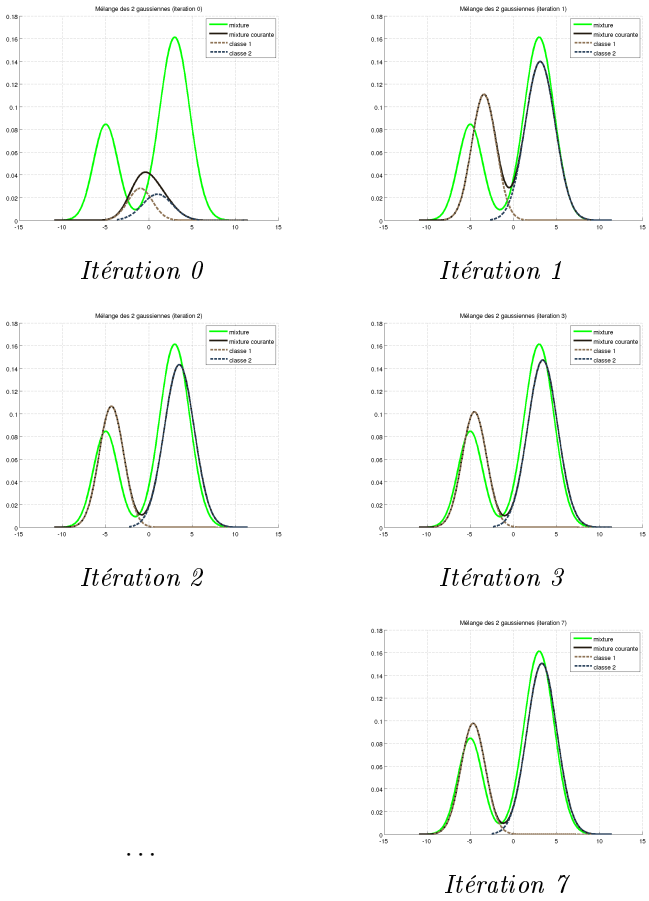
Dans un algorithme EM généralisé (GEM), l'étape **M** consiste simplement à améliorer la fonction Q au lieu de la maximiser, car il est trop difficile de calculer Q. Les propriétés d'un algorithme EM et GEM sont les mêmes.

2.3 Exemple sur un mélange de deux gaussiennes

On génère aléatoirement un ensemble de 1000 observations monodimensionnelles issues d'une mixture de deux gaussiennes, dont les caractéristiques sont :

$$\begin{aligned}K :&\quad \{[2]\}\\ p :&\quad \{[1]\}\\ N :&\quad \{[300][700]\}\\ \pi :&\quad \{[0.300000][0.700000]\}\\ \mu :&\quad \{[-5.000000][3.000000]\}\\ \Sigma :&\quad \{[2.000000][3.000000]\}\end{aligned}$$

On initialise l'algorithme EM avec les faux paramètres : $\pi = \{[0.100000][0.100000]\}$ et $\mu = \{[-1.000000][1.000000]\}$. *[La courbe verte est le mélange gaussien optimal.]*



Itération 0 *Itération 1*

Itération 2 *Itération 3*

... *Itération 7*

Figure 3: *Itérations de l'algorithme EM appliqué à un mélange de deux gaussiennes.*

L'algorithme converge au bout de sept itérations en utilisant la stabilisation de la vraisemblance comme test de convergence (avec $\epsilon = 10^{-4}$).

3 Analyse en variables latentes indépendantes

On suppose que les observations sont conditionnellement indépendantes entre elles. Les variables latentes (desquelles sont issues les variables observées) sont supposées identiquement et indépendamment distribuées.
Nous allons modifier la notation précédemment employée. Désormais, Y désignera les variables latentes, Z les classes de ces variables mais X indique toujours les observations.

3.1 Cadre du problème et notations

Chaque observation est un vecteur de D composantes et chaque composante est de dimension p, où $(p, D) \in \mathbb{N}^2$.

Soit $\mathbb{X}$, un ensemble composé de N observations : $\mathbb{X} = \{x_1, \dots, x_i, \dots, x_N\}$, où $x_i \in \mathbb{R}^{p \times D}$.
On admet alors l'existence de $p \times D$ *variables observées* pour chaque i^{e} observation : $X_i = \big(X_{i1}, \dots, X_{i(p \cdot D)}\big)$.

Pour une certaine observation x_i, le vecteur X_i de ses variables observées résulte d'un vecteur Y_i composé de $p \times L$ variables non observées ($D > L$). On appelle $Y_i = \big(Y_{i1}, \dots, Y_{i(p \cdot L)}\big)$, le vecteur des *variables latentes* continues (ou *sources*) pour la i^{e} observation.
Soit $\mathbb{Y}$, l'ensemble des N réalisations pour les N vecteurs de variables latentes associées à chaque observation : $\mathbb{Y} = \{y_1, \dots, y_i, \dots, y_N\}$, où $y_i \in \mathbb{R}^{p \times L}$ est la réalisation continue du vecteur Y_i.

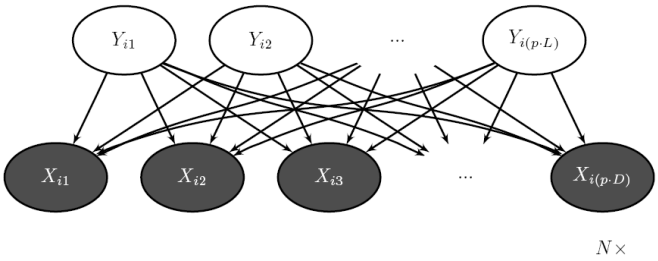


Figure 4: *Modèle graphique de génération des données.*

⚠ On cherche à caractériser un modèle génératif capable de modéliser les distributions des variables latentes. Ainsi, on sera capable de caractériser les variables observées à partir des variables latentes.

On étudie le modèle linéaire suivant :

$$X_i{}^T = A \cdot Y_i{}^T + \varepsilon^T \tag{3.1}$$

où X_i $_{(1 \times (p \cdot D))}$ et Y_i $_{(1 \times (p \cdot L))}$ représentent respectivement les variables observées et les variables latentes continues pour la i^{e} observation, A $_{((p \cdot D) \times (p \cdot L))}$ est la *matrice de mixage* et ε $_{(1 \times (p \cdot D))}$ est un bruit indépendant de Y_i.

$$\iff \quad X = Y \cdot A^T + \xi \tag{3.2}$$

où X $_{(N \times (p \cdot D))}$ représente la matrice des variables observées, Y $_{(N \times (p \cdot L))}$ représente la matrice des variables latentes (sources), A $_{((p \cdot D) \times (p \cdot L))}$ est la *matrice de mixage* et ξ $_{(N \times (p \cdot D))}$ est un bruit indépendant de Y.

3.2 Analyse en Composantes Indépendantes (ICA)

L'ICA (ou ACI) est un modèle probabiliste linéaire issu de la communauté du Traitement du signal, né il y a une vingtaine d'années (Hérault, 1985). On considère ici que les variables observées sont générées par une combinaison linéaire des variables latentes (ou sources), lesquelles sont supposées être indépendantes et non gaussiennes.

Usuellement, on utilise une Analyse en Composantes Principales comme prétraitement pour retirer le bruit ξ de la matrice des observations : on extrait les $(p \cdot L)$ meilleures composantes (ou facteurs) au sens de la variance expliquée.
Autrement dit : $\dim(\text{ACP}(\mathbb{X})) = (p \cdot L) \neq (p \cdot D)$.

⚠ Pour simplifier la notation dans le cadre d'une Analyse sans bruit, on introduit la nouvelle variable S.
On posera à l'avenir : $S = p \cdot L$.

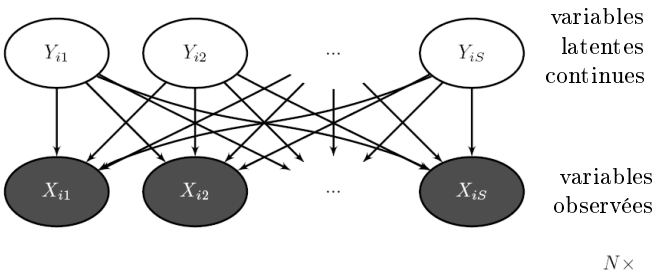


Figure 5: *Modèle graphique de l'Analyse en Composantes Indépendantes sans bruit.*

On suppose que A $_{(S \times S)}$, la nouvelle matrice de mixage est inversible et unique.

On pose alors :

$$\begin{aligned} & X = Y \cdot A^T \\ \Leftrightarrow \quad & Y = X \cdot {}^T\!A^{-1} = X \cdot W^T \end{aligned} \tag{3.3}$$

On appelle W $_{(S \times S)}$, la *matrice de démixage*.

Comme A est une matrice non-singulière et la transformation entre X et Y est linéaire, on applique le théorème "Transformation linéaire d'une densité" extrait du livre *Analyse en Composantes Indépendantes* (Hyvärinen, 2001 [15] page $35 - 36$) :

$$\forall x_i \in \mathbb{X}, \qquad f_{X_i}(x_i; A) = \frac{1}{|\det(A)|} \cdot f_{Y_i}(x_i \cdot {}^T\!A^{-1}) \tag{3.4}$$

$$\text{où} \begin{cases} f_{X_i} \text{ est la densité de } X_i \\ f_{Y_i} \text{ est la densité de } Y_i \end{cases}$$

La vraisemblance des données observées (conditionnellement indépendantes) est le produit des densités de $\mathbb{X}$:

$$L(A; \mathbb{X}) \;=\; \frac{1}{|\det(A)|^N} \cdot \prod_{i=1}^{N} f_{Y_i}(x_i \cdot {}^T\!A^{-1}) \tag{3.5}$$

La Log-vraisemblance des données observées est alors :

$$\mathcal{L}(A;\mathbb{X}) = -N\cdot\log|\det(A)| + \sum_{i=1}^{N}\log f_{Y_i}(x_i\cdot{}^TA^{-1}) \quad (3.6)$$

Or, les variables latentes sont supposées indépendantes entre elles. On peut donc décomposer la densité de Y_i comme suit :

$$\forall i\in\{1,\dots,N\},\quad f_{Y_i} = \prod_{s=1}^{S} f_{Y_{is}}(y_{is}) \quad (3.7)$$

On réécrit alors la Log-vraisemblance exacte :

$$\mathcal{L}(A;\mathbb{X}) = -N\cdot\log|\det(A)| + \sum_{i=1}^{N}\sum_{s=1}^{S}\log f_{Y_{is}}(x_i\cdot({}^TA^{-1})_{.s}) \quad (3.8)$$

Un des objectifs est d'estimer la meilleure matrice de mixage A^* au sens de la vraisemblance :

$$A^* = \arg\max_A \mathcal{L}(A;\mathbb{X}) \quad (3.9)$$

Par conséquent, nous allons chercher à calculer le gradient de la Log-vraisemblance par rapport à A :

$$\nabla A = \frac{\partial}{\partial A}\mathcal{L}(A;\mathbb{X})$$

On détaille l'ensemble des calculs du gradient ∇A en annexe (voir Annexe A.5). On introduit alors la fonction de décorrélation sur les sources :

$$g : \begin{array}{ccc} \mathbb{R}^S & \longrightarrow & \mathbb{R}^S \\ y_i & \longmapsto & \begin{pmatrix} -\dfrac{\partial}{\partial y_{i1}}\log f_{Y_{i1}}(y_{i1}) \\ \vdots \\ -\dfrac{\partial}{\partial y_{iS}}\log f_{Y_{iS}}(y_{iS}) \end{pmatrix} \end{array} \quad (3.10)$$

Dans le cadre de l'ICA, on choisit g selon que les densités marginales aux sources g_s sont super/subgaussiennes. Un critère nous permet d'associer chaque densité marginale g_s à une densité super/subgaussienne (Hyvärinen, 2001 [15] pages 208 – 210). Cette fonction est une approximation de la densité des sources.

Le gradient "classique" de la Log-vraisemblance est alors :

$$\begin{aligned} \nabla_{cla}\,A &= -N\cdot{}^TA^{-1} + {}^TA^{-1}\cdot g(\mathbb{Y})\cdot\mathbb{Y} \\ &= {}^TA^{-1}\cdot(g(\mathbb{Y})\cdot\mathbb{Y} - N\cdot \mathrm{I}_S) \end{aligned} \quad (3.11)$$

Aussi, on détermine le gradient naturel (Hyvärinen, 2001 [15] pages 67 – 68) :

$$\begin{aligned} \nabla_{nat}\,A &= A\cdot{}^TA\cdot\nabla_{cla}\,A \\ &= A\cdot(g(\mathbb{Y})\cdot\mathbb{Y} - N\cdot\mathrm{I}_S) \end{aligned} \quad (3.12)$$

La mise à jour de la matrice de mixage A se fera par montée de gradient : $A^{(q+1)} = A^{(q)} + \tau^{(q)}\cdot\nabla A^{(q)}$

– dans la direction du gradient classique :

$$A^{(q+1)} = A^{(q)} + \tau^{(q)}\cdot{}^TA^{(q)^{-1}}\cdot\left(g^{(q)}(\mathbb{Y}^{(q)})\cdot\mathbb{Y}^{(q)} - N\cdot\mathrm{I}_S\right) \quad (3.13)$$

– dans la direction du gradient naturel :

$$A^{(q+1)} = A^{(q)} + \tau^{(q)}\cdot A^{(q)}\cdot\left(g^{(q)}(\mathbb{Y}^{(q)})\cdot\mathbb{Y}^{(q)} - N\cdot\mathrm{I}_S\right) \quad (3.14)$$

Le paramètre $\tau^{(q)}$ est fixé manuellement ou par *backtracking* (voir Annexe A.4).

De manière analogue, on pourrait chercher à estimer la meilleure matrice de démixage W^*. Il suffirait alors de reformuler quelques équations (voir Annexe A.5). Cependant, nous verrons dans la dernière section qu'il est nécessaire de travailler sur la matrice de mixage pour imposer certains a priori sur la structure de l'application.

3.3 Analyse en Facteurs Indépendants (IFA)

L'IFA est un modèle probabiliste linéaire qui généralise l'Analyse Factorielle (FA), l'Analyse en Composantes Principales (PCA) et l'Analyse en Composantes Indépendantes (ICA). Toutes ces méthodes sont dites *factorielles* : elles créent des "facteurs" qui synthétisent les caractéristiques initiales.

3.3.1 Principe et notations

Ce modèle a été introduit dès 1998 (Moulines, 1998), puis un autre modèle d'IFA "sans bruit" a été proposé par H. Attias (Attias, 1999 [1]). C'est ce modèle qui a été choisi pour nos travaux.

Soit $\mathbb{C}l$, une partition sur les variables latentes continues, de taille K (ensemble de K classes) :
$\mathbb{C}l = \{C_1,\dots,C_k,\dots,C_K\}$. En général, $C_k\in\{1,\dots,K\}$ ou $C_k\in\{0,1\}^K$. Chaque réalisation y_{is} d'une variable latente Y_{is} est associée à une classe z_{is}.
Aussi, l'ensemble des variables aléatoires discrètes des classes (sur les sources) est noté : $Z = \{Z_1,\dots,Z_i,\dots,Z_N\}$, où $Z_i = (Z_{i1},\dots,Z_{is},\dots,Z_{iS})$.
Enfin, on note l'ensemble des classes des N variables latentes : $\mathbb{Z} = \{z_1,\dots,z_i,\dots,z_N\}$,où $z_i\in\mathbb{C}l^S$.

L'IFA généralise l'ICA dans le sens où la densité de chaque variable latente Y_{is} est associée à un mélange de K_s densités gaussiennes (voir la sous-section 2.1) :

$\forall(i,s)\in\{1,\dots,N\}\times\{1,\dots,S\}$,

$$\begin{aligned} f_{Y_{is}}(y_{is}) &= \sum_{k_s=1}^{K_s}\pi_{k_s}\cdot f_{k_s}(y_{is};\underbrace{\mu_{k_s},\Sigma_{k_s}}_{\theta_{k_s}}) \\ &= \sum_{k_s=1}^{K_s}\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s}) \end{aligned} \quad (3.15)$$

On considère le modèle graphique suivant et on suppose que chaque variable observée X_{is} est une combinaison linéaire de S variables latentes continues Y_i (3.3) :

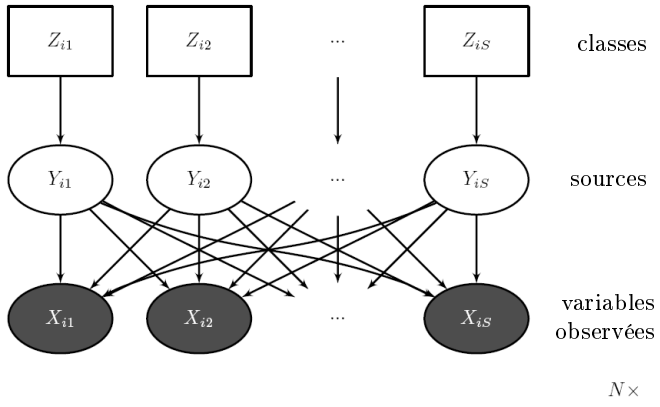


Figure 6: *Modèle graphique de l'Analyse en Facteurs Indépendants sans bruit.*

On note : $\pi_s = (\pi_{1_s}, \dots, \pi_{K_s})$ et $\theta_s = (\theta_{1_s}, \dots, \theta_{K_s})$, où $s \in \{1 \dots, S\}$.

⚠ Notre objectif est encore d'estimer le meilleur paramètre global Φ^* au sens de la vraisemblance :

$$\Phi^* = \arg\max_{\Phi} \mathcal{L}(\Phi; \mathbb{X}) \tag{3.16}$$

où $\Phi = (A, \pi_1, \dots, \pi_S, \theta_1, \dots, \theta_S)$ et
la Log-vraisemblance des données observées est :

$$\mathcal{L}(\Phi; \mathbb{X}) = -N \cdot \log|\det(A)| + \sum_{i=1}^{N}\sum_{s=1}^{S} \log \sum_{k_s=1}^{K_s} \pi_{k_s} \cdot \mathcal{N}(y_{is}; \theta_{k_s}) \tag{3.17}$$

avec $y_{is} = x_i \cdot \left({}^TA^{-1}\right)_{.s}$.

Le problème d'optimisation sous-jacent nous amène à utiliser un algorithme GEM (Generalized Expectation Maximisation) au cours duquel on alternera entre paramétrisation des mélanges gaussiens (densités des sources) et montée de gradient pour la matrice de mixage.

La première phase correspond aux étapes E & M (voir la section 2) mettant à jour les paramètres des mélanges gaussiens ; la seconde phase consiste à mettre à jour la matrice de mixage par montée de gradient (3.13) & (3.14) ; la dernière phase de l'algorithme retire une indétermination sur le paramètre global en normalisant les sources.

Tout d'abord, on détermine la fonction g (voir partie A.5 & Attias, 1999 [1] page 24 - équation 69) :

$$\begin{aligned} g_s(y_i) &= -\frac{\partial}{\partial y_{is}} \log f_{Y_{is}}(y_{is}) \\ &= \sum_{k_s=1}^{K_s} t_{ik_s} \cdot \frac{y_{is} - \mu_{k_s}}{\Sigma_{k_s}} \end{aligned} \tag{3.18}$$

où t_{ik_s} représente la loi a posteriori de la classe C_{k_s} pour le i^{e} vecteur de sources (2.16) :

$$t_{ik_s} = \frac{\pi_{k_s} \cdot \mathcal{N}(y_{is}; \theta_{k_s})}{\sum_{l_s=1}^{K_s} \pi_{l_s} \cdot \mathcal{N}(y_{is}; \theta_{l_s})} \tag{3.19}$$

Après avoir estimé le paramètre global Φ, nous allons lever certaines ambiguités dûes au modèle initial (3.3). Il existe un phénomène de *facteur d'échelle* à cause duquel tous les paramètres de Φ sont estimés à un multiple près.

Pour remédier à ce problème, on normalise le paramètre global (Attias, 1999 [1] page 9/22, équation 31/67) en calculant la matrice de variance-covariance pour chaque source s :

$$\begin{aligned} \Sigma_s &= \mathbb{E}(Y_{.s}{}^2) - \mathbb{E}(Y_{.s})^2 \\ &= \left(\sum_{k_s=1}^{K_s} \pi_{k_s} \cdot \left(\Sigma_{k_s} + \mu_{k_s}{}^2\right)\right) - \left(\sum_{k_s=1}^{K_s} \pi_{k_s} \cdot \mu_{k_s}\right)^2 \end{aligned} \tag{3.20}$$

On normalise alors la matrice de mixage/démixage[3] et les paramètres $(\mu_{k_s}, \Sigma_{k_s})_{1 \leqslant k_s \leqslant K_s}$:

$$\begin{aligned} A_{.s} &\leftarrow A_{.s} \cdot \sqrt{\Sigma_s} \\ \mu_{k_s} &\leftarrow \mu_{k_s} / \sqrt{\Sigma_s} \\ \Sigma_{k_s} &\leftarrow \Sigma_{k_s} / \Sigma_s \end{aligned} \tag{3.21}$$

⚠ Enfin, on modélise S variables latentes continues (ou sources) pour S variables observées. Il est important de noter que seules L d'entre elles sont *pertinentes* ; les $(S - L)$ autres sont des *sources-bruits*. Ces sources sont monoclasses et suivent une loi normale centrée-réduite :

$$\forall s \in \textit{sources}-\textit{bruits}, \quad K_s = 1 \;\Rightarrow\; t_{ik_s} = 1 \quad \text{et} \quad s \rightsquigarrow \mathcal{N}(0, 1) \tag{3.22}$$

Remarque 3.1 : Rôle de la fonction g

La fonction g représente une alternative à des méthodes de décorrélation non linéaire comme celles qui utilisent l'"astuce du noyau". En effet, quel que soit le gradient mis à jour (classique (3.13) ou naturel (3.14)), la matrice de mixage se stabilise à l'optimum :

$$\begin{aligned} & A^{(q+1)} \propto A^{(q)} \\ \Leftrightarrow\; & g^{(q)}(\mathbb{Y}^{(q)}) \cdot \mathbb{Y}^{(q)} - N \cdot \mathrm{I}_S \propto 0 \\ \Leftrightarrow\; & \tfrac{1}{N} \cdot g^{(q)}(\mathbb{Y}^{(q)}) \cdot \mathbb{Y}^{(q)} \propto \mathrm{I}_S \end{aligned} \tag{3.23}$$

Les sources $\mathbb{Y}$ et l'approximation des densités de ces sources $g(\mathbb{Y})$ sont donc décorrélées de façon non linéaire.

⇛ On résume le problème d'estimation pour l'Analyse en Facteurs Indépendants sans bruit en sept étapes :

1. Mise à jour des sources (3.3)
2. Mise à jour des paramètres des sources (des mélanges gaussiens) par l'algorithme EM (section 2)
3. Mise à jour de la fonction g (3.18)
4. Calcul du gradient classique (3.11) / naturel (3.12)
5. Recherche linéaire du paramètre τ (cf. Annexe A.4)
6. Mise à jour de la matrice de mixage (3.13) & (3.14)
7. Normalisation des sources (3.20) & (3.21)

[3] L'équation de normalisation de la matrice de démixage est : $W_{.s} \leftarrow W_{.s}/\sqrt{\Sigma_s}$

3.3.2 Extension au mode semi-supervisé

En général, l'IFA est mise en oeuvre dans un cadre non supervisé. Ce contexte entraîne une indétermination appelée *ordre de permutation* (les sources sont permutées). Nous verrons qu'il est possible d'ajouter de l'information à la matrice de mixage dans le cas des Circuits de Voie (suite à des contraintes spatiales) pour atténuer le phénomène de permutation des sources (section 4.2).

On peut aussi améliorer l'identification des sources en introduisant des données d'apprentissage labellisées. De plus, on sait que plus la taille de l'échantillon est grande, plus l'estimation obtenue est fiable. On se place alors dans le cadre d'un problème *semi-supervisé*. L'étiquette de chaque observation sera associée à une classe ou sera inconnue, respectivement comme pour le mode supervisé et le mode non supervisé. Ainsi, l'ensemble des données d'apprentissage participera à l'estimation des paramètres. On se place même dans le cadre *transductif* (Vapnik, 1999 [25]) dans l'exemple jouet des Crabes, où l'on cherche à labelliser uniquement les étiquettes des observations non labellisées.

On partitionne l'ensemble des N données observées en un ensemble labellisé (de taille M) et un ensemble non labellisé (de taille $N-M$).
La fonction de vraisemblance des données observées $\mathbb{X}$ en mode semi-supervisé s'écrit alors :

$$\begin{aligned} L(\Phi;\mathbb{X}) &= \prod_{i=1}^{N}\prod_{s=1}^{S} f_{X_{is}}(x_{is};\Phi) \\ &= \frac{1}{|\det(A)|^N}\cdot\prod_{i=1}^{N}\prod_{s=1}^{S} f_{Y_{is}}(y_{is};\Phi) \\ &= \frac{1}{|\det(A)|^N}\cdot\underbrace{\prod_{i=1}^{M}\prod_{s=1}^{S}\prod_{k_s=1}^{K_s}\left(\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right)^{\mathbb{1}_{Z_{is}=C_{k_s}}}}_{\text{labellisé}} \\ &\quad\cdot\underbrace{\prod_{i=M+1}^{N}\prod_{s=1}^{S}\sum_{k_s=1}^{K_s}\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})}_{\text{non labellisé}} \end{aligned}$$

avec $y_{is} = x_i\cdot\left({}^TA^{-1}\right)_{.s}$.

En passant au logarithme, on obtient la log-vraisemblance "exacte" des données observées en mode semi-supervisé (*formulation* 1) :

$$\begin{aligned} \mathcal{L}_1(\Phi;\mathbb{X}) &= -N\cdot\log|\det(A)| \\ &\quad+\sum_{i=1}^{M}\sum_{s=1}^{S}\sum_{k_s=1}^{K_s}\mathbb{1}_{Z_{is}=C_{k_s}}\cdot\log\left(\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right) \\ &\quad+\sum_{i=M+1}^{N}\sum_{s=1}^{S}\log\left(\sum_{k_s=1}^{K_s}\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right) \end{aligned} \tag{3.24}$$

Dans l'algorithme GEM modélisant l'IFA, on intègre cette contrainte sur les labels lors du calcul des lois a posteriori en modifiant l'étape d'*Estimation*.

Par rapport au pseudo-code, les lois a posteriori des observations labellisées resteront inchangées :

$$t_{ik_s} = \mathbb{1}_{Z_{is}=C_{k_s}} \quad ,\forall i\in\{1,\dots,M\} \tag{3.25}$$

Remarque 3.2 : Log-vraisemblance approchée

Nous avons essayé une nouvelle approche pour calculer la vraisemblance. On pourrait estimer la log-vraisemblance des données non labellisées par la fonction Q (2.2) :

$$Q(\Phi,\Phi^{(q)}) = \sum_{i=M+1}^{N}\sum_{s=1}^{S}\sum_{k_s=1}^{K_s}\underbrace{\mathbb{E}_Z\left(\mathbb{1}_{Z_{is}=C_{k_s}}|\mathbb{X},\Phi^{(q)}\right)}_{t_{ik_s}{}^{(q)}} \cdot\log\left(\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right)$$

$$\text{avec}\quad t_{ik_s}{}^{(q)} = \frac{\pi_{k_s}{}^{(q)}\cdot\mathcal{N}\left(y_{is};\theta_{k_s}{}^{(q)}\right)}{\sum_{l_s=1}^{K_s}\pi_{l_s}{}^{(q)}\cdot\mathcal{N}\left(y_{is};\theta_{l_s}{}^{(q)}\right)} \quad\text{et}\quad y_{is} = x_i\cdot\left({}^TA^{-1}\right)_{.s}$$

On calcule alors la log-vraisemblance "approchée" des données observées (*formulation* 2) :

$$\begin{aligned} \mathcal{L}_2(\Phi;\mathbb{X}) &= -N\cdot\log|\det(A)| \\ &\quad+\sum_{i=1}^{M}\sum_{s=1}^{S}\sum_{k_s=1}^{K_s}\mathbb{1}_{Z_{is}=C_{k_s}}\cdot\log\left(\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right) \\ &\quad+Q(\Phi,\Phi^{(q)}) \\ &= -N\cdot\log|\det(A)| \\ &\quad+\sum_{i=1}^{M}\sum_{s=1}^{S}\sum_{k_s=1}^{K_s}t_{ik_s}\cdot\log\left(\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right) \\ &\quad+\sum_{i=M+1}^{N}\sum_{s=1}^{S}\sum_{k_s=1}^{K_s}t_{ik_s}{}^{(q)}\cdot\log\left(\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right) \end{aligned}$$

$$\begin{aligned} \Leftrightarrow\quad \mathcal{L}_2(\Phi;\mathbb{X}) &= -N\cdot\log|\det(A)| \\ &\quad+\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{k_s=1}^{K_s}t_{ik_s}{}^{(q)}\cdot\log\left(\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right) \end{aligned} \tag{3.26}$$

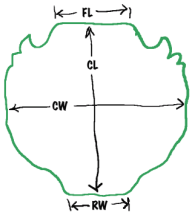
$$\text{car}\quad t_{ik_s}{}^{(q+1)} = t_{ik_s}{}^{(q)} = \mathbb{1}_{Z_{is}=C_{k_s}} \quad ,\ \forall i\in\{1,\dots,M\}.$$

On remarque que cette 2$^{\text{de}}$ formulation est égale à la 1$^{\text{re}}$ à la fonction H (des données non labellisées) près (2.2).

L'ensemble des pseudo-codes nécessaires à l'Analyse en Facteurs Indépendants en mode semi-supervisé se trouvent en annexe (voir Annexe A.6).

Exemple jouet : Crabes

Historiquement, les modèles de mélange ont été introduits par Pearson en 1894 dans un contexte d'identification de deux populations de crabes.



On considère ici un ensemble de 200 observations de crabes. Chaque individu est caractérisé par 5 variables observées (en mm) : longueur de la carapace avant (FL), longueur de la carapace arrière (RW), longueur de la ligne médiane de la carapace (CL), largeur maximale de carapace (CW), et, profondeur du corps (BD).

Pour plus d'informations, cette base de données est fournie et décrite à la page ici.

On peut partitionner la population de deux manières : sexe (male/femelle) et espèce (bleu/orange). Nous allons représenter chaque partition par 1 variable latente.
On identifie 5 variables latentes (3 représentent du bruit) à partir des 5 variables observées. On modèlise alors le problème par un modèle graphique composé de 5 variables observées et 2 sources pertinentes (plus 3 *sources-bruit* non représentées) :

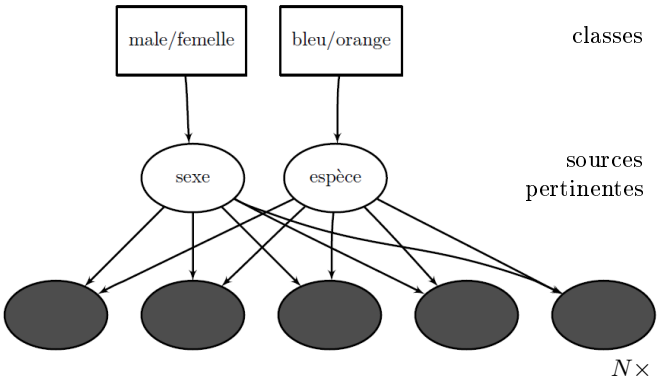


Figure 7: *Modèle graphique de l'IFA sans bruit pour l'exemple des Crabes.*

Pour valider notre modèle, nous allons le tester dans un cadre *transductif*. On procède à 10 lancements successifs de l'Analyse en Facteurs Indépendants sur l'exemple des Crabes : N individus servent à l'apprentissage dont M sont labellisés ($\frac{M}{4}$ de chaque classe). On sauvegarde uniquement le modèle-solution qui maximise l'estimation du maximum de vraisemblance.

Formellement, on a :

N	$=$	200	$\equiv$	nombre d'individus
M	$\in$	$\{0,\ldots,180\}$	$\equiv$	nombre d'individus labellisés
S	$=$	5	$\equiv$	nombre de variables observées
L	$=$	2	$\equiv$	nombre de sources pertinentes
K	$=$	2	$\equiv$	nombre de classes par source pertinente

Voici quelques figures obtenues après apprentissage d'une base avec 140 données labellisées et avec une montée de gradient dans le sens du gradient naturel :

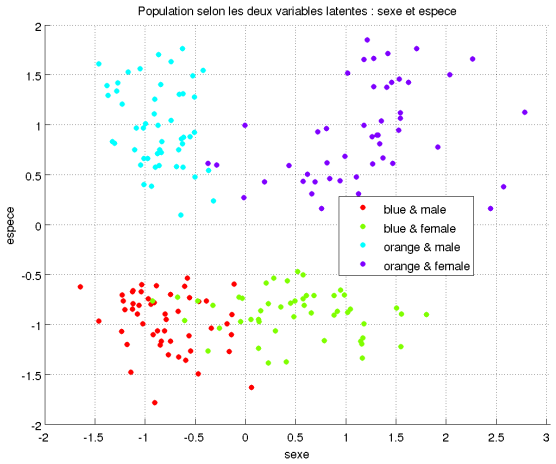
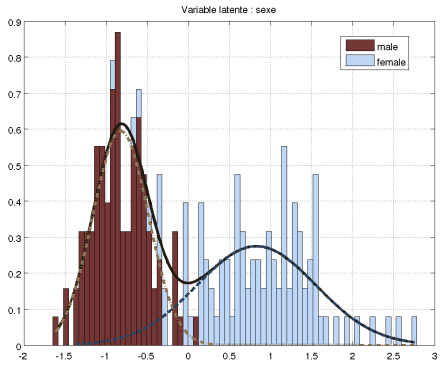
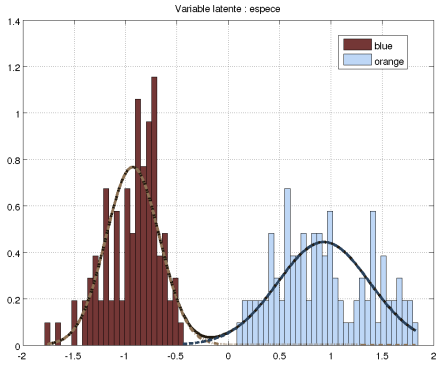


Figure 8: *Population des crabes selon les deux variables latentes pertinentes (avec gradient naturel et $M = 140$).*



[I]



[II]

Figure 9: *Histogrammes des crabes selon chaque variable latente pertinente (gradient naturel et $M = 140$).*

Validation du modèle et sélection des paramètres

Le comportement attendu est :

- une augmentation de la Log-vraisemblance à mesure que le nombre de données labellisées augmente
- une convergence plus rapide lorsque l'on fournit plus de données labellisées

Nous présentons ici les résultats les plus significatifs avec le gradient naturel et la 1re formulation de la Log-vraisemblance (3.24) (et $\epsilon = 10^{-8}$ (2.14)) :

M	N	q	Log-vraisemblance		Taux de bonne classif	
			$\mathcal{L}_1$	std	sexe	espèce
0	200	111	126.47	32.36	87.50%	100.00%
20	200	45	212.24	49.08	91.67%	100.00%
40	200	35	294.75	48.35	91.88%	100.00%
60	200	37	384.55	53.85	92.86%	100.00%
80	200	31	438.14	60.29	90.00%	100.00%
100	200	30	590.96	29.00	89.00%	100.00%
120	200	29	622.95	45.33	90.00%	100.00%
140	200	28	754.42	23.24	96.67%	100.00%
160	200	24	924.38	55.01	92.50%	100.00%
180	200	22	1089.31	124.90	90.00%	100.00%

Table 1: *Principaux résultats sur la base des Crabes.*

Remarque 3.3 : Expériences menées

Expérimentalement, nous avons constaté que le gradient classique (3.13) était instable, tout comme la 2e formulation de la Log-vraisemblance (3.26). En effet, on observe que la 2e Log-vraisemblance n'augmente pas systématiquement au cours des itérations de l'algorithme.

Les résultats complets de nos expériences se trouvent en annexe (voir Annexe A.6).

Nos expériences sont résumées par le graphique suivant :

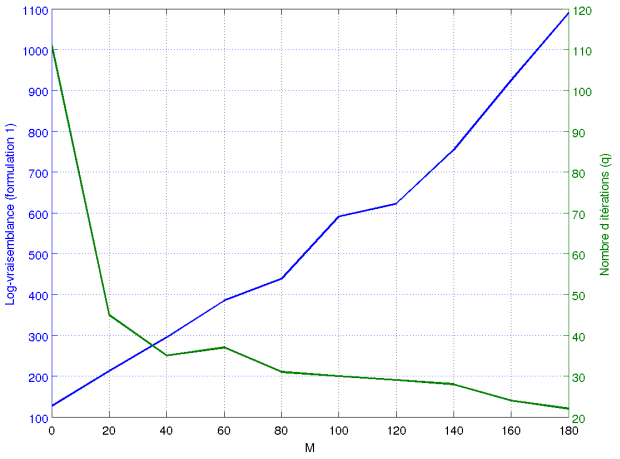


Figure 10: *Croissance de la Log-vraisemblance $\mathcal{L}_1$ et décroissance du nombre d'itérations (q) en fonction de la taille de la base labellisée M.*

Par conséquent, dans la suite de nos travaux, nous utiliserons exclusivement un modèle d'IFA avec une montée de gradient dans la direction du gradient naturel et avec une stabilisation de la vraisemblance basée sur la 1re formulation de la Log-vraisemblance.

⚠ Nous avons fait le choix d'étudier le modèle d'IFA sans bruit en mode semi-supervisé. Cependant, une extension vers le mode de supervision douce a été envisagée : les justifications théoriques et un pseudo-code se trouvent en annexe (voir Annexe A.7).

Extension au mode de supervision douce

Un étiquetage par des *labels doux*, basé sur les fonctions de croyance, a été proposé au cours de la thèse (Côme, 2009 [9] pages 123-127). Ce cadre de supervision généralise l'extension au semi-supervisé. Il s'appuye sur la *Théorie de l'évidence*[4] (Shafer, 1976 [23]). Ce cadre permet de traiter des données incertaines et imprécises. Historiquement, P. Dempster et G. Shafer sont à l'origine de cette théorie puis P. Smets en propose une axiomatique (Smets, 1990 [24]). Aussi, ce dernier propose un modèle de croyance transférable, ce qui permet de définir le conditionnement (Bloch, 1995 [4]).

Comme dans la partie précédente, on cherche à ajouter de l'information aux étiquettes des observations dans le but de nous affranchir de l'*ordre de permutation* des sources. On considère ici qu'une distribution de probabilité est une fonction de masse particulière. On fait quelques rappels sur les principaux objets de la Théorie de l'évidence en annexe (voir Annexe A.7). L'étiquette de chaque i^{e} observation sera modélisée par une plausibilité pl_{ik_s} (issue de la fonction de masse m_i).

De plus, il existe une analogie entre la fonction de vraisemblance et la fonction de plausibilité des observations conditionnellement au paramètre global. L'objectif est alors d'estimer le paramètre global en maximisant cette fonction de plausibilité :

$$\Phi^* = \arg\max_{\Phi} Pl(\Phi; \mathbb{X}) \tag{3.27}$$

En passant au logarithme, on obtient le critère en supervision douce :

$$\begin{aligned} \mathcal{P}l(\Phi; \mathbb{X}) &= -N \cdot \log|\det(A)| \\ &+ \sum_{i=1}^{N}\sum_{s=1}^{S} \log\left(\sum_{k_s=1}^{K_s} pl_{ik_s} \cdot \pi_{k_s} \cdot \mathcal{N}(y_{is}; \theta_{k_s})\right) + Cste \end{aligned} \tag{3.28}$$

[4] "Théorie des possibilités et théorie des probabilités apparaissent comme descendantes d'un ancêtre commun, la théorie de l'évidence", (Bouchon-Meunier, 2007 [5] page 56).

4 Application au diagnostic des Circuits de Voie

4.1 Présentation du problème

Le circuit de voie est un élément essentiel de la sécurité sur les Lignes à Grande Vitesse (LGV) des réseaux ferrés français. Sur ce type de lignes, la circulation des trains peut atteindre jusqu'à 300 km/h. Un circuit de voie permet de détecter la présence d'un train sur une portion de voie ; il se compose d'un émetteur, d'un récepteur et d'une ligne de transmission. Le système Transmission Voie-Machine (TVM) assure la transmission d'information de signalisation (vitesse maximale autorisée, pente moyenne sur la voie,...) en continu entre la voie et les véhicules.

Un seul type de signal concerne nos travaux :
Icc (Ampère) : intensité du courant électrique (amplitude de la porteuse injectée dans les rails)

Ces données sont échantillonnées le long de circuits de voie avec un pas d'environ $1\,m$ par un véhicule, nommé "IRIS". L'INRETS reçoit des enregistrements de ces échantillons tous les 15 jours environ.

Le courant Icc circule entre un *émetteur* et un *récepteur* posés sur les rails. Un signal fort mesuré au niveau du récepteur (supérieur à un seuil d'environ 800 mA) signifie que la voie est libre : pas de véhicule sur la voie. A l'inverse, un signal trop faible s'explique par la présence d'un véhicule sur la voie ferrée (*shunt* : court-circuite les rails), ou par l'existence d'un *défaut* sur le courant Icc. Un *circuit de voie* (CdV) mesure entre 800 et 2500 mètres.

Du fait de sa longueur, un circuit de voie subit un *affaiblissement linéique* : phénomène physique qui fait qu'un signal donné ne peut être transmis que sur une distance limitée. Autrement dit, la valeur de l'*inductance* (= indice de résistance au signal) est élevée sur un circuit de voie. Pour améliorer la transmission du signal, des *condensateurs de compensation*[5] sont installés tout le long du circuit de voie ; ceux-ci sont espacés de 60 à 80 mètres. Un condensateur peu être comparé à un réservoir d'énergie : d'une borne, il charge une quantité électrique et de l'autre, il décharge l'énergie accumulée.

Un circuit de voie peut être défectueux pour plusieurs raisons : interférences entre zones du CdV (diaphonie transversale/longitudinale), connectique de l'émetteur/-récepteur, condensateurs arrachés/mal fixés/viellissants (pertes par conduction/pertes par hystérésis diélectrique)... Les condensateurs feront l'objet du diagnostic dont les données observées seront extraites du signal Icc.

De plus amples informations d'ordre technique (sur la structure physique des circuits de voie), se trouvent dans la thèse de A. Debiolles (Debiolles, 2007 [13]).

[5] Les *condensateurs au polypropylène*, reconnus pour leur fiabilité, compensent l'affaiblissement du signal.

Chaque *arche* du signal Icc représente la compensation de l'affaiblissement linéique par un condensateur.

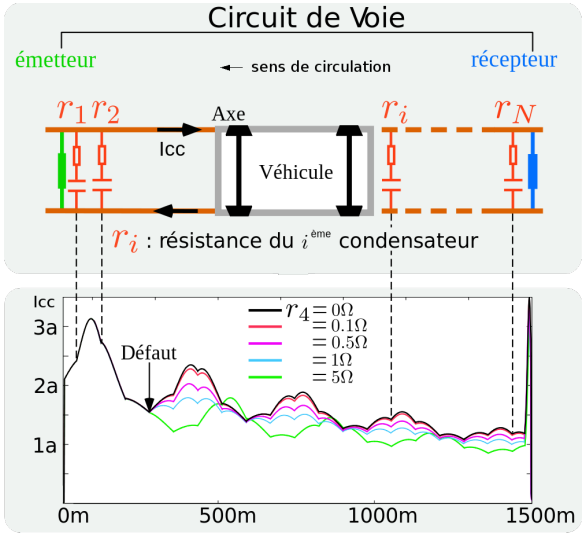


Figure 11: *Schéma d'un circuit de voie ferroviaire.*

Par rapport au modèle graphique qui modélise l'IFA (sous-section 3.3.1), on reprend les notations :

N ≡ nombre d'exemples de circuits de voie
L ≡ nombre de condensateurs (pour chaque circuit de voie)
p ≡ nombre de coefficients nécessaires pour caractériser chaque arche du signal Icc
Y_{is} ≡ capacité d'un condensateur du i^e circuit de voie (la capacité d'un condensateur sans défaut est de $22\mu F \pm 10\%$)
Z_{is} ≡ état d'un condensateur du i^e circuit de voie (3 états possibles : pas de défaut, petit défaut et gros défaut)

On rappelle que : $S = p \cdot L$. Il y aura donc S variables latentes mais seules L seront des sources pertinentes, les $(p-1) \cdot L$ autres seront considérées comme du bruit.

Les mesures du signal Icc sont entachées d'un *bruit additif*. Il s'agit en réalité de la somme de deux types de bruits : bruit de mesure (bruit blanc issu d'une loi normale) et bruit dû à la qualité du contact roue/rail (bruit asymétrique).

Chaque arche de la courbe Icc est estimée par un polynôme de degré 2 (méthode des splines) du type :
$\forall (i,s) \in \{1,\dots,N\} \times \{1,\dots,L\}$,

$$a_{is} \cdot t^2 + b_{is} \cdot t + c_{is} \tag{4.1}$$

où $\begin{cases} a_{is},\ b_{is},\ \text{et } c_{is} \text{ sont les paramètres à estimer} \\ t \in \mathbb{R} \text{ est un échantillon du signal Icc sans bruit} \end{cases}$

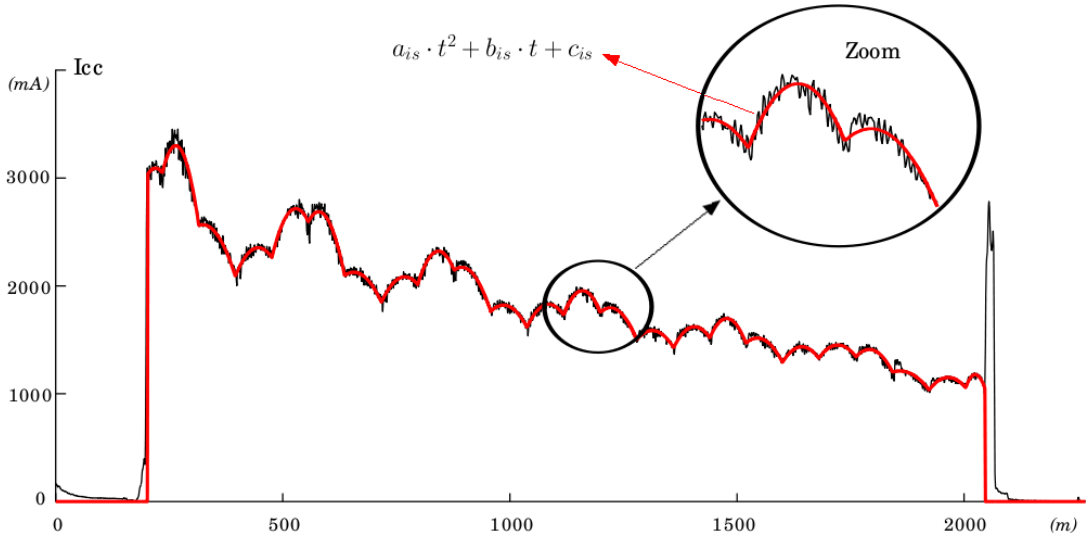


Figure 12: *Extraction des données observées à partir du signal Icc.*

Il a été montré que deux coefficients suffisent à l'estimation d'une arche (Côme, 2009 [9]).
En effet, une relation continue relie approximativement le coefficient $c_{i,s+1}$ au triplet (a_{is}, b_{is}, c_{is}) :
$a_{is} \cdot t^2 + b_{is} \cdot t + c_{is} = c_{i,s+1}$, où t est le dernier échantillon de la s^{e} arche. On choisit de décrire la s^{e} arche du i^{e} circuit de voie par le couple (b_{is}, c_{is}), car le coefficient a_{is} est le coefficient le plus bruité parmi les trois.

⚠ Ces coefficients correspondront aux données observées mises en entrée du modèle d'apprentissage.

Quatre fréquences différentes sont utilisées de façon alternée pour éliminer la diaphonie longitudinale et transversale. On remarque que les condensateurs sont placés à différentes distances ($\pm$ 4 m) selon la fréquence du signal employée sur le circuit de voie :

$$\begin{array}{lcl} 60\ m & : & \text{fréquences 1700 } Hz \text{ et 2000 } Hz \\ 80\ m & : & \text{fréquences 2300 } Hz \text{ et 2600 } Hz \end{array} \qquad (4.2)$$

Par conséquent, on est capable de déterminer le nombre de condensateurs (nombre de sources pertinentes) pour chaque circuit de voie en fonction de sa longueur. C'est une hypothèse forte de notre modélisation.

Le second a priori est motivé par la structure même du Circuit de voie : un condensateur défectueux va dégrader l'ensemble des signatures (arches) des condensateurs en aval de celui-ci (en direction du récepteur) (voir figure 11). Il en résulte une indépendance statistique entre certaines variables observées et certaines variables latentes.

⚠ Pour prendre en considération cette hypothèse physique, nous nous plaçons dans le cadre d'une Analyse en variables latentes indépendantes sans bruit. Autrement dit, on étudie directement (sans prétraitements pour ne pas détruire la structure du problème) le modèle linéaire suivant : $X = Y \cdot A^T$ (3.3).

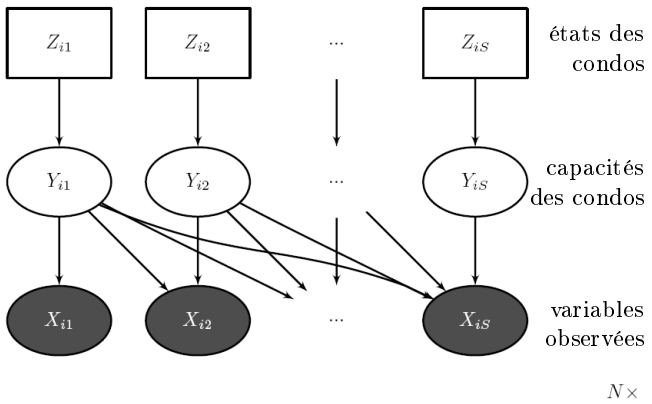


Figure 13: *Modèle graphique de l'IFA sans bruit appliquée aux Circuits de Voie avec contraintes spatiales.*

4.2 Modélisation avec contraintes spatiales

Deux extensions majeures de l'IFA ont été proposées pour cette application dans la thèse (Côme, 2009 [9]).

La première extension contraint les labels (étiquettes) en changeant le mode d'apprentissage (semi-supervisé 3.3.2, supervision douce 3.3.2), la deuxième extension porte sur le processus de mixage (contraintes sur la matrice de mixage A).

Concrètement, on va appliquer un masque sur la matrice de mixage pour annuler certains de ses coefficients. Cette propriété vient de la notion même d'indépendance (Côme, 2009 [9] pages 116-118) :

$$\begin{array}{l} \text{Si } g > h, \text{ alors } X_{ih} \perp\!\!\!\perp Y_{ig} \\ \iff \quad A^T_{gh} = 0 \quad (\text{ou } A_{hg} = 0\ ) \end{array} \qquad (4.3)$$

Ainsi, nous allons utiliser le produit d'Hadamard $\odot$ (ou "produit composante par composante") entre la matrice de mixage $A_{(S \times S)}$ et une matrice binaire $M_{(S \times S)}$,

$$\text{avec } M_{hg} = \begin{cases} 0, & \text{si } X_{ih} \perp\!\!\!\perp Y_{ig} \\ 1, & \text{sinon} \end{cases} :$$

$$B = M \odot A \quad , \text{ où } B_{hg} = \begin{cases} 0, & \text{si } X_{ih} \perp\!\!\!\perp Y_{ig} \ (\text{où } g > h) \\ A_{hg}, & \text{sinon} \end{cases} \tag{4.4}$$

La matrice M est triangulaire inférieure dans l'hypothèse où les composantes des observations sont de dimension 1 ($\Leftrightarrow \ p = 1$).

Dans l'application, les composantes des observations sont modélisées par 2 coefficients ($\Leftrightarrow \ p = 2$).
La matrice masque $M_{(S \times S)}$ est une matrice par blocs :

$$M = \begin{pmatrix} 1 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 & \cdots & 1 \end{pmatrix} \tag{4.5}$$

Il a été choisi de contraindre la matrice de mixage A et non la matrice de démixage W. Concrètement, cet a priori sera ajouté au moment de l'initialisation et lors du calcul du gradient.

Remarque 4.1 : Contraintes sur la matrice de démixage

La matrice de démixage W ne pourrait pas être contrainte de manière analogue. Une matrice ayant certains de ses coefficients nuls n'implique pas toujours que l'inverse de cette matrice ait des zéros aux mêmes places. D'un point de vue statistique, l'annulation de certains coefficients dans la matrice de démixage se traduit par une indépendance conditionnelle (Côme, 2009 [9] pages 116-118) :

Si $g > h$, alors $(X_{ih} \perp\!\!\!\perp Y_{ig}) | X_{i1}, \ldots, X_{i,h-1}, X_{i,h+1}, \ldots, X_{iS}$

$$\iff \quad W_{hg} = 0 \tag{4.6}$$

Expériences sur données théoriques

Nous allons tester le modèle IFA en mode semi-supervisé avec les informations a priori sur la structure de l'application décrites précédemment (algorithme 8). A partir d'un programme de génération de CdV théoriques, nous avons créé une base de 4000 circuits de voie d'une longueur de 1500 m contenant 18 condensateurs et dont la fréquence du signal Icc est de 2300 Hz.

Formellement, on a :

$\#BD$	$=$	4000	$\equiv$ nombre d'individus
M	$\in$	$\{0, \ldots, 500\}$	$\equiv$ nombre d'individus labellisés
N	$=$	500	$\equiv$ taille de la base d'apprentissage
T	$=$	2000	$\equiv$ taille de la base de test
S	$=$	36	$\equiv$ nombre de variables observées
L	$=$	18	$\equiv$ nombre de sources pertinentes
K	$=$	3	$\equiv$ nombre de classes par source pertinente

Nous avons testé cette base avec/sans les contraintes[6]. L'ensemble des calculs a duré environ trois jours.

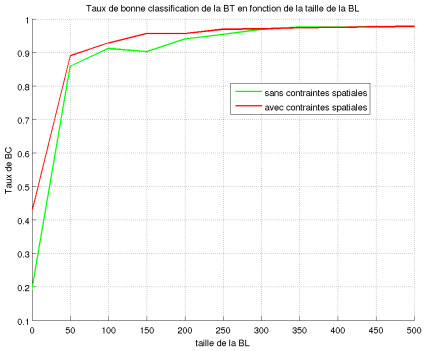


Figure 14: *Taux de bonne classification* TBC_1 *(A.8) sur l'état de défaut des circuits de voie de la base de test.*

L'apport des contraintes spatiales (a priori sur la matrice de mixage) améliore les performances de classification des circuits de voie mais aussi la corrélation entre les capacités estimées et réelles des condensateurs :

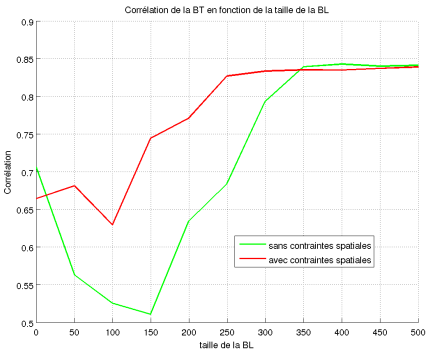


Figure 15: *Corrélation entre les capacités estimées et les capacités réelles des circuits de voie de la base de test.*

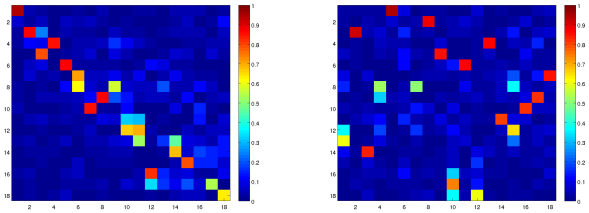


Figure 16: *Matrice de corrélation des capacités lorsqu'aucune des données d'apprentissage n'est labellisée (avec/sans contraintes spatiales).*

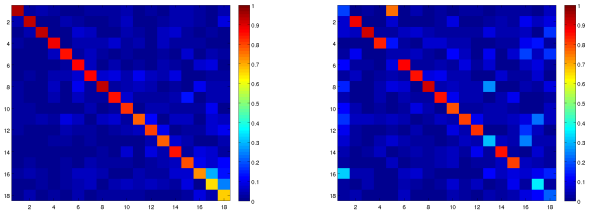


Figure 17: *Matrice de corrélation des capacités lorsque la moitié des données d'apprentissage est labellisée (avec/sans contraintes spatiales).*

[6] L'ensemble des fichiers résumant ces expériences se trouve à la page http://nicolas.cheifetz.free.fr/stages/Figures/IFA/.

4.3 Système complexe avec nombre variable de sources

Dans la thèse (Côme, 2009 [9]), la taille de chaque circuit de voie est supposée constante (même nombre de condensateurs pour tous les CdVs). Or, il est fréquent de voir le nombre de condensateurs varier selon la longueur du circuit de voie. Le cardinal avoisine toujours la vingtaine d'individus.

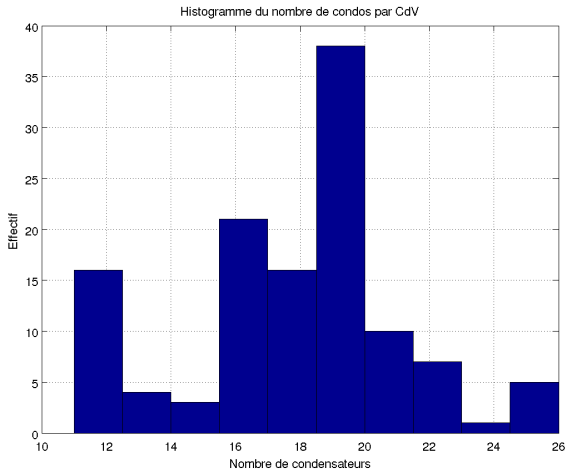


Figure 18: *Histogramme du nombre de condensateurs par circuit de voie, dans une base réelle de* 121 *CdVs.*

De plus, il est facile de déterminer le nombre de condensateurs à partir des signaux réels. On connaît le nombre de condensateurs (sources pertinentes) pour chaque circuit de voie (observation) ; c'est une hypothèse forte (4.1). Nous allons donc développer une extension de l'outil de diagnostic pour prendre en compte l'information disponible de tous les circuits de voie, quelle que soit leur taille.

Pour développer cette extension, nous avons repris les motivations de l'a priori sur le processus de mixage entre les variables observées et les sources. Le point de départ de la réflexion est l'équation : $X = Y \cdot A^T$ (3.3).

- Tout d'abord, on détaille un exemple.

Considérons l'observation x_i d'un circuit de voie composé de 3 condensateurs ($L = 3/S = 6$) :

$$x_i = y_i \cdot A^T$$

$$\Leftrightarrow \quad A^T = \begin{array}{|cccccc|}\hline \star & \star & \star & \star & \star & \star \\ 0 & \star & \star & 0 & \star & \star \\ 0 & 0 & \star & 0 & 0 & \star \\ \star & \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star & \star \\ \hline \end{array}$$

$$\begin{array}{|cccccc|} \hline y_{i1} & y_{i2} & y_{i3} & y_{i4} & y_{i5} & y_{i6} \\ \hline \end{array} \quad \begin{array}{|ccc|ccc|} \hline b_{i1} & b_{i2} & b_{i3} & c_{i1} & c_{i2} & c_{i3} \\ \hline \end{array}$$

$$\begin{array}{c} || \\ y_i \end{array} \qquad \begin{array}{c} || \\ x_i \end{array}$$

Le modèle graphique correspondant à cet exemple, est le suivant[7] :

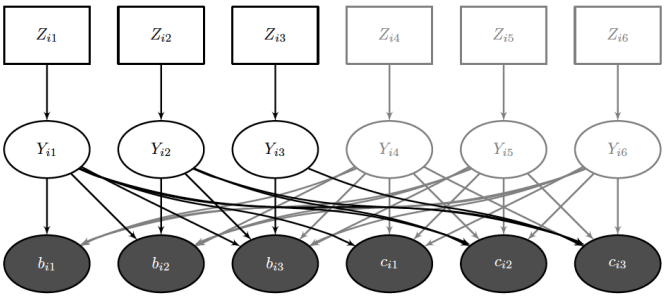


Figure 19: *Modèle graphique de l'IFA appliquée à un Circuit de Voie de trois condensateurs.*

Dans cet exemple, avec un nombre constant de sources, le paramètre global à estimer est :

$$\Phi = (A, (\pi_s, \theta_s)_{s\in\{1,\dots,6\}})$$

Cas général pour un apprentissage sur trois bases

Considérons trois bases d'apprentissage contenant chacune des circuits de voie d'une taille différente :

- $\mathcal{B}_1$ contient D_1 circuits de voie composés de L_1 condensateurs chacun
- $\mathcal{B}_2$ contient D_2 circuits de voie composés de L_2 condensateurs chacun
- $\mathcal{B}_3$ contient D_3 circuits de voie composés de L_3 condensateurs chacun

où $L_1 > L_2 > L_3$ et $S_v = 2 \cdot L_v,\ \forall v \in \{1, 2, 3\}$.

L'objectif est d'estimer le paramètre global suivant :

$$\Phi = \left(A_1, A_2, A_3, (\pi_s, \theta_s)_{s\in\{1,\dots,S_3,\dots,S_2,\dots,S_1\}}\right)$$

L'idée est qu'un circuit de voie pourra servir à l'apprentissage d'autres circuits de voie de même taille ou plus petits ; ainsi, on pourra augmenter la fiabilité de l'estimation du paramètre global. Cette hypothèse est dûe aux relations de dépendances/indépendances qui unissent sources et variables observées.

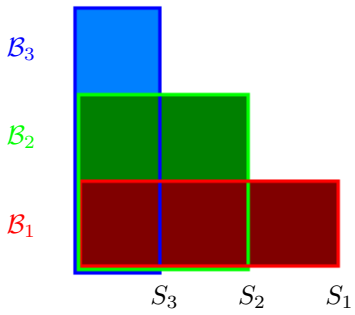


Figure 20: *Trois bases d'apprentissage pour un seul apprentissage.*

[7] Les sources-bruits Y_{i4}, Y_{i5} et Y_{i6}, leurs dépendances avec les variables observées, et leurs classes, ont un teint grisé.

On construit alors trois nouveaux ensembles d'apprentissage $\mathbb{X}_1$, $\mathbb{X}_2$ et $\mathbb{X}_3$, où $\mathbb{X}_v$ est la matrice des observations dont les variables observées sont issues d'au moins D_v sources pertinentes.

Autrement dit, on a :
- $\mathbb{X}_1$ $_{(N_1 \times S_1)}$ est obtenue à partir de $\mathcal{B}_1$
- $\mathbb{X}_2$ $_{(N_2 \times S_2)}$ est obtenue à partir de $\mathcal{B}_1$ et $\mathcal{B}_2$
- $\mathbb{X}_3$ $_{(N_3 \times S_3)}$ est obtenue à partir de $\mathcal{B}_1$, $\mathcal{B}_2$ et $\mathcal{B}_3$

avec $N_1 = D_1$, $N_2 = D_1 + D_2$, et $N_3 = D_1 + D_2 + D_3$.

L'algorithme GEM alterne entre paramétrisation des densités des sources et mise à jour de la matrice de mixage :

- La densité de chaque source pertinente sera paramètrée par la matrice d'observations contenant suffisamment de sources pertinentes et le plus d'individus :

$$\underbrace{1, \dots, L_3}_{\mathbb{X}_3}, \underbrace{L_3+1, \dots, L_2}_{\mathbb{X}_2}, \underbrace{L_2+1, \dots, L_1}_{\mathbb{X}_1} \tag{4.7}$$

  On rappelle que les densités des sources-bruits sont des densités normales centrées-réduites.
- Chaque matrice de mixage est déterminée à partir des sources associées et de l'ensemble des paramètres des densités de mélange. Les matrices de mixage sont mises à jour indépendamment car elles sont calculées à partir de sources différentes. On sauve ainsi les relations de dépendance/indépendance existant entre les variables observées et les sources de chaque ensemble d'apprentissage.

D'une certaine manière, les matrices de mixages sont liées par construction : on obtient les paramètres gaussiens sur l'ensemble des observations et les matrices de mixage sont calculées à partir de ces mêmes paramètres gaussiens. On considère donc que le paramètre global est estimé au moyen d'un unique apprentissage.

Le pseudo-code correspondant est en annexe (voir algorithme 9).

Enfin, il est nécessaire de se doter d'un critère pour détecter la convergence et sélectionner le meilleur paramètre global parmi tous les lancements de l'IFA. Pour agréger l'information de toutes les données observées, nous allons utiliser une variante de la Log-vraisemblance (3.24) :

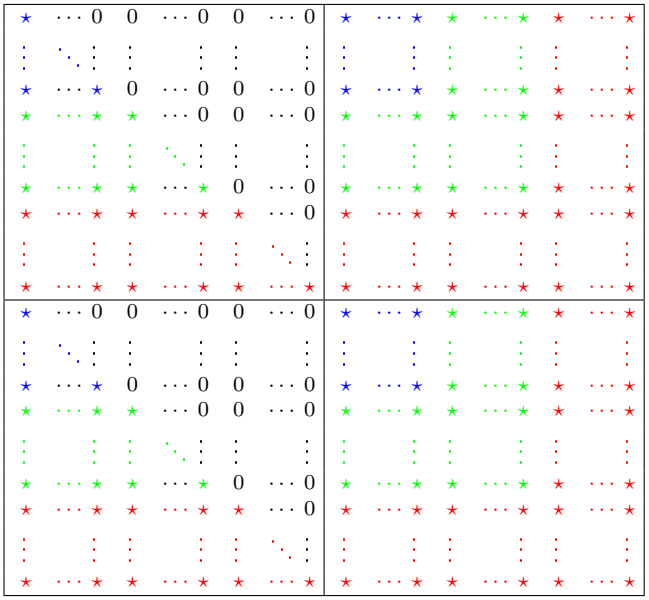
$$\begin{aligned}\mathcal{L}_3(\Phi; \mathbb{X}_1, \dots, \mathbb{X}_V) \;=\; & \sum_{v=1}^{V} -N_v \cdot \log|\det(A_v)| \\ & + \sum_{i=1}^{M_v}\sum_{s=1}^{S_v}\sum_{k_s=1}^{K_s} \mathbb{1}_{Z_{is}=C_{k_s}} \cdot \log\left(\pi_{k_s} \cdot \mathcal{N}(x_i \cdot {}^{T}A_v^{-1}{}_{(.s)}; \theta_{k_s})\right) \\ & + \sum_{i=M_v+1}^{N_v}\sum_{s=1}^{S_v} \log\left(\sum_{k_s=1}^{K_s} \pi_{k_s} \cdot \mathcal{N}(x_i \cdot {}^{T}A_v^{-1}{}_{(.s)}; \theta_{k_s})\right)\end{aligned} \tag{4.8}$$

où A_v $_{(S_v \times S_v)}$ est la matrice de mixage associée à la matrice d'observations $\mathbb{X}_v$.

Remarque 4.2 : Apprentissage du paramètre global avec une seule matrice de mixage ?

Nous avons testé une autre extension pour laquelle les paramètres gaussiens sont déterminés de la même manière mais une seule matrice de mixage est apprise pour des données observées de tailles différentes. On extrait toutes les matrices de l'unique matrice de mixage apprise.

Dans le cas de trois bases, la matrice de mixage $_{(S_1 \times S_1)}$ serait de la forme suivante :



Lors de l'étape de la mise à jour de la fonction g (3.18), on construit une matrice g de taille $_{(S_1 \times (N_1+N_2+N_3))}$.
A l'étape suivante, on calcule le gradient en multipliant cette matrice par les sources $\mathbb{Y}_v$ mises côte à côte en ligne. On obtient une matrice $_{((N_1+N_2+N_3) \times S_1)}$.
Cette mise en oeuvre, bien que plus simple à l'exécution (une seule montée de gradient à effectuer), a l'inconvénient d'ajouter des relations de dépendance/ indépendance entre les variables observées et les sources.
En effet, la mise à jour de la matrice de mixage intégrerait des liaisons supplémentaires pour les individus de petite taille.

Expériences sur données théoriques

Nous allons tester le modèle IFA avec contraintes spatiales en mode semi-supervisé dont l'apprentissage sera réalisé à partir de trois bases de tailles différentes. A partir d'un programme de génération de CdV théoriques, nous avons créé trois bases (fréquence de 2300 Hz) :

- $\mathcal{B}_1$ contient 3000 circuits de voie d'une longueur de 2000 m, et composés de 24 condensateurs chacun
- $\mathcal{B}_2$ contient 3000 circuits de voie d'une longueur de 1700 m, et composés de 20 condensateurs chacun
- $\mathcal{B}_3$ contient 3000 circuits de voie d'une longueur de 1500 m, et composés de 18 condensateurs chacun

Etant donné la durée importante des calculs, nous avons pu tester l'extension uniquement dans un cadre supervisé. Nous présenterons l'ensemble des résultats en semi-supervisé au cours de la soutenance. Formellement, on a :

$$
\begin{array}{rcll}
\#\mathcal{B}_v & = & 3000 & \equiv \text{nombre d'individus de } \mathcal{B}_v \\
N_1 & = & 500 & \equiv \text{nombre d'individus labellisés} \\
N_2 & = & 300 & \equiv \cdots \\
N_3 & = & 300 & \equiv \cdots \\
T & = & 1500 & \equiv \text{taille de la base de test } \forall\ \mathcal{B}_v \\
L_1 & = & 24 & \equiv \text{nombre de sources pertinentes pour chaque individu de } \mathcal{B}_1 \\
L_2 & = & 20 & \equiv \text{nombre de sources pertinentes pour chaque individu de } \mathcal{B}_2 \\
L_3 & = & 18 & \equiv \text{nombre de sources pertinentes pour chaque individu de } \mathcal{B}_3 \\
K & = & 3 & \equiv \text{nombre de classes par source pertinente} \\
 & & 20 & \equiv \text{nombre de lancements}
\end{array}
$$

avec $v \in \{1, 2, 3\}$.

Nous présentons ici les performances du modèle "extension" appris sur trois bases de tailles différentes, contre un modèle "classique" (algorithme 8) pour chaque base.

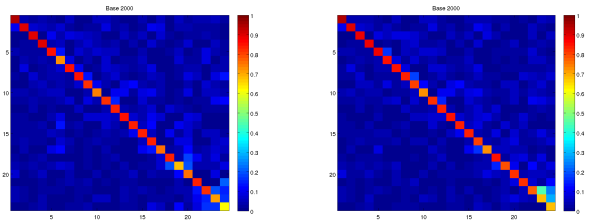


Figure 21: *Matrice de corrélation des capacités des* 24 *condensateurs lorsque toutes les données d'apprentissage sont labellisées pour la base* $\mathcal{B}_1$ *(extension/classique).*

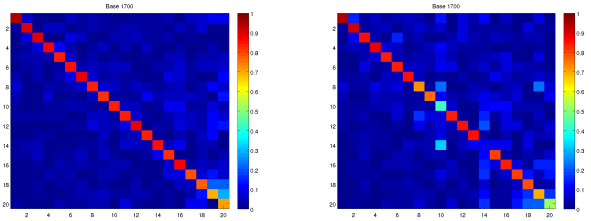


Figure 22: *Matrice de corrélation des capacités des* 20 *condensateurs lorsque toutes les données d'apprentissage sont labellisées pour la base* $\mathcal{B}_2$ *(extension/classique).*

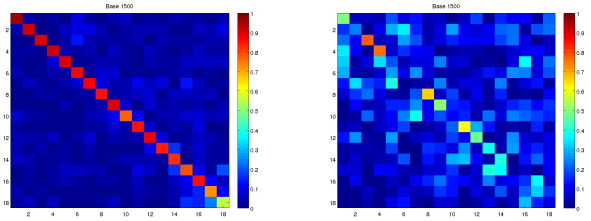


Figure 23: *Matrice de corrélation des capacités des* 18 *condensateurs lorsque toutes les données d'apprentissage sont labellisées pour la base* $\mathcal{B}_3$ *(extension/classique)).*

On constate expérimentalement que l'apport de l'extension est de plus en plus significatif à mesure que la taille des exemples de circuits de voie diminue.

	Corrélation			TBC_1		
	$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$	$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$
Extension	81.3	83.7	84.3	97.21	97.67	97.95
Classique	82.0	75.7	41.90	97.21	96.35	95.17

	TBC_2		
	$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$
Extension	78.40	82.26	83.91
Classique	81.49	72.31	50.90

Table 2: Performance (en %) du modèle classique/extension en mode supervisé avec des circuits de voie de taille variable.

5 Conclusion et perspectives

Nous avons étudié dans ce rapport différentes variantes d'une méthode générative, l'Analyse en Facteurs Indépendants. Nous avons présenté ce modèle en détail tout en le justifiant. A partir de cette base théorique, nous avons testé plusieurs modèles dédiés au diagnostic d'un système complexe : le circuit de voie. Notre modélisation exploite les caractéristiques de ce système complexe, comme la relation conditionnelle qui unit chacun de ses sous-systèmes. Nos résultats démontrent la pertinence de cette méthode dans un cadre semi-supervisé. Nous avons également proposé une extension capable d'apprendre sur des circuits de voie dont le nombre de sous-systèmes est variable, et un critère à partir de ce modèle génératif.

Cependant, l'extension n'améliore pas les performances, en terme de durée de l'exécution, bien qu'un unique apprentissage soit réalisé. Cette "lenteur" est due à la longueur de la recherche linéaire pour fixer le pas de la montée de gradient. Nous avons alors envisagé de remplacer la mise à jour du gradient par la méthode de Quasi-Newton à mémoire limitée LBFGS[8] (Zhu et al., 1994 [26]). De plus, le modèle étudié admet une relation linéaire entre les variables observées et les variables latentes continues. Ce modèle pourrait être étendue au cas non linéaire[9]. Enfin, on teste la convergence de notre algorithme et on sélectionne le meilleur paramètre global par un seul critère : la vraisemblance. Or, il existe plusieurs critères issus de l'Analyse en Composantes Indépendantes (décorrélation entre $\mathbb{Y}$ et $g(\mathbb{Y})$, critère sur les densités de $\mathbb{Y}$). On pourrait alors agréger plusieurs de ces critères.

[8] Une implémentation Matlab est disponible à l'adresse `http://www.cs.toronto.edu/~liam/software.shtml`.

[9] Analyse Factorielle Non Linéaire (Valpola, 2002).

A Annexes

A.1 Représentation graphique des modèles

Pour donner une description synthétique des modèles, nous les présentons sous la forme de modèles graphiques. Apparus dans les années 80, un modèle graphique se caractérise par un graphe orienté acyclique et une méthode de calcul pour l'inférence. Les modèles graphiques sont au confluent de la théorie des graphes et de la théorie des probabilités. Cependant, les modalités d'inférence sur les modèles graphiques/réseaux bayésiens ne sont pas exploités dans ces travaux. On utilise simplement la représentation des modèles graphiques pour faciliter la compréhension des modèles étudiés en explicitant la nature des variables aléatoires et les dépendances entre chacune d'elles.

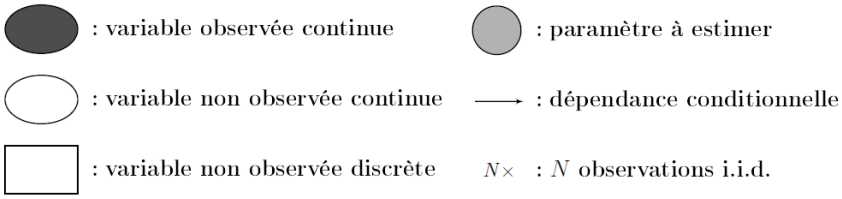


Figure 24: *Structure d'un modèle graphique.*

Nous reprenons ici les conventions décrites dans l'*Introduction aux modèles graphiques* (Jordan, 1997 [17]).

A.2 Comparaison mélange gaussien standard/parcimonieux

Dans le cas d'un mélange gaussien parcimonieux avec l'hypothèse d'*homoscedasticité* (indépendance entre les observations d'une même classe), la matrice de variance-covariance est diagonale. En supposant que l'on considère un problème à K classes, le nombre total de coefficients des paramètres à estimer (Celeux et al, 1995 [6]) est :

- $\frac{K\cdot(p+1)\cdot(p+2)}{2} - 1$ pour un modèle de mélange standard
- $K \cdot (2p+1) - 1$ pour un modèle de mélange parcimonieux

où p est la dimension de chaque observation.

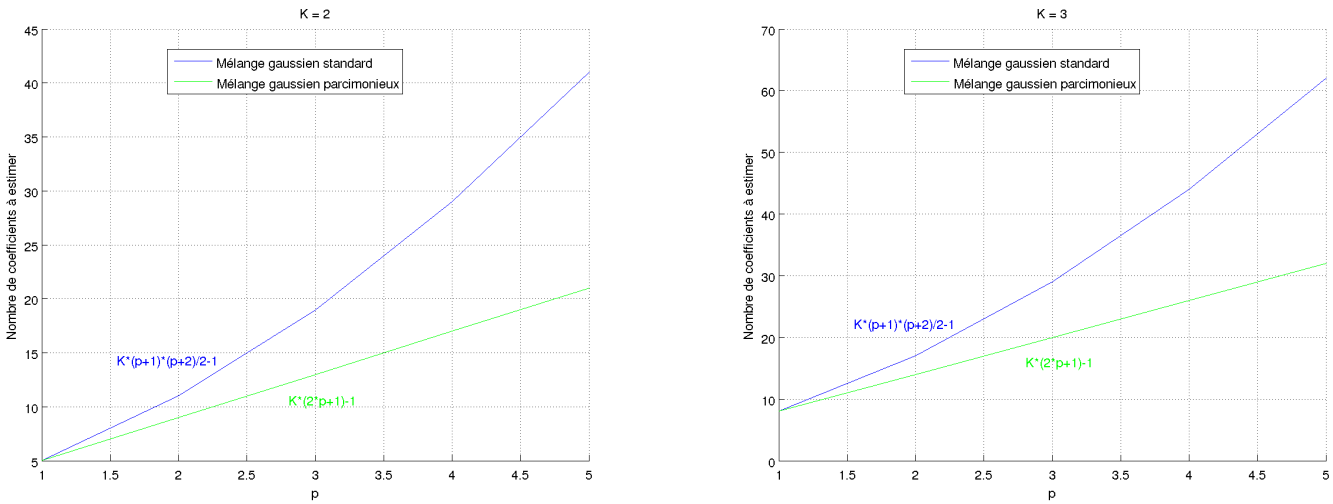


Figure 25: *Nombre de coefficients à estimer pour un modèle de mélange gaussien standard/parcimonieux.*

De manière générale, on constate qu'un modèle de mélange gaussien parcimonieux est moins coûteux (en nombre de coefficients à estimer) qu'un modèle de mélange gaussien standard.

Toutefois, on remarque que le nombre d'estimation est identique lorsque les données sont monodimensionnelles ($p = 1$) pour un modèle parcimonieux ou non ; dans nos travaux, toutes les données mises en entrée de l'algorithme EM sont supposées monodimensionnelles. On calcule 5 coefficients dans le cas biclasse et 8, dans le cas triclasse.

A.3 Croissance de la vraisemblance au cours de l'algorithme EM

Nous allons montrer que la Log-vraisemblance des données observées $\mathcal{L}(\Phi^{(q)};\mathbb{X})$ augmente à chaque itération (q) de l'algorithme EM. Pour cela, nous allons prouver l'inégalité :

$$\begin{aligned} & \mathcal{L}(\Phi^{(q+1)};\mathbb{X}) - \mathcal{L}(\Phi^{(q)};\mathbb{X}) \geq 0 \\ \Leftrightarrow \quad & \left(Q(\Phi^{(q+1)},\Phi^{(q)}) - Q(\Phi^{(q)},\Phi^{(q)})\right) - \left(H(\Phi^{(q+1)},\Phi^{(q)}) - H(\Phi^{(q)},\Phi^{(q)})\right) \geq 0 \end{aligned} \tag{A.1}$$

- Par construction, l'étape M consiste à maximiser la fonction Q :

$$\begin{aligned} \Phi^{(q+1)} &= \arg\max_{\Phi}\ Q(\Phi,\Phi^{(q)}) \\ &= \arg\max_{\Phi} \left\{ \mathbb{E}_{X,Y}\left(\log p(X,Y;\Phi)|\mathbb{X},\Phi^{(q)}\right)\right\} \\ &= \arg\max_{\Phi} \left\{ \mathbb{E}_{Y}\left(\log p(\mathbb{X},Y;\Phi)|\mathbb{X},\Phi^{(q)}\right)\right\} \\ &= \arg\max_{\Phi} \left\{ \sum_{y\in\mathbb{C}l^N} \log p(\mathbb{X},y;\Phi)\cdot p\left(y|\mathbb{X},\Phi^{(q)}\right)\right\} \end{aligned} \tag{A.2}$$

  Donc, on sait par avance que :

$$Q(\Phi^{(q+1)},\Phi^{(q)}) - Q(\Phi^{(q)},\Phi^{(q)}) \geq 0 \tag{A.3}$$

- L'algorithme EM maximise la fonction Q mais on ne cherche pas à minimiser la fonction H.
  On calcule la fonction H :

$$\begin{aligned} H(\Phi,\Phi^{(q)}) &= \mathbb{E}_{X,Y}\left(\log p(\cancel{X},Y|\mathbb{X};\Phi)|\mathbb{X},\Phi^{(q)}\right) \\ &= \sum_{y\in\mathbb{C}l^N} \log p(y|\mathbb{X};\Phi)\cdot p\left(y|\mathbb{X},\Phi^{(q)}\right) \end{aligned} \tag{A.4}$$

  Montrons que H se dégrade naturellement au cours des itérations de l'algorithme EM :

$$\begin{aligned} H(\Phi^{(q+1)},\Phi^{(q)}) - H(\Phi^{(q)},\Phi^{(q)}) &= \sum_{y\in\mathbb{C}l^N} \log p(y|\mathbb{X};\Phi^{(q+1)})\cdot p\left(y|\mathbb{X},\Phi^{(q)}\right) - \log p(y|\mathbb{X};\Phi^{(q)})\cdot p\left(y|\mathbb{X},\Phi^{(q)}\right) \\ &= \sum_{y\in\mathbb{C}l^N} \log\left(\frac{p(y|\mathbb{X};\Phi^{(q+1)})}{p(y|\mathbb{X};\Phi^{(q)})}\right)\cdot p\left(y|\mathbb{X},\Phi^{(q)}\right) \\ &\leq \log\left(\sum_{y\in\mathbb{C}l^N} \frac{p(y|\mathbb{X};\Phi^{(q+1)})}{p(y|\mathbb{X};\Phi^{(q)})}\cdot p\left(y|\mathbb{X},\Phi^{(q)}\right)\right) \\ &= \log\left(\sum_{y\in\mathbb{C}l^N} p\left(y|\mathbb{X},\Phi^{(q+1)}\right)\right) \\ &= \log 1 \\ &= 0 \end{aligned} \tag{A.5}$$

  Cette inégalité est basée sur le fait que le logarithme est une fonction concave. On applique alors l'*inégalité de Jensen* : $f\left(\sum_i \lambda_i\cdot x_i\right) \leq \sum_i \lambda_i\cdot f(x_i)$, où f est une fonction convexe. Plus d'explications se trouvent dans le livre *Elements of Information Theory* de T. Cover et J. Thomas [8].
  Au final, on trouve bien : $H(\Phi^{(q+1)},\Phi^{(q)}) - H(\Phi^{(q)},\Phi^{(q)}) \leq 0$.

□

On en conlut que $\mathcal{L}(\Phi^{(q+1)};\mathbb{X}) \geq \mathcal{L}(\Phi^{(q)};\mathbb{X})$ et maximiser la fonction Q suffit à garantir la croissance de la vraisemblance au cours de l'algorithme EM.

A.4 Recherche linéaire du paramètre τ par backtracking

Cet algorithme est extrait de la thèse de E. Côme (Côme, 2009 [9] Annexe A.5 page 168).

Fonction `LinearSearch`($\mathbb{X}$, ∇A, $maxIter$, ρ, Φ)

Entrées : données observées centrées réduites $\mathbb{X}$, gradient de la Log-vraisemblance ∇A, nombre maximal d'itérations $maxIter$, paramètre ρ, et, paramètre global $\Phi = (A, \pi_1, \ldots, \pi_S, \theta_1, \ldots, \theta_S)$

 # Initialisation
τ , $it = 0$

Tant que $\left(\mathcal{L}(A + \tau \cdot \nabla A; \mathbb{X}, \Phi) \leq \mathcal{L}(A; \mathbb{X}, \Phi) + \tau \cdot \| \nabla A \|_2{}^2\right)$ && $(it < maxIter)$ **faire**

 # Diminution du pas
$\tau \leftarrow \tau \cdot \rho$
$it \leftarrow it + 1$

Sorties : estimation du meilleur pas τ^*

où, $\| \nabla A \|_2{}^2 = \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} \nabla A_{s_1 s_2}{}^2$.

Concrètement, le paramètre ρ est instancié à 0.1 et $maxIter$ est fixé à 20. Nous allons donc diminuer le pas jusqu'à ce qu'il améliore suffisamment la Log-vraisemblance.

A.5 Calcul du gradient sur la matrice de mixage et mise à jour de la matrice de démixage

- **Calcul du gradient de la Log-vraisemblance par rapport à la matrice de mixage**

◦ $\frac{\partial}{\partial A} \log |\det(A)| = \frac{\frac{\partial}{\partial A} |\det(A)|}{|\det(A)|} = {}^{10} \frac{{}^T A^{-1} \cdot \cancel{|\det(A)|}}{\cancel{|\det(A)|}} = {}^T A^{-1}$

◦ $\frac{\partial}{\partial A} \sum_{i=1}^{N} \log f_{Y_i}(x_i \cdot {}^T A^{-1}) = \sum_{i=1}^{N} \frac{\partial}{\partial y_i} \log f_{Y_i}(x_i \cdot {}^T A^{-1}) \cdot \frac{\partial}{\partial A} y_i = - {}^T A^{-1} \cdot \sum_{i=1}^{N} \frac{\partial}{\partial y_i} \log f_{Y_i}(y_i) \cdot x_i \cdot {}^T A^{-1} = {}^T A^{-1} \cdot \sum_{i=1}^{N} g(y_i) \cdot y_i$

où la fonction g est définie telle que $g : \begin{array}{ccc} \mathbb{R}^S & \longrightarrow & \mathbb{R}^S \\ y_i & \longmapsto & \left(-\frac{\partial}{\partial y_{i1}} \log f_{Y_{i1}}(y_{i1}) \quad , \cdots , \quad -\frac{\partial}{\partial y_{iS}} \log f_{Y_{iS}}(y_{iS}) \right)^T \end{array}$

- **Mise à jour de la matrice de démixage par montée de gradient**

On se place dans le cadre d'une transformation linéaire sans bruit entre les variables observées X et les variables latentes continues Y, avec la matrice de démixage W (3.3) : $Y = X \cdot W^T$.

De manière analogue à la recherche de la meilleure matrice de mixage (3.9), on cherche ici à estimer la meilleure matrice de démixage :

$$\begin{aligned} W^* &= \arg\max_W \mathcal{L}(W; \mathbb{X}) \\ &= \arg\max_W \left\{ N \cdot \log |\det(W)| + \sum_{i=1}^{N} \log f_{Y_i}(x_i \cdot W^T) \right\} \end{aligned} \tag{A.6}$$

◦ $\frac{\partial}{\partial W} N \cdot \log |\det(W)| = N \cdot {}^T W^{-1}$

◦ $\frac{\partial}{\partial W} \sum_{i=1}^{N} \log f_{Y_i}(x_i \cdot W^T) = \sum_{i=1}^{N} \frac{\partial}{\partial y_i} \log f_{Y_i}(x_i \cdot W^T) \cdot \frac{\partial}{\partial W}(x_i \cdot W^T) = -\sum_{i=1}^{N} g(y_i) \cdot x_i = - g(\mathbb{Y}) \cdot \mathbb{X}$

On calcule alors le gradient de la Log-vraisemblance :

$$\frac{\partial}{\partial W} \mathcal{L}(W; \mathbb{X}) = N \cdot {}^T W^{-1} - g(\mathbb{Y}) \cdot \mathbb{X} = (N \cdot \mathrm{I}_S - g(\mathbb{Y}) \cdot \mathbb{Y}) \cdot {}^T W^{-1} \tag{A.7}$$

Enfin, on obtient la montée de gradient dans la direction du gradient naturel (Hyvärinen, 2001 [15] pages 67 – 68) :

$$\begin{aligned} W^{(q+1)} &= W^{(q)} + \tau^{(q)} \cdot \frac{\partial}{\partial W} \mathcal{L}(W^{(q)}; \mathbb{X}) \cdot {}^T W^{(q)} \cdot W^{(q)} \\ &= W^{(q)} + \tau^{(q)} \cdot \left(N \cdot \mathrm{I}_S - g^{(q)}(\mathbb{Y}^{(q)}) \cdot \mathbb{Y}^{(q)}\right) \cdot W^{(q)} \end{aligned} \tag{A.8}$$

[10] Extraits de l'*Analyse en Composantes Indépendantes* (Hyvärinen, 2001 [15] page 61).

A.6 Pseudo-codes pour l'Analyse en Facteurs Indépendants en mode semi-supervisé

- Pseudo-code de la Log-vraisemblance (*formulation* 1) des données observées en mode semi-supervisé (3.24) :

Fonction `LogVraisemblance1(`$\mathbb{X}$, Φ`)`

Entrées : données observées centrées réduites $\mathbb{X}$, et, paramètre global $\Phi = (A, \pi_1, \ldots, \pi_S, \theta_1, \ldots, \theta_S)$
avec A, la matrice de mixage et $(\pi_1, \ldots, \pi_S, \theta_1, \ldots, \theta_S)$, les paramètres des mélanges gaussiens

 # Initialisation
N = nombre d'individus
S = nombre de sources
K = nombre de classes

 # mise à jour des variables latentes continues
$\mathbb{Y} = \mathbb{X} \cdot {}^T A^{-1}$

$L = -N \cdot \log|\det(A)|$
pour chaque $(i, s) \in \{1, \ldots, N\} \times \{1, \ldots, S\}$ **faire**
  si *cas des données labellisées* **alors**
    pour chaque $k_s \in \{1, \ldots, K_s\}$ **faire**
      $L \leftarrow L + \mathbb{1}_{Z_{is}=C_{k_s}} \cdot \log(\pi_{k_s} \cdot \mathcal{N}(y_{is}; \theta_{k_s}))$
  sinon si *cas des données non labellisées* **alors**
    $$L \leftarrow L + \log\left(\sum_{k_s=1}^{K_s} \pi_{k_s} \cdot \mathcal{N}(y_{is}; \theta_{k_s})\right)$$

Sorties : approximation de la log-vraisemblance (*formulation* 1) en mode semi-supervisé

- Pseudo-code de la Log-vraisemblance (*formulation* 2) des données observées en mode semi-supervisé (3.26) :

Fonction `LogVraisemblance2(`$\mathbb{X}$, Φ, $(t_{ik_s})_{1 \leq i \leq N, 1 \leq k_s \leq K_s}$`)`

Entrées : données observées centrées réduites $\mathbb{X}$, paramètre global $\Phi = (A, \pi_1, \ldots, \pi_S, \theta_1, \ldots, \theta_S)$, et,
lois a posteriori t_{ik_s}

 # Initialisation
N = nombre d'individus
S = nombre de sources
K = nombre de classes

 # mise à jour des variables latentes continues
$\mathbb{Y} = \mathbb{X} \cdot {}^T A^{-1}$

$L = -N \cdot \log|\det(A)|$
pour chaque $(i, s, k_s) \in \{1, \ldots, N\} \times \{1, \ldots, S\} \times \{1, \ldots, K_s\}$ **faire**
  $$L \leftarrow L + \sum_{i=1}^{N} \sum_{s=1}^{S} \sum_{k_s=1}^{K_s} t_{ik_s} \cdot \log(\pi_{k_s} \cdot \mathcal{N}(y_{is}; \theta_{k_s}))$$

Sorties : approximation de la log-vraisemblance (*formulation* 2) en mode semi-supervisé

- Pseudo-code de l'Analyse en Facteurs Indépendantrs sans bruit (3.3) sur la matrice de mixage dans un cadre *semi-supervisé* [11] :

[11] On rappelle que le test de convergence est la stabilisation de la Log-vraisemblance (2.14).

Algorithme 6 : Pseudo-code IFA sans bruit sur la matrice de mixage dans un cadre *semi-supervisé*

Entrées : données observées centrées-réduites $\mathbb{X}$, classes des observations labellisées $\mathbb{Z}$, nombre de sources (variables latentes continues) pertinentes, nombre de classes K pour chaque source pertinente, les *sources-bruits* sont mono-classes et suivent une loi normale centrée-réduite $\mathcal{N}(0,1)$, nombre maximal d'itérations, type de gradient à calculer (classique/naturel), et, éventuellement le paramètre global à l'itération 0

 # Initialisation
$q = 0$
$\Phi^{(q)} = \left(A^{(q)}, {\pi_1}^{(q)}, \dots, {\pi_S}^{(q)}, {\theta_1}^{(q)}, \dots, {\theta_S}^{(q)}\right)$
 # plusieurs matrices $A^{(q)}$ sont générées aléatoirement ; on sauve celle qui a le meilleur conditionnement
 # les π_s et θ_s des "sources pertinentes" sont déterminés par un K-means sur les s^{e} composantes de $\mathbb{X}$

Tant que *non convergence* **faire**

*# **1.** Mise à jour des sources (3.3)*
$\mathbb{Y}^{(q)} = \mathbb{X} \cdot {}^{T}{A^{(q)}}^{-1}$

*# **2.** Mise à jour des paramètres des sources (paramètres des mélanges gaussiens) par algorithme EM* (section 2)
*# **Etape E** (Estimation) : calcul des lois a posteriori (3.19) & (3.25)*
pour chaque $(s, k_s) \in \{1, \dots, S\} \times \{1, \dots, K_s\}$ **faire**

$${t_{ik_s}}^{(q)} = \mathbb{1}_{Z_i = C_{k_s}} \qquad ,\forall i \in \{1, \dots, M\}$$

$${t_{ik_s}}^{(q)} = {\pi_{k_s}}^{(q)} \cdot \mathcal{N}\left({y_{is}}^{(q)}; {\theta_{k_s}}^{(q)}\right) \Big/ \sum_{l_s=1}^{K_s} {\pi_{l_s}}^{(q)} \cdot \mathcal{N}\left({y_{is}}^{(q)}; {\theta_{l_s}}^{(q)}\right) \qquad ,\forall i \in \{M+1, \dots, N\}$$

*# **Etape M** (Maximisation) : calcul de* $\left({\pi_1}^{(q+1)}, \dots, {\pi_S}^{(q+1)}, {\theta_1}^{(q+1)}, \dots, {\theta_S}^{(q+1)}\right)$
pour chaque $(s, k_s) \in \{1, \dots, S\} \times \{1, \dots, K_s\}$ **faire**

$${\pi_{k_s}}^{(q+1)} = \sum_{i=1}^{N} {t_{ik_s}}^{(q)} / N$$

$${\mu_{k_s}}^{(q+1)} = \sum_{i=1}^{N} {t_{ik_s}}^{(q)} \cdot {y_{is}}^{(q)} \Big/ \sum_{j=1}^{N} {t_{jk_s}}^{(q)}$$

$${\Sigma_{k_s}}^{(q+1)} = \sum_{i=1}^{N} {t_{ik_s}}^{(q)} \cdot \left({y_{is}}^{(q)} - {\mu_{k_s}}^{(q+1)}\right)^2 \Big/ \sum_{j=1}^{N} {t_{jk_s}}^{(q)}$$

*# **3.** Mise à jour de la fonction g (3.18)*
pour chaque $(i, s) \in \{1, \dots, N\} \times \{1, \dots, S\}$ **faire**

$$g_s^{(q)}({y_i}^{(q)}) = \sum_{k_s=1}^{K_s} {t_{ik_s}}^{(q)} \cdot \frac{{y_{is}}^{(q)} - {\mu_{k_s}}^{(q+1)}}{{\Sigma_{k_s}}^{(q+1)}}$$

*# **4.** Calcul du gradient*

$$\begin{array}{ll} & \nabla A^{(q)} = {}^{T}{A^{(q)}}^{-1} \cdot \left(g^{(q)}(\mathbb{Y}^{(q)}) \cdot \mathbb{Y}^{(q)} - N \cdot \mathrm{I}_S\right) \quad \text{\#"classique" (3.11)} \\ \text{\# ou} & \nabla A^{(q)} = \quad A^{(q)} \quad \cdot \left(g^{(q)}(\mathbb{Y}^{(q)}) \cdot \mathbb{Y}^{(q)} - N \cdot \mathrm{I}_S\right) \quad \text{\#"naturel" (3.12)} \end{array}$$

*# **5.** Recherche linéaire du paramètre τ par backtracking* (cf. Annexe A.4)
$\tau^{(q)} = \text{RechercheLineaire}(\mathbb{X},\ A^{(q)},\ \nabla A^{(q)})$

*# **6.** Mise à jour de la matrice de mixage par montée de gradient (3.13) & (3.14)*
$A^{(q+1)} = A^{(q)} + \tau^{(q)} \cdot \nabla A^{(q)}$

*# **7.** Normalisation des sources (3.20) & (3.21)*
pour chaque $s \in \{1, \dots, S\}$ **faire**

$${\Sigma_s}^{(q)} = \left(\sum_{k_s=1}^{K_s} {\pi_{k_s}}^{(q+1)} \cdot \left({\Sigma_{k_s}}^{(q+1)} + {\mu_{k_s}}^{(q+1)^2}\right)\right) - \left(\sum_{k_s=1}^{K_s} {\pi_{k_s}}^{(q+1)} \cdot {\mu_{k_s}}^{(q+1)}\right)^2$$

$${A_{.s}}^{(q+1)} \leftarrow {A_{.s}}^{(q+1)} \cdot \sqrt{{\Sigma_s}^{(q)}}$$

pour chaque $k_s \in \{1, \dots, K_s\}$ **faire**

$${\mu_{k_s}}^{(q+1)} \leftarrow {\mu_{k_s}}^{(q+1)} \Big/ \sqrt{{\Sigma_s}^{(q)}}$$

$${\Sigma_{k_s}}^{(q+1)} \leftarrow {\Sigma_{k_s}}^{(q+1)} \Big/ {\Sigma_s}^{(q)}$$

$q \leftarrow q + 1$

Sorties : estimation du paramètre global Φ^*, et, une approximation de la Log-vraisemblance sur l'ensemble des données

Expériences sur la base des Crabes

Le nombre maximal d'itérations est fixé à 1000 itérations et le seuil de stabilisation (2.14) est de 10^{-8}.

M	N	gradient	Nombre de lancements	q	Log-vraisemblance $\mathcal{L}_1$	std	Log-vraisemblance $\mathcal{L}_2$	std	Taux de bonne classif sexe	Taux de bonne classif espèce
0	200	naturel	10	111	126.47	32.36	83.46	75.28	87.50%	100.00%
0	200	classique	10	1000	87.69	128.04	37.57	–	50.00%	50.00%
20	200	naturel	10	45	212.24	49.08	88.03	114.52	91.67%	100.00%
20	200	classique	10	403	129.82	155.96	−62.77	162.54	58.89%	100.00%
40	200	naturel	10	35	294.75	48.35	93.97	100.96	91.88%	100.00%
40	200	classique	10	197	234.96	265.92	−48.88	267.78	63.75%	100.00%
60	200	naturel	10	37	384.55	53.85	84.18	99.30	92.86%	100.00%
60	200	classique	10	338	272.58	243.91	−58.86	259.76	60.00%	100.00%
80	200	naturel	10	31	438.14	60.29	89.77	109.25	90.00%	100.00%
80	200	classique	10	231	371.75	221.18	−57.37	216.19	48.33%	100.00%
100	200	naturel	10	30	590.96	29.00	58.58	72.72	89.00%	100.00%
100	200	classique	10	222	190.21	239.63	−235.16	249.18	74.00%	100.00%
120	200	naturel	10	29	622.95	45.33	52.57	60.96	90.00%	100.00%
120	200	classique	10	258	156.09	296.82	−320.71	332.28	85.00%	100.00%
140	200	naturel	10	28	754.42	23.24	−30.68	49.90	96.67%	100.00%
140	200	classique	10	297	439.02	400.75	−135.18	383.19	61.67%	100.00%
160	200	naturel	10	24	924.38	55.01	−139.99	120.22	92.50%	100.00%
160	200	classique	10	117	394.91	460.82	−365.79	551.28	72.50%	90.00%
180	200	naturel	10	22	1089.31	124.90	−428.51	242.49	90.00%	100.00%
180	200	classique	10	272	458.93	268.51	−236.15	284.76	65.00%	100.00%

Table 3: Résultats complets sur la base des Crabes

On constate expérimentalement que le gradient dans la direction du gradient classique est instable. La convergence de l'algorithme vers un maximum local n'est donc pas garantie.
D'autre part, la seconde formulation de la Log-vraisemblance en mode semi-supervisé n'est pas monotone croissante en fonction des itérations de l'algorithme. Ceci s'explique par un manque suffisant d'information et parce que le modèle choisi est un algorithme GEM qui améliore simplement la fonction Q au lieu de la maximiser.

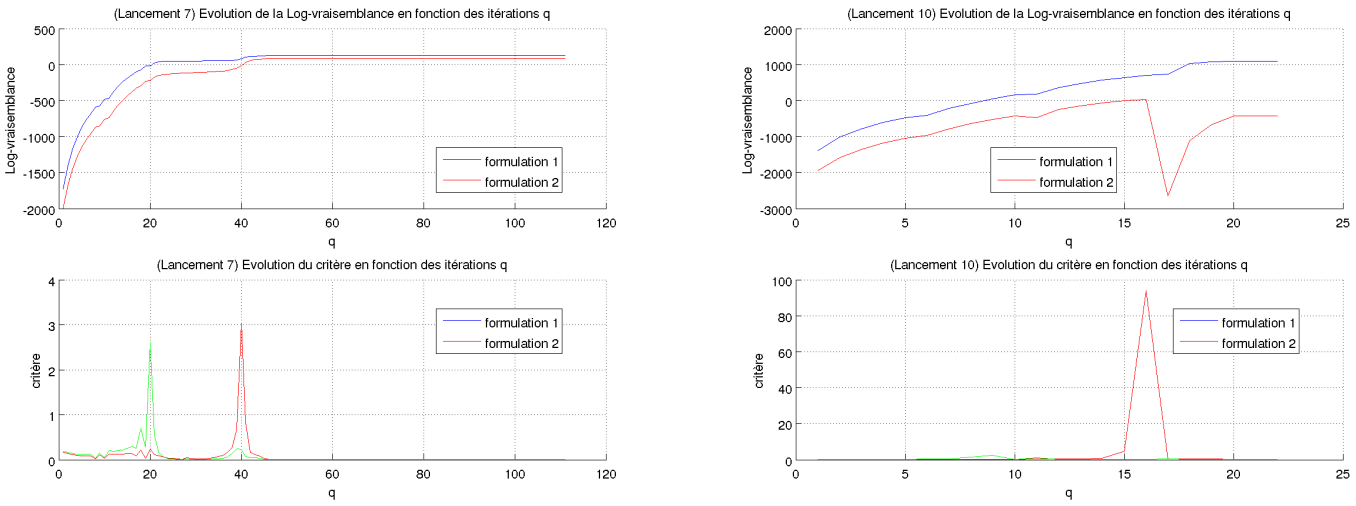


Figure 26: *Deux exemples sur l'allure des courbes de la Log-vraisemblance et du critère de stabilisation en fonction du nombre d'itérations (respectivement $M = 0$ et $M = 180$ avec le gradient naturel).*

Chaque expérience a été suivie par plusieurs "contôles" développés au cours de ce stage. Un fichier `log` regroupe l'ensemble des événements survenus en temps réel sur une expérience ; on génère les courbes des Log-vraisemblances $\mathcal{L}_1$ et $\mathcal{L}_2$, et la courbe de l'évolution du critère sur la stabilisation de la vraisemblance. L'ensemble de ces fichiers résumant ces expériences se trouve à la page `http://nicolas.cheifetz.free.fr/stages/Figures/Crabs/`.

A.7 Eléments de la Théorie de l'évidence et Supervision douce

- **Quelques définitions sur la Théorie de l'évidence**

La Théorie de l'évidence (Shafer, 1976 [23]) englobe la théorie des probabilités et celle des possibilités. On reprend ici la notation et certains concepts présents dans le Que sais-je ? sur la *Logique floue* (Bouchon-Meunier, 2007 [5]).

Soit U, un univers de référence fini (avec extension possible au cas infini).

Définition A.7.1 *Une* ***masse***[12] *(ou fonction de masse de croyance) est une fonction telle que :*

$$m : \begin{array}{ccc} \mathcal{P}(U) & \longrightarrow & [0,1] \end{array} \quad , \text{où } \sum_{A \subseteq U} m(A) = 1 \tag{A.9}$$

- *"Condition de normalisation" (ou "Hypothèse du monde clos") :* $m(\emptyset) = 0$.
  Si cette condition n'est pas vérifiée, on l'appelle "Hypothèse du monde ouvert".

Définition A.7.2 *Un* ***élément focal*** *est un ensemble F tel que :* $\emptyset \neq F \subseteq U$, et, $m(F) \neq 0$. (A.10)

Définition A.7.3 *Le* ***degré de croyance*** *(ou fonction de croyance) est une fonction telle que :*

$$bel : \begin{array}{ccc} \mathcal{P}(U) & \longrightarrow & \mathbb{R}^+ \\ A & \longmapsto & \sum_{B/\emptyset \neq B \subseteq A} m(B) \end{array} \tag{A.11}$$

- *Propriété :* $\forall A_1, \ldots, A_n \subseteq U, \quad bel(\bigcup_{i=1}^{n} A_i) \geq \sum_{I \subseteq \{1,\ldots,n\}, I \neq \emptyset} (-1)^{|I|+1} bel(\bigcap_{i=1}^{n} A_i)$
- *Une fonction de croyance est une capacité de Choquet monotone d'ordre infini.*

Définition A.7.4 *Le* ***degré de plausibilité*** *(ou fonction de plausibilité) est une fonction telle que :*

$$pl : \begin{array}{ccc} \mathcal{P}(U) & \longrightarrow & \mathbb{R}^+ \\ A & \longmapsto & \sum_{B/A \cap B \neq \emptyset} m(B) \end{array} \tag{A.12}$$

- *Propriété :* $\forall A_1, \ldots, A_n \subseteq U, \quad pl(\bigcup_{i=1}^{n} A_i) \leq \sum_{I \subseteq \{1,\ldots,n\}, I \neq \emptyset} (-1)^{|I|+1} pl(\bigcap_{i=1}^{n} A_i)$

Propriétés :
- $bel(A) \leq pl(A)$
- $bel(A) = 1 - pl(A^c)$
- $bel(\emptyset) = 0$, et, $bel(U) = 1$
- Si $A \cap B = 0$, alors $bel(A \cup B) = bel(A) + bel(B)$
- (Lien avec les probabilités) $\forall A \subseteq U, \ P(A) \in [bel(A), pl(A)]$

- **Extension au mode de supervision douce**

L'objectif est d'estimer le paramètre global en maximisant la fonction de plausibilité sur les observations :

$$\Phi^* = \arg\max_{\Phi} \ Pl(\Phi; \mathbb{X}) \tag{A.13}$$

D'après le Théorème de Bayes Généralisé (Smets, 1993) et comme les observations sont conditionnellement indépendantes, on écrit :

$$\begin{array}{rcl} Pl(\Phi; \mathbb{X}) & = & pl(X_1 = x_1, \ldots, X_N = x_N | \Phi) \\ & = & \prod_{i=1}^{N} pl(x_i | \Phi) \end{array} \tag{A.14}$$

Théorème A.1 *des* ***plausibilités totales*** *(cas d'un espace produit)*

$$\forall x_i \in \mathbb{X}, \quad pl(x_i) = \sum_{C_k \in Cl} m(C_k) \cdot pl(x_i | C_k) \tag{A.15}$$

où, $m(C_k)$ est la fonction de masse de croyance sur l'ensemble $Cl \cap C_k$, et, $pl(x_i|C_k)$ représente le degré de plausibilité que la i^{e} observation appartienne à la classe C_k.

[12] $m(A)$ représente le degré de croyance d'un groupe d'observateurs en un événement A sans accréditer les sous-ensembles de A.

On détermine le nouveau critère à maximiser :

$$\begin{aligned} Pl(\Phi;\mathbb{X}) &= \prod_{i=1}^{N} pl(x_i|\Phi) = \prod_{i=1}^{N}\left(\sum_{k=1}^{K} \underbrace{m(C_k|\Phi)}_{pl_{ik}\cdot\pi_k} \cdot \underbrace{pl(x_i|C_k,\Phi)}_{f_{X_i}(x_i|C_k,\Phi)\cdot dx_i}\right) \\ &= \prod_{i=1}^{N}\frac{1}{|\det(A)|}\cdot\prod_{s=1}^{S}\left(\sum_{k_s=1}^{K_s} pl_{ik_s}\cdot\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right)\cdot dx_i \end{aligned} \tag{A.16}$$

où pl_{ik_s} est la plausibilité (label doux) que la classe associée à la s^{e} source de la i^{e} observation soit C_{k_s}.

On obtient alors le critère en supervision douce :

$$\mathcal{P}l(\Phi;\mathbb{X}) = -N\cdot\log|\det(A)| + \sum_{i=1}^{N}\sum_{s=1}^{S}\log\left(\sum_{k_s=1}^{K_s} pl_{ik_s}\cdot\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})\right) + Cste \tag{A.17}$$

Comme pour le mode semi-supervisé, on intègre la connaissance des labels doux lors du calcul des lois a posteriori en modifiant uniquement l'étape d'*Estimation*.
On présente ici le pseudo-code de l'IFA sans bruit pour la labellisation douce :

Algorithme 7 : Pseudo-code IFA sans bruit sur la matrice de mixage et gradient naturel avec labels doux

Entrées : données observées centrées-réduites $\mathbb{X}$, labels doux $(pl_{ik_s})_{i\in\{1,\dots,N\},k_s\in\{1,\dots,K_s\}}$, nombre de sources pertinentes, nombre de classes K pour les sources pertinentes, et nombre maximal d'itérations

 # Initialisation

Tant que *non convergence* **faire**

*# **1.** Mise à jour des sources (3.3)*

*# **2.** Mise à jour des paramètres des sources (paramètres des mélanges gaussiens) par algorithme EM* (section 2)
*# **Etape E** (Estimation) : calcul des lois a posteriori* (3.19)
pour chaque $(s,k_s)\in\{1,\dots,S\}\times\{1,\dots,K_s\}$ **faire**

$$t_{ik_s} = \frac{pl_{ik_s}\cdot\pi_{k_s}\cdot\mathcal{N}(y_{is};\theta_{k_s})}{\sum_{l_s=1}^{K_s} pl_{il_s}\cdot\pi_{l_s}\cdot\mathcal{N}(y_{is};\theta_{l_s})} \quad ,\forall i\in\{1,\cdots,N\}$$

*# **Etape M** (Maximisation) : calcul de* $\left(\pi_1{}^{(q+1)},\dots,\pi_S{}^{(q+1)},\theta_1{}^{(q+1)},\dots,\theta_S{}^{(q+1)}\right)$

*# **3.** Mise à jour de la fonction g* (3.18)

*# **4.** Calcul du gradient naturel* (3.12)

*# **5.** Recherche linéaire du paramètre τ par backtracking* (cf. Annexe A.4)

*# **6.** Mise à jour de la matrice de mixage par montée de gradient* (3.14)

*# **7.** Normalisation des sources* (3.20) & (3.21)

$q \leftarrow q+1$

Sorties : estimation du paramètre global Φ^*, et, une approximation de la Log-vraisemblance sur l'ensemble des données

A.8 Pseudo-code pour l'Analyse en Facteurs Indépendants avec contraintes spatiales en mode semi-supervisé

Nous avons choisi d'étudier un modèle d'Analyse en Facteurs Indépendants avec une montée de gradient dans la direction du gradient naturel (3.14) et, dont la stabilisation de la vraisemblance est basée sur la 1re formulation de la Log-vraisemblance (3.24) en mode semi-supervisé, avec les contraintes spatiales décrites dans la sous-section 4.2.

Pseudo-code de l'Analyse en Facteurs Indépendantrs sans bruit (3.3) avec contraintes spatiales sur la matrice de mixage et gradient naturel dans un cadre *semi-supervisé* :

Algorithme 8 : Pseudo-code IFA sans bruit avec contraintes spatiales sur la matrice de mixage et gradient naturel dans un cadre semi-supervisé

Entrées : données observées centrées-réduites $\mathbb{X}$, classes des observations labellisées $\mathbb{Z}$, matrice masque M, nombre de sources (variables latentes continues) pertinentes, nombre de classes K pour les sources pertinentes, les *sources-bruits* sont mono-classes et suivent une loi normale centrée-réduite $\mathcal{N}(0,1)$, nombre maximal d'itérations, et, éventuellement le paramètre global à l'itération 0

 # Initialisation
$q = 0, \quad A^{(q)} \leftarrow M \odot A^{(q)}$
$\Phi^{(q)} = \left(A^{(q)}, {\pi_1}^{(q)}, \dots, {\pi_S}^{(q)}, {\theta_1}^{(q)}, \dots, {\theta_S}^{(q)}\right)$
 # plusieurs matrices $A^{(q)}$ sont générées aléatoirement ; on sauve celle qui a le meilleur conditionnement
 # les π_s et θ_s des "sources pertinentes" sont déterminés par un K-means sur les s^{e} composantes de $\mathbb{X}$

Tant que *non convergence* **faire**

*# **1.** Mise à jour des sources (3.3)*
$\mathbb{Y}^{(q)} = \mathbb{X} \cdot {}^{T}{A^{(q)}}^{-1}$

*# **2.** Mise à jour des paramètres des sources (paramètres des mélanges gaussiens) par algorithme EM* (section 2)
*# **Etape E** (Estimation) : calcul des lois a posteriori (3.19) & (3.25)*
pour chaque $(s, k_s) \in \{1, \dots, S\} \times \{1, \dots, K_s\}$ **faire**

$${t_{ik_s}}^{(q)} = \mathbb{1}_{Z_i = C_{k_s}} \qquad ,\forall i \in \{1, \dots, M\}$$

$${t_{ik_s}}^{(q)} = {\pi_{k_s}}^{(q)} \cdot \mathcal{N}\left({y_{is}}^{(q)}; {\theta_{k_s}}^{(q)}\right) \Big/ \sum_{l_s=1}^{K_s} {\pi_{l_s}}^{(q)} \cdot \mathcal{N}\left({y_{is}}^{(q)}; {\theta_{l_s}}^{(q)}\right) \qquad ,\forall i \in \{M+1, \dots, N\}$$

*# **Etape M** (Maximisation) : calcul de* $\left({\pi_1}^{(q+1)}, \dots, {\pi_S}^{(q+1)}, {\theta_1}^{(q+1)}, \dots, {\theta_S}^{(q+1)}\right)$
pour chaque $(s, k_s) \in \{1, \dots, S\} \times \{1, \dots, K_s\}$ **faire**

$${\pi_{k_s}}^{(q+1)} = \sum_{i=1}^{N} {t_{ik_s}}^{(q)} / N$$

$${\mu_{k_s}}^{(q+1)} = \sum_{i=1}^{N} {t_{ik_s}}^{(q)} \cdot {y_{is}}^{(q)} \Big/ \sum_{j=1}^{N} {t_{jk_s}}^{(q)}$$

$${\Sigma_{k_s}}^{(q+1)} = \sum_{i=1}^{N} {t_{ik_s}}^{(q)} \cdot \left({y_{is}}^{(q)} - {\mu_{k_s}}^{(q+1)}\right)^2 \Big/ \sum_{j=1}^{N} {t_{jk_s}}^{(q)}$$

*# **3.** Mise à jour de la fonction g (3.18)*
pour chaque $(i, s) \in \{1, \dots, N\} \times \{1, \dots, S\}$ **faire**

$$g_s^{(q)}({y_i}^{(q)}) = \sum_{k_s=1}^{K_s} {t_{ik_s}}^{(q)} \cdot \frac{{y_{is}}^{(q)} - {\mu_{k_s}}^{(q+1)}}{{\Sigma_{k_s}}^{(q+1)}}$$

*# **4.** Calcul du gradient naturel (3.12)*
$$\nabla_{nat} A^{(q)} = M \odot \left(A^{(q)} \cdot \left(g^{(q)}(\mathbb{Y}^{(q)}) \cdot \mathbb{Y}^{(q)} - N \cdot \mathrm{I}_S\right)\right)$$

*# **5.** Recherche linéaire du paramètre τ par backtracking* (cf. Annexe A.4)
$\tau^{(q)} = \text{RechercheLineaire}(\mathbb{X},\ A^{(q)},\ \nabla_{nat} A^{(q)})$

*# **6.** Mise à jour de la matrice de mixage par montée de gradient (3.14)*
$A^{(q+1)} = A^{(q)} + \tau^{(q)} \cdot \nabla_{nat} A^{(q)}$

*# **7.** Normalisation des sources (3.20) & (3.21)*
pour chaque $s \in \{1, \dots, S\}$ **faire**

$${\Sigma_s}^{(q)} = \left(\sum_{k_s=1}^{K_s} {\pi_{k_s}}^{(q+1)} \cdot \left({\Sigma_{k_s}}^{(q+1)} + {\mu_{k_s}}^{(q+1)2}\right)\right) - \left(\sum_{k_s=1}^{K_s} {\pi_{k_s}}^{(q+1)} \cdot {\mu_{k_s}}^{(q+1)}\right)^2$$

$${A_{.s}}^{(q+1)} \leftarrow {A_{.s}}^{(q+1)} \cdot \sqrt{{\Sigma_s}^{(q)}}$$

pour chaque $k_s \in \{1, \dots, K_s\}$ **faire**

$${\mu_{k_s}}^{(q+1)} \leftarrow {\mu_{k_s}}^{(q+1)} \Big/ \sqrt{{\Sigma_s}^{(q)}}$$
$${\Sigma_{k_s}}^{(q+1)} \leftarrow {\Sigma_{k_s}}^{(q+1)} \Big/ {\Sigma_s}^{(q)}$$

$q \leftarrow q + 1$

Sorties : estimation du paramètre global Φ^*, et, une approximation de la Log-vraisemblance sur l'ensemble des données

Expériences sur données théoriques

Le nombre maximal d'itérations est fixé à 1000, le seuil de stabilisation (2.14) est de 10^{-5}, et, la montée de gradient se fait dans la direction du gradient naturel (3.12).

M	N	contraintes spatiales	Nombre de lancements	q	Log-vraisemblance $\mathcal{L}_1$	#BT ($=T$)	Performance TBC$_1$	TBC$_2$	Corrélation
0	500	sans	20	430	−5901.656708	2000	19.29%	33.41%	70.6%
0	500	avec	20	328	−7366.941402	2000	42.51%	36.55%	66.4%
50	500	sans	20	298	−5664.430980	2000	85.84%	45.36%	56.3%
50	500	avec	20	250	−6487.245797	2000	88.97%	52.75%	68.1%
100	500	sans	20	322	−5967.822591	2000	91.17%	44.04%	52.5%
100	500	avec	20	241	−5574.100874	2000	92.83%	55.76%	62.9%
150	500	sans	20	313	−4965.694211	2000	90.19%	43.32%	51.1%
150	500	avec	20	303	−3038.418960	2000	95.64%	69.94%	74.4%
200	500	sans	20	78	−2805.104273	2000	94.02%	54.44%	63.5%
200	500	avec	20	244	−1914.253642	2000	95.64%	71.91%	77.1%
250	500	sans	20	542	−622.827765	2000	95.37%	63.70%	68.4%
250	500	avec	20	339	−110.624043	2000	96.91%	78.61%	82.7%
300	500	sans	20	329	1921.022811	2000	96.92%	80.33%	79.3%
300	500	avec	20	208	1270.143412	2000	97.02%	77.98%	83.3%
350	500	sans	20	106	3422.155276	2000	97.50%	83.27%	83.9%
350	500	avec	20	380	2490.461940	2000	97.36%	81.30%	83.5%
400	500	sans	20	405	4773.463326	2000	97.49%	81.88%	84.3%
400	500	avec	20	498	3872.753148	2000	97.44%	81.37%	83.5%
450	500	sans	20	114	6050.172412	2000	97.49%	81.92%	84.0%
450	500	avec	20	550	5188.814277	2000	97.64%	81.25%	83.7%
500	500	sans	20	130	7451.934589	2000	97.81%	82.89%	84.1%
500	500	avec	20	178	6603.944147	2000	97.78%	82.45%	83.9%

Table 4: Résultats complets sur circuits de voie théoriques avec/sans contraintes spatiales

Les critères de performance sur la base de test, sont :

– indice de corrélation : $\frac{1}{L}\cdot\sum_{s=1}^{L} r^2_{y_s,\hat{y_s}}$, où $r_{y_s,\hat{y_s}} = \dfrac{\sum_{i=1}^{T}(y_{is}-\bar{y_s})\cdot(\hat{y_{is}}-\bar{\hat{y}_s})}{\sqrt{\sum_{i=1}^{T}(y_{is}-\bar{y_s})^2}\cdot\sqrt{\sum_{i=1}^{T}(\hat{y_{is}}-\bar{\hat{y}_s})^2}}$

– 1er taux de bonne classification : $\text{TBC}_1 = \sum_{k=1}^{K} mc_{k,k} \;/\; \sum_{k=1}^{K}\sum_{l=1}^{K} mc_{k,l}$

– 2nd taux de bonne classification : $\text{TBC}_2 = \frac{1}{K}\cdot\sum_{l=1}^{K} tbc_{k=l}$

avec $mc \equiv$ matrice de confusion sur toutes les sources
$tbc_{k=l} \equiv$ moyenne du taux de bonne classification moyen pour la classe l, sur toutes les sources

↪ Par exemple, après l'apprentissage (avec contraintes spatiales) de 500 CdVs, dont 250 sont labellisés, composés chacun de 18 condensateurs, on obtient en classification sur une base de test de 2000 CdVs :

- $mc = \begin{pmatrix} 33635 & 448 & 127 \\ 107 & 811 & 163 \\ 24 & 242 & 443 \end{pmatrix}$ [13] $\Longrightarrow \quad \mathrm{TBC}_1 = \dfrac{33635+811+443}{36000} = 0.9691$

- $\left.\begin{array}{rcl} tbc_{k=1} & = & 0.983192 \\ tbc_{k=2} & = & 0.750231 \\ tbc_{k=3} & = & 0.624824 \end{array}\right\} \Longrightarrow \quad \mathrm{TBC}_2 = \dfrac{0.983192+0.750231+0.624824}{3} = 0.7861$

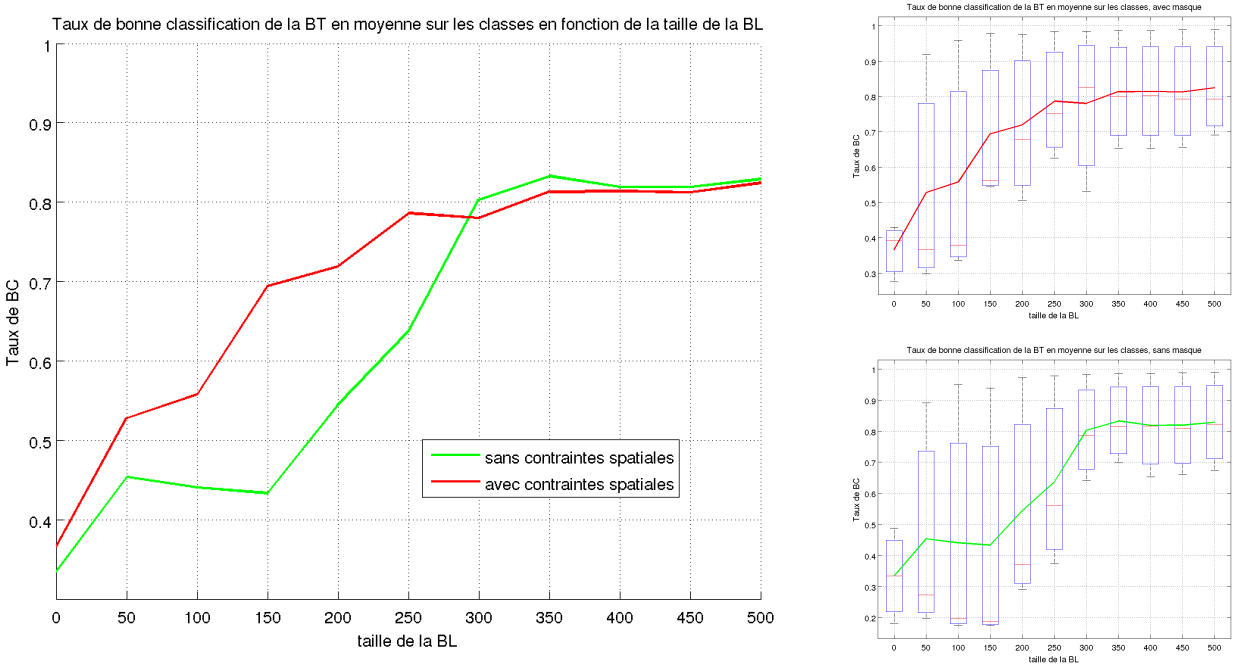


Figure 27: *Taux de bonne classification* TBC_2 sur l'état de défaut des circuits de voie de la base de test.

A.9 Pseudo-code pour l'Analyse en Facteurs Indépendants en mode semi-supervisé avec nombre variable de sources

Pour apprendre sur des observations de dimensions différentes, nous avons choisi de déterminer les densités des sources à partir de l'ensemble de observations mises en apprentissage ; ces sources sont calculées suivant différentes matrices de mixage.

Pseudo-code de l'Analyse en Facteurs Indépendantrs sans bruit (3.3) avec contraintes spatiales sur la matrice de mixage, gradient naturel, et nombre variable de sources dans un cadre semi-supervisé :

[13] On a bien : $\dfrac{1}{18} \cdot \sum_{k=1}^{3} \sum_{l=1}^{3} mc_{(k,l)} = \dfrac{36000}{18} = 2000 \ (= \#BT)$

Algorithme 9 : Pseudo-code IFA sans bruit avec contraintes spatiales sur la matrice de mixage, gradient naturel et nombre variable de sources dans un cadre semi-supervisé

Entrées : nombre de bases V, bases des données observées centrées-réduites $(\mathbb{X}_v)_{v\in\{1,\dots,V\}}$, classes des données labellisées $(\mathbb{Z}_v)_{v\in\{1,\dots,V\}}$, matrices masque $(M_v)_{v\in\{1,\dots,V\}}$, nombre de sources (variables latentes continues) pertinentes, nombre de classes K pour les sources pertinentes, nombre maximal d'itérations, et, éventuellement le paramètre global à l'itération 0

 # Initialisation

$q = 0, \quad A_v{}^{(q)} \leftarrow M_v \odot A_v{}^{(q)}, \ \forall v \in \{1,\dots,V\}$

$\Phi^{(q)} = \left((A_v{}^{(q)})_{1\le v\le V}, \pi_1{}^{(q)}, \dots, \pi_{S_1}{}^{(q)}, \theta_1{}^{(q)}, \dots, \theta_{S_1}{}^{(q)}\right)$, où $S_V \le \dots \le S_1$

Tant que *non convergence* **faire**

*# **1.** Mise à jour des sources (3.3)*

pour chaque $v \in \{1,\dots,V\}$ **faire**

$\mathbb{Y}_v{}^{(q)} = \mathbb{X}_v \cdot {}^T A_v{}^{(q)-1}$

*# **2.** Mise à jour des paramètres des sources (paramètres des mélanges gaussiens) par algorithme EM* (section 2)

*# **Etape E** (Estimation) : calcul des lois a posteriori (3.19) & (3.25)*

pour chaque $(v,s,k_s) \in \{1,\dots,V\}\{1,\dots,S_v\}\times\{1,\dots,K_s\}$ **faire**

$t_{i_v k_s}{}^{(q)} = \mathbb{1}_{Z_{i_v}=C_{k_s}} \qquad ,\forall i_v \in \{1,\dots,M_v\}$

$t_{i_v k_s}{}^{(q)} = \pi_{k_s}{}^{(q)} \cdot \mathcal{N}\left(y_{i_v s}{}^{(q)}; \theta_{k_s}{}^{(q)}\right) \ / \ \sum_{l_s=1}^{K_s} \pi_{l_s}{}^{(q)} \cdot \mathcal{N}\left(y_{i_v s}{}^{(q)}; \theta_{l_s}{}^{(q)}\right) \qquad ,\forall i_v \in \{M_v+1,\dots,N_v\}$

*# **Etape M** (Maximisation) : calcul de* $\left(\pi_1{}^{(q+1)}, \dots, \pi_{S_1}{}^{(q+1)}, \theta_1{}^{(q+1)}, \dots, \theta_{S_1}{}^{(q+1)}\right)$

pour chaque $(s,k_s) \in \{1,\dots,S_1\}\times\{1,\dots,K_s\}$ **faire**

$\pi_{k_s}{}^{(q+1)} = \sum_{i_v=1}^{N_v} t_{i_v k_s}{}^{(q)} / N_v$

$\mu_{k_s}{}^{(q+1)} = \sum_{i_v=1}^{N_v} t_{i_v k_s}{}^{(q)} \cdot y_{i_v s}{}^{(q)} \ / \ \sum_{j_v=1}^{N_v} t_{j_v k_s}{}^{(q)}$, où $v = \underset{v\in\{1,\dots,V\}}{\arg\min}\{S_v - s \,/\, S_v > s\}$

$\Sigma_{k_s}{}^{(q+1)} = \sum_{i=1}^{N_v} t_{i_v k_s}{}^{(q)} \cdot \left(y_{i_v s}{}^{(q)} - \mu_{k_s}{}^{(q+1)}\right)^2 \ / \ \sum_{j_v=1}^{N_v} t_{j_v k_s}{}^{(q)}$

pour chaque $v \in \{1,\dots,V\}$ **faire**

*# **3.** Mise à jour de la fonction g_v (3.18)*

pour chaque $(i_v,s) \in \{1,\dots,N_v\}\times\{1,\dots,S_v\}$ **faire**

$g_{vs}{}^{(q)}(y_{i_v}{}^{(q)}) = \sum_{k_s=1}^{K_s} t_{i_v k_s}{}^{(q)} \cdot \dfrac{y_{i_v s}{}^{(q)} - \mu_{k_s}{}^{(q+1)}}{\Sigma_{k_s}{}^{(q+1)}}$

*# **4.** Calcul du gradient naturel (3.12)*

$\nabla_{nat} A_v{}^{(q)} = M_v \odot \left(A_v{}^{(q)} \cdot \left(g_v{}^{(q)}(\mathbb{Y}_v{}^{(q)}) \cdot \mathbb{Y}_v{}^{(q)} - N_v \cdot \mathrm{I}_S\right)\right)$

*# **5.** Recherche linéaire du paramètre τ par backtracking* (cf. Annexe A.4)

$\tau_v^{(q)} = \text{RechercheLineaire}(\mathbb{X}_v, \ A_v{}^{(q)}, \ \nabla_{nat} A_v{}^{(q)})$

*# **6.** Mise à jour de la matrice de mixage par montée de gradient (3.14)*

$A_v{}^{(q+1)} = A_v{}^{(q)} + \tau_v{}^{(q)} \cdot \nabla_{nat} A_v{}^{(q)}$

*# **7.** Normalisation des sources (3.20) & (3.21)*

pour chaque $s \in \{1,\dots,S\}$ **faire**

$\Sigma_s{}^{(q)} = \left(\sum_{k_s=1}^{K_s} \pi_{k_s}{}^{(q+1)} \cdot \left(\Sigma_{k_s}{}^{(q+1)} + \mu_{k_s}{}^{(q+1)2}\right)\right) - \left(\sum_{k_s=1}^{K_s} \pi_{k_s}{}^{(q+1)} \cdot \mu_{k_s}{}^{(q+1)}\right)^2$

$(A_v)_{.s}{}^{(q+1)} \leftarrow (A_v)_{.s}{}^{(q+1)} \cdot \sqrt{\Sigma_s{}^{(q)}}$

pour chaque $k_s \in \{1,\dots,K_s\}$ **faire**

$\mu_{k_s}{}^{(q+1)} \leftarrow \mu_{k_s}{}^{(q+1)} \ / \ \sqrt{\Sigma_s{}^{(q)}}$

$\Sigma_{k_s}{}^{(q+1)} \leftarrow \Sigma_{k_s}{}^{(q+1)} \ / \ \Sigma_s{}^{(q)}$

$q \leftarrow q + 1$

Sorties : estimation du paramètre global Φ^*, et, une approximation de la Log-vraisemblance sur l'ensemble des données

Liste des figures

Liste des tableaux

Mise en oeuvre

Ces travaux ont été menés à l'aide de plusieurs outils informatiques :

MATLAB	:	logiciel propriétaire pour le développement d'applications scientifiques utilisé pour une capacité de calculs formels optimisés et de génération de graphiques
LaTeX 2_ε	:	logiciel libre de composition de documents
Graphviz	:	logiciel libre de génération de graphes
GIMP	:	logiciel libre de traitement d'images

Bibliographie

[1] H. Attias. Independent factor analysis. *Neural Computation*, 11 :803–851, 1999.

[2] J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report TR-97-021, ICSI, 1997.

[3] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] I. Bloch. *Fondements des probabilités et des croyances : une discussion des travaux de Cox et Smets*. 15^{e} colloque GRETSI, 1995.

[5] B. Bouchon-Meunier. *La Logique Floue*. Number 2702 in Que Sais-je ? Presses Universitaires de France, Paris, France, 2007.

[6] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28 :781–793, 1995.

[7] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

[8] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley InterScience, 2006.

[9] 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 & INRETS-LTN, 2009.

[10] E. Côme, L. Oukhellou, P. Aknin, and T. Denoeux. Partially-supervised learning in independent factor analysis. ESANN, 2009.

[11] E. Côme, L. Oukhellou, T. Denoeux, and P. Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3) :334 – 348, 2009.

[12] E. Côme, L. Oukhellou, T. Denoeux, and P. Aknin. Noiseless independent factor analysis with mixing constraints in a semi-supervised framework. application to railway device fault diagnosis. ICANN, 2009.

[13] A. Debiolles. *Diagnostic de systèmes complexes à base de modèle interne, reconnaissance des formes et fusion d'informations. Application au diagnostic des Circuits de Voie ferroviaires*. PhD thesis, Université de Technologie de Compiègne, 2007.

[14] A. Dempster, N. Laird, and D. Rubin. *Maximum-likelihood from incomplete data via the em algorithm*, 1977. J. Royal Statist. Soc. Ser. B.

[15] A. Hyvärinen, J. Karhunen, and E. Oja. *Independant Component Analysis*. Wiley InterScience, 2001.

[16] T. Jebara. *Machine Learning : Discriminative and Generative*. Springer, 2004.

[17] M. Jordan. *An introduction to graphical models*. Berkeley, U. C., 1997.

[18] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley InterScience, 1997.

[19] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley InterScience, 2000.

[20] A. Samé, L. Oukhellou, E. Côme, and P. Aknin. Mixture-model-based signal denoising. *Advances in Data Analysis and Classification*, 1(1) :39–51, 2007.

[21] A. Samé. *Modèles de mélange et classification de données acoustiques en temps réel*. PhD thesis, Université de Technologie de Compiègne, 2004.

[22] G. Saporta and J.-M. Bouroche. *L'Analyse des Données*. Number 1854 in Que Sais-je ? Presses Universitaires de France, Paris, France, 2005.

[23] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

[24] P. Smets. The combination of evidence in the transferrable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5) :447–458, 1990.

[25] V. N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, 1999.

[26] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. L-bfgs-b - fortran subroutines for large-scale bound constrained optimization. Technical report, ACM Trans. Math. Software, 1994.