UNIVERSITÉ DE TECHNOLOGIE DE COMPIÈGNE
HEUDIASYC, INRETS-LTN

THÈSE

présentée en première version en vu d'obtenir le grade de Docteur,
spécialité « « Technologies de l'Information et des Systèmes » »

par

Etienne Côme

APPRENTISSAGE DE MODÈLES GÉNÉRATIFS POUR LE DIAGNOSTIC DE SYSTÈMES COMPLEXES AVEC LABELLISATION DOUCE ET CONTRAINTES SPATIALES

Thèse soutenue le 16/01/2009 devant le jury composé de :

M^{r}	CHRISTOPHE AMBROISE	Université d'Evry	(Rapporteur)
M^{r}	MICHEL VERLEYSEN	Université Catholique de Louvain	(Rapporteur)
M^{r}	PATRICE AKNIN	INRETS	(Examinateur)
M^{r}	THIERRY DENŒUX	UTC	(Examinateur)
M^{r}	GERARD GOVAERT	UTC	(Examinateur)
M^{lle}	LATIFA OUKHELLOU	Université Paris 12	(Examinateur)

À Clélia,

Remerciements

Je voudrais tout d'abord exprimer mes plus profonds remerciements à mes deux directeurs de thèse M^r^ Patrice Aknin et M^r^ Thierry Denœux pour leurs précieux conseils et leurs encouragements. J'aimerais également remercier mon encadrante au sein de l'institut national de recherche sur les transports et leur sécurité M^lle^ Latifa Oukhellou pour son écoute attentive et sa disponibilité.

Les rapporteurs M^r^ Michel Verleysen et M^r^ Christophe Ambroise reçoivent également mes remerciements les plus sincères pour leur lecture attentive de la thèse et les remarques constructives qu'ils m'ont faites.

Je tiens aussi à adresser mes remerciements les plus chaleureux à l'ensemble des membres de l'équipe de recherche du LTN, avec une pensée particulière pour Roland, Allou, Laurent, Cyril, Faicel et Zohra avec qui j'ai eu le plaisir de travailler, d'échanger et aussi de me détendre...

Les différents projets industriels auxquels j'ai pu participer durant cette thèse m'ont aussi permis de rencontrer des personnes ouvertes et intéressantes que je tiens à remercier : M^r^ Michel Marot, M^r^ Marc Anthony, M^r^ Michel Ducloux et M^r^ Guillaume Feuillet.

Enfin, bien évidement, il reste à remercier toutes les personnes qui ont participé indirectement à la réussite de cette longue épreuve que constitue la rédaction d'une thèse. En particulier Matthieu, Alain, Eustache, Aurélien, Julien (B), Fetiha, Claire, Patrick, Johann, Eloïse, Julien (C), François, Katel, Florence, Camille...

Mes parents et mon frère doivent également être remerciés pour leurs encouragements, leur soutien et leur curiosité. Merci enfin à tous ceux qui ont contribué à me faire avancer, les nombreux professeurs qui ont marqué ma scolarité et m'ont fait choisir cette direction.

Paris, le 10 août 2011.

TABLE DES MATIÈRES

LISTE DES FIGURES

1 Introduction

Penser, c'est oublier des différences, c'est généraliser, abstraire.
Jorge Luis Borges, **Funes ou la mémoire, dans Fictions (1942)**

Sommaire

Dans ce chapitre, nous évoquons le contexte applicatif qui a sous-tendu cette thèse. Nous introduisons pour cela la problématique du diagnostic et présentons les principales étapes de sa résolution dans le cadre de l'apprentissage statistique. Nous présentons succinctement différents paradigmes d'apprentissage mis

en jeu pour la résolution de problèmes classiques tels que la classification, la régression, le regroupement automatique ; ceux-ci pourront être utilisés de manière pertinente dans le contexte applicatif. Plusieurs points essentiels à la compréhension et à la mise en œuvre pratique de ce type d'approche sont également abordés, à savoir les concepts de risque empirique, de vraisemblance et de régularisation. Nous présentons ensuite la sélection de modèle et l'extension aux méthodes non-linéaires grâce à l'utilisation de l'astuce du noyau. Enfin, différents problèmes encore ouverts dans ce champ disciplinaire, comme la gestion de labels imparfait en classification et l'introduction d'a priori de structure, seront explorés.

1.1 Diagnostic de systèmes industriels

1.1.1 Contexte des travaux

diagnostic

Cette thèse a pour objectif de travailler à la mise au point de méthodes innovantes de reconnaissances des formes dans un contexte particulier, celui de la surveillance de composants de l'infrastructure ferroviaire. Les travaux effectués visaient à résoudre des problèmes concrets de diagnostic automatique et à fournir des outils théoriques adaptés aux particularités de cette application, laquelle est abordée dans cette section.

mode de fonctionnement

La problématique du diagnostic automatique est la suivante : estimer automatiquement la classe de fonctionnement d'un composant d'un système industriel (dégradé ou non, état 1, 2,... ou K,...). Cette décision doit être prise à la lumière d'observations du composant, effectuées à l'aide d'un ou de plusieurs capteurs. Cette problématique est donc celle de l'association automatique d'un ensemble de mesures issues de capteurs à un ensemble de modes de fonctionnement, (Dubuisson 1990; 2001a;b).

maintenance préventive conditionnelle

Le diagnostic automatique prend généralement tout son sens dans une démarche industrielle globale visant à mettre en place des procédures de maintenance préventive conditionnelle permettant d'intervenir avant que le système surveillé ne tombe en panne et d'éviter ainsi la rupture de service. C'est un enjeu majeur pour des entreprises telles que les gestionnaires d'infrastructure ferroviaire qui ont en charge la maintenance d'un réseau étendu et complexe truffé d'équipements.

Ce problème peut être résolu par différentes méthodes. Les systèmes experts et les arbres de défaillances ont été utilisés pour modéliser des connaissances expertes sur des systèmes complexes en vue de mettre au point des systèmes de diagnostic automatique (Zwingelstein 2002). Des modèles d'état du système à diagnostiquer, basés sur une modélisation physique peuvent aussi être pertinents. En effet, en comparant les mesures effectuées sur le système à celles prédites par le modèle, il est possible de déterminer si le système est ou non dans un mode de fonctionnement dégradé (Dubuisson 1990).

Enfin, l'apprentissage statistique permet aussi de résoudre ce type de problème. Une telle approche vise à apprendre à partir d'un ensemble d'exemples une fonction de classification ou de régression permettant de fournir ultérieurement une affectation à l'une des classes (correspondant aux modes de fonctionnement) pour n'importe quelle nouvelle mesure fournie en entrée. On parle aussi de diagnostic par reconnaissance des formes (RdF) dans la mesure où le traitement mis au point cherchera à reconnaître la « forme » de l'observation pour ensuite l'affecter. L'ensemble de N exemples utilisés pour apprendre cette relation est appelé ensemble d'apprentissage, chaque exemple est constitué de variables ou descripteurs généralement à valeur dans $\mathbb{R}^P$ décrivant les signaux de mesures. Dans ce cadre, un problème de diagnostic peut être formalisé comme un problème de classification lorsque l'ensemble des modes de fonctionnement est fini, ou comme un problème de régression lorsque le système présente un continuum de modes de fonctionnement. C'est ce type d'approche qui a été choisi dans le cadre de cette thèse et qui sera développé dans ce mémoire.

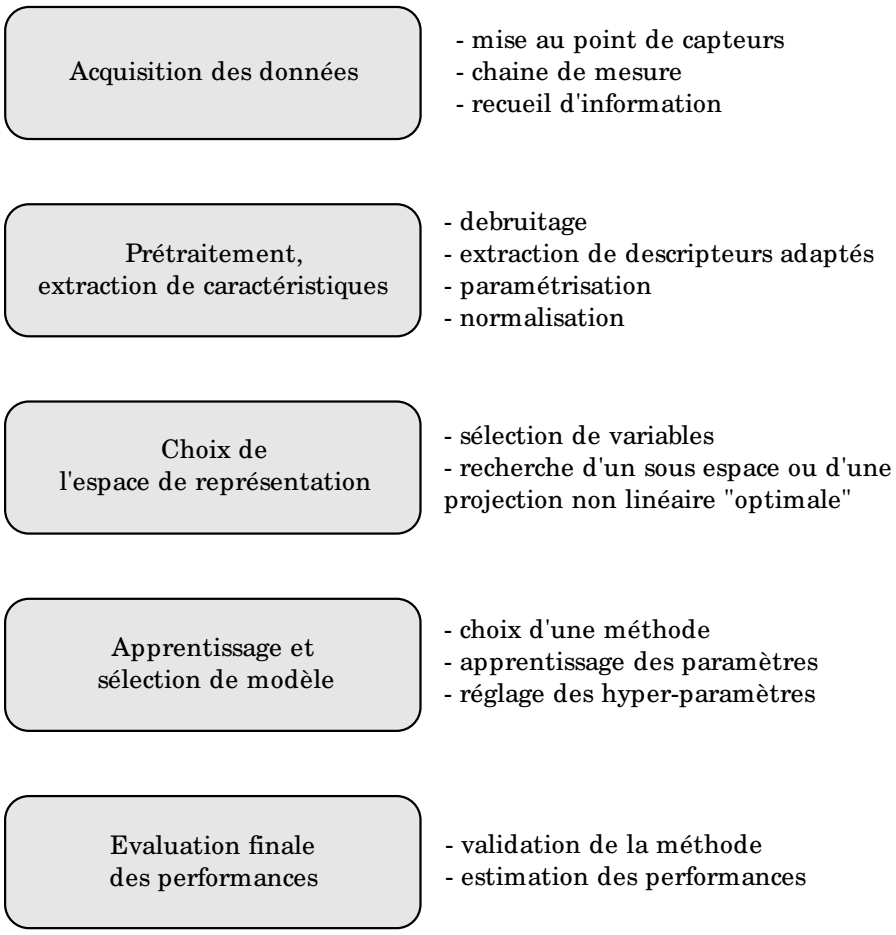


FIG. 1.1 – *Les différentes étapes de la mise en place d'un système de diagnostic automatique à base de reconnaissance des formes*

La mise en place d'un système de diagnostic automatique par reconnaissance des formes nécessite de répondre à différente questions : qu'elles sont les mesures ou caractéristiques pertinentes pour décrire l'état du système à diagnostiquer ? Quels prétraitements effectuer sur ces mesures afin de simplifier le problème décisionnel ? Quels sont les sorties attendues du système de diagnostic ? Lorsque ces différents éléments ont été correctement spécifiés, il reste à constituer la base d'apprentissage, à choisir une méthode de classification adaptée, à effectuer l'apprentissage et à valider le système de diagnostic automatique obtenu. L'ensemble des différentes étapes nécessaires à la mise en place d'une méthode de diagnostic automatique est présenté sur la figure 1.1. Les applications pratiques d'une telle démarche dans un objectif de diagnostic sont nombreuses ; nous pouvons citer par exemple :

- diagnostic de machines outils à partir de mesures vibratoires (Saravanan et al. 2008) ;
- diagnostic de procédés chimiques de traitement de l'eau à partir de mesures physico-chimiques (Valentin et Denœux 2001) ;
- diagnostic de cuves de stockage à partir de mesures d'émissions acoustiques (Samé 2004) ;

- détection et classification de fissure de rails à partir de mesures ultrasons (Aknin et Cygan 2004).

1.1.2 L'apprentissage statistique

L'apprentissage statistique est situé à l'intersection de nombreux champs disciplinaires. La théorie des probabilités, les statistiques et les théories de l'incertain en général, sont ainsi au centre de ce domaine de recherche car elles offrent des cadres théoriques adaptés à l'analyse des méthodes proposées et permettent d'en construire de nouvelles. Mais, l'informatique et l'optimisation sont elles aussi devenues indispensable. En effet, dans la majorité des cas de figures un problème d'apprentissage peut être formulé comme un problème d'optimisation : maximisation d'une probabilité ou d'une vraisemblance, minimisation d'une énergie, d'un risque ou d'un côut. Finalement, l'intelligence artificielle, la théorie de l'information et la théorie de la décision font aussi partie des disciplines qui ont influencé ce domaine.

L'apprentissage statistique repose sur l'analyse d'exemples de réalisations d'un phénomène, cette analyse devant permettre de tirer des conclusions quand à la nature de la population dont ils sont issus et d'utiliser celles-ci pour comprendre le phénomène étudié ou classer de nouvelles réalisations. La démarche générale
induction repose donc sur le principe d'induction à partir d'exemples[1].

La croissance actuelle du nombre d'informations créées, stockées, manipulées chaque jour a favorisé la reconnaissance de l'apprentissage statistique comme l'une des sciences fondamentales de notre société de l'information. Les problèmes qu'elle aborde sont en nombre croissant et ils ne cessent de se diversifier. La liste suivante donne quelques exemples d'applications :

- prédire si un patient risque une rechute au vu d'informations biologiques ;
- ordonner les résultats d'une requête internet en fonction de leur pertinence ;
- trouver un ensemble de pixels dans une image qui soient cohérents et délimitent des objets (segmentation d'images) ;
- trouver des archétypes de clients à partir d'analyse de tickets de caisses ;
- détecter un changement de tendance dans une série temporelle, valeurs boursières ou météorologiques par exemple ;
- reconnaître une adresse postale manuscrite automatiquement ;
- établir les interactions biologiques probables entre gènes à partir de relevés de puces ADN.

L'apprentissage statistique regroupe différentes méthodes visant à analyser des données en vue de mieux comprendre celles-ci ou de disposer d'une méthode de prédiction d'une ou de plusieurs variables d'intérêt. L'ensemble de données à analyser ainsi que l'objectif final de l'analyse peuvent prendre différentes formes. Cependant plusieurs problèmes de référence, d'intérêts pratique et théorique évidents, ont été formalisés et des solutions à ceux-ci ont été proposées. Pour une introduction générale au problème de l'apprentissage statistique, (Duda et al. 2000, Hastie et al. 2006, Vapnik 1999, Bishop 2006) constituent des ouvrages de référence en anglais ; en français il est possible de lire (Saporta 1990, Govaert 2003).

[1] L'induction est classiquement opposée à la déduction.

Même si cette discipline est actuellement très vivante, elle n'en trouve pas moins ses racines dans des travaux plus anciens tel que ceux des statisticiens du début du vingtième siècle K. Pearson et R. Fisher. La figure 1.2 présente une frise chronologique regroupant différentes avancées majeures de ce domaine depuis les premiers travaux sur l'analyse en composantes principales en 1901 jusqu'aux machines à vecteurs supports en 1993.

Nous présentons maintenant les deux contextes classiques de ce domaine, celui de l'apprentissage supervisé et celui de l'apprentissage non supervisé. En effet, il est nécessaire de présenter en détail ces deux domaines dans la mesure où une partie de nos travaux a portée sur des solutions se trouvant à mis chemin de ces deux extrêmes.

1.2 L'APPRENTISSAGE SUPERVISÉ

L'apprentissage supervisé se focalise sur la recherche de fonctions permettant de prédire la valeur prise par une variable d'intérêt à partir de l'observation d'un second ensemble de variables appelées variables explicatives, la prédiction devant bien sûr être la meilleure possible. L'apprentissage supervisé est riche d'un grand nombre de méthodes, certaines s'appuient sur une formalisation probabiliste du problème, d'autres non. Nous insisterons en particulier sur les méthodes utilisant une formulation probabiliste du problème qui font intervenir les éléments suivants :

- Des entrées : des variables explicatives $X \in \mathcal{X}$. En statistique, ces variables sont appelées covariables ou régresseurs. Elles sont généralement à valeur dans $\mathbb{R}^P$.

- Une sortie : une variable d'intérêt $Y \in \mathcal{Y}$, cette variable peut être de différente nature, continue $\mathcal{Y} = \mathbb{R}$, auquel cas on parle de régression, ou catégorielle $\mathcal{Y} = \{c_1, \ldots, c_K\}$, c'est le problème de la classification.

- Un ensemble de réalisations indépendantes et identiquement distribuées (*i.i.d*) du couple $(X, Y) : \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ supposées issues d'une loi de probabilité jointe inconnue.

- Une fonction de prédiction $f(\mathbf{x}) : \mathcal{X} \to \mathcal{Y}$ reliant les entrées et la sortie ; c'est l'inconnue du problème à estimer à partir de l'ensemble de données évoquées ci-dessus.

- Une fonction de coût $R(f(\mathbf{x}), y) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, traduisant les coûts associés à chaque type d'erreur de prédiction.

- De futures réalisations de X qui seront observées ; associées à d'autres réalisations de Y, quant à elle inobservées et devant être prédites. C'est sur ces nouveaux individus que la règle de prédiction sera testée.

L'objectif de l'apprentissage supervisé est d'apprendre à partir des réalisations entrées-sorties disponibles, la fonction f permettant d'obtenir un coût minimal de

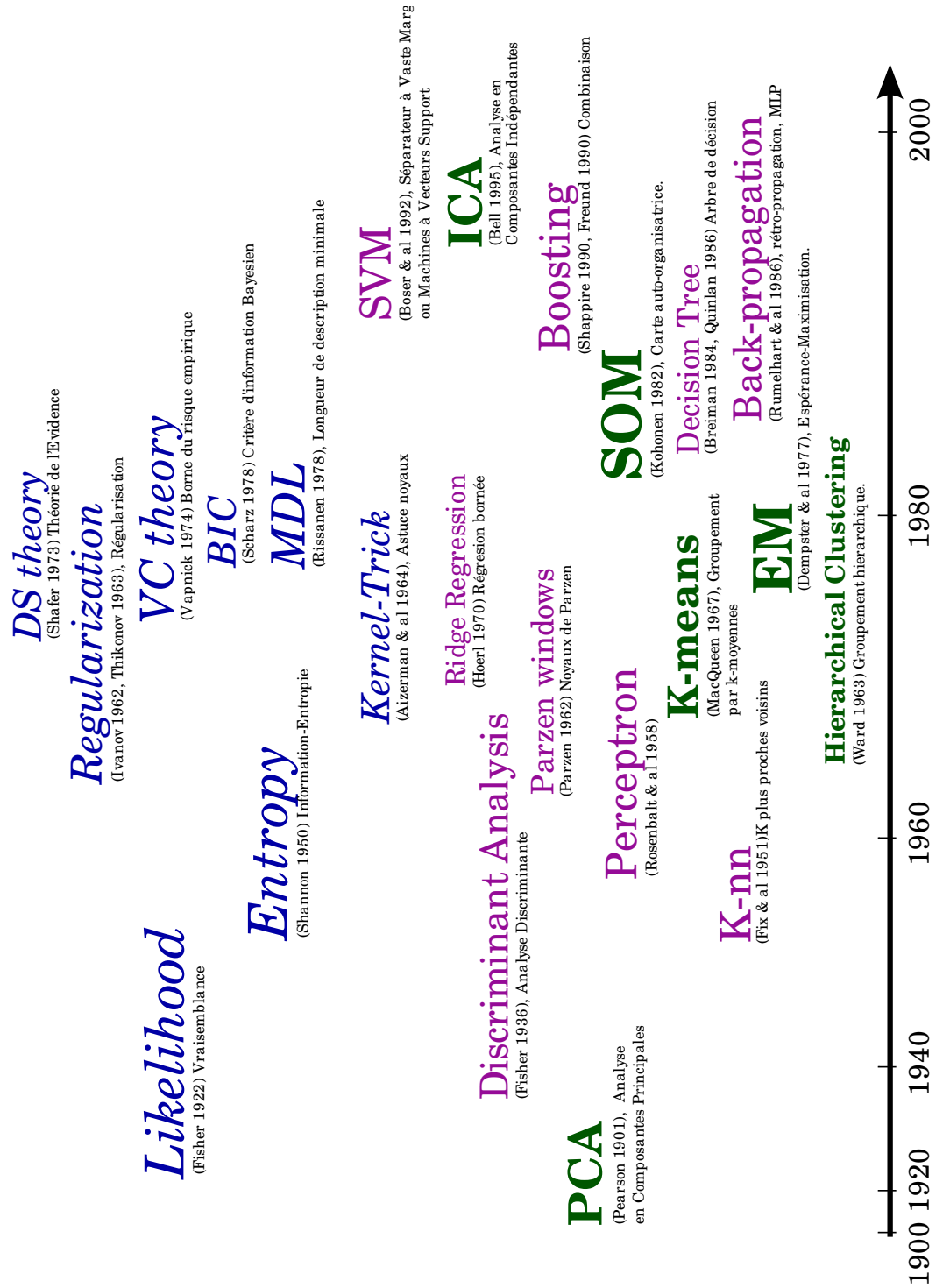


FIG. 1.2 – *Un siècle d'apprentissage statistique. Frise chronologique présentant les avancées théoriques en bleu italique, les avancés plus pratiques en mauve lorsqu'elles concernent l'apprentissage supervisé et en vert gras lorsqu'elles concernent l'apprentissage non supervisé.*

la fonctionnelle pour les futures réalisations du couple. Si la loi jointe des données était parfaitement connue le problème de l'apprentissage supervisé se résumerait à un problème d'optimisation : trouver la fonction f qui minimise l'espérance du coût par rapport à la mesure définie par la loi jointe du couple (X, Y) :

fonction de coût

Définition 1.1 (Espérance du risque, coût)

$$\mathbb{E}\left[R(f)\right] = \int_{\mathcal{X}} \int_{\mathcal{Y}} R(f(\mathbf{x}), y).p(\mathbf{x}, y).d\mathbf{x}.dy \ . \tag{1.1}$$

Malheureusement, cette loi jointe est inconnue et seul un ensemble de réalisations issues de celle-ci est disponible. Il est alors naturel de s'intéresser au risque empirique qui, en remplaçant l'espérance par une moyenne empirique, permet de définir un critère utilisant les données disponibles :

Définition 1.2 (Risque empirique)

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} R(f(\mathbf{x}_i), y_i). \tag{1.2}$$

Lors de la résolution d'un problème concret, il est nécessaire de définir l'espace fonctionnel considéré pour f ainsi que la fonction de coût utilisée pour déterminer la solution. La résolution convenable du problème dépend grandement de ces deux choix.

Exemple 1.1 (Minimisation d'un risque empirique en régression) :

Supposons le jeu de données d'apprentissage $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ tel que $\mathcal{X} = \mathbb{R}^P, \mathcal{Y} = \mathbb{R}$ et une fonction de coût $R(f(\mathbf{x}), y) = (y - f(\mathbf{x}))^2$. De plus, faisons l'hypothèse que f est linéaire, $f(\mathbf{x}) = \mathbf{w}^t.\mathbf{x}$. Le risque empirique s'écrit alors :

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} (f(\mathbf{x}_i) - y_i)^2, \tag{1.3}$$

en notant $\mathbf{X}$ la matrice de taille $N \times P$ contenant les réalisations de X et $\mathbf{y}$ le vecteur de taille $N \times 1$ contenant les réalisations de Y associées, nous obtenons :

$$\begin{aligned} \hat{\mathbf{w}} &= \arg\min_{\mathbf{w} \in \mathbb{R}^P} \frac{1}{N} \sum_{i=1}^{N} (\mathbf{w}^t.\mathbf{x}_i - y_i)^2 \\ \hat{\mathbf{w}} &= \arg\min_{\mathbf{w} \in \mathbb{R}^P} ||\mathbf{X}.\mathbf{w} - \mathbf{y}||^2. \end{aligned}$$

On montre (Saporta 1990, pages 376) que le vecteur de coefficient $\hat{\mathbf{w}}$ définissant la solution s'exprime matriciellement de la façon suivante :

$$\hat{\mathbf{w}} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{y}). \tag{1.4}$$

C'est le problème bien connu de l'estimation de la droite des moindres carrés si $p = 2$.

régularisation

Lorsque l'espace fonctionnel considéré pour f est de grande taille (c'est le cas par

exemple des extensions non-linéaires à base de noyaux qui seront présentées en section 1.4), le problème de minimisation du risque empirique peut être mal posé et plusieurs solutions optimales coexister. Dans ce cas de figure, le risque empirique est remplacé par un autre critère d'évaluation des solutions : le risque empirique régularisé.

Définition 1.3 (Risque empirique régularisé)

$$R_{reg}(f) = \frac{1}{N}\sum_{i=1}^{N} R(f(\mathbf{x}_i), y_i) + \lambda\Gamma(f), \tag{1.5}$$

où Γ est une fonctionnelle de régularisation qui pénalise les solutions trop complexes. Cette approche introduite dans le cadre des problèmes mal posés, (Ivanov 1962, Thikonov 1963) peut souvent être interprétée d'un point de vue bayésien comme un a priori sur les paramètres définissant la solution (Marín et Robert 2007). Comme on peut le constater (1.5), un risque empirique régularisé fait intervenir deux termes : un risque empirique évidemment et un deuxième terme dépendant uniquement de la fonction f et favorisant les solutions simples dans un certain sens (nombre limité de coefficients non-nuls pour les pénalisations de type L_1 (Efron et al. 2004), faible énergie de la fonction (Wahba 1990),...) Ce second terme peut généralement être associé à une contrainte sur les valeurs pouvant être prises par les paramètres, c'est pourquoi on parle également d'estimateur restreint en statistique.

approche bayésienne

Exemple 1.2 (Minimisation d'un risque empirique régularisé en régression) :

Supposons le jeu de données d'apprentissage $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ avec $\mathcal{X} = \mathbb{R}^P, \mathcal{Y} = \mathbb{R}$ et la fonction de coût $R(f(\mathbf{x}), y) = (y - f(\mathbf{x}))^2$. De plus, faisons l'hypothèse que f est linéaire, $f(\mathbf{x}) = \mathbf{w}^t.\mathbf{x}$ et que la norme de $\mathbf{w}$ est pertinente pour régulariser le problème, nous obtenons :

$$R_{reg}(f) = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i; \mathbf{w}) - y_i)^2 + \lambda||\mathbf{w}||^2. \tag{1.6}$$

La minimisation du risque empirique régularisé conduit à :

$$\begin{aligned}\hat{\mathbf{w}} &= \arg\min_{\mathbf{w}\in\mathbb{R}^P} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^t.\mathbf{x}_i - y_i)^2 + \lambda||\mathbf{w}||^2 \\ \hat{\mathbf{w}} &= \arg\min_{\mathbf{w}\in\mathbb{R}^P} ||\mathbf{X}.\mathbf{w} - \mathbf{y}||^2 + \lambda||\mathbf{w}||^2\end{aligned}$$

On montre (Hastie et al. 2006, pages 59-60) que le vecteur de coefficient $\hat{\mathbf{w}}$ définissant la solution s'exprime matriciellement de la façon suivante :

$$\hat{\mathbf{w}} = (\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^t\mathbf{y}). \tag{1.7}$$

Nous reviendrons sur ce problème dans la section 1.5 consacrée à la sélection de modèle.

Nous précisons maintenant, comment ce type de formulation peut être appliqué aux problèmes de la classification et de la régression. Nous étudions aussi, durant

les paragraphes qui suivent, le lien existant entre cette formulation et l'estimation au sens du maximum de vraisemblance de paramètres décrivant des modèles statistiques. La notion de vraisemblance introduite par Fisher (1922), postule que toute l'information pouvant être extraite d'un jeu de données pour estimer les paramètres d'un modèle statistique est contenue dans la fonction de vraisemblance, celle-ci étant définie comme la probabilité des observations connaissant les valeurs des paramètres. Cette fonction devant être maximisée par rapport aux paramètres pour obtenir leurs estimés.

vraisemblance

Définition 1.4 (Vraisemblance) *Soit un ensemble de données* $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ *distribuées suivant des variables aléatoires i.i.d. de variable associée* X*. Si la densité de probabilité sur* X *est donnée par* $p(\mathbf{x}, \boldsymbol{\psi})$*, la fonction de vraisemblance* $L(\boldsymbol{\psi}; \mathbf{X})$ *est définie par :*

$$L(\boldsymbol{\psi}; \mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{x}_i; \boldsymbol{\psi}). \tag{1.8}$$

Le principe du maximum de vraisemblance conduit à sélectionner le vecteur de paramètre solution du problème :

$$\widehat{\boldsymbol{\psi}}_{ml} = \arg\max_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} L(\boldsymbol{\psi}; \mathbf{X}). \tag{1.9}$$

Nous allons voir que cette démarche possède des lien étroits avec la minimisation du risque empririque aussi bien dans le cadre de la régression que dans le cadre de la classification.

1.2.1 Le problème de la régression

La formalisation du problème de l'apprentissage supervisé sous la forme de la minimisation d'un risque empirique possède des liens importants avec la problématique de l'estimation statistique comme le montre l'exemple de la régression. Pour la régression la fonction de coût la plus classique (utilisées dans les exemples 1.1 et 1.2) est la fonction de coût L_2 :

$$R_{L_2}(f) = (y - f(\mathbf{x}))^2. \tag{1.10}$$

Il est intéressant de noter que cette fonction peut être obtenue à l'aide d'un modèle probabiliste de génération des données. En effet, supposons que les données soient de la forme :

$$Y \quad = \quad f(X, \boldsymbol{\psi}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \nu), \tag{1.11}$$

où ψ est un vecteur de paramètres, et ϵ un bruit aléatoire gaussien indépendant de X. Ce modèle suppose donc l'existence d'une relation entrées-sortie déterministe entachée d'un bruit de mesure additif gaussien.

vraisemblance conditionnelle

Dans le cadre de la régression, nous nous intéressons à la loi conditionnelle de la variable d'intérêt connaissant les variables explicatives, dans ce cas de figure il est usuel d'utiliser la vraisemblance conditionnelle[2] définie par :

[2] pour plus de détails sur cette notion importante dans le cadre discriminatif en classification voir (Hastie et al. 2006, page 31),(Bouchard 2005, page 32), (Jebara 2001, pages 32-36).

Définition 1.5 (Vraisemblance conditionnelle) *Soit un ensemble de données i.i.d.* $(\mathbf{y}, \mathbf{X}) = \{(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)\}$ *et un modèle paramétrique de la loi conditionnelle de* Y *sachant* X, $p(y|\mathbf{x}, \boldsymbol{\psi})$. *La fonction de vraisemblance* $L_{cond}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{X})$ *est définie par :*

$$L_{cond}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i; \boldsymbol{\psi}), \tag{1.12}$$

Dans le cadre de la régression, la loi conditionnelle des Y définie par l'équation (1.11) est donnée par :

$$Y|X = \mathbf{x} \sim \mathcal{N}(f(\mathbf{x}, \boldsymbol{\psi}), \nu), \tag{1.13}$$

et la vraisemblance conditionnelle devient :

$$L_{cond}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \varphi(y_i; f(\mathbf{x}_i, \boldsymbol{\psi}), \nu), \tag{1.14}$$

avec $\varphi(., \mu, \nu)$ la densité d'une loi normale de moyenne μ et de variance ν. En passant au logarithme de cette expression (ce qui ne modifie pas la valeur pour laquelle cette fonction est maximisée), nous obtenons :

$$\begin{aligned} \mathcal{L}_{cond}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{X}) &= \log(L_{cond}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{X})) \\ &= -\frac{1}{2\nu} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i, \boldsymbol{\psi}))^2 + N. \log(\sqrt{2.\pi}\nu). \end{aligned} \tag{1.15}$$

La maximisation de la vraisemblance conditionnelle dans ce modèle conduit à la même solution que la minimisation du risque empirique (1.2) en utilisant la fonction de coût des moindre carrés (1.10). Un parallèle peut ainsi être établi entre fonction de coût et modèle statistique de bruit pour le problème de la régression (Pontil et al. 1998). La figure 1.3 présente différentes fonctions de coût utilisées dans le cadre de la régression et les densités de bruit auxquels elles correspondent.

Nous venons de voir au travers de l'exemple de la régression que la maximisation d'une vraisemblance pouvait conduire à un critère identique à celui obtenu en utilisant la notion de risque empirique. Cependant ces deux démarches sont philosophiquement différentes. Dans le cadre d'une approche de type vraisemblance, on spécifie préalablement un modèle et les hypothèses faites sur la nature des données au travers du modèle déterminent la solution obtenue. Dans une approche basée sur la notion de risque empirique, aucune hypothèse n'est faite sur la nature des données ; c'est la définition des coûts associés à une mauvaise prédiction qui détermine la solution.

1.2.2 Le problème de la classification

Dans le cadre de la classification, qui cherche à prédire une variable catégorielle $\mathcal{Y} = \{c_1, \ldots, c_K\}$, il est aussi possible de trouver un lien très fort entre estimation de modèle statistique et apprentissage supervisé. En effet, la fonction de coût naturelle du problème a la forme suivante :

$$R_{\{0,1\}}(f) = \begin{cases} 0 \text{ si } f(\mathbf{x}) = y, \\ 1 \text{ si } f(\mathbf{x}) \neq y. \end{cases} \tag{1.16}$$

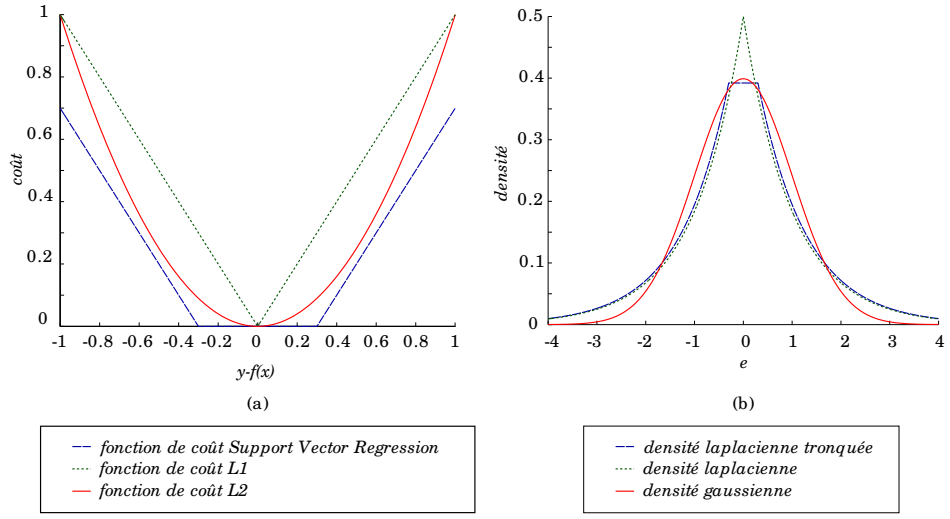


FIG. 1.3 – *(a) Fonctions de coût correspondant à différentes méthodes de régression : Fonction de coût « support vectors regression » (SVR) :* $max(|f(x)-y|-t,0)$ *(avec* $t=0.3$*), fonction de coût* $L_1 : |y-f(\mathbf{x})|$*, fonction de coût* $L_2 : (y-f(x))^2$*, en fonction de l'erreur de prédiction* $y-f(x)$*. (b) Modèles de densité de bruit équivalent, laplacien tronqué, laplacien et gaussien.*

Celle-ci comptabilise les erreurs de classement commises par la fonction f. Si nous observons l'espérance de celle-ci, nous obtenons :

$$\mathbb{E}\left[R_{\{0,1\}}(f)\right] = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} R_{\{0,1\}}(f(\mathbf{x}),y).p(\mathbf{x},y).d\mathbf{x} \tag{1.17}$$

$$= \int_{\mathcal{X}} \left(\sum_{y\in\mathcal{Y}} R_{\{0,1\}}(f(\mathbf{x}),y).p(y|\mathbf{x}) \right) p(\mathbf{x}).d\mathbf{x}\ . \tag{1.18}$$

Pour minimiser ce risque, il suffit de minimiser pour toutes les valeurs x la fonction :

$$f(\mathbf{x}) = \arg\min_{f(\mathbf{x})\in\mathcal{Y}} \sum_{y\in\mathcal{Y}} R_{\{0,1\}}(f(\mathbf{x}),y).p(y|\mathbf{x}), \tag{1.19}$$

ce qui donne :

$$f(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} p(y|\mathbf{x}), \forall \mathbf{x}\in\mathcal{X}. \tag{1.20}$$

maximum a posteriori

Ceci correspond au classifieur de Bayes. Le problème de la classification peut ainsi être ramené au problème de l'estimation de la loi conditionnelle $p(y|\mathbf{x})$ d'appartenance aux classes sachant les observations ; c'est l'approche adoptée par exemple par la régression logistique. Avec ce type de méthode, la décision finale quand à la classe de l'individu est prise à partir de ces probabilités en utilisant (1.20) ; une telle démarche est nommée classification par maximum a posteriori.

Cependant, toutes les méthodes de classification n'ont pas pour objectif d'estimer la loi conditionnelle $p(y|\mathbf{x})$ pour toute observation x. Certaines méthodes visent plus humblement à estimer les frontières de décision entre les classes et ne fournissent que la classe la plus probable pour chaque valeur de x. C'est le cas par exemple des Machines à Vecteurs Supports (SVM en anglais) (Boser et al. 1992, Vapnik 1999). En conséquence, suivant la méthode de classification choisie, les résultats obtenus peuvent être plus ou moins informatifs.

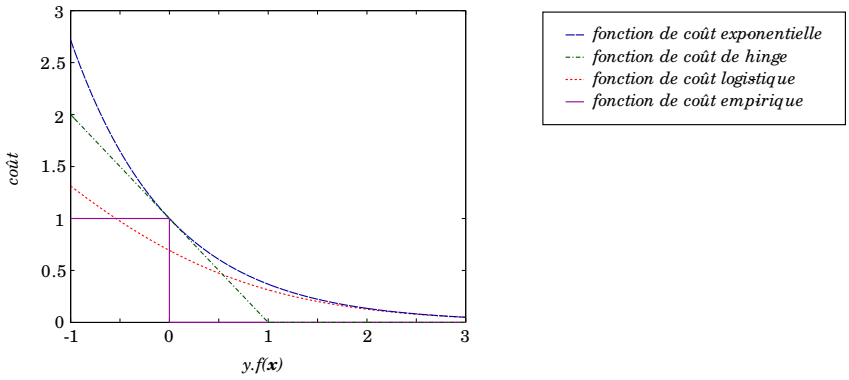


FIG. 1.4 – *Fonctions de coût correspondant à différentes méthodes de classification binaire, fonction de coût exponentielle (Boosting), fonction de coût de hinge utilisée par les SVM, fonction de coût logistique utilisée par la régression logistique, et fonction de coût empirique. Les labels sont supposés être de la forme* $\mathcal{Y} = \{-1; +1\}$ *et les coûts sont représentés en fonction de la marge* $y.f(\mathbf{x})$.

Le problème de l'estimation de $p(y|\mathbf{x})$ est cependant au centre du problème de la classification supervisée. D'un point de vue statistique, deux types d'approches peuvent être envisagés pour estimer cette loi conditionnelle ; les méthodes discriminatives et les méthodes génératives.

1.2.3 Approche discriminative / générative

approche discriminative

Les méthodes discriminatives, recherchent un modèle de $p(y|\mathbf{x})$. C'est le cas par exemple de la régression logistique qui postule un modèle linéaire généralisé de la forme suivante pour les densités conditionnelles :

$$p(Y = c_k|\mathbf{x};\boldsymbol{\psi}) = \frac{\exp(\beta_k^t\mathbf{x})}{1+\sum_{k'=1}^{K-1}\exp(\beta_{k'}^t\mathbf{x})}, \forall k \in \{1,\ldots,K-1\} \quad (1.21)$$

$$p(Y = c_K|\mathbf{x};\boldsymbol{\psi}) = \frac{1}{1+\sum_{k'=1}^{K-1}\exp(\beta_{k'}^t\mathbf{x})}, \quad (1.22)$$

où les paramètres à estimer sont $\boldsymbol{\psi} = (\beta_1,\ldots,\beta_K)$. Afin d'estimer ces paramètres, une vraisemblance conditionnelle est utilisée. Celle-ci est donnée par :

$$L_{cond}(\boldsymbol{\psi};\mathbf{y},\mathbf{X}) = \prod_{i=1}^{N}\prod_{k=1}^{K} p(Y = c_k|\mathbf{x}_i;\boldsymbol{\psi})^{z_{ik}}, \quad (1.23)$$

où z_{ik} est un vecteur binaire indiquant l'appartenance de l'individu i à la classe k : $z_{ik} = 1$ si $y_i = c_k$, $z_{ik} = 0$ sinon.

Pour construire le classifieur cette vraisemblance doit être maximisée par rapport aux paramètres sur le jeu de données d'apprentissage. Un algorithme de type Newton-Raphson, peut par exemple être utilisé, après avoir calculé les dérivés de la fonction de log-vraisemblance par rapport aux paramètres et la matrice hessienne correspondante (Hastie et al. 2006, pages 98-99).

approche générative

Les méthodes génératives, proposent quant à elles un modèle de la loi jointe des données $p(y, \mathbf{x})$, et en déduisent ensuite la distribution conditionnelle des Y grâce à l'application de la règle de Bayes : $p(y|\mathbf{x}) = p(y, \mathbf{x})/p(\mathbf{x})$ où $p(\mathbf{x})$ est la densité marginale sur les observations. Les paramètres du modèle sont dans ce cas de figure estimés grâce à une vraisemblance jointe. Ce type d'approche peut par exemple utiliser des lois normales pour définir les densités conditionnelles aux classes ; c'est l'approche (QDA) pour « analyse discriminante quadratique », ce nom venant de la nature quadratique de la frontière de décision induite par ce type de modèle. Dans ce cas, la densité de chacune des classes est gaussienne et le modèle est défini par :

$$p(Y = c_k) = \pi_k \tag{1.24}$$
$$p(\mathbf{x}|Y = c_k; \boldsymbol{\psi}) = \boldsymbol{\varphi}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1.25}$$

avec $\boldsymbol{\varphi}(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ la densité d'une variable aléatoire normale multivariée de moyenne $\boldsymbol{\mu}$ et de matrice de variance covariance $\boldsymbol{\Sigma}$. La vraisemblance des paramètres $\boldsymbol{\psi} = (\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$ prend alors la forme suivante :

$$L(\boldsymbol{\psi}; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left(\pi_k \boldsymbol{\varphi}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}}. \tag{1.26}$$

Avec ce modèle, le problème de la maximisation de la vraisemblance a une solution analytique ; les proportions sont estimées à l'aide des proportions empiriques des différentes classes, les moyennes par les moyennes empiriques et les matrices de variance-covariance par les matrices de variance-covariance empirique. Ce modèle peut cependant être encore simplifié en imposant des contraintes sur les matrices de variances-covariances. En imposant par exemple que celles-ci soient identiques pour toutes les classes, nous obtenons l'Analyse Discriminante Linéaire (LDA pour Linear Discriminant Analysis en anglais), qui induit une frontière de décision linéaire entre les classes.

Les méthodes génératives peuvent aussi utiliser des mélanges de gaussiennes pour définir des densités conditionnelles aux classes de forme plus complexe. Le modèle s'écrit alors :

$$p(Y = c_k; \boldsymbol{\psi}) = \pi_k \tag{1.27}$$
$$p(\mathbf{x}|Y = c_k; \boldsymbol{\psi}) = \sum_{g=1}^{G_k} \alpha_{kg} \boldsymbol{\varphi}(\mathbf{x}; \boldsymbol{\mu}_{kg}, \boldsymbol{\Sigma}_{kg}), \tag{1.28}$$

La frontière de décision construite par un tel modèle est plus complexe, mais le nombre de paramètres devient aussi plus important :

$$\boldsymbol{\psi} = \left(\pi_k, \alpha_{kg}, \boldsymbol{\mu}_{kg}, \boldsymbol{\Sigma}_{kg}\right), \forall k \in \{1, \ldots, K\}, \forall g \in \{1, \ldots, G_k\}. \tag{1.29}$$

Et la vraisemblance prend la forme suivante :

$$L(\boldsymbol{\psi}; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left(\pi_k \sum_{g=1}^{G_k} \alpha_{kg} \boldsymbol{\varphi}\left(\mathbf{x}_i, \boldsymbol{\mu}_{kg}, \boldsymbol{\Sigma}_{kg}\right) \right)^{z_{ik}}. \tag{1.30}$$

L'estimation des paramètres d'un tel modèle, n'est pas aussi aisée que précédemment. Il est nécessaire d'avoir recours à un algorithme d'optimisation spécifique,

l'algorithme EM, sur lequel nous reviendrons amplement au chapitre 2 de cette thèse.

La figure 1.5, présente à titre d'illustration les résultats de différents modèles génératifs et de la régression logistique sur un jeu de données de dimension deux (problème des crabes) permettant une représentation graphique. Nous pouvons observer sur celle-ci les formes des frontières de décisions obtenues suivant les densités conditionnelles postulées, de la forme la plus simple : frontière de décision linéaire pour la régression logistique (figure 1.5 (a)) et l'analyse linéaire discriminante (figure 1.5 (b)), à des formes plus complexes pour les autres modèles. Nous pouvons également observer qu'à complexité identique les approches discriminative et générative fournissent des frontière voisines (figures 1.5 (a),1.5 (b)) mais que les probabilités a posteriori peuvent être, quant à elles, nettement différentes.

Les méthodes discriminatives ont comme avantage de répondre directement au problème posé. En effet, les paramètres du modèle sont tous utilisés pour estimer la frontière de classification et pas autre chose. Les approches discriminatives sont surtout efficaces lorsque les données X ont une distribution inconnue et difficilement modélisable. Dans ce contexte, les résultats théoriques sur les bornes du risque empirique sont intéressants (Vapnik 1999). Cependant, ces méthodes incorporent difficilement des invariances ou des structures, comme des indépendances conditionnelles ou des variables latentes, et même si elles sont optimales asymptotiquement, elles peuvent être dépassées par des approches génératives pour des tailles finies d'échantillons d'apprentissage (Ng et Jordan 2001).

D'un point de vue asymptotique, les approches génératives sont optimales au sens où elles fournissent l'estimateur de plus faible variance, lorsque le modèle de densité postulé est exact. Les méthodes discriminatives leurs sont supérieures lorsque le modèle ne correspond pas à la réalité, (Bouchard 2005, pages 32-36). L'une des voies de recherche actuelle vise à combiner ces deux approches afin de trouver un compromis permettant d'améliorer les performances (Bouchard 2005, Lasserre et al. 2006, Bishop et Lasserre 2007, Druck et al. 2007).

1.2.4 Une solution discriminative classique, les SVM

Les Séparateurs à Vaste Marge ou Machine à Vecteurs Supports (SVM), (Boser et al. 1992) sont actuellement la solution classique au problème de la classification par approche discriminante. Ils proposent une solution au problème de la classification binaire reposant sur la recherche d'un hyperplan particulier pour séparer les deux classes : l'hyperplan séparateur de marge maximale. Celui-ci est formellement défini, dans le cas où les données sont linéairement séparables, comme l'hyperplan le plus éloigné de tous les points de l'ensemble d'apprentissage. Dans le cas de figure où les données ne sont pas linéairement séparables la recherche de l'hyperplan optimal est remplacée par un problème de minimisation de risque empirique régularisé, utilisant la fonction de coût hinge présenté sur la figure 1.4. Les SVM possèdent différentes propriétés attractives : tout d'abord l'hyperplan recherché est unique, les SVM ne souffrent donc pas du problème des minima locaux lorsque les paramètres du noyau sont fixés. D'autre part, la règle de décision issue de l'hyperplan obtenu ne fait intervenir dans le calcul qu'un nombre réduit de points appelés vecteurs supports, ce qui rend l'évaluation de la fonction de décision

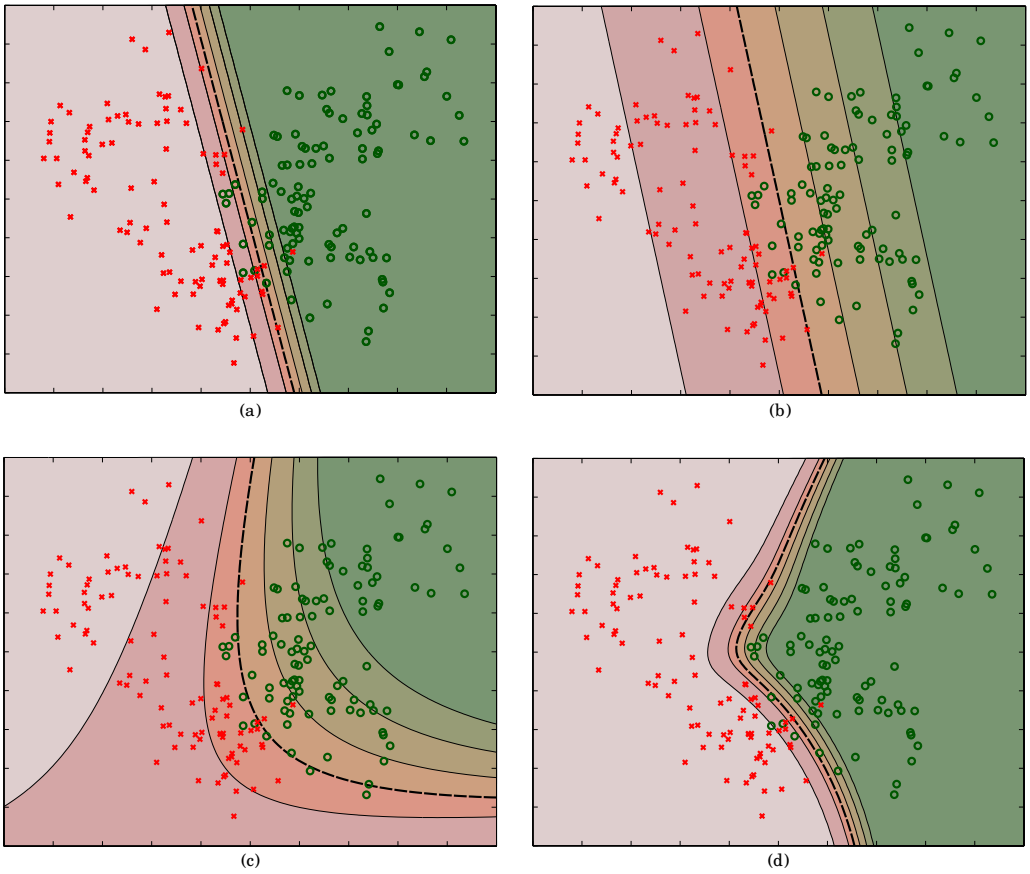


FIG. 1.5 – *Exemple de résultats de classification sur un problème à deux classes. Les deux variables utilisées correspondent aux deuxième et troisième axes principaux du jeu de données crabes (http :/ /rweb.stat.umn.edu /R /library /MASS /html /crabs.html), la classification vise à reconnaître deux espèces différentes de crabes. Les différentes couleurs correspondent aux courbes de niveaux de* $p(Y = c_1|\mathbf{x}) \in \{0,\ 0.1,\ 0.35,\ 0.5,\ 0.65,\ 0.8,\ 0.9\}$. *La frontière de décision est tracée en trait plus épais et pointillé. La figure (a) présente les résultats de la régression logistique. La figure (b) donne les résultats d'un modèle génératif utilisant une gaussienne par classe, celles-ci ayant la même matrice de variance covariance (LDA). La figure (c) correspond au résultat d'un second modèle génératif avec une gaussienne par classe ayant chacune leurs propre matrice de variance-covariance (QDA). La figure (d) présente les résultats d'un modèle génératif utilisant un mélange de gaussiennes, avec deux composantes par classes.*

rapide pour diagnostiquer de nouvelles mesures ; on dit que la solution est parcimonieuse. Finalement, les SVM se sont véritablement imposés lorsqu'ils ont été associés aux méthodes à noyaux que nous décrirons dans la section 1.4. En effet, cette association à permis d'étendre cette approche pour produire une frontière de décision non linéaire.

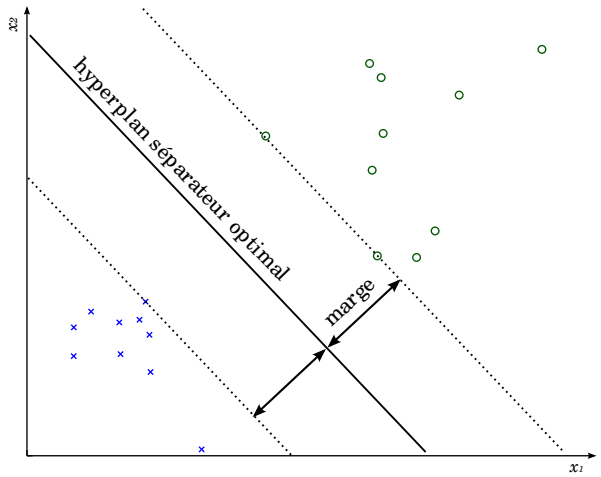


FIG. 1.6 – *Séparateur à vaste marge, exemple de problème de classification linéairement séparable, hyperplan séparateur optimal et marge*

D'autres solutions aux problèmes de la classification existent. On peut citer l'algorithme des k plus proches voisins, les arbres de décision, les réseaux de neurones ou bien encore les méthodes de boosting. Ces dernières ne seront pas présentés dans ce mémoire dans la mesure où elles ne font pas l'objet de développements. De très bonnes introductions à ces méthodes peuvent être trouvées dans les ouvrages Hastie et al. (2006), Bishop (2006).

1.2.5 Autres problèmes d'intérêt pour le diagnostic

Lors de cette introduction aux problématiques de l'apprentissage supervisé, deux problèmes classiques ont été exposés, la classification et la régression. Cependant, l'apprentissage supervisé ne se résume pas à ces deux problèmes, d'autres alternatives telles que la régression ordinale et la détection de nouveauté sont possibles et peuvent s'avérer intéressante en particulier lors de la résolution d'un problème de diagnostic.

régression ordinale

La régression ordinale, se situe à mi-chemin entre régression et classification. Elle suppose en effet, que la variable d'intérêt est catégorielle comme en classification mais elle postule aussi que ces catégories possèdent un ordre naturel. Dans le cadre du diagnostic, ce type de variable est particulièrement bien adapté à la description de mode de fonctionnement dégradés ordonnés suivant leur gravité. Pour plus de détails sur cette approche, il est possible de lire (Chu et Ghahramani 2004).

L'autre problème pouvant être cité est celui de la détection de nouveautés qui

peut être formalisé comme un problème d'estimation de quantile (Schölkopf et al. 2001). Cette approche présente un intérêt majeur pour le diagnostic car elle permet de traiter en pratique les cas où aucune observation n'est disponible sur certaines classes de défauts. Ceci est particulièrement intéressant dans les applications sécuritaire où l'observation de défauts graves est rare voir inexistante (e.g. nucléaire). Cette méthode permet la mise au point de système de surveillance automatique pouvant détecter un comportement atypique du système à surveiller (Aregui et Denœux 2006). Une fois encore, les méthodes à noyau peuvent être utilisées pour résoudre ce problème (one class SVM en anglais) (Schölkopf et al. 2001).

détection de nouveauté

Nous allons maintenant quitter le domaine de l'apprentissage supervisé pour nous tourner vers une autre branche très active du domaine de la reconnaissance des formes, l'apprentissage non-supervisé, appelé aussi fouille de données.

1.3 L'APPRENTISSAGE NON SUPERVISÉ

L'apprentissage non supervisé rentre dans le cadre de la statistique dite descriptive. Il vise à trouver une structure cohérente au sein d'un ensemble de données susceptible d'en faciliter l'interprétation, l'analyse, la représentation. Les problèmes de référence de l'apprentissage non supervisé sont le problème du regroupement automatique et le problème de la réduction de la dimension, avec généralement pour objectif final la visualisation des données en deux dimensions. Nous allons présenter succinctement ces deux problématiques ainsi que les solutions qui ont été proposées dont certaines s'appuient sur des arguments géométriques et d'autres sur une modélisation probabiliste. Dans ce second cas de figure, le modèle postulé pour expliquer les données fait généralement intervenir des variables inobservées appelées variables latentes auxquelles le second chapitre de cette thèse fera une place importante.

1.3.1 Le regroupement automatique

regroupement automatique

Le problème du regroupement automatique, version non supervisée de la classification, vise à trouver des groupes d'individus homogènes au sein des données. Les approches hiérarchique (Ward 1963) (Classification Ascendante Hiérarchique, CAH), l'algorithme des « k-means » (MacQueen 1967), les modèles de mélange (Mclachlan et Peel 2000), et les approches spectrales (Ng et al. 2002) basées sur la théorie des graphes sont des solutions classiques à ce problème. Tout comme dans le cadre de la classification supervisée les sorties fournies par les différentes méthodes de regroupement automatique peuvent être de différentes natures. Certaines méthodes fournissent uniquement une partition « dure » des données en groupes, d'autres méthodes fournissent une partition probabiliste des données. Enfin, certaines méthodes comme les approches hiérarchiques fournissent un ensemble de partitions dures imbriquées formant une hiérarchie. Nous nous intéresserons plus particulièrement dans ce mémoire aux modèles de mélange qui utilisent une formulation probabiliste du problème du regroupement automatique. D'un point de vue statistique le problème du regroupement automatique peut

être formulé comme un problème d'estimation de densité. L'existence de différents groupes dans les données est pour cela introduit grâce à une variable discrète $Y, \mathcal{Y} = \{c_1, \ldots, c_K\}$ codant l'appartenance aux différents groupes. Cette variable n'étant pas observée, le modèle de densité devant être estimé correspond à la densité marginale sur les observations :

$$p(\mathbf{x}; \boldsymbol{\psi}) = \sum_{k=1}^{K} p(Y = c_k; \boldsymbol{\psi}) p(\mathbf{x}|Y = c_k; \boldsymbol{\psi}) \tag{1.31}$$

$$p(\mathbf{x}; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}; \boldsymbol{\theta}_k). \tag{1.32}$$

L'estimation des différents paramètres définissant cette densité

$$\boldsymbol{\psi} = (\pi_1, , \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K), \tag{1.33}$$

par une approche de type maximum de vraisemblance permet de retrouver les groupes associés aux différentes valeurs possibles de la variable latente. En appliquant simplement la règle de Bayes, il vient :

$$p(Y = c_k|\mathbf{x}) = \frac{\pi_k f_k(\mathbf{x}; \boldsymbol{\theta}_k)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(\mathbf{x}; \boldsymbol{\theta}_{k'})}. \tag{1.34}$$

Il est donc possible avec ce type d'approche d'obtenir une classification probabiliste des données, c'est-à-dire d'obtenir la probabilité d'appartenance à chacun des groupes pour tous les individus de l'ensemble d'apprentissage. La forme des différentes sous populations recherchées est ajustée dans ce modèle à l'aide des densités conditionnelles aux classes $f_k(\mathbf{x}; \boldsymbol{\theta}_k), k = 1, \ldots, K$, qui peuvent être de différentes formes afin de prendre en compte les particularité du jeu de données.

Il est intéressant de remarquer que cette formulation du problème du regroupement automatique est équivalente à la formalisation du problème de la classification dans un cadre génératif excepté que l'information sur les classes d'origine des différents individus est manquante. Nous verrons, après avoir développé en détails les différents problèmes et enjeux gravitant autour des modèles de mélange dans le chapitre 2 de cette thèse, comment il est possible d'utiliser le même type d'approche pour traiter des cas de figure où cette information n'est plus manquante, mais imparfaite.

1.3.2 La réduction de la dimension

Le problème de réduction de la dimension vise à extraire du jeu de données initial de nouvelles variables peu nombreuses portant le plus d'information possible sur le jeu de données initial et permettant par exemple la représentation en deux dimension du jeu de données. C'est ici que nous retrouvons les méthodes telles que l'analyse en composantes principales, les cartes auto-organisatrices ou bien encore l'analyse en composantes indépendantes. La première de ces méthodes est aussi la plus ancienne (Pearson 1901, Hotelling 1933) ; elle vise à trouver des directions orthogonales maximisant la variance expliquée et minimisant l'erreur de reconstruction. Cette méthode linéaire très largement utilisée permet de réduire la dimension de l'espace de travail.

Les cartes auto-organisatrice, (Kohonen 1982), permettent de projeter de manière non-linéaire les individus de l'ensemble d'apprentissage sur un espace topologique structuré appelé carte. Cet espace, est généralement défini par un maillage régulier en deux dimensions où chaque nœud peut être associé à une position dans l'espace des descripteurs. Cette méthode permet de représenter de manière pertinente les données en deux dimensions et est donc souvent utilisée en analyse de données exploratoire. D'autres méthodes telles que le positionnement multidimensionnel (multidimensional scaling, MDS, en anglais) tentent elles aussi de trouver une représentation en faible dimensions du jeu de données.

L'analyse en composantes indépendantes, quant à elle, vise à extraire de nouvelles variables qui soient statistiquement indépendantes entre elles. Cet objectif est justifié par la volonté de trouver les différentes « causes » de variation au sein du jeu de données. Ces trois méthodes peuvent être vues comme des modèles statistiques à variables latentes, elle seront présentées plus amplement dans le chapitre 2 de cette thèse.

Après ce rapide aperçu des problématiques associées à l'apprentissage non supervisé et supervisé et à leur solutions courantes, nous présentons deux éléments important qui traversent ces deux domaines, les approches noyaux et la problématique de la sélection de modèles.

1.4 Les méthodes à noyaux

Les différentes méthodes que nous venons de présenter dans le contexte de l'apprentissage supervisé ou de l'apprentissage non supervisé, sont des méthodes linéaires. L'astuce du noyau que nous allons présenter maintenant, permet de travailler dans des espaces résultants de transformations non linéaires des variables initiales. Grâce à celles-ci, il devient possible de construire aisément des frontières de décisions non linéaires dans l'espace d'origine ou bien encore d'étendre l'analyse en composantes principales afin d'extraire de nouvelles variables pertinentes qui dépendent de manière non linéaire des variables initiales.

1.4.1 L'astuce du noyau

L'astuce du noyau (« kernel trick » en anglais) est à l'origine de l'engouement pour les méthodes à noyau. Elle tire partie de la possibilité de calculer directement des produits scalaires dans des espaces de grandes dimensions, résultant de transformations non linéaires des variables initiales, sans projeter explicitement les points de l'ensemble d'apprentissage dans ces espaces. Grâce à cette astuce, n'importe quel algorithme reposant sur la notion de produit scalaire peut être étendu pour traiter des cas non linéaires, en remplaçant le produit scalaire usuel par un noyau dont la définition est donnée ci-après.

Définition 1.6 (Noyau) *Un noyau κ est une fonction de $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ symétrique, telle que :*

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = < \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) >, \quad \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}, \tag{1.35}$$

avec $\mathbf{\Phi} : \mathcal{X} \rightarrow F$, où F est un espace muni d'un produit scalaire.

Un noyau est donc un produit scalaire, c'est-à-dire une forme bi-linéaire symétrique définie positive. Pour manipuler un nuage de points, il n'est pas forcément nécessaire de connaître la position de chacun d'entre eux dans l'espace, il suffit souvent de pouvoir calculer leurs produits scalaires. Le noyau polynomial permet par exemple de travailler dans un espace résultant de transformations polynomiales des variables initiales (cf. exemple 1.3).

Exemple 1.3 (Noyau polynomial de degré deux) :

Soit $\mathbf{x} = (x_1, x_2)$ et $\mathbf{\Phi}(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1.x_2)$ on peut définir :

$$\begin{aligned} \kappa(\mathbf{x}_i, \mathbf{x}_j) &= < \mathbf{x}_i, \mathbf{x}_j >^2 \quad (1.36) \\ &= (x_{i1}.x_{j1} + x_{i2}.x_{j2})^2 \\ &= x_{i1}^2.x_{j1}^2 + x_{i2}^2.x_{j2}^2 + 2.x_{i1}.x_{j1}.x_{i2}.x_{j2} \\ &= < \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) > . \quad (1.37) \end{aligned}$$

Grâce aux noyaux il est donc possible de travailler dans des espaces résultants de transformations non linéaires des variables initiales ; cela peut avoir un impact important sur la solution trouvée, comme le montre la figure 1.7. Nous pouvons observer sur celle-ci qu'un problème de discrimination qui n'était pas linéaire dans l'espace initiale peut le devenir dans l'espace transformé.

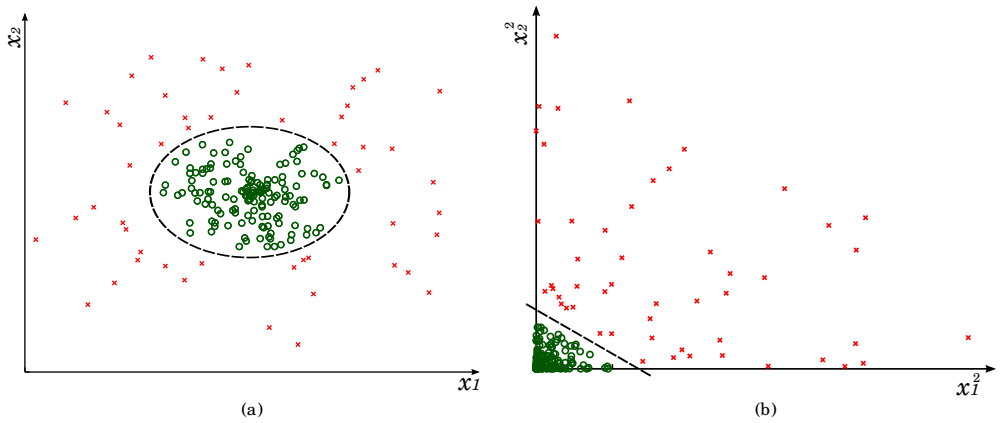


FIG. 1.7 – *Exemple de changement de représentation transformant une frontière de décision non linéaire en frontière de décision linéaire. La frontière de décision dans l'espace d'origine (a) correspond à une ellipse. La figure de droite présente la projection des mêmes individus dans le plan définis par les deux premières composantes de l'espace noyau généré par le noyau polynomial de degré 2, la frontière de décision est linéaire dans ce plan (b).*

De nombreuses méthodes linéaires peuvent être reformulées de manière à ne faire intervenir que des produits scalaires entre individus et celles-ci peuvent donc être étendues pour traiter des données projetées dans un espace résultant de transformations non-linéaires des variables initiales grâce à l'astuce du noyau. Les SVM (Boser et al. 1992) utilisent ainsi l'astuce du noyau pour produire des frontières de décision non-linéaires, de même que la régression logistique (Zhu et Hastie 2001) qui peut elle aussi être « kernélisée ».

Les méthodes d'apprentissage non supervisées ont aussi profité de l'astuce du

noyau, l'analyse en composantes principales ou l'analyse canonique des corrélations peuvent par exemple être écrites en ne faisant intervenir que des produits scalaires entre individus et peuvent donc être étendues pour produire des composantes dépendant de manière non linéaire des variables initiales, (Shawe-Taylor et Cristianini 2004, pages 143-155). Nous présenterons d'ailleurs dans le dernier chapitre de cette thèse une application de l'analyse canonique des corrélations en utilisant l'astuce du noyau qui permettra la mise au point d'un prétraitement non-linéaire.

Dans le cadre des méthodes à noyaux les données ne sont plus manipulées à l'aide de la matrice d'observation $\mathbf{X}$ contenant les différentes variables pour chacun des individus, mais grâce à la matrice contenant les différentes valeurs du noyau pour tous les couples d'individus. Cette matrice, appelée matrice de Gram, est définie formellement de la manière suivante :

Définition 1.7 (Matrice de Gram) *Soit un jeu de données* $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ *et un noyau* κ*, la matrice de Gram* $\mathbf{G}$ *associée, est la matrice telle que :*

$$\mathbf{G}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j), \quad \forall i, j \in \{1, \ldots, N\}. \tag{1.38}$$

Beaucoup de problèmes linéaires fournissent une solution $\mathbf{w}$ qui peut aussi s'écrire comme une combinaison linéaire des points de l'ensemble d'apprentissage :

$$\mathbf{w} = \mathbf{X}^t \boldsymbol{\alpha}. \tag{1.39}$$

En remplaçant $\mathbf{w}$ par $\mathbf{X}^t\boldsymbol{\alpha}$, il est alors possible de reformuler le problème d'estimation en fonction de $\boldsymbol{\alpha}$ et non plus de $\mathbf{w}$, nous allons voir que cette reformulation ne fait intervenir que des produits scalaires entre individus dans l'exemple qui suit.

Exemple 1.4 (Reformulation de la « Ridge » régression) :

En reprenant l'exemple (1.2), nous obtenons (Shawe-Taylor et Cristianini 2004, pages 31-32) :

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} (\mathbf{X}\mathbf{X}^t\boldsymbol{\alpha} - \mathbf{y})^t(\mathbf{X}.\mathbf{X}^t\boldsymbol{\alpha} - \mathbf{y}) + \lambda\boldsymbol{\alpha}^t\mathbf{X}\mathbf{X}^t\boldsymbol{\alpha} \\ \hat{\boldsymbol{\alpha}} &= \mathbf{y}/(\mathbf{X}.\mathbf{X}^t + \lambda\mathbf{I}) = \mathbf{y}/(\mathbf{G} + \lambda\mathbf{I}), \end{aligned} \tag{1.40}$$

avec $\mathbf{G} = \mathbf{X}\mathbf{X}^t$. Cette solution ne fait intervenir que des produits scalaires entre individus au travers la matrice de Gram $\mathbf{G} = \mathbf{X}\mathbf{X}^t$ calculée en utilisant le produit scalaire usuel ; cette méthode peut donc être être étendue en remplaçant cette matrice par une matrice de Gram calculée en utilisant un noyau non-linéaire.

Les algorithmes associés aux méthodes à noyaux, comme le montre l'exemple précédent, utilisent en général en entrée une matrice de Gram décrivant les données grâce aux valeurs prisent par le noyau lorsqu'il est évalué sur les différents exemples de l'ensemble d'apprentissage. En utilisant cette nouvelle représentation du jeu de données, les méthodes à noyaux dissocient clairement représentation des données et exploitation des données. Cela a permis d'étendre l'apprentissage statistique au traitement des données structurées grâce à la définition de noyau sur celles-ci. Des noyaux ont par exemple été définis pour travailler sur des chaînes

de caractères dans le cadre de la classification de texte (Lodhi et al. 2000), pour l'analyse de séquences biologiques (Qiu et al. 2007, Vert et al. 2006),...

Il est intéressant de noter que le noyau peut aussi être un outil utile pour encoder une information experte disponible sur le problème (Schölkopf et Smola 2001, Chap. 11),(Lauer et Bloch 2008).

1.4.2 Espace de Hilbert à noyau reproduisant (RKHS)

Lors de l'émergence des méthodes à noyaux, une propriété a été particulièrement étudiée, la propriété de définie positivité d'un noyau. En effet, tout noyau défini positif peut être vu comme un produit scalaire dans un espace fonctionnel appelé espace de Hilbert à noyau reproduisant (RKHS). C'est donc la propriété que doit posséder un noyau pour être valide. Dans cette section nous développons le lien qui peut être établi entre cette propriété et la définition des espaces de Hilbert à noyaux reproduisant.

Définition 1.8 (Noyau défini positif) *Un fonction* $\kappa(\mathbf{x}_j, \mathbf{x}_j) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ *est défini positive si* $\forall T \in \mathbb{N}, \quad \forall \alpha_1, \ldots, \alpha_T \in \mathbb{R}, \quad \forall \mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathcal{X}$*, on a :*

$$\sum_{i=1}^{T} \sum_{j=1}^{T} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0. \tag{1.41}$$

La matrice de Gram associée à un noyau défini positif est donc semi-définie positive. Nous allons voir qu'il est toujours possible de construire un espace fonctionnel associé à un noyau défini positif tel que ce noyau définisse un produit scalaire entre fonctions. Commençons pour cela par définir l'espace fonctionnel considéré et le produit scalaire utilisé :

Définition 1.9 (Produit scalaire et norme associé à un RKHS) *Soit un noyau défini positif* κ*, et l'espace fonctionnel* $\mathcal{H}$ *:* $f \in \mathcal{H}$ *si et seulement si* f *peut s'écrire :*

$$f(.) = \sum_{i=1}^{T_f} \alpha_i \kappa(., \mathbf{x}_i), \tag{1.42}$$

avec $T_f \in \mathbb{N}$ *et* $\alpha_i \in \mathbb{R}, \forall i \in \{1, \ldots, T_f\}$*. Soit* $f(.)$ *et* $g(.)$ *deux fonctions appartenant à* $\mathcal{H}$*, on à alors :*

$$f(.) = \sum_{i=1}^{T_f} \alpha_i \kappa(., \mathbf{x}_i), \quad g(.) = \sum_{j=1}^{T_g} \beta_j \kappa(., \mathbf{x}_j).$$

Le produit scalaire dans le RKHS $\mathcal{H}$ *associé à* κ *est défini par :*

$$< f(.), g(.) >_{\mathcal{H}} = \sum_{i=1}^{T_f} \sum_{j=1}^{T_g} \alpha_i \beta_j \kappa(\mathbf{x}_i, \mathbf{x}_j), \tag{1.43}$$

la norme d'une fonction étant classiquement définie par $||f|| = \sqrt{< f(.), f(.) >_{\mathcal{H}}}$.

Si le noyau utilisé pour définir cet espace fonctionnel est défini positif, le produit scalaire possède les propriétés nécessaire de symétrie, bi-linéarité, et définie positivité. De plus, le noyau joue un rôle particulier. Nous avons en effet d'après cette définition :

$$< \kappa(., \mathbf{x}), f(.) >_{\mathcal{H}} = f(\mathbf{x}), \tag{1.44}$$

et plus particulièrement :

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = < \kappa(., \mathbf{x}_i), \kappa(., \mathbf{x}_j) >_{\mathcal{H}} . \tag{1.45}$$

Un noyau peut donc être vue de deux manières : tout d'abord comme un produit scalaire entre deux individus projetés dans un espace résultant de transformations non linéaires des variables initiales mais aussi comme le produit scalaire entre deux fonctions définies à partir de ces mêmes points (c'est l'approche fonctionnelle des méthodes à noyaux). Ce sont ces propriétés qui ont donné le nom d'espace de Hilbert à noyau reproduisant. Les noyaux définis positifs peuvent ainsi être vus comme des produits scalaires entre deux fonctions. Nous allons voir que grâce à cette représentation fonctionnelle des noyaux il est possible de travailler dans des espaces fonctionnels de dimension infinie pour résoudre des problèmes d'apprentissage. Le théorème du représentant que nous présentons succinctement joue un rôle important dans cette possibilité.

1.4.3 Le théorème du représentant

Le théorème du représentant (Kimeldorf et Wahba 1971), (Wahba 1990, page 19) stipule qu'il est possible de résoudre des problèmes d'optimisation dont l'argument est une fonction vivant dans un RKHS même si celui-ci est de dimension infinie pourvu que la fonction à optimiser dépende des valeurs prises par la fonction recherchée sur un ensemble fini de points et de la norme de cette fonction dans le RKHS.

Théorème 1.1 (Le théorème du représentant) *Étant donné N points dans $\mathcal{X}$: $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, un noyau $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ défini positif, et un espace de Hilbert à noyaux reproduisant associé $\mathcal{H}$, une fonction $C : \mathbb{R}^N \rightarrow \mathbb{R}$, une fonction strictement monotone croissante $\Gamma(.) : \mathbb{R} \rightarrow \mathbb{R}$ et λ un réel positif. La solution au problème d'optimisation suivant :*

$$\arg\min_{f \in \mathcal{H}} C(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)) + \lambda . \Gamma(||f||_{\mathcal{H}}), \tag{1.46}$$

peut s'écrire sous la forme :

$$f(.) = \sum_{i=1}^{N} \alpha_i \kappa(., \mathbf{x}_i). \tag{1.47}$$

Trouver la meilleure fonction dans $\mathcal{H}$ est donc équivalent à trouver les poids $\alpha_1, \ldots, \alpha_N$ définissant la solution. Ces poids étant en nombre fini, le problème d'optimisation peut donc être résolu à l'aide de méthodes d'optimisation classiques. Le théorème du représentant, de par sa formulation très générale, permet de prendre en compte la majorité des problèmes d'optimisation rencontrés dans le cadre de l'apprentissage statistique.

Il est aussi intéressant de noter que la norme d'une fonction dans le RKHS permet d'introduire une pénalisation intelligente des solutions. L'inégalité de Cauchy-Swartz permet en effet d'écrire :

$$\begin{aligned} \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X} : |f(\mathbf{x}_i) - f(\mathbf{x}_j)| &= | < f(.), \kappa(., \mathbf{x}_i) >_{\mathcal{H}} - < f(.), \kappa(., \mathbf{x}_j) >_{\mathcal{H}} | \\ &= | < f(.), \kappa(., \mathbf{x}_i) - \kappa(., \mathbf{x}_j) >_{\mathcal{H}} | \\ &\leq ||f||_{\mathcal{H}}.||\kappa(., \mathbf{x}_i) - \kappa(., \mathbf{x}_j)||_{\mathcal{H}}. \end{aligned} \tag{1.48}$$

La norme de f dans $\mathcal{H}$ contrôle donc les variations de f. En pénalisant les solutions ayant une norme trop grande dans $\mathcal{H}$, on favorise les solutions « douces ».

Nous avons vu dans cette section comment travailler dans des espaces fonctionnels de grande dimension en utilisant des noyaux, ce qui ouvre la possibilité de traiter des problèmes non linéaires. Nous abordons maintenant une autre problématique importante dans le cadre de l'apprentissage statistique : la sélection de modèle.

1.5 La sélection de modèle

La sélection de modèle concerne aussi bien l'apprentissage supervisé que l'apprentissage non-supervisé. Il s'agit de trouver parmi un ensemble de modèles possibles pour résoudre une tâche donnée celui le plus à même de l'effectuer correctement. La formulation du problème diffère légèrement dans les deux cas (supervisé et non supervisé) et les solutions aussi.

1.5.1 La sélection de modèle dans le cadre supervisé

Lors de la résolution d'un problème de classification ou de régression, il est nécessaire de trouver un modèle adapté en terme de complexité au problème à traiter. En effet, il est important que le classifieur obtenu ait de bons résultats sur les données qui lui seront présentées ultérieurement, on parle de capacité de généralisation (Hastie et al. 2006).

Le compromis biais / variance

Depuis la « découverte » du compromis biais variance, nous savons que cette capacité de généralisation est liée à la complexité du classifieur. Un classifieur très compliqué (faisant intervenir un nombre important de paramètres) pourra apprendre par cœur l'ensemble d'apprentissage, mais généralisera très mal sur de nouvelles données. La variance des performances d'un tel classifieur sera très importante suivant le jeu de données utilisé. A l'inverse un classifieur utilisant peu de paramètres obtiendra peut être de moins bons résultats sur l'ensemble d'apprentissage, mais de par sa simplicité ne pourra pas apprendre par cœur celui-ci. Il sera donc moins affecté par de petites variations de l'ensemble d'apprentissage et généralisera mieux. On dit qu'il aura une variance plus faible mais son biais sera plus important. La figure 1.8 schématise ce point en présentant l'erreur sur l'ensemble d'apprentissage et l'erreur de généralisation en fonction de la complexité du modèle. La question lors de la phase de sélection de modèle est de trouver le modèle

ayant le bon nombre de paramètres pour obtenir le compromis biais variance optimal. Chaque méthode de classification dispose d'un ou de plusieurs paramètres permettant de contrôler ce compromis biais / variance, on parle d'hyper paramètres de la méthode. Pour trouver le modèle idéal, il est nécessaire de disposer d'estimateurs efficaces de l'erreur de généralisation. En comparant ces estimés pour chacun des modèles en concurrence il est alors possible de choisir celui qui obtient les meilleures performances sur les nouvelles données qui lui seront présentées lors de son exploitation.

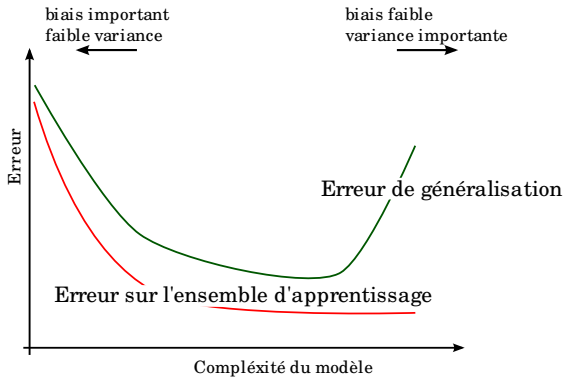


FIG. 1.8 – *Le compromis biais / variance : erreur de généralisation et erreur d'apprentissage en fonction de la complexité du classifieur*

L'estimation de l'erreur de généralisation

La solution la plus simple pour estimer l'erreur de généralisation repose sur un ensemble de données différent de l'ensemble d'apprentissage, appelé ensemble de validation. L'estimation des performances du classifieur ou du régresseur sur cet ensemble fournit une indication sur les capacités de généralisation de celui-ci. Cet estimateur est non biaisé car il utilise des données différentes de celles utilisées lors de l'apprentissage.

Lorsque le nombre d'exemples disponibles est trop faible et qu'une telle approche est impossible à mettre en oeuvre, il est envisageable de se tourner vers des approches de type ré-échantillonnage, (Efron et Tibshirani 1994), telle que la validation croisée qui utilise un découpage du jeu de données en blocs, ou le bootstrap qui utilise des échantillons tirés aléatoirement avec remise au sein du jeu de données. Ces solutions bien qu'efficaces et largement employées peuvent cependant avoir un coût calculatoire important. C'est pourquoi, d'autre solutions se basent sur des critères plus faciles à calculer et faisant intervenir un terme d'attache aux données et un terme de pénalisation dépendant de la complexité (du nombre de paramètres). Ces critères, le plus souvent rapides à calculer, permettent d'accélérer la phase de sélection de modèle. C'est dans ce cadre que l'on retrouve les méthodes telles que la minimisation du risque structurel introduite par (Vapnik 1999) (SRM en anglais), le critère d'information bayésienne (BIC en anglais) (Schwarz 1978), ou bien encore la méthode de la longueur de description minimale (MDL en anglais) (Rissanen 1978).

Pour ajuster le compromis biais/variance, plusieurs solutions de complexité diffé-

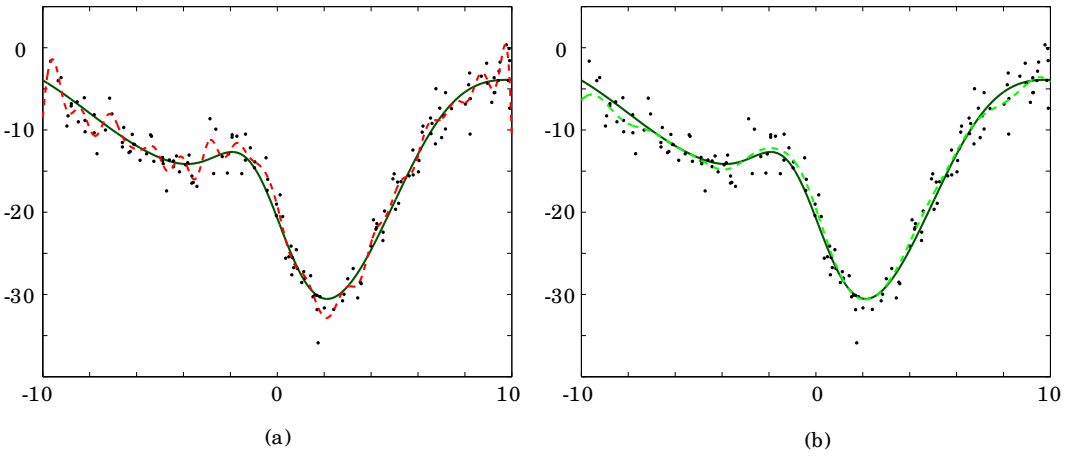


FIG. 1.9 – *Problème de régression non linéaire, contrôle de la complexité par régularisation. Points de l'ensemble d'apprentissage (noir), fonction cible (vert), estimation par ridge régression, noyau gaussien : (a) pénalisation trop faible :* $\lambda = 1.10^{-6}$ *(pointillés rouge), (b) meilleur compromis* $\lambda = 0.005$ *(pointillés vert clair).*

rente doivent être mises en compétition et la meilleure d'entre elles, au sens de l'estimateur de l'erreur de généralisation choisi, est conservée. Chaque méthode construit généralement des solutions vivant dans des espaces de plus en plus complexes (nombre de neurones cachés pour les perceptrons multi-couches, profondeur de l'arbre pour les arbres de décision, structure de la matrice de covariance pour les classifieurs basés sur les mélanges de gaussiennes, paramètres du noyau et coefficient de régularisation pour les SVM,...). Ces paramètres discrets ou continus peuvent prendre part directement au problème d'optimisation en modifiant par exemple la valeur du compromis λ dans un problème de minimisation d'un risque empirique régularisé, ou bien participer à la définition de l'espace de solutions envisagé.

Comme nous l'avons dit les procédures de sélection de modèles peuvent être gourmandes en temps de calcul. En effet, elles nécessitent de résoudre le problème de l'apprentissage pour un ensemble important de valeurs des hyperparamètres et d'estimer pour chacune d'entre elles un indicateur de l'erreur de généralisation.

chemin de régularisation

En ce qui concerne les problèmes de minimisation de risque empirique, où le contrôle de la complexité est réalisé grâce à une régularisation, il est intéressant de noter l'émergence d'algorithmes efficaces de calcul de l'ensemble des solutions correspondant à différentes valeurs de compromis entre les deux termes du critère. Ces méthodes tirent partie des liens existants entre les différents problèmes d'optimisation ainsi définis pour limiter le temps de calcul. C'est le principe des chemins de régularisation, popularisés par l'algorithme LARS, (Efron et al. 2004).

1.5.2 La sélection de modèle dans le cadre non-supervisé

Dans le cadre non supervisé, le problème de la sélection de modèle se pose de manière légèrement différente dans la mesure où il est difficile de définir un critère tel que l'erreur de généralisation. Cependant, la sélection du nombre de classes d'un modèle de mélange ou bien encore la problématique du choix du nombre de composantes principales significatives peuvent être formalisées comme des problèmes de sélection de modèles. Dans le cadre supervisé, nous avons vu que l'erreur sur l'ensemble d'apprentissage ne constituait pas un bon critère pour sélectionner le modèle optimal ; il en va de même dans le cadre non supervisé. L'utilisation du simple critère de vraisemblance pour la sélection du nombre de composantes dans les modèles de mélange conduit par exemple à sélectionner le plus grand nombre de classes car la vraisemblance croît avec le nombre de composants du modèle. En effet, lorsque le nombre de classes augmente, le modèle devient plus riche et s'adapte mieux aux données. Les critères de choix du nombre de classes se basent sur la pénalisation de la vraisemblance par un terme dépendant de la complexité du modèle. Comme dans le cadre d'un risque empirique régularisé, les critères utilisés font intervenir deux terme ; le premier est généralement une vraisemblance, le second pénalise les modèles trop complexes. Les critères BIC (Schwarz 1978), AIC (Akaike 1973), ICL (Biernacki et al. 2000) utilisés pour déterminer le nombre de composants d'un modèle de mélange font partie des solutions couramment utilisées. Des critères obtenus par des approches similaires (Seghouane et Cichocki 2007) existent pour la sélection du nombre de composantes principales devant être conservées pour obtenir un compromis idéal entre compression et représentation.

1.6 Problèmes ouverts et objectifs de la thèse

Le chemin parcouru dans le domaine de l'apprentissage statistique est déjà important ; beaucoup de choses ont été faites d'un point de vue pratique comme théorique (cf. figure 1.2). De mon point de vue, les problèmes ouverts se concentrent sur l'utilisation de toute l'information à disposition pour la résolution d'un cas concret. Dans le domaine du diagnostic d'infrastructure ferroviaire il est légitime de s'interroger sur les questions suivantes :

- Comment prendre en compte des hypothèses sur le mécanisme de génération des données ?
- Comment rendre la solution obtenue invariante par rapport à certains phénomènes / paramètres de nuisances ?
- Comment utiliser à la fois des données labellisées et non labellisées ?
- Comment traiter les informations imparfaites (imprécises ou incertaines) obtenues lors de la phase d'étiquetage des données ?

Pour répondre à ces différentes questions nous nous sommes naturellement tournés vers une approche plutôt paramétrique et générative du problème de diagnostic. En effet, les modèles génératifs en posant la question du mécanisme de génération des données offrent un point de vue pertinent pour prendre en compte des hypothèses sur ce mécanisme. Ils permettent également de prendre en compte l'existence de paramètres de nuisances si ceux-ci sont intégrés à la modélisation du problème. Enfin, comme le montre la principale contribution de cette thèse, ceux-ci

peuvent parfaitement prendre en compte des données imparfaites lors de la phase d'apprentissage.

Les modèles paramétriques ne dominent cependant pas le domaine de l'apprentissage automatique. Les modèles non paramétriques qui ne reposent sur aucune hypothèse quant à la distribution des données sont largement utilisés, en particulier les machines à vecteurs supports. Cependant, le plus grand avantage des méthodes d'estimations non-paramétriques, est aussi leur principal défaut : ne reposer sur aucune hypothèse c'est probablement ce priver de l'avantage de pouvoir effectuer l'apprentissage dans de bonnes conditions avec des jeux de données de taille limité ou lorsque la qualité des labels utilisés lors de l'apprentissage peut être remise en cause. Faire des hypothèses, c'est prendre en considération des informations supplémentaires qui peuvent conduire à une amélioration des performances lorsque ces hypothèses sont en adéquation avec la réalité. Le domaine de l'apprentissage automatique a été largement influencé par les travaux de V. Vapnik et la devise de cet auteur à donc eu un impact important sur ce domaine :

When solving a given problem, try to avoid solving a more general problem as an intermediate step.[3] (Vapnik 1999)

Cette maxime a conduit à une prédominance des méthodes de classification discriminative avec comme algorithme de référence les machines à vecteurs supports. Cependant, la classification supervisée peut aussi bénéficier d'informations supplémentaires, qu'il est difficile de prendre en compte avec une approche purement discriminitive comme les SVM. Ainsi, la formulation du problème de la classification dans un cadre strictement supervisé ne convient pas toujours à la résolution de problématiques pratiques où l'information disponible peut prendre des formes beaucoup plus variées qu'un ensemble de points parfaitement labellisés. L'article polémique de Hand (2006), *« Classifier technology and the illusions of progress »* va dans se sens et remet en question la course à l'élaboration de nouvelle méthodes améliorant à la marge les performances, pour plutôt proposer de travailler à la mise au point de méthodes s'appuyant sur une formalisation plus spécifique du problème prenant en compte ses particularités.

Ce mémoire aborde deux questions fondamentales :

1. la question de la prise en compte d'une information de plus ou moins bonne qualité sur les étiquettes lors de la phase d'apprentissage d'un classifieur ;
2. la question de la prise en compte d'information a priori sur la structuration des données.

Les différentes contributions de cette thèse se sont essentiellement appuyées sur la théorie des fonctions de croyance, sur une démarche générative et sur l'utilisation de modèles paramétriques définis à l'aide de modèles graphiques ;

Pourquoi la théorie des fonctions de croyance ?

La théorie des fonctions de croyance est une théorie de l'incertain, alternative à la théorie des probabilités et qui offre des possibilités de modélisations intéres-

[3] Lors de la résolution d'un problème essayer d'éviter d'en résoudre un autre plus général comme étape intermédiaire

santes. Cette théorie permet en particulier de séparer, lors de la représentation d'une information, l'imprécision de l'incertitude que nous avons sur celle-ci. Elle offre un moyen d'expression très riche en ce qui concerne la modélisation des informations qu'un expert peut apporter sur les états de fonctionnement associés à un ensemble d'observations du système. Le doute et l'absence d'information peuvent par exemple être modélisés de manière simple. Cette théorie nous a semblé être un bon point de départ pour attaquer les problèmes relatifs à la nature parfois incomplète des données rencontrées lors de la mise au point d'un système de diagnostic industriel.

Pourquoi une approche générative ?

L'approche discriminative et l'approche générative sont deux approches complémentaires du problème d'estimation des paramètres d'un modèle. Lorsqu'un grand nombre de données d'apprentissage parfaitement labellisées est disponible, certaines méthodes discriminatives non linéaires peuvent être très performantes. Les méthodes génératives offrent quant à elle des outils de modélisation élaborés qui obtiennent de bon résultats lorsque les données sont de moins bonne qualité. Nous verrons en particulier qu'elles peuvent se montrer extrêmement pertinentes lorsque les données sont labellisés de manière imprécises, incertaines. L'approche générative pose de plus la question de la génération des données et imaginer un processus de simulations des données observées est un exercice naturel qui aide à définir des hypothèses sur la forme de la loi jointe du couple (X, Y). Enfin, les modèles génératifs possèdent de forts liens avec les modèles à variables latentes rencontrés dans le cadre de l'apprentissage non supervisé. Ils peuvent donc naturellement prendre en compte une information incomplète.

Enfin, la supériorité des méthodes discriminatives dans un contexte de données abondantes et parfaitement labellisées semble acquise, mais il en est tout autre dans les contextes d'étiquettes imparfaites qui nous intéressent plus particulièrement.

Pourquoi les modèles paramétriques ?

Les modèles paramétriques proposent des modèles rigides comportant un nombre fini de paramètres. La famille de probabilités qu'ils définissent peut être paramétrée par un vecteur ψ dans un espace de dimension finie. Lorsque des informations sur la structure des données sont disponibles, celles-ci peuvent être prises en compte lors de la définition de la famille de densités considérées. Se contenter de l'information issue du jeu de données d'apprentissage seul est souvent sous optimal. La prise en compte d'informations extérieures, notamment sous la forme d'hypothèses sur la structure des données, peut être extrêmement pertinent pour construire des systèmes efficaces. Les modèles graphiques (Jordan 2006), en tant qu'outil d'aide à la modélisation, peuvent se révéler très pertinents dans ce cadre.

Plan de la thèse

Le chapitre 2 de ce mémoire est consacré à la présentation de différents outils utilisés et manipulés dans cette thèse. Ce chapitre détaillera des modèles statistiques à variables latentes qui jouent un rôle important dans le cadre non-supervisé et que nous étendrons pour prendre en compte des informations plus riches grâce à la théorie des fonctions de croyance. Ce chapitre sera donc également consacré à la description de cette théorie. Pour ce faire, nous reviendrons sur ses principales définitions et nous évoquerons l'extension de cette théorie au départ mise au point pour travailler dans un cadre discret, à la gestion des variables définies sur un référentiel continu.

Le chapitre 3 de cette thèse regroupe nos travaux sur l'utilisation de la théorie des fonctions de croyance pour résoudre le problème de l'apprentissage des paramètres des modèles de mélange, lorsque l'information sur les labels des individus de l'ensemble d'apprentissage présente des imperfections.

L'utilisation d'une démarche similaire, dans le cadre de l'analyse en facteurs indépendants, sera développée dans le chapitre 4. Ce chapitre proposera une solution pour la prise en compte d' informations supplémentaires sur le processus de génération des données. Ces informations prendront la forme d'hypothèses d'indépendances entre variables du modèle pouvant être faites par exemple à partir d'informations sur la physique du système.

Enfin, le dernier chapitre de cette thèse exposera les résultats obtenus sur un problème pratique de diagnostic d'un élément essentiel de la chaîne de contrôle-commande des trains grande vittesse, le circuit de voie TVM.

2 Concepts et outils

L'homme n'a de connaissance des choses naturelles que par les moyens de la correspondance avec ce qui tombe sous les sens.
René Descartes, **Cogitationes Privatae (1859)**

Sommaire

Les travaux regroupés au sein de cette thèse sont au confluent de deux domaines : les modèles à variables latentes et la théorie des fonctions de croyance. Les modèles à variables latentes sont au cœur de l'apprentissage statistique et ont pour particularité d'intégrer au modèle statistique des variables non observées. La théorie des fonctions de croyance est un cadre théorique permettant de représenter et de manipuler des informations pouvant être aussi bien imprécises qu'incertaines. Ces deux outils nous ont permis de proposer différentes solutions pour prendre en compte des informations imprécises et incertaines lors de la résolution de problèmes de classification.

La première partie de ce chapitre est consacrée aux modèles à variables latentes et à l'algorithme EM, qui joue un rôle clef dans l'estimation des paramètres de ces modèles. Après une présentation des modèles graphiques, utiles tout au long de ce chapitre, nous détaillerons différents modèles à variables latentes. Les modèles de mélange, qui intègrent une variable latente discrète, seront abordés ainsi que différents modèles intégrants des variables latentes continues tels que l'analyse factorielle et l'analyse en composantes indépendantes.

Dans un second temps, nous introduirons les éléments essentiels à la théorie des fonctions de croyance en illustrant cette approche par des exemples. Nous aborderons ensuite des éléments plus avancés comme la notion d'indépendance, celle du cadre de discernement continu et celle du théorème de Bayes, mais revus et transposés dans le cadre de cette théorie. Finalement, nous conclurons ce chapitre en évoquant le point de jonction entre ces deux outils introduisant ainsi nos contributions.

2.1 Les modèles à variables latentes

Les modèles à variables latentes jouent un rôle très important dans le domaine de l'apprentissage principalement non supervisé. Comme nous l'avons dit, l'apprentissage non supervisé vise à trouver des structures pertinentes au sein des données, (par exemple l'existence de sous populations homogènes) et les modèles à variables latentes offrent une solution élégante à ce problème. Leur introduction est déjà ancienne et ceux-ci sont sont utilisés dans de très nombreux domaines applicatifs : traitement de la parole et chaîne de Markov cachée (Rabiner et Juang 1993, chap. 6), traitement des images et champ de Markov caché (Besag 1974), psychométrie et analyse factorielle Spearman (1904), ... Cette section sera en particulier consacrée aux modèles faisant intervenir une unique variable latente discrète, appelés modèles de mélange, et aux modèles intégrant des variables latentes continues, qui nous ont été particulièrement utiles lors de la résolution de problèmes de diagnostic.

Les variables latentes postulées par ces différentes méthodes ne sont pas observées lors de la phase d'estimation des paramètres ; l'information est manquante, d'où le nom de variables latentes. Le statut de celles-ci peut varier, elles peuvent correspondre à une réalité inobservée, comme s'avérer être une simple astuce de modélisation. Avant de présenter différents modèles à variables latentes, nous allons étudier l'algorithme EM qui est une solution générale au problème de l'estimation dans le cadre des modèles à variables latentes et constitue un point de passage obligatoire dans ce domaine.

2.1.1 L'algorithme EM

L'algorithme EM, (Baum et al. 1970, Dempster et al. 1977), est la solution classique aux problèmes d'apprentissage faisant intervenir des variables latentes. De par sa simplicité et sa capacité à résoudre nombre de problèmes d'estimations différents, cet algorithme est devenu incontournable. Il repose sur le concept de données manquantes et utilise la structure particulière du problème d'estimation. En effet, les variables latentes n'étant pas observées, la vraisemblance utilisée pour estimer les paramètres est une vraisemblance marginale, obtenue en sommant ou en intégrant sur le domaine des variables latentes (voir définition 2.1).

vraisemblance marginale

Définition 2.1 (Vraisemblance marginale) *Supposons deux ensembles de variables : X représente le groupe de variables observées et Y les variables inobservées. Le modèle est défini sur $\mathcal{X} \times \mathcal{Y}$ par les densités $p(y;\boldsymbol{\psi})$ et $p(\mathbf{x}|y;\boldsymbol{\psi})$ paramétrées par $\boldsymbol{\psi}$. La densité marginale sur $\mathcal{X}$ est alors donnée par :*

$$p(\mathbf{x};\boldsymbol{\psi}) = \sum_{y\in\mathcal{Y}} p(y;\boldsymbol{\psi})p(\mathbf{x}|y;\boldsymbol{\psi}), \tag{2.1}$$

et lorsque seules des réalisations i.i.d. de X (notées $\mathbf{X}$) sont disponibles, la fonction à maximiser pour obtenir une estimation des paramètres $\hat{\boldsymbol{\psi}}_{ml}$, au sens du maximum de vraisemblance, est la vraisemblance marginale :

$$L(\boldsymbol{\psi};\mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{x}_i;\boldsymbol{\psi}) = \prod_{i=1}^{N}\left(\sum_{y\in\mathcal{Y}} p(y;\boldsymbol{\psi})p(\mathbf{x}_i|y;\boldsymbol{\psi})\right). \tag{2.2}$$

L'algorithme EM a été conçu pour tirer partie de cet état de fait et exploiter la structure de la fonction de vraisemblance. Il utilise les méthodes d'estimation généralement employées lorsque toutes les données sont observées, alors même que celles-ci ne le sont pas. L'algorithme EM construit, pour palier ce problème, une distribution de probabilité sur les valeurs pouvant être prises par la ou les variables latentes se sert de celles-ci pour mettre à jour les paramètres utilisés dans les expressions de la densité marginale (2.1). C'est un algorithme itératif simple constitué de deux étapes d'où il tire son nom :

- *étape E, (Espérance) :* lors de cette étape, l'information actuellement disponible, c'est-à-dire les données observées mais aussi l'estimé courant des paramètres est utilisée pour construire une distribution de probabilité sur les valeurs pouvant être prises par les variables latentes pour chacun des individus : $p(y|\mathbf{x}_i; \boldsymbol{\psi}^{(q)})$;
- *étape M, (Maximisation) :* lors de cette étape, les paramètres du modèle sont estimés en utilisant les distributions de probabilité définies à l'étape précédente. Cette étape peut généralement être traitée en ayant recours aux méthodes d'estimation utilisées lorsque toutes les données sont observées, comme le maximum de vraisemblance.

Nous allons dans cette section présenter les éléments nécessaires à la compréhension de l'algorithme et de quelques unes de ses propriétés, telles que la garantie de croissance de la vraisemblance au cours des itérations. Le point de départ de cette présentation est la décomposition de la loi jointe entre variables observées et variables latentes sous la forme d'un produit entre une probabilité conditionnelle et une probabilité a priori.

$$p(\mathbf{x}, y; \boldsymbol{\psi}) = p(y|\mathbf{x}; \boldsymbol{\psi}) p(\mathbf{x}; \boldsymbol{\psi}). \tag{2.3}$$

En passant au logarithme, nous obtenons l'égalité suivante :

$$\log\left(p(\mathbf{x}, y; \boldsymbol{\psi})\right) = \log\left(p(y|\mathbf{x}; \boldsymbol{\psi})\right) + \log\left(p(\mathbf{x}; \boldsymbol{\psi})\right), \tag{2.4}$$

ce qui mène à l'expression suivante pour la log-vraisemblance marginale :

$$\log\left(p(\mathbf{x}; \boldsymbol{\psi})\right) = \log\left(p(\mathbf{x}, y; \boldsymbol{\psi})\right) - \log\left(p(y|\mathbf{x}; \boldsymbol{\psi})\right). \tag{2.5}$$

En prenant l'espérance de cette expression par rapport à la loi conditionnelle des variables latentes connaissant les données observées et la valeur courante des paramètres, nous obtenons :

$$\begin{aligned}\mathbb{E}[\log\left(p(\mathbf{x}; \boldsymbol{\psi})\right) | X = \mathbf{x}, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}] = \\ \mathbb{E}[\log\left(p(\mathbf{x}, Y; \boldsymbol{\psi})\right) | X = \mathbf{x}, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}] \\ - \mathbb{E}[\log\left(p(Y|\mathbf{x}; \boldsymbol{\psi})\right) | X = \mathbf{x}, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}],\end{aligned} \tag{2.6}$$

or $\log\left(p(\mathbf{x}; \boldsymbol{\psi})\right)$ ne dépend pas de Y et donc :

$$\mathbb{E}[\log\left(p(\mathbf{x}; \boldsymbol{\psi})\right) | X = \mathbf{x}, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}] = \log\left(p(\mathbf{x}; \boldsymbol{\psi})\right) = \mathcal{L}(\boldsymbol{\psi}; \mathbf{x}). \tag{2.7}$$

Ce qui nous permet de décomposer la log-vraisemblance marginale :

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{x}) = Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) - H(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}), \tag{2.8}$$

avec :

$$\begin{aligned} Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \mathbb{E}[\log\left(p(\mathbf{x}, Y; \boldsymbol{\psi})\right) | X = \mathbf{x}, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}] \\ &= \sum_{y \in \mathcal{Y}} p(y|\mathbf{x}, \boldsymbol{\psi}^{(q)}) \log\left(p(\mathbf{x}, y; \boldsymbol{\psi})\right) \\ H(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \mathbb{E}[\log\left(p(Y|\mathbf{x}; \boldsymbol{\psi})\right) | X = \mathbf{x}, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}] \\ &= \sum_{y \in \mathcal{Y}} p(y|\mathbf{x}, \boldsymbol{\psi}^{(q)}) \log\left(p(y|\mathbf{x}; \boldsymbol{\psi})\right). \end{aligned}$$

vraisemblance complétée

Le premier terme Q de cette décomposition correspond à l'espérance conditionnelle de la log-vraisemblance des données complétées de leur partie non-observée, on parle de vraisemblance complétée. Nous noterons une telle vraisemblance $\mathcal{L}_c$. La maximisation de cette quantité est au cœur de l'algorithme EM car celle-ci est suffisante pour faire croître la vraisemblance. En effet, la différence entre la log-vraisemblance de deux estimés des paramètres est donnée par :

$$\begin{aligned} \mathcal{L}(\boldsymbol{\psi}^{(q+1)}; \mathbf{x}) - \mathcal{L}(\boldsymbol{\psi}^{(q)}; \mathbf{x}) = &\left(Q(\boldsymbol{\psi}^{(q+1)}, \boldsymbol{\psi}^{(q)}) - Q(\boldsymbol{\psi}^{(q)}, \boldsymbol{\psi}^{(q)})\right) \\ &+ \left(H(\boldsymbol{\psi}^{(q)}, \boldsymbol{\psi}^{(q)}) - H(\boldsymbol{\psi}^{(q+1)}, \boldsymbol{\psi}^{(q)})\right). \end{aligned} \quad (2.9)$$

Or le deuxième terme de cette équation a une forme intéressante :

$$\begin{aligned} &H(\boldsymbol{\psi}^{(q)}, \boldsymbol{\psi}^{(q)}) - H(\boldsymbol{\psi}^{(q+1)}, \boldsymbol{\psi}^{(q)}) = \\ &\quad \sum_{y \in \mathcal{Y}} p(y|\mathbf{x}; \boldsymbol{\psi}^{(q)}) \log\left(p(y|\mathbf{x}; \boldsymbol{\psi}^{(q)})\right) - \sum_{y \in \mathcal{Y}} p(y|\mathbf{x}; \boldsymbol{\psi}^{(q)}) \log\left(p(y|\mathbf{x}; \boldsymbol{\psi}^{(q+1)})\right) \\ &\quad = -\sum_{y \in \mathcal{Y}} p(y|\mathbf{x}; \boldsymbol{\psi}^{(q)}) \log\left(\frac{p(y|\mathbf{x}; \boldsymbol{\psi}^{(q+1)})}{p(y|\mathbf{x}; \boldsymbol{\psi}^{(q)})}\right), \end{aligned} \quad (2.10)$$

et donc

$$H(\boldsymbol{\psi}^{(q)}, \boldsymbol{\psi}^{(q)}) - H(\boldsymbol{\psi}^{(q+1)}, \boldsymbol{\psi}^{(q)}) \geq 0, \quad (2.11)$$

en vertu de l'inégalité de Jensen (Cover et Thomas 1991, pages 25-30).

Ce deuxième terme étant forcément positif, l'algorithme EM s'attache à maximiser la fonction auxiliaire Q durant l'étape M de chacune de ses itérations :

$$\boldsymbol{\psi}^{(q+1)} = \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}), \quad (2.12)$$

ce qui garantit

$$Q(\boldsymbol{\psi}^{(q+1)}, \boldsymbol{\psi}^{(q)}) - Q(\boldsymbol{\psi}^{(q)}, \boldsymbol{\psi}^{(q)}) \geq 0, \quad (2.13)$$

en utilisant (2.9,2.11) on obtient finalement :

$$\mathcal{L}(\boldsymbol{\psi}^{(q+1)}, \mathbf{x}) - \mathcal{L}(\boldsymbol{\psi}^{(q)}, \mathbf{x}) \geq 0. \quad (2.14)$$

L'algorithme EM fait donc croître la vraisemblance des paramètres à chacune de ses itérations (cf. algorithme 1). En ajoutant des conditions assez générales sur la fonction de vraisemblance, il est possible de démontrer qu'il converge vers un maximum local de celle-ci, (Dempster et al. 1977, Wu 1983, Xu et Jordan 1996). Cette formulation générale mène à différents algorithmes dédiés à l'apprentissage des paramètres des modèles de mélange, de l'analyse en composantes principales probabiliste ou bien encore de l'analyse en facteurs indépendants. Par conséquent cette solution peut être vue comme un principe de conception d'algorithmes plutôt que comme un simple algorithme.

Algorithme 1: pseudo-code de l'algorithme EM « générique ».

Données : Matrice des données : $\mathbf{X}$
 # Initialisation
$\boldsymbol{\psi}^{(0)}, q = 0$

tant que *test de convergence* **faire**
 # Etape E
 # calcul de Q

$Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) = \mathbb{E}[\log\left(p(\mathbf{x}, Y; \boldsymbol{\psi})\right) | X = x, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}]$

 # Etape M
 # maximisation de la fonction auxiliaire

$\boldsymbol{\psi}^{(q+1)} = \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)})$

$q = q + 1$

Résultat : Paramètres estimés : $\widehat{\boldsymbol{\psi}}^{ml}$

L'algorithme EM vu comme un algorithme d'optimisation alternée

La présentation précédente de l'algorithme EM, dans l'esprit des travaux de Dempster et al. (1977), n'est pas la seule possible. En effet, Neal et Hinton (1998) ont proposé un autre point de vue sur cet algorithme qui a ouvert la voie aux approches variationnelles très utilisées dans le cadre bayésien (Jordan et al. 1999, Jaakkola et Jordan 2000). Dans cette présentation de l'algorithme, on constate principalement que la vraisemblance marginale peut être bornée. En effet, en vertu de l'inégalité de Jensen nous pouvons écrire :

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\psi}; \mathbf{x}) &= \log(p(\mathbf{x}|\boldsymbol{\psi})) = \log\left(\sum_{y \in \mathcal{Y}} p(\mathbf{x}, y|\boldsymbol{\psi})\right) \\
&= \log\left(\sum_{y \in \mathcal{Y}} h(y|\mathbf{x}) \frac{p(\mathbf{x}, y|\boldsymbol{\psi})}{h(y|\mathbf{x})}\right) \\
&\geq \sum_{y \in \mathcal{Y}} h(y|\mathbf{x}) \log\left(\frac{p(\mathbf{x}, y|\boldsymbol{\psi})}{h(y|\mathbf{x})}\right), \qquad (2.15)
\end{aligned}$$

où $h(y|\mathbf{x})$ est une densité[1] quelconque sur $\mathcal{Y}$. L'algorithme EM peut alors être présenté comme un algorithme d'optimisation alternée. Définissons pour cela la fonction :

$$\mathcal{B}(h, \boldsymbol{\psi}) = \sum_{y \in \mathcal{Y}} h(y|\mathbf{x}) \log\left(\frac{p(\mathbf{x}, y|\boldsymbol{\psi})}{h(y|\mathbf{x})}\right). \qquad (2.16)$$

Nous allons voir qu'en maximisant alternativement cette fonction par rapport à h et $\boldsymbol{\psi}$ nous obtenons l'algorithme EM. En effet, si nous posons :

$$h(y|\mathbf{x}) = p(y|\mathbf{x}; \boldsymbol{\psi}), \qquad (2.17)$$

[1] Si la variable latente est discrète cette fonction correspondra à une fonction de probabilité, dans le cas continue elle correspondra à une densité.

nous obtenons en utilisant (2.8) :

$$\mathcal{B}(p(y|\mathbf{x};\boldsymbol{\psi}),\boldsymbol{\psi}) = \mathcal{L}(\boldsymbol{\psi};\mathbf{x}). \tag{2.18}$$

Or, $\mathcal{L}(\boldsymbol{\psi};\mathbf{x})$ est une borne supérieure de $\mathcal{B}(h,\boldsymbol{\psi})$ (2.15) nous pouvons donc conclure que $p(y|\mathbf{x},\boldsymbol{\psi}^{(q)})$ est la densité qui maximise $\mathcal{B}(h,\boldsymbol{\psi}^{(q)})$ par rapport à h. L'étape E de l'algorithme peut ainsi être vue comme une maximisation.

Cette maximisation permet de plus de combler le « fossé » séparant la Borne $\mathcal{B}$ de la log-vraisemblance $\mathcal{L}$ et d'obtenir l'égalité entre ces deux fonctions au point $\boldsymbol{\psi}^{(q)}$, ce qui est essentiel pour démontrer la convergence de l'algorithme.

En ce qui concerne l'étape M, nous pouvons observer que :

$$\mathcal{B}(h,\boldsymbol{\psi}) = \sum_{y\in\mathcal{Y}} h(y|\mathbf{x}) \log\left(p(\mathbf{x},y|\boldsymbol{\psi})\right) - \sum_{y\in\mathcal{Y}} h(y|\mathbf{x}) \log\left(h(y|\mathbf{x})\right). \tag{2.19}$$

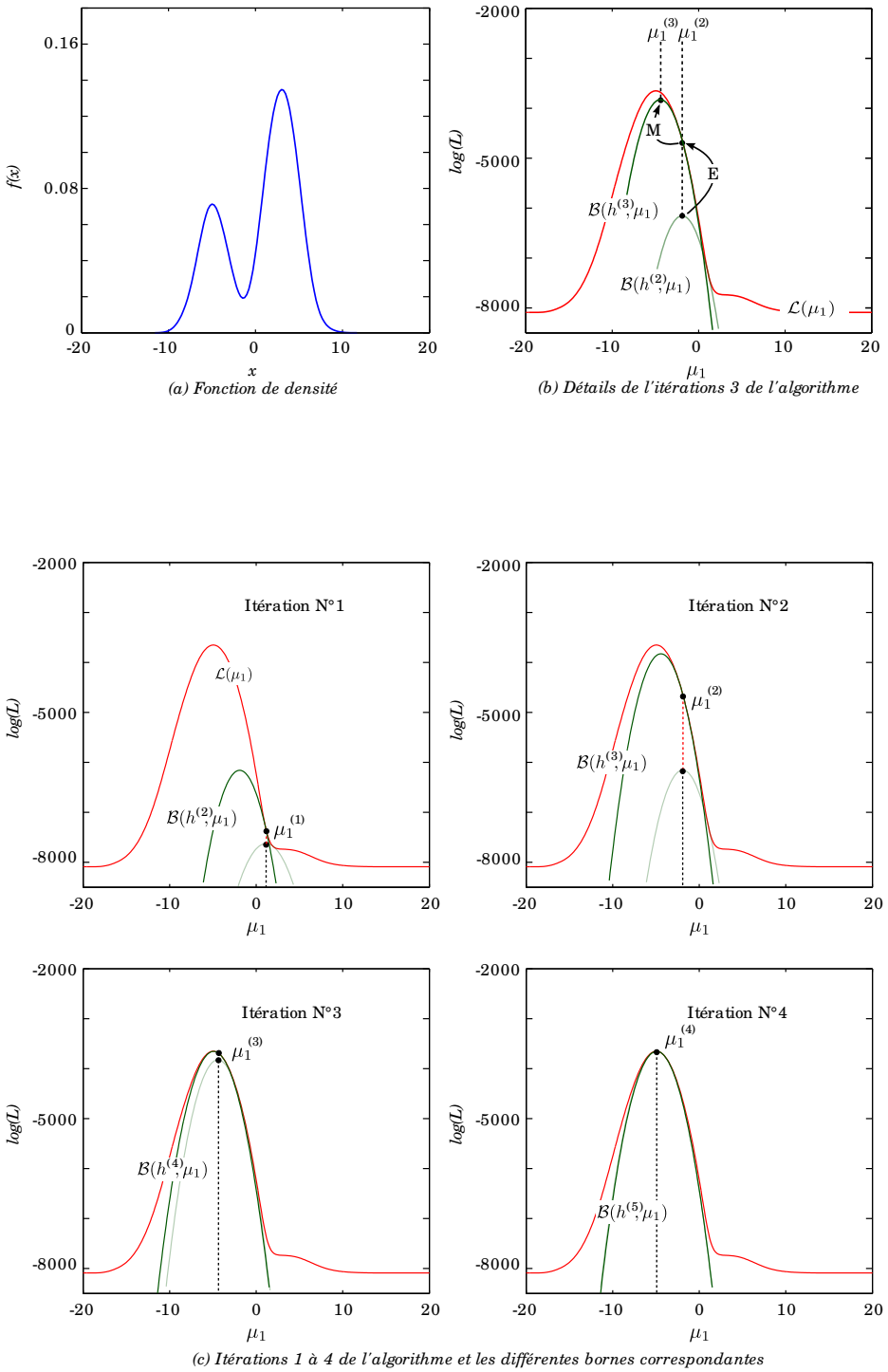
La maximisation de $\mathcal{B}(h,\boldsymbol{\psi})$ par rapport à $\boldsymbol{\psi}$ est donc équivalente à la maximisation du premier terme de (2.19) car le second terme ne dépend pas de $\boldsymbol{\psi}$. La maximisation de cette borne par rapport à $\boldsymbol{\psi}$ est par conséquent équivalente à la maximisation de l'espérance conditionnelle de la log-vraisemblance des données complétées ; nous retrouvons donc l'étape M de l'algorithme.

optimisation variationnelle

Les approches variationnelles suivent un cheminement similaire à cette présentation de l'algorithme. Elles s'appuient sur une fonction de deux arguments pour laquelle la fonction à maximiser constitue une borne supérieure, et optimisent alternativement cette fonction par rapport à chacun de ses deux arguments.

Exemple 2.1 (Illustration de l'algorithme EM) :

> Un jeu de données de 1000 observations a été simulé suivant un modèle de mélange gaussien à deux composantes dont la densité est présentée sur la figure 2.1 (a). Les paramètres de cette densité sont les suivant :
>
> $$\begin{aligned} \pi_1 &= 0.3, & \pi_2 &= 0.7 \\ \mu_1 &= -5, & \mu_2 &= 3 \\ \nu_1 &= 2, & \nu_2 &= 3 \end{aligned}$$
>
> Le problème d'estimation « jouet » étudié dans cet exemple concerne l'estimation de la moyenne μ_1 d'une composante du mélange, les autres paramètres étant connus. Ce choix a été motivé par la possibilité de représenter dans ce cas de figure le fonction de log-vraisemblance en dimension 1. La figure 2.1 (b) présente la fonction de log-vraisemblance à maximiser ainsi que deux bornes consécutives construites par l'algorithme, finalement la figure 2.1 (c) fournit les bornes des itérations 1 à 4 de l'algorithme. Nous pouvons clairement observer sur ces figures le fonctionnement de l'algorithme :
> - étape E : construction d'une borne égalant la log-vraisemblance pour la valeur courante du paramètre et strictement inférieure à la log-vraisemblance pour toutes les autres valeurs du paramètre ;
> - étape M : maximisation de la borne courante par rapport au paramètre.
>
> Ces deux étapes étant répétées jusqu'à la convergence.



(a) *Fonction de densité*

(b) *Détails de l'itérations 3 de l'algorithme*

(c) *Itérations 1 à 4 de l'algorithme et les différentes bornes correspondantes*

FIG. 2.1 – *Illustration de l'algorithme EM : l'algorithme est utilisé pour estimer la moyenne d'une composante d'un mélange de gaussiennes, tous les autres paramètres étant connus. La figure (a) présente la densité ayant servie à simuler les données. La figure (b) présente la fonction de vraisemblance ainsi que les différentes bornes des itérations 2 et 3 de l'algorithme en fonction de μ_1 et détaille le passage de l'une à l'autre. La figure (c) fournit les bornes des itérations 1 à 4.*

Remarques pratiques quant à la mise en œuvre d'un algorithme EM

Dans la pratique, la mise en œuvre d'un algorithme EM nécessite de prendre quelques précautions. Lorsque les variables latentes sont à valeurs discrètes, les probabilités a posteriori peuvent par exemple être stockées avantageusement sur une échelle logarithmique afin d'éviter les problèmes de précision machine. D'autre part, comme tout algorithme itératif, l'algorithme EM doit être arrêté ; il est donc nécessaire de mettre en place un test de convergence. Différentes solutions sont envisageables pour cela ; il est par exemple possible de surveiller l'évolution de la fonction de vraisemblance (Mclachlan et Peel 2000, page 52), l'évolution des probabilités a posteriori, ou bien encore celle des paramètres.

Un autre point mérite d'être vérifié lors de la mise en œuvre de cet algorithme : la nature non singulière de la fonction de log-vraisemblance au voisinage de la solution obtenue. Ce cas peut ce présenter lorsque l'algorithme est utilisé pour estimer les paramètres d'un modèle de mélange gaussien. Dans ce cadre, la vraisemblance est singulière sur les bords de l'espace des paramètres car il est possible de faire tendre celle-ci vers l'infini en faisant tendre la variance d'une des composantes vers zéro, pourvu que le centre de cette composante soit identique à la position d'un point de l'ensemble d'apprentissage (Mclachlan et Peel 2000, page 99), (Archambeau et al. 2003).

maximum local

maximum global

Enfin, comme l'algorithme EM converge vers un maximum local de la vraisemblance, l'initialisation de l'algorithme a un rôle très important . Il est courant d'utiliser différentes initialisations aléatoires de l'algorithme et de retenir la meilleure d'entre elles, en espérant que la convergence qui s'en suit corresponde au maximum global de la vraisemblance. Des stratégies d'initialisation plus sophistiquées peuvent être mises au point suivant le problème traité (Biernacki et al. 2003). Il est aussi intéressant de noter l'existence de test permettant de savoir si la solution obtenue par l'algorithme correspond au maximum global de la vraisemblance (Biernacki 2005).

Extensions de l'algorithme EM

L'algorithme EM a été très étudié et de nombreuses extensions ont été proposées pour répondre au problème des maxima locaux, afin adapter celui-ci aux situations ou l'étape M peut être délicate. Nous présentons ici quelques unes d'entre elles. L'algorithme EM peut tout d'abord être utilisé « en ligne », c'est-à-dire lorsque les échantillons d'apprentissage se présentent un à un. Dans ce cas de figure, il est possible de trouver des formules récursives de mise à jour des paramètres faisant intervenir l'estimé de l'itération précédente et une statistique associée au nouveau point d'apprentissage (Titterington 1984, Cappé et Moulines 2007). Il est aussi envisageable, de mettre à jour les paramètres de manière asynchrone, lorsque la taille du vecteur de paramètres est importante ou lorsque le problème s'y prête, (Meng et Rubin 1993). D'autre part, lorsque l'étape de maximisation est difficile à réaliser, c'est-à-dire lorsqu'il n'existe pas de solution analytique, celle-ci peut être effectuée partiellement dans la mesure ou la croissance de Q est suffisante pour garantir la croissance de la vraisemblance et conserver les bonnes propriétés de l'algorithme (Dempster et al. 1977). Cette variante est nommée GEM pour algo-

algorithme GEM

rithme EM Généralisé. Enfin, de nombreuses extensions ont été proposées dans le cas particulier des modèles de mélange, nous reviendrons sur celles-ci au paragraphe 2.1.3 après avoir présenté les modèles graphiques et les modèles de mélange.

Pour un tour d'horizon des différentes applications envisageables de l'algorithme EM et de ses nombreuses extensions on pourra se reporter à l'ouvrage de Mclachlan et Krishnan (1996).

2.1.2 Les modèles graphiques

Les premiers travaux sur les modèles graphiques datent du début des années 1980 (Pearl 1988), avec l'introduction des réseaux bayésiens et leur intérêt ne s'est pas démenti depuis. Les possibilités offertes par ces modèles sont importantes et les ouvrages de Jordan (1998; 2006) en donnent un large aperçu.

Les modèles graphiques permettent de représenter un ensemble d'hypothèses d'indépendances et d'indépendances conditionnelles sur une loi jointe faisant intervenir différentes variables aléatoires. Ils associent deux éléments clefs : un graphe orienté acyclique représentant les hypothèses d'indépendance et un modèle de calcul et d'inférence au sein du graphe. Leurs principaux avantages sont les suivants :

- la visualisation simple de la structure d'un modèle probabiliste ;
- les différentes propriétés du modèle comme les hypothèse d'indépendance conditionnelle peuvent être extraites du graphe ;
- les calculs dans les modèles complexes peuvent être effectués efficacement grâce à des méthodes de calcul locales qui tirent parti de la structure du réseau ; l'algorithme de l'arbre de jonction permet ainsi de déterminer les lois marginales de tous les nœuds du réseau de façon exacte en utilisant des calculs locaux (Lauritzen et Spieger 1988).

Le graphe orienté servant à représenter les hypothèses faites sur la loi jointe décrit une factorisation de celle-ci. Chaque sommet du graphe représente une variable et chaque arc représente une relation de dépendance conditionnelle entre la variable fille et la variable parente. La loi jointe sur toutes les variables du graphe, c'est-à-dire sur tous ses sommets, peut être reliée aux lois conditionnelles de toutes les variables connaissant leurs parents, grâce à la définition suivante :

Définition 2.2 (Loi jointe et factorisation décrite par un graphe) *Soit un graphe orienté acyclique $G = \{S, A\}$ où $S = \{X_1, \ldots, X_L\}$ est l'ensemble des sommets et A l'ensemble des arcs. Nous notons $par(x), X \in S$ l'application qui fait correspondre à chaque nœud du graphe l'ensemble de ses parents dans G. La loi jointe décrite par le graphe G est définie par :*

$$p(x_1, x_2, \ldots, x_L) = \prod_{l=1}^{L} p(x_l | par(x_l)). \tag{2.20}$$

Les modèles graphiques qui marient la théorie des graphes et la théorie des probabilités mettent en lumière l'équivalence qu'il est possible d'établir entre une propriété d'un graphe appelée D-séparation et la notion d'indépendance conditionnelle

en probabilité (Shenoy 1994; 1992). D'autres théories de l'incertain possédant une notion équivalente à la notion d'indépendance conditionnelle, tel que la théorie des fonctions de croyance qui sera détaillée à la fin de ce chapitre, peuvent tirer partie de cette relation et proposent donc des outils similaires aux modèles graphiques probabilistes (Xu et Smets 1994, Shenoy et Kohlas 2000).

Les conventions utilisées dans cette thèse pour représenter les divers modèles graphiques que nous rencontrerons sont détaillées en figure 2.2. Nous avons utilisé des couleurs différentes pour représenter les variables observées (fond noir) et les variables non observées (fond blanc). La forme carrée sera associée à une variable discrète et une forme ronde à une variable continue. Un ensemble de variables i.i.d. sera représenté graphiquement grâce à un rectangle entourant la variable parente. Enfin, nous ferons parfois apparaître explicitement les paramètres du modèle ; dans ce cas de figure ceux-ci seront représentés à l'aide de ronds plus petits et de couleur gris claire.

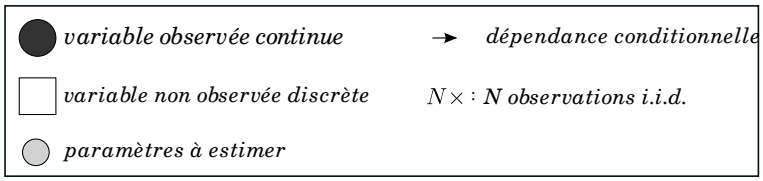


FIG. 2.2 – *Modèle graphique : conventions de représentation*

2.1.3 Les modèles de mélange

Les modèles de mélange sont adaptés à la résolution du problème du regroupement automatique, c'est-à-dire la recherche de groupes homogènes au sein des données. En effet, ceux-ci traduisent l'hypothèse même à la base de ce concept : l'existence de différentes sous-populations ou classes d'individus dans le jeu de données. L'existence de ces sous populations est introduite dans le modèle par l'intermédiaire d'une variable latente Y à valeur dans un ensemble discret $\mathcal{Y} = \{c_1, \ldots, c_K\}$. De manière plus concrète, les modèles de mélange supposent un modèle de génération des données de la forme suivante :

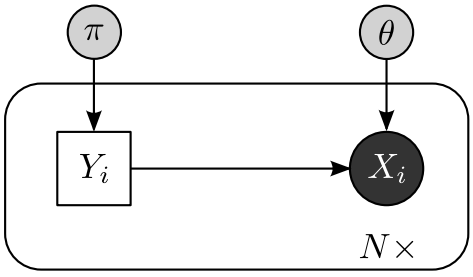


FIG. 2.3 – *Modèle graphique de génération des données d'un modèle de mélange.*

- La classe (sous population) d'origine de chaque observation Y_i est tirée aléatoirement parmi un ensemble de classes possibles : $\mathcal{Y} = \{c_1, \ldots, c_K\}$. Plus précisément, la classe des observations est une réalisation d'une variable aléatoire i.i.d.

de loi multinomiale $\mathcal{M}(1, \pi_1, \ldots, \pi_K)$. Les π_k sont les proportions de chacune des classes dans la population globale et vérifient $\sum_{k=1}^{K} \pi_k = 1$. Pour simplifier les équations, il est possible de coder l'information sur la classe de chacun des points à l'aide d'une variable binaire $\mathbf{z}_i \in \{0,1\}^K$, tel que : $z_{ik} = 1$ si $y_i = c_k$, et $z_{ik'} = 0$ sinon.

- Les valeurs observées $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ sont ensuite tirées en utilisant une loi conditionnelle qui dépend de la classe de l'individu. Plus formellement, $X_1, \ldots, X_N \sim X$ sont des variables aléatoires à valeur dans $\mathcal{X}$, telles que $f(\mathbf{x}|Y = c_k) = f(\mathbf{x}; \boldsymbol{\theta}_k), \quad \forall k \in \{1, \ldots, K\}$.

Les paramètres de ce modèle sont finalement les proportions des classes et les paramètres des densités conditionnelles utilisées pour tirer les variables observées. L'ensemble de ces paramètres sera noté :

$$\boldsymbol{\psi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K). \tag{2.21}$$

La figure 2.3 présente le modèle graphique représentant un modèle de mélange générique intégrant les différents paramètres intervenant dans sa définition. Conformément a ce qui a été dit, l'estimation des paramètres des modèles de mélange s'effectue classiquement au travers de la maximisation de la log-vraisemblance marginale des données observées :

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k) \right). \tag{2.22}$$

Suivant les lois conditionnelles utilisées pour tirer les observations, les modèles de mélanges peuvent être adaptés à différents problèmes. Nous présentons ici quelques exemples.

Exemple 2.2 (Modèle de mélange gaussien pour la classification automatique) :

Le choix le plus classique pour la forme des lois conditionnelles quand les observations sont à valeur dans $\mathbb{R}^P$, est sans aucun doute celui correspondant à la loi normale multivariée :

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\theta}_k) &= \boldsymbol{\varphi}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \frac{1}{(2.\pi)^{\frac{P}{2}} |\det(\boldsymbol{\Sigma}_k)|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right), \end{aligned} \tag{2.23}$$

où $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, contient le vecteur moyenne et la matrice de variance-covariance de la sous-population k. Un exemple de densité associée à un modèle de mélange gaussien unidimensionnel est présenté sur la figure 2.4.

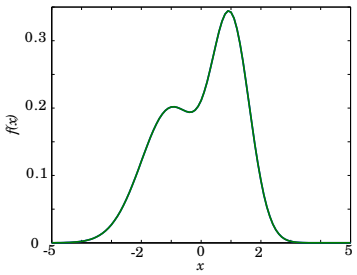


FIG. 2.4 – *Exemples de modèles de mélange : densité d'un mélange de gaussiennes monodimensionnel à 2 composantes.*

Exemple 2.3 (Modèle de mélange de lois de Weibull pour l'étude de durée de vie) :

Les modèles de mélange peuvent aussi être utilisés dans le cadre de l'étude de durée de vie pour modéliser l'hétérogénéité de la population (Marín et al. 2005). Cela peut par exemple être intéressant lorsque des composants sont fabriqués par différentes entreprises. Dans ce cas de figure, les observations x sont à valeur dans $\mathbb{R}^+$ et les lois conditionnelles sont de la forme :

$$f(x;\boldsymbol{\theta}_k) = \left(\frac{\alpha_k}{\lambda_k}\right)\left(\frac{x}{\lambda_k}\right)^{(\alpha_k-1)} \exp\left(-\left(\frac{x}{\lambda_k}\right)^{\alpha_k}\right), \tag{2.24}$$

où $\boldsymbol{\theta}_k = (\alpha_k, \lambda_k)$ représente les paramètres de forme et d'échelle de la sous-population k. Un exemple de fonction de répartition associée à un mélange de Weibull est présenté sur la figure 2.5.

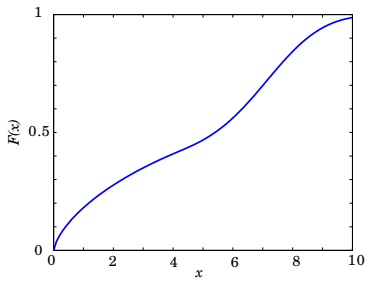


FIG. 2.5 – *Exemples de modèles de mélange : fonction de répartition d'un mélange de lois de Weibull.*

Exemple 2.4 (Modèle de mélange de lois multinomiales pour la classification automatique de texte) :

Lorsque les observations correspondent à un vecteur de variables catégorielles $X = (X^1, \ldots, X^P), X^1 \in \{0,1\}^{nb_1}, \ldots, X^P \in \{0,1\}^{nb_p}$, il est possible de définir une loi conditionnelle à chacune des classes à partir d'un produit de lois multinomiales.

$$f(\mathbf{x}; \boldsymbol{\theta}_k) = \prod_{p=1}^{P} \prod_{c=1}^{nb_p} \left(\rho_k^{pc}\right)^{\mathbf{x}^{pc}}, \tag{2.25}$$

avec $\boldsymbol{\theta}_k = \left(\rho_k^{pc}\right),\ \forall p \in \{1, \ldots, P\}, \forall c \in \{1, \ldots, nb_p\}$ les probabilités de chacune des catégories pour chacune des variables. L'hypothèse d'indépendance conditionnelle des différentes variables observées connaissant la classe d'origine est réalisée, comme le montre le modèle graphique présenté en figure 2.6. Ce type de modèle peut par exemple être utilisé pour classifier automatiquement un ensemble de documents par rapport à leurs données textuelles. Chaque variable catégorielle définit dans ce cas le nombre de fois où un mot apparaît dans un texte (Nigam et al. 2000).

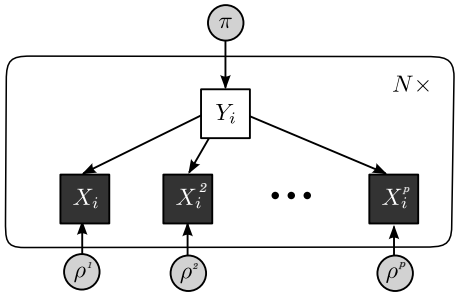


FIG. 2.6 – *Modèle graphique associé au modèle de mélange de lois multinomiales sous hypothèse d'indépendance conditionnelle des différentes variables observables.*

L'algorithme EM dans le contexte des modèles de mélange

En ce qui concerne l'adaptation de l'algorithme EM au contexte des modèles de mélange, il est intéressant de noter que l'étape E prend une forme générale quelque soit les formes paramétriques de densités conditionnelles postulées. Dans tous les cas de figure, il est nécessaire de calculer les probabilités a posteriori $p(c_k|\mathbf{x}_i; \boldsymbol{\psi}^{(q)})$ c'est-à-dire la probabilité que l'observation i provienne de la classe k connaissant les variables observées et la valeur courante des paramètres. Pour alléger les notations, nous noterons t_{ik} ces probabilités a posteriori. Quelque soit la forme paramétrique postulée pour les lois conditionnelles, ces probabilités a posteriori peuvent être calculées en utilisant l'expression suivante :

$$t_{ik}^{(q)} = p\left(c_k|\mathbf{x}_i; \boldsymbol{\psi}^{(q)}\right) = \frac{\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} \pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})}, \tag{2.26}$$

en utilisant ces notations, la fonction auxiliaire Q prend la forme suivante dans le cadre des modèles de mélange :

$$Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik}^{(q)} \log(\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)). \tag{2.27}$$

L'étape de maximisation de l'algorithme est en revanche affectée par le choix des densités conditionnelles, excepté en ce concerne les proportions. En effet, pour

maximiser la fonction auxiliaire Q par rapport à celles-ci, il suffit d'utiliser la formule suivante (cf. annexe .1) :

$$\pi_k^{(q+1)} = \sum_{i=1}^{N} t_{ik}^{(q)} / N. \tag{2.28}$$

La mise à jour des proportions lors de l'étape M correspond donc à un simple calcul de fréquence pondéré.

En ce qui concerne les autres paramètres, qui eux diffèrent suivant les lois conditionnelles postulées, il est nécessaire de dériver les formules de mises à jour au cas par cas. Cependant, lorsqu'il existe une solution analytique au problème d'estimation dans le cas de données complètes (x et y connues), la maximisation de Q conduit à une solution similaire, mais ou la contribution de chacun des individus est pondérée par les probabilités a posteriori d'appartenance aux classes. Par exemple, en ce qui concerne les modèles de mélanges gaussiens, en calculant la dérivée de Q par rapport aux paramètres et en annulant celle-ci, nous obtenons les formules analytiques suivantes :

$$\begin{aligned} \boldsymbol{\mu}_k^{(q+1)} &= \frac{1}{\sum_{i=1}^{N} t_{ik}^{(q)}} \sum_{i=1}^{N} t_{ik}^{(q)} \mathbf{x}_i \\ \boldsymbol{\Sigma}_k^{(q+1)} &= \frac{1}{\sum_{i=1}^{N} t_{ik}^{(q)}} \sum_{i=1}^{N} t_{ik}^{(q)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})^t (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)}). \end{aligned}$$

L'algorithme EM prend alors une forme simple lorsque des gaussiennes sont utilisées pour modéliser chacun des groupes (cf. algorithme 2).

Les modèles de mélange gaussien sont particulièrement intéressantes dans le contexte du diagnostic. L'hypothèse de normalité des distributions conditionnelles est en effet légitime pour représenter la dispersion des mesures de contrôle au sein de chacun des modes de fonctionnement. Nous profitons donc de ce chapitre pour décrire quelques extensions des modèles de mélanges gaussiens qui peuvent être pertinentes lors de la résolution pratique d'un problème de diagnostic.

Variantes autour des mélanges gaussiens

De nombreuses extensions ont été proposées aux modèles de mélange gaussien ; celles-ci concernent généralement la forme du modèle postulé. Les modèles de mélange parcimonieux visent par exemple à réduire le nombre de paramètres, (Banfield et Raftery 1993, Celeux et Govaert 1995) et avancent pour cela différentes hypothèses sur la forme des matrices de variance-covariance. Il est par exemple possible de contraindre celles-ci à être identiques, c'est l'hypothèse d'homoscedasticité. Il est aussi courant de forcer ces matrices à être diagonales, ce qui est équivalent à supposer que les différentes variables sont indépendantes conditionnellement aux classes, ou bien encore de forcer ces matrices à être proportionnelles à la matrice identité. Ces différentes paramétrisation ainsi que d'autres peuvent être obtenues à partir de la décomposition spectrale des matrices de variance-covariance :

$$\boldsymbol{\Sigma} = \alpha \, A \, D \, A^t, \tag{2.29}$$

Algorithme 2: pseudo-code de l'algorithme EM pour les modèles de mélange gaussien.

Données : Matrice des données : $\mathbf{X}$

 # Initialisation

$$\boldsymbol{\psi}^{(0)} = \left(\pi_1^{(0)}, \ldots, \pi_K^{(0)}, \boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_K^{(0)}\right), q = 0$$

tant que *test de convergence* **faire**

 # Etape E

 # calcul des probabilités a posteriori

pour tous les $k \in \{1, \ldots, K\}$ **faire**

$$t_{ik}^{(q)} = p\left(c_k|\mathbf{x}_i; \boldsymbol{\psi}^{(q)}\right) = \frac{\pi_k^{(q)} \boldsymbol{\varphi}(\mathbf{x}_i; \boldsymbol{\mu}_k^{(q)}, \boldsymbol{\Sigma}_k^{(q)})}{\sum_{k'=1}^{K} \boldsymbol{\varphi}(\mathbf{x}_i; \boldsymbol{\mu}_{k'}^{(q)}, \boldsymbol{\Sigma}_{k'}^{(q)})}, \quad \forall i \in \{1, \ldots, N\}$$

 # Etape M

 # maximisation de la fonction auxiliaire

pour tous les $k \in \{1, \ldots, K\}$ **faire**

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

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{(q)}} \sum_{i=1}^{N} t_{ik}^{(q)} \mathbf{x}_i$$

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{(q)}} \sum_{i=1}^{N} t_{ik}^{(q)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})^t (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})$$

$q = q + 1$

Résultat : Paramètres estimés : $\widehat{\boldsymbol{\psi}}^{ml}$, probabilités a posteriori : t_{ik}

avec D une matrice diagonale et A une matrice orthogonale. Dans ce contexte α peut être interprété comme un volume, A comme une matrice d'orientation et D comme une « forme ». En imposant que certains de ces éléments soient communs aux différentes classes ou non, ou même fixé a priori, il est possible de définir un ensemble de modèles plus ou moins parcimonieux. Ce type d'approche peut être particulièrement intéressant pour éviter des problèmes d'instabilité lorsque les données vivent dans des espaces de dimensions importantes. D'un point de vue algorithmique les modèles de mélange parcimonieux peuvent être estimés à partir d'un algorithme EM très proche de l'algorithme utilisé dans le cadre du modèle générique, avec quelques modifications de l'étape de maximisation afin de prendre en compte les particularités de chaque modèle (Celeux et Govaert 1995).

Cette extension des modèles de mélange gaussien, n'est pas la seule qui puisse être rencontrée dans la littérature. Des approches bayésiennes peuvent par exemple être utilisées pour résoudre les problèmes d'instabilités de ces modèles dans les espaces de grande dimension (Fraley et Raftery 2002; 2007). Enfin, il est possible de coupler les modèles de mélange gaussien à un modèle de réduction de dimension, c'est le point de vue des mélanges d'analyse en composantes principales (Tipping et Bishop 1997a), et de la méthode HDDC (Bouveyron et al. 2007) ou bien encore des mélanges d'analyses en composantes principales robustes (Archambeau et al. 2008).

Remarque 2.1 (Nombre de classes et sélection de modèle) *Lorsqu'un modèle de mélange est utilisé pour analyser un jeu de données, il est nécessaire de déterminer le*

nombre de sous populations à représenter. Différentes situations peuvent être rencontrées ; il est tout d'abord possible que les classes correspondent à une réalité physique. Dans ce cas de figure, le nombre de sous populations est fixé a priori en utilisant cette information. Mais il est également possible qu'aucune information ne permettent de fixer le nombre de groupes a priori. Dans ce contexte, les critères pénalisant les modèles trop complexes tels que AIC, BIC, ICL ... sont utilisés pour déterminer un nombre de groupes optimal. Le choix d'un modèle parcimonieux particulier peut aussi être guidé par de tels critères, car, comme nous l'avons dit lors du chapitre introductif, ces critères en pénalisant les modèles complexes, permettent de trouver un compromis intéressant entre « attache au données » et « simplicité du modèle ».

Après cette description des modèles de mélange gaussien et de certaines de leurs extensions, nous nous intéressons maintenant aux différentes variantes ayant été proposés autour de l'algorithme EM dans le cadre des modèles de mélanges.

Extensions de l'algorithme EM dans le cadre des modèles de mélange

L'algorithme EM a donné naissance à de nombreuse variantes dans le contexte des modèles de mélanges. L'une d'entre elle présente un intérêt particulier car elle permet de faire le lien entre l'algorithme EM pour les modèle de mélange gaussien et les algorithmes de type k-means : c'est l'algorithme Classification EM (CEM), (Celeux et Govaert 1992). Dans cette variante une étape supplémentaire de classification est effectuée entre l'étape E et l'étape M. Cette étape transforme les probabilités a posteriori sur $\mathcal{Y}$ calculées lors de l'étape E pour tous les individus par des distributions certaines imposant une masse de 1 sur l'hypothèse la plus probable. Cet algorithme correspond lui aussi à un algorithme d'optimisation alternée, mais le critère utilisé n'est pas une vraisemblance ; il conduit donc à des estimés biaisés des paramètres du modèles. Cependant lorsque l'entropie de la distribution a posteriori sur les variables cachées est faible, ce biais reste peu important et cet algorithme présente l'avantage de converger plus vite que l'algorithme EM classique. De plus, celui-ci trouve une interprétation toute naturelle lorsque l'objectif n'est pas l'estimation des paramètres mais la recherche d'une classification optimale des données.

D'autres extensions se sont tournées vers le problème de la convergence locale de l'algorithme, c'est le cas de l'algorithme SEM développé par Celeux et Diebolt (1988). Dans celui-ci, une étape supplémentaire de simulation appelée étape S (pour stochastique) est rajoutée entre les étapes E et M. Dans cette étape, une classe est tirée pour chacun des individus en utilisant les probabilités a posteriori calculées dans l'étape E. Cette étape vise à empêcher l'algorithme de rester bloqué dans le bassin d'attraction d'un minimum local de la vraisemblance. D'autres algorithmes utilisent des stratégies différentes pour parvenir au maximum global, l'algorithme DA-EM, (Ueda et Nakano 1995), s'inspire par exemple de la méthode du recuit simulé. L'algorithme SM-EM, Ueda et al. (2000) effectue des mouvements aléatoires de type séparation, fusion entre composantes du mélange, afin d'améliorer la solution obtenue.

Nous allons maintenant, nous tourner vers d'autres modèles statistiques faisant intervenir des variables latentes désormais continues.

2.1.4 Les modèles à variables latentes continues gaussiennes

Les modèles de mélanges, que nous venons de présenter, supposaient l'existence de différentes sous-populations aux propriétés différentes pour décrire le jeu de données et modélisait l'appartenance à celles-ci par l'intermédiaire d'une variable latente discrète. Dans le cas où le système présente un continuum d'états, la notion de groupe ou classe s'avère mal adaptée et l'utilisation de variables latentes continues devient nécessaire. L'estimation de ces variables permet alors de mieux comprendre les données, de mieux les représenter et de mieux les analyser. Formellement, en supposant que les données ont été centrées, ces modèles dans leur version linéaire sont définis par l'équation suivante :

$$\mathbf{x} = A\mathbf{z} + \xi, \tag{2.30}$$

où ξ est une réalisation de Ξ un bruit indépendant de $\mathbf{z}$; la matrice A est une matrice de taille $P \times S$ qui lie les variables latentes représentées dans cette équation par le vecteur $\mathbf{z}$ de dimension S et les variables observées, représentées par le vecteur $\mathbf{x}$ de taille P.

L'objectif des modèles à variables latentes continues est d'estimer la matrice de mixage A, mais aussi de fournir un estimé des variables latentes $\mathbf{Z}$ qui sont elles aussi des inconnues dans ce problème, à partir des seules observations de X.

Ce problème semble a priori très difficile. Cependant, le nombre de variables latentes étant généralement inférieur au nombre de variables observées une certaine redondance doit exister dans les données. En exploitant cette redondance il doit être possible de remonter aux différentes causes « naturelles » de variations des données (à quelques indéterminations près), ou tout au moins de trouver une représentation plus compacte de celles-ci.

Comme les modèles de mélange, les modèles à variables latentes continues peuvent être interprétés d'un point de vue génératif. Les données sont alors supposées être issues du processus suivant : les variables latentes sont tout d'abord tirées suivant une distribution, les données étant ensuite obtenues en transformant linéairement ces variables et en ajoutant à ce résultat un bruit, qui le plus souvent sera supposé gaussien. La figure 2.7 représente le modèle graphique associé à ce processus génératif.

Pour rendre les choses plus concrètes l'exemple de la soirée cocktail peut être utilisé : supposons que P micros sont placés à différents endroits d'une pièce où S personnes sont en discussion. L'enregistrement obtenu à chaque instant par chacun des micros contient un mélange des voix des différentes personnes présentes. Ce mélange peut être supposé linéaire, les coefficients du mélange étant alors directement reliés aux caractéristiques physiques de la pièce dans laquelle a lieu l'expérience et aux positions des micros et locuteurs dans la pièce. Dans cet exemple, l'objectif des modèles à variables latentes continues est de retrouver les différents signaux de paroles des différents locuteurs dans les enregistrements (Haykin et Chen 2005). Les exemples d'application où les modèles de ce type peuvent être

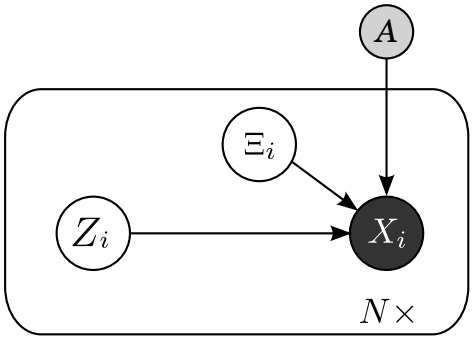


FIG. 2.7 – *Modèle graphique de génération des données d'un modèle à variable latentes continues bruité*

utiles sont nombreux : l'enregistrement de propriétés atmosphériques en différentes localisations peut par exemple tirer partie de telles méthodes pour parvenir à séparer et à estimer l'influence de différents foyers de pollutions (Khlaifi et al. 2005). Dans le cadre du traitement d'images médicales (Vigàrio 1997, Makeig et al. 1997, Biswal et Ulmer 1999) ces modèles sont aussi couramment utilisés, ainsi qu'en économie (Kiviluoto et Oja 1998, Back et Weigend 1997). Nous verrons également dans le dernier chapitre de cette thèse comment un modèle de ce type peut servir à mettre en place un système de diagnostic d'un système multi-composants.

Dans le cadre des modèles à variables latentes continues, il est important de noter que certaines indéterminations sont inévitables. La première indétermination concerne l'échelle des variables latentes. En observant l'équation (2.30), il apparait clairement qu'un changement d'échelle d'une variable latente peut être compensé en divisant la colonne de A correspondante par le facteur d'échelle utilisé. Tous les modèles à variables latentes ne permettent donc d'estimer celles-ci qu'à un facteur d'échelle près. Pour résoudre ce problème, il est courant de contraindre la variance des variables latentes à 1. La deuxième indétermination qui affecte les modèles à variables latentes continues concerne l'ordre de celles-ci ; en effet, il est possible de les permuter tout en conservant le même résultat, si les colonnes de A sont permutées elles aussi.

Les méthodes que nous allons présenter ici diffèrent par les hypothèses qu'elles postulent quant à la distribution des variables latentes. Les principales d'entre elles sont l'analyse factorielle (Factor Analysis en anglais, FA) et l'analyse en composantes principales (ACP) qui postulent des variables latentes gaussiennes et l'analyse en composantes indépendantes (ACI) qui ne fait pas l'hypothèse de normalité des variables latentes et utilise une hypothèse plus forte : l'indépendance de celles-ci. Nous présentons tout d'abord dans cette section l'analyse factorielle et l'ACP probabiliste. L'analyse en composantes indépendantes fera l'objet de la section suivante.

L'analyse factorielle (FA)

L'analyse factorielle est le premier modèle statistique à variables latentes continues qui a été introduit. Cette méthode a principalement été développée dans le

cadre de la psychométrie, (Spearman 1904, Thurstone 1947). Dans ce contexte, l'objectif est d'estimer une ou plusieurs variables non mesurables car mal définies, telle que l'intelligence, aux travers de leurs influences sur des variables observables, par exemple les résultats obtenus sur un panel d'exercices.

Pour extraire de telles variables, l'analyse factorielle repose sur des hypothèses fortes : les P variables observées sont supposées dépendre de manière linéaire d'un ensemble de S causes regroupées dans le vecteur $\mathbf{z} \in \mathbb{R}^S$ lesquelles sont sous-jacentes aux variations observées et distribuées suivant une loi normale multivariée. De plus, le nombre de causes est supposé inférieur au nombre de variables observées ($S < P$) ce qui permet à ce modèle de proposer une représentation plus parcimonieuse des données et plus particulièrement de la structure de la matrice de variance-covariance empirique. Ce type de modèle a trouvé de nombreuses applications en sciences sociales (Bartholomew et Martin 1999) son domaine d'origine mais aussi dans d'autres domaines. Pour spécifier entièrement le modèle défini par l'équation (2.30), les hypothèses suivantes sont faites dans le cadre de l'analyse factorielle :

$$\begin{aligned} Z &\sim \mathcal{N}(0, \mathbf{I}) \\ \Xi &\sim \mathcal{N}(0, \mathbf{\Lambda}), \quad \mathbf{\Lambda} \text{ diagonale} \\ Z &\perp\!\!\!\perp \Xi. \end{aligned}$$

La première hypothèse implique une décorrélation des variables latentes, ce qui est souhaitable pour en faciliter l'interprétation. Cette hypothèse est équivalente à l'hypothèse d'indépendance pour des variables gaussiennes. Le choix d'une matrice de variance-covariance égale à l'identité permet également de lever le problème de l'indétermination du modèle par rapport aux changements d'échelle. Enfin, il est intéressant de noter que dans ce modèle, où $\mathbf{\Lambda}$ est diagonale, les variables observées sont indépendantes conditionnellement aux variables latentes : $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_P | Z$. Les variables latentes servent donc à expliquer les corrélations entre les différentes variables observées, tandis que le bruit modélise leurs variabilités intrinsèques.

De plus cette contrainte de diagonalité est nécessaire faute de quoi le maximum de vraisemblance serait atteint par une solution triviale. Pour mettre en évidence ce fait, il suffit d'observer la loi des variables observées qui est donnée par :

$$X \sim \mathcal{N}(0, AA^t + \mathbf{\Lambda}). \tag{2.31}$$

L'analyse factorielle peut ainsi être vue comme un modèle gaussien avec une matrice de variance-covariance de forme particulière. Si la matrice de variance-covariance du bruit $\mathbf{\Lambda}$ n'est pas contrainte il est possible de poser :

$$\begin{aligned} \mathbf{\Lambda} &= \frac{1}{N}\mathbf{X}^t\mathbf{X} \\ A &= 0, \end{aligned} \tag{2.32}$$

ce qui permet d'atteindre le maximum de vraisemblance mais ne présente pas d'intérêt car dans ce cas toutes les variations des données sont expliquées par le bruit.

L'estimation des paramètres du modèle, c'est-à-dire l'estimation des matrices A et $\mathbf{\Lambda}$ peut être effectuée à l'aide d'un algorithme de type EM, (Rubin et Thayer

1982), qui repose sur la maximisation de l'espérance conditionnelle de la log-vraisemblance des données complétées : $(\mathbf{x}_1, \mathbf{z}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)$. Cette vraisemblance est égale à :

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{Z}) &= -\frac{N}{2}\log(|\det(\boldsymbol{\Lambda})|) - \frac{1}{2}\sum_{i=1}^{N}\mathbf{z}_i^t\mathbf{z}_i - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - A\mathbf{z}_i)^t\boldsymbol{\Lambda}^{-1}(\mathbf{x}_i - A\mathbf{z}_i) \\ &= -\frac{N}{2}\log(|\det(\boldsymbol{\Lambda})|) - \frac{N}{2}\mathrm{tr}(V\boldsymbol{\Lambda}^{-1}), \end{aligned} \tag{2.33}$$

avec $V = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - A.\mathbf{z}_i)(\mathbf{x}_i - A.\mathbf{z}_i)^t$.

L'espérance conditionnelle de celle-ci, par rapport à un jeu de paramètres courant et aux données observées, permet de définir la fonction auxiliaire maximisée lors de l'étape M de l'algorithme. Soit :

$$\begin{aligned}Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \mathbb{E}[\mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{Z})|\mathbf{X}, \boldsymbol{\psi}^{(q)}] \\ &= -\frac{N}{2}\log(|\det(\boldsymbol{\Lambda})|) - \frac{N}{2}\mathrm{tr}(\mathbb{E}[V|\mathbf{X}, \boldsymbol{\psi}^{(q)}]\boldsymbol{\Lambda}^{-1}). \end{aligned} \tag{2.34}$$

Cette fonction dépend de deux statistiques qui doivent être calculées lors de l'étape E pour tous les individus :

$$\widehat{\mathbf{z}_i} = \mathbb{E}[Z|X = \mathbf{x}_i, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}] = A^t(AA^t + \boldsymbol{\Lambda})^{-1}\mathbf{x}_i \tag{2.35}$$

$$\widehat{\mathbf{z}_i\mathbf{z}_i^t} = \mathbb{E}[ZZ^t|X = \mathbf{x}_i, \boldsymbol{\psi} = \boldsymbol{\psi}^{(q)}] = (\mathbf{I} - A^t\boldsymbol{\Lambda}^{-1}A)^{-1} + \widehat{\mathbf{z}_i}\widehat{\mathbf{z}_i}^t. \tag{2.36}$$

L'étape M de l'algorithme correspond quant à elle à la maximisation de (2.34) par rapport aux paramètres. Une solution analytique peut être trouvée à cette maximisation. La matrice de mélange est pour cela simplement mise à jour en effectuant une régression linéaire multiple, la variance du bruit étant quant à elle estimée à partir des résidus de cette régression. L'algorithme 3 récapitule ces différents éléments. Il est intéressant de noter que, comme dans le cas des modèles de mélange, l'interprétation des différentes étapes de cet algorithme est aisée.

L'une des principales critiques faites à l'encontre de l'analyse factorielle est son indétermination par rapport à une multiplication de la matrice A par une matrice orthogonale R, c'est-à-dire par rapport aux rotations. En effet, dans ce cas de figure la distribution des données observées s'écrit :

$$\begin{aligned}X &\sim \mathcal{N}(0, ARR^tA^t + \boldsymbol{\Lambda}) \\ X &\sim \mathcal{N}(0, AA^t + \boldsymbol{\Lambda}).\end{aligned}$$

Les vraisemblances de A et de AR sont donc identiques si R est orthogonale. Le problème d'estimation associé à l'analyse factorielle possède donc une infinité de solutions. Pour résoudre ce problème et proposer une solution unique à l'utilisateur, différentes approches ont été proposées, elles visent généralement à trouver une rotation menant à une structure simple pour la matrice A. La méthode varimax (Kaiser 1958), est la solution de ce type la plus utilisée. Il est important de remarquer que cette indétermination découle directement de l'utilisation d'une distribution gaussienne pour modéliser les variables latentes. Nous verrons lorsque nous aborderons l'analyse en composantes indépendantes que cette indétermination peut être levée lorsque l'hypothèse de normalité des variables latentes n'est pas faite.

Algorithme 3: pseudo-code de l'algorithme EM pour l'analyse factorielle.

Données : Matrice des données : $\mathbf{X}$

 # Initialisation

$\boldsymbol{\psi}^{(0)} = \left(A^{(0)}, \mathbf{\Lambda}^{(0)}\right), q = 0$

tant que *test de convergence* **faire**

 # Etape E

 # Estimation de l'espérance des variables latentes

$\widehat{\mathbf{z}_i} = A^t(AA^t + \mathbf{\Lambda})^{-1}\mathbf{x}_i, \quad \forall i \in \{1, \ldots, N\}$

 # Estimation de la variance-covariance des variables latentes

$\widehat{\mathbf{z}_i\mathbf{z}_i^t} = (\mathbf{I} - A^t\mathbf{\Lambda}^{-1}A)^{-1} + \widehat{\mathbf{z}}_i\widehat{\mathbf{z}}_i{}^t, \quad \forall i \in \{1, \ldots, N\}$

 # Etape M

 # maximisation de la fonction auxiliaire

$A^{(q+1)} = \left(\sum_{i=1}^N \mathbf{x}_i\widehat{\mathbf{z}}_i{}^t\right)\left(\sum_{i=1}^N \widehat{\mathbf{z}_i\mathbf{z}_i^t}\right)$

$\mathbf{\Lambda}^{(q+1)} = diag\left(\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^t - A^{(q+1)}\widehat{\mathbf{z}}_i\mathbf{x}_i^t\right)$

$q = q + 1$

Résultat : Paramètres estimés : $\widehat{\boldsymbol{\psi}}^{ml}$, variables latentes estimées : $\widehat{\mathbf{Z}}^{ml}$

L'analyse en composantes principales (ACP)

L'origine de l'analyse en composantes principales peut être trouvée dans les travaux précurseurs de Pearson (1901) pour le cas bi-dimensionnel et dans ceux de Hotelling (1933) qui a étendu cette méthode pour une dimensions quelconque. C'est une méthode de réduction de la dimension très utilisée. L'objectif classique est de trouver une projection des données sur un sous-espace qui conserve le maximum d'information. Cette méthode possède un statut particulier dans le cadre des modèles à variables latentes car les composantes ne sont généralement pas assimilées à des causes naturelles de variation, l'objectif est plus humblement de trouver une représentation plus compacte des données.

La popularité de cette méthode provient de différentes propriétés. Tout d'abord, elle est doublement optimale : elle maximise la variance expliquée (Hotelling 1933), ce qui est équivalent à dire qu'elle minimise l'erreur de reconstruction après réduction de dimension (cf. figure 2.8).

La présentation classique de cette méthode repose sur la résolution d'un de ces deux problèmes d'optimisation. Prenons par exemple celui du maximum de variance expliquée. Dans ce cas l'objectif est de trouver une direction w telle que la combinaison linéaire $\mathbf{X}$.w soit de variance maximale. En supposant que les données $\mathbf{X}$ sont centrées, ce problème d'optimisation peut être écrit sous la forme ma-

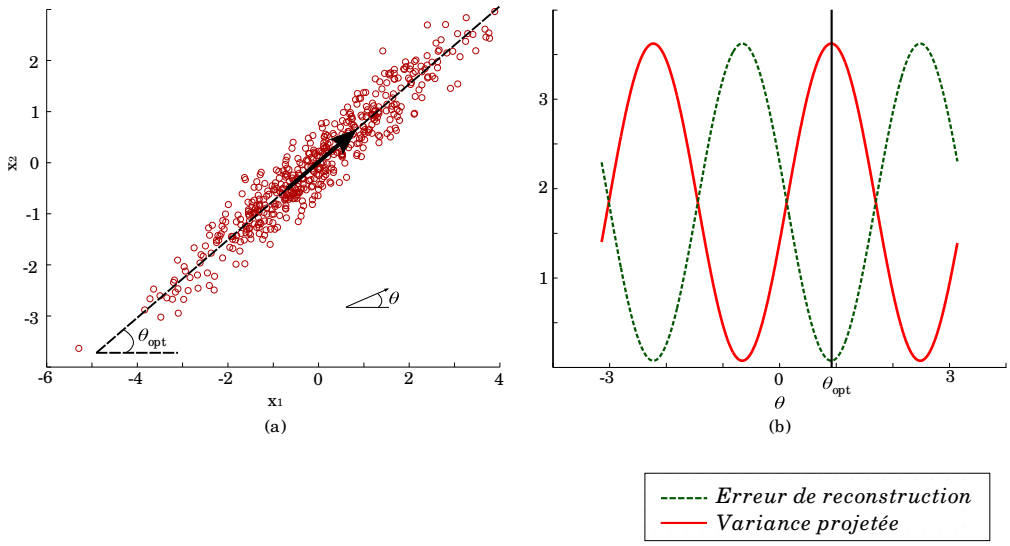


FIG. 2.8 – *(a) Nuage en dimension 2 simulé suivant une loi normale ; la première composante principale de ce jeu de donnée est représentée à l'aide d'une flèche, (b) erreur de reconstruction quadratique moyenne et variance projetée en fonction de θ, lorsque $\mathbf{w} = (sin(\theta),\ cos(\theta))^t$. Le minimum de l'erreur de reconstruction coïncide bien avec le maximum de la variance projetée.*

tricielle suivante :

$$\arg\max_{\mathbf{w}} \frac{1}{N} \frac{\mathbf{w}^t \mathbf{X}^t \mathbf{X} \mathbf{w}}{\mathbf{w}^t \mathbf{w}}, \text{avec } ||\mathbf{w}||^2 = 1. \tag{2.37}$$

quotient de Rayleigh

Ce type de problème d'optimisation associé au quotient de deux formes quadratique (ici $\mathbf{X}^t\mathbf{X}$ et $\mathbf{I}$), appelé quotient de Rayleigh (Borga et al. 1995), possède une solution particulière. Il s'agit de la valeur propre dominante de la matrice :

$$\frac{1}{N}\mathbf{I}^{-1}.\mathbf{X}^t.\mathbf{X} = \frac{1}{N}\mathbf{X}^t.\mathbf{X}, \tag{2.38}$$

valeurs propres

qui n'est autre que la matrice de variance-covariance des données. La direction solution du problème d'optimisation (2.37) est donc le vecteur propre dominant de la matrice de variance-covariance empirique. De plus, la valeur propre associée à cette direction est égale à la variance expliquée par ce premier axe principal. L'analyse en composantes principales utilise ce résultat pour construire un sous espace de rang $S < p$ qui soit optimal. En effet, la recherche de la deuxième direction de variance maximale dans l'espace orthogonal au premier axe trouvé donne, par un raisonnement similaire, le second vecteur propre de la matrice de variance-covariance empirique. Trouver le sous espace de dimension S qui maximise la variance projetée est donc équivalent à trouver les S premiers couples de (valeurs, vecteurs) propres de la matrice de variance-covariance empirique. La solution de ce problème est donc analytique. Les composantes principales peuvent êtres déterminées directement à partir des données sans avoir recourt à un algorithme d'optimisation itératif, c'est la seconde propriété très intéressante de cette méthode. Finalement, cette méthode étant linéaire, les opérations de projections (compression) sur les composantes principales, ou inversement de reconstruction (décompression) sont aisées car elles ne nécessitent que des produits matriciels. Après cette présentation classique, de l'analyse en composantes principales nous

développons les liens qu'entretient cette méthode avec l'analyse factorielle. Nous présentons pour cela le modèle à variables latentes continues qui lui est associé.

Remarque 2.2 (Quotient de rayleigh et analyse de données) *De nombreuses méthodes d'analyse de données telle que l'analyse canonique des corrélations, l'analyse factorielle discriminante,... peuvent être présentées dans des formulations similaires, en utilisant un quotient de Rayleigh, (Borga et al. 1995).*

Le modèle à variables latentes associé à l'ACP

Différents travaux (Tipping et Bishop 1997b, Roweis et Ghahramani 1997) ont permis de mettre en exergue le lien fort qui unit l'ACP à l'analyse factorielle, et ont permis de décrire cette méthode comme un modèle statistique à variables latentes, avec comme conséquence directe la possibilité d'utiliser un algorithme de type EM pour en estimer les paramètres. Cet algorithme est particulièrement intéressant du point de vue de la complexité, lorsque les données sont nombreuses, de grande dimension et quand seules quelques composantes principales doivent être extraites. De plus, une telle approche permet de traiter élégamment le problème des données manquantes. Pour ce faire, un modèle probabiliste est associé à l'analyse en composantes principales ; ce modèle est de forme identique à celui de l'analyse factorielle, les variables observées étant supposées être des transformations linéaires d'un ensemble restreint de variables latentes auxquelles un bruit additif gaussien indépendant est ajouté. Nous retrouvons donc le modèle défini par l'équation (2.30).

L'analyse en composantes principales se singularise cependant de l'analyse factorielle de par le modèle de bruit postulé. Dans le cadre de l'analyse factorielle, le bruit était supposé gaussien centré de matrice de variance-covariance diagonale alors qu'ici la matrice de variance-covariance est proportionnelle à la matrice identité, le bruit est donc supposé isotrope. Pour simplifier les notations, nous supposons que les données ont été centrées, ce qui évite d'alourdir le modèle d'un paramètre supplémentaire. Formellement le modèle est défini par les hypothèses suivantes :

$$\begin{aligned} Z &\sim \mathcal{N}(0, \mathbf{I}) \\ \Xi &\sim \mathcal{N}(0, \nu\mathbf{I}) \\ Z &\perp\!\!\!\perp \Xi. \end{aligned}$$

La loi des variables observées peut être déduite de ce modèle. Nous obtenons :

$$X \sim \mathcal{N}(0, AA^t + \nu\mathbf{I}). \tag{2.39}$$

Comme le montre les travaux de Tipping et Bishop (1997b), l'estimation des paramètres de ce modèle par maximum de vraisemblance conduit à une solution analytique, tout comme l'ACP, celle-ci fait intervenir les S premiers couples (valeurs propres, vecteurs propres) de la matrice de variance-covariance :

$$\hat{\nu} = \frac{1}{P-S} \sum_{j=S+1}^{P} \lambda_j \tag{2.40}$$

$$\hat{A} = V_s(L_s - \hat{\nu}\mathbf{I})^{1/2}R, \tag{2.41}$$

où V_s est la matrice contenant les vecteurs propres associées aux S premières valeurs propres de la matrice de variance-covariance empirique des données, L_s une matrice diagonale contenant ces S premières valeurs propres et R une matrice de rotation de taille $S \times S$ quelconque qui rend compte de l'indétermination du modèle par rapport aux rotations. Ce modèle, bien qu'il fournisse une solution très proche de l'ACP ne permet pas de retrouver exactement l'analyse en composantes principales, Il est nécessaire pour cela de contraindre la variance du bruit à tendre vers 0.

Après avoir présenté l'analyse factorielle et l'analyse en composantes principales nous nous tournons maintenant vers un autre modèle à variables latentes qui lui ne postule pas une forme gaussienne pour les variables latentes mais utilise l'hypothèse plus forte d'indépendance de ces différentes variables.

Remarque 2.3 (Indéterminations) *Tout comme l'analyse factorielle, l'analyse en composantes principales peut être considérée comme indéterminée par rapport aux rotations de la matrice de mixage ; cependant elle propose une solution « canonique » donnée par les vecteurs propres de la matrice de variance-covariance. De plus, le problème de l'ordre des variables latentes est lui aussi levé par ce biais, celles-ci pouvant être ordonnées suivant la variance projetée, c'est-à-dire suivant leur énergie.*

2.2 L'ANALYSE EN COMPOSANTES INDÉPENDANTES (ACI)

L'analyse en composantes indépendantes est apparue dans la communauté traitement du signal et analyse de données française dans les années 80 avec les travaux fondateurs de Hérault et al. (1985). Cette méthode à ensuite connue une diffusion importante au sein de la communauté internationale dans les années 90 avec des travaux tels que (Comon 1994, Bell et Sejnowski 1995, Cardoso 1997). Ces travaux ont permis une meilleure compréhension des bases théoriques de cette méthode et la mise au point d'algorithmes efficaces, (Amari et al. 1996, Cardoso et Laheld 1996, Cardoso 1999, Hyvärinen 1999). Différents ouvrages ont depuis lors été consacrés à cette méthode (Hyvärinen 2001, Roberts et Everson 2001, Jutten et Comon 2007a;b).

Remarque 2.4 (sources ou variables latentes) *Du fait de l'origine de l'analyse en composantes indépendantes dans la communauté traitement du signal et en référence au problème de la séparation aveugle de source, le terme de source est plus fréquemment utilisé que celui de variable latente. Nous emploierons indistinctement ces deux termes dans cette section.*

2.2.1 Principe

Le principe de base de cette approche est de rechercher des variables latentes qui soient indépendantes entre elles. Quelque soit le couple de variables latentes considéré, l'observation de l'une ne doit donc apporter aucune information sur la seconde. Cette hypothèse a un double effet : tout d'abord la solution obtenue n'est plus, comme dans l'analyse factorielle, indéterminée par rapport aux rotations de

la matrice de mixage, si les variables latentes ne sont pas gaussiennes. En contre partie, les statistiques d'ordre 1 et 2 qui seules étaient utilisées dans le cadre de l'analyse factorielle et de l'analyse en composantes principales, ne suffisent plus à déterminer la solution. D'autres statistiques d'ordres supérieurs doivent être prises en compte pour estimer les paramètres du modèle.

Exemple 2.5 (Illustration de l'ACI) :

> La figure 2.9 permet de visualiser, sur un problème jouet, l'apport de l'analyse en composantes indépendantes par rapport à une analyse en composantes principales. Nous présentons sur cette figure des données simulées suivant deux variables indépendantes correspondant chacune à un modèle de mélange gaussien à deux composantes, les densités de ces deux variables aléatoires sont également représentées sur la figure 2.9 (a). Ces mêmes données transformées linéairement (simulant ainsi un processus de mesure imparfaite couplant ces deux variables) sont ensuite représentés ainsi que les résultats d'une ACP et d'une ACI. Nous pouvons observer sur cette figure que l'ACP ne permet pas de retrouver les deux variables latentes ayant servi à la simulation, les densités marginales ne correspondent pas aux densités des variables latentes (indétermination par rapport aux rotations), contrairement à l'ACI qui permet de retrouver celles-ci aux indétermination du modèle près. La deuxième composante indépendante est très fortement corrélée avec Z_1 (coefficient de corrélation de -0.99), la première est quand à elle fortement corrélé avec Z_2 (coefficient de corrélation de -0.99), les densités marginales correspondent. Les variables latentes ont donc bien été retrouvées, aux facteurs d'échelles et aux permutations près. Cet exemple est bien sûr illustratif et ne rend pas compte de la complexité rencontrée dans les espace de grande dimension.

L'analyse en composantes indépendantes peut être présentée tout comme l'analyse en composantes principales, à partir de différents principes :

- minimisation de l'information mutuelle (Amari et al. 1996) ;
- maximisation de la vraisemblance (Moulines et al. 1997, Attias 1999) ;
- décorrélation non linéaire (Hérault et al. 1985, Bach et Jordan 2003) ;
- maximisation de l'écart à la normalité (Hyvärinen 2001, chap. 8) ;
- diagonalisation de tenseurs (Cardoso 1999) ;

Cependant, à l'inverse de l'analyse en composantes principales, ces différents principes peuvent conduire dans la pratique à différents algorithmes. Nous présentons dans ce paragraphe quelques uns de ces principes, et nous insistons sur les liens qui peuvent être établis entre eux. Mais avant cela, nous revenons sur le modèle de
bruit l'analyse en composantes indépendantes, en particulier sur le traitement du bruit au sein de celui-ci. Certaines méthodes (Moulines et al. 1997, Attias 1999, Ikeda 2000) s'appuient sur le modèle défini par l'équation (2.30) ; cependant le modèle le plus utilisé est légèrement plus simple. Celui-ci suppose une relation déterministe entre variables latentes et variables observées avec de plus un nombre identique de variables observées et latentes. Le modèle est alors de la forme :

$$\mathbf{x} = A\,\mathbf{z}, \tag{2.42}$$

avec A de taille $S \times S$. Le bruit, n'est cependant pas oublié dans cette modélisation et différentes solutions existent pour prendre celui-ci en compte. Il est tout d'abord

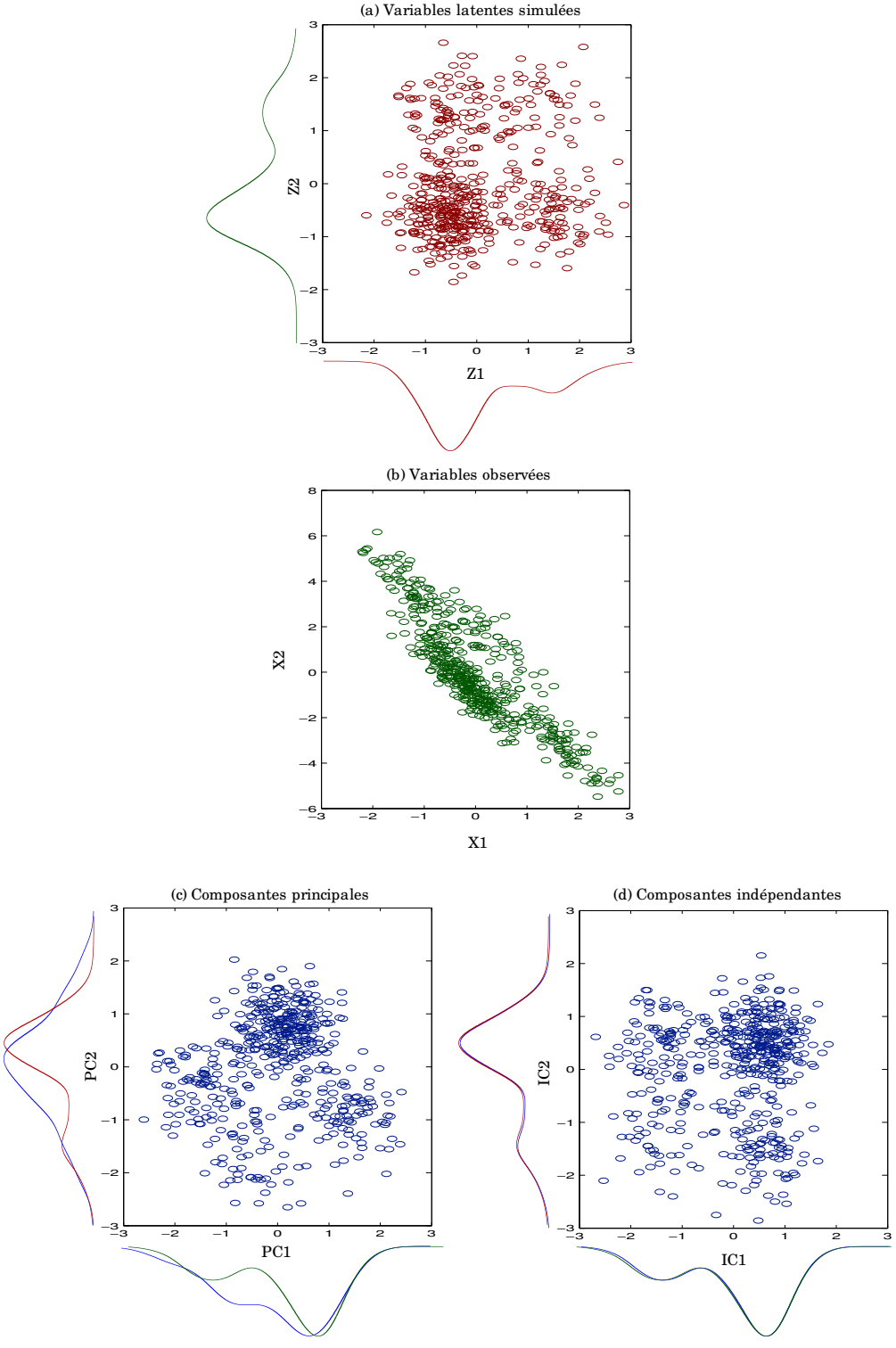


FIG. 2.9 – *Comparaison de l'analyse en composantes principales et de l'analyse en composantes indépendantes : (a) nuage de 500 points correspondant à des réalisations de deux variables aléatoires indépendantes suivant chacune un modèle de mélange gaussien à 2 composantes et densités de ces deux variables, (b) nuage de point correspondant à une transformation linéaire de ces données par une matrice de mixage aléatoire, (c) projection sur les deux axes principaux normalisés et densités marginales sur chacun des axes, (d) projection sur les deux composantes indépendantes normalisés et densités marginales sur ces deux composantes.*

possible d'utiliser une analyse en composantes principales pour séparer l'information utile du bruit, en ne conservant que les S premières composantes principales significatives. Une autre solution consiste à extraire autant de composantes indépendantes que de variables observées et d'identifier ensuite certaines d'entre elles comme des composantes de bruit. L'intérêt du modèle défini par l'équation (2.42) est tout d'abord que si la matrice A est non-singulière, l'équation peut être inversée :

$$\mathbf{z} = A^{-1}\,\mathbf{x} = W\,\mathbf{x}. \tag{2.43}$$

Ceci conduit à rechercher non plus la matrice de mixage A, mais la matrice de démixage W qui, à partir des données observées, permet d'obtenir les variables latentes. De plus, il est aussi possible d'établir une relation déterministe entre la distribution de Z et la distribution de X, comme le montre le théorème suivant.

Théorème 2.1 (Densité d'une transformation linéaire) *Soient deux variables aléatoires telles que* $\mathbf{x} = A.\mathbf{z}$*, avec* A *une matrice inversible. Il existe alors une relation déterministe entre la densité de* $\mathbf{x}$, $f^{\mathcal{X}}$ *et celle de* $\mathbf{z}$, $g^{\mathcal{Z}}$ *qui est donnée par :*

$$f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(A)|} g^{\mathcal{Z}}(A^{-1}\,\mathbf{x}), \tag{2.44}$$

(cf. annexe .4).

Cette propriété permet, par exemple, d'établir facilement la forme de la vraisemblance de ce modèle et simplifie grandement le problème d'estimation de la matrice de mixage et des sources. Nous allons voir maintenant comment, en utilisant ce modèle, le problème de l'analyse en composantes indépendantes peut être formaliser dans le cadre de la théorie de l'information.

2.2.2 ACI et théorie de l'information

La théorie de l'information introduite par Shannon (1948) a conduit à la définition de différents concepts dont celui d'information mutuelle. Ce dernier permet de mesurer la dépendance entre un ensemble de variables et offre donc des opportunités pour construire différents critères efficaces pour effectuer une analyse en composantes indépendantes.

entropie La théorie de l'information s'appuie sur le concept d'entropie qui mesure le degré d'incertitude quant au résultat d'une expérience aléatoire. Lorsque la variable aléatoire définissant cette expérience est définie sur un référentiel continu, l'entropie est nommée entropie différentielle et est définie par :

Définition 2.3 (Entropie différentielle) *L'entropie d'une variable aléatoire* Z *définie sur* $\mathcal{Z}$*, de densité* $f^{\mathcal{Z}}(\mathbf{z})$*, est donnée par :*

$$H(Z) = -\mathbb{E}[\log(f^{\mathcal{Z}}(Z))] = -\int_{\mathcal{Z}} f^{\mathcal{Z}}(\mathbf{z}) \log(f^{\mathcal{Z}}(\mathbf{z})).\mathrm{d}\mathbf{z}\,. \tag{2.45}$$

Cette définition permet de dériver différentes propriétés intéressantes. En particulier, de par l'utilisation de la fonction logarithme qui transforme les produits en sommes, cette mesure est additive pour des variables aléatoires indépendantes :

Propriété 2.1 (Entropie et indépendance) *Soit Z_1 et Z_2 deux variables aléatoires indépendantes : $Z_1 \perp\!\!\!\perp Z_2$, et soit $Z = (Z_1, Z_2)$ la variable aléatoire jointe ; nous avons alors :*

$$H(Z) = H(Z_1) + H(Z_2)\,. \tag{2.46}$$

La divergence de Kullback-Leibler est un autre outil de la théorie de l'information qui permet de quantifier l'écart entre deux densités de probabilités et qui nous permettra de définir le concept d'information mutuelle.

Définition 2.4 (Divergence de Kullback-Leibler) *Soit, deux densités de probabilité $f^{\mathcal{Z}}(\mathbf{z})$ et $g^{\mathcal{Z}}(\mathbf{z})$ définies sur un même espace $\mathcal{Z}$; la divergence de Kullback-Leibler entre ces deux densités est données par :*

$$KL(f^{\mathcal{Z}}||g^{\mathcal{Z}}) = \int_{\mathcal{Z}} f^{\mathcal{Z}}(\mathbf{z}) \log\left(\frac{f^{\mathcal{Z}}(\mathbf{z})}{g^{\mathcal{Z}}(\mathbf{z})}\right).\mathrm{d}\mathbf{z}\,. \tag{2.47}$$

information mutuelle

L'information mutuelle d'un ensemble de variables est définie comme la divergence entre leur loi jointe et la loi construite en multipliant les marginales, ce qui permet de mesurer la redondance entre ces différentes variables.

Définition 2.5 (Information mutuelle) *Soit une variable aléatoire Z définie sur un espace produit $\mathcal{Z} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_S$. En notant la densité de probabilité jointe sur $\mathcal{Z}$: $f^{\mathcal{Z}}$ et $f^{\mathcal{Z}_s}, s \in \{1, \ldots, S\}$ les densités marginales, l'information mutuelle de cette variable aléatoire est donnée par :*

$$IM(Z) = KL(f^{\mathcal{Z}}||\prod_{s=1}^{S} f^{\mathcal{Z}_s})\,. \tag{2.48}$$

L'information mutuelle peut aussi être définie grâce à l'entropie :

$$IM(Z) = \sum_{s=1}^{S} H(Z_s) - H(Z)\,. \tag{2.49}$$

L'information mutuelle mesure donc l'écart entre une densité de probabilité jointe et la densité construite en multipliant ses densités marginales, c'est-à-dire la densité jointe lorsque l'hypothèse d'indépendance est faite. Elle possède des propriétés très intéressantes qui découlent des définitions (2.4,2.5), en particulier :

Propriété 2.2 (Information mutuelle et indépendance)

$$IM(Z) \;\geq\; 0, \quad \forall Z \tag{2.50}$$

$$IM(Z) = 0 \;\Leftrightarrow\; Z_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp Z_S\,. \tag{2.51}$$

L'information mutuelle permet donc de mesurer l'indépendance entre différentes variables aléatoires et fournit un principe d'estimation dans le contexte de l'analyse en composantes indépendantes. Celui-ci correspond à rechercher la matrice de démixage W telle que :

$$\widehat{W} = \arg\min_{W} \widehat{IM}(W.\mathbf{X}^t), \tag{2.52}$$

où $\widehat{IM}(W.\mathbf{X}^t)$ est un estimé de l'information mutuelle des variables latentes. Pour obtenir un critère implémentable, à partir de ce principe l'information mutuelle

des variables latentes est tout d'abord reformulée en prenant en considération les particularités du modèle. En effet, dans le cadre des modèles de la forme (2.43), il est possible, en utilisant l'équation (2.44) reliant la densité de Z à la densité de X et les définitions données précédemment, d'exprimer l'information mutuelle sous la forme suivante :

$$IM(Z) = \sum_{s=1}^{S} H(Z_s) - H(X) - \log(|\det(W)|), \tag{2.53}$$

En prenant en considération le fait que les variables latentes sont décorrélées ($\mathbb{E}[ZZ^t] = \mathbf{I}$), puisqu'elles sont supposées indépendantes, on peut écrire :

$$IM(Z) = \sum_{s=1}^{S} H(Z_s) - H(X) + cste, \tag{2.54}$$

car

$$\det(W.\Sigma_x.W^t) = \det(W).\det(\Sigma_x).\det(W^t) = \det(\mathbf{I}) = 1 \tag{2.55}$$

$$\Rightarrow \det(W) = cste\ . \tag{2.56}$$

La relation (2.54) montre que les variations de l'information mutuelle en fonction de la matrice de démixage W choisie, dépendent uniquement des variations des entropies marginales, car $H(X)$ ne dépend pas de la matrice de démixage. Cette réécriture permet de mettre en place différents algorithmes pour minimiser l'information mutuelle par rapport à cette matrice.

L'information mutuelle écrite sous cette forme fait intervenir les densités des variables latentes au travers de l'entropie de chacune d'entre elles ; or celles-ci sont inconnues. Pour parvenir à un critère utilisable en pratique, il est donc nécessaire d'estimer ces densités. Différentes solutions ont été proposées dans la littérature pour cela, certaines d'entre elles utilisent un modèle non paramétrique. Les densités marginales peuvent alors être estimées par une méthode non paramétrique de type noyaux ; l'information mutuelle en est ensuite déduite à l'aide des relations (2.54,2.3), (Jutten et Comon 2007b, chap. 2). Ces méthodes souffrent d'une complexité importante et d'un manque de robustesse.

D'autres solutions utilisent des estimations de l'entropie basée sur les cumulants, c'est-à-dire sur les statistiques d'ordre supérieurs. Les développements en séries de Gram-Charlier permettent d'approximer des densités grâce aux cumulants et permettent donc de construire dans le contexte de l'analyse en composante indépendantes des estimateurs extrêmement simples de l'entropie, (Hyvärinen 2001, p. 113-115). En remplaçant dans cette approximation les cumulants par leurs estimés empiriques, il est possible de construire un critère simple mesurant l'indépendance d'un ensemble de variables.

maximum de non gaussianité

Remarque 2.5 (Maximisation de l'écart à la normalité) *Cette réécriture de l'information mutuelle permet aussi de mettre en évidence la relation entre cette approche et des méthodes plus anciennes tel que la recherche de directions maximalement non gaussiennes (Friedman et Tukey 1974). En effet, l'entropie d'une variable aléatoire centrée réduite est maximale lorsque cette variable aléatoire suit une loi normale, Minimiser l'entropie c'est donc en quelque sorte maximiser l'écart à cette distribution de référence. Cette relation permet d'éclairer les relations existantes entre*

l'analyse en composantes indépendantes et la méthode des projections révélatrices citée précédemment, qui recherche elle aussi des directions de projections permettant de révéler des structures non gaussiennes. La différence majeure entre ces deux approches réside dans la nature séquentielle de l'approche de type poursuite, les composantes sont en effet dans ce cas extraites une à une, alors qu'elles sont extraites simultanément dans le cadre de l'analyse en composantes indépendantes.

Nous allons voir finalement, que l'information mutuelle peut aussi être mise en relation avec le principe du maximum de vraisemblance.

2.2.3 IFA et maximum de vraisemblance

Si l'on considère connues les densités des variables latentes, il est possible d'approximer l'information mutuelle par :

$$\widehat{IM}(Z) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{S}\log\left(f^{\mathcal{Z}_s}\left((W\mathbf{x}_i)_s\right)\right) + \log(|det(W)| + cste\ . \tag{2.57}$$

Or la vraisemblance du modèle de l'analyse en composantes indépendantes s'écrit en utilisant (2.44) et l'hypothèse d'indépendance des variables latentes :

$$L(W;\mathbf{X}) = \prod_{i=1}^{N}|\det(W)|\left(\prod_{s=1}^{S}f^{\mathcal{Z}_s}\left((W\mathbf{x}_i)_s\right)\right). \tag{2.58}$$

En passant au logarithme, nous obtenons :

$$\mathcal{L}(W;\mathbf{X}) = \sum_{i=1}^{N}\sum_{s=1}^{S}\log\left(f^{\mathcal{Z}_s}\left((W\mathbf{x}_i)_s\right)\right) + N\log(\det(W)), \tag{2.59}$$

ce qui en divisant par N donne une forme identique à (2.57) au signe près. La minimisation de l'information mutuelle des différentes variables latentes et la maximisation de la vraisemblance sont donc équivalentes lorsque les densités des sources sont fixées.

Cependant, les algorithmes construits à partir de ces deux principes sont généralement différents. Lorsque le critère à optimiser est dérivé de l'information mutuelle, des estimations non paramétriques des différentes densités marginales sont utilisées. Dans le contexte d'une estimation par maximum de vraisemblance, il est par contre logique de postuler une forme semi-paramétrique pour chacune d'entre elles et d'intégrer ces paramètres supplémentaires au problème d'estimation.

Cette approche, introduite dans (Moulines et al. 1997, Attias 1999), propose de modéliser les densités marginales à l'aide de modèles de mélange. Cette spécialisation de l'ACI a été dénommée analyse en facteurs indépendants par Attias (Independant Factor Analysis en anglais, IFA). Chaque densité marginale prend alors la forme suivante :

$$f^{\mathcal{Z}_s}(z_s) = \sum_{k=1}^{K_s}\pi_k^s\varphi(z_s;\mu_k^s,\nu_k^s), \tag{2.60}$$

Le vecteur des paramètres du modèle est donc complété des paramètres supplémentaires servant à modéliser les différentes sources :

$$\boldsymbol{\psi} = (W, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^S, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^S, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^S), \tag{2.61}$$

et la log-vraisemblance du modèle devient :

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}) = N \log(|\det(W)|) + \sum_{i=1}^{N} \sum_{s=1}^{S} \log \left(\sum_{k=1}^{K_s} \pi_k^s \varphi \left((W\mathbf{x}_i)_s, \mu_k^s, \nu_k^s \right) \right). \tag{2.62}$$

gradient naturel

Pour estimer les différents paramètres du modèle par maximum de vraisemblance, il est possible de se tourner vers une stratégie d'optimisation alternée, le problème s'y prêtant bien. En effet, l'algorithme du gradient naturel permet d'optimiser la log-vraisemblance par rapport à W lorsque les paramètres des densités marginales sont fixés. Inversement, lorsque la matrice de démixage est gelée, l'algorithme EM permet de maximiser cette vraisemblance par rapport aux paramètres des densités des sources. Ces constatations conduisent naturellement à la mise au point d'un algorithme GEM (cf. algorithme 4) pour maximiser conjointement la vraisemblance par rapport à tous les paramètres. Celui-ci alterne simplement la mise à jour des sources à l'aide de l'algorithme EM et une montée de gradient pour optimiser la log-vraisemblance par rapport à la matrice de démixage.

Nous donnons tout d'abord quelques éléments en ce qui concerne l'estimation de la matrice de démixage lorsque les densités sont fixées. Les solutions classiques s'appuient pour cela sur l'algorithme du gradient naturel (Amari et al. 1996). Cet algorithme est un algorithme de type gradient et utilise donc la dérivée de la log-vraisemblance par rapport à la matrice démixage. Celle-ci est donnée par :

$$\frac{\partial \mathcal{L}(W; \mathbf{X})}{\partial W} \propto (W^{-1})^t - \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(W\mathbf{x}_i)\mathbf{x}_i^t, \tag{2.63}$$

(cf. annexe .2) avec :

$$\mathbf{g}(\mathbf{z}) = \left[\frac{-\partial \log(f^{\mathcal{Z}_1}(z_1))}{\partial z_1}, \ldots, \frac{-\partial \log(f^{\mathcal{Z}_S}(z_S))}{\partial z_S} \right]^t. \tag{2.64}$$

Lorsque les densités des différentes variables latentes sont fixées, il est donc possible d'utiliser ce gradient pour maximiser la vraisemblance par rapport à W. La règle de mise à jour de la matrice de démixage étant alors simplement donnée par :

$$W^{(q+1)} = W^{(q)} + \tau \left(\left((W^{(q)})^{-1} \right)^t - \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(W^{(q)}\mathbf{x}_i)\mathbf{x}_i^t \right), \tag{2.65}$$

où τ est le pas de gradient qui doit être ajusté, ce qui peut être fait manuellement ou en utilisant des méthodes de recherche linéaire (cf. annexe .5).

Dans la pratique, l'algorithme du gradient naturel est cependant préféré (Amari et al. 1996, MacKay 1996). Cet algorithme tire partie de la structure particulière du problème d'optimisation pour trouver une direction de descente de meilleure qualité. Celle-ci est obtenue en multipliant le gradient à droite par W^tW, ce qui conduit à la règle de mise à jour suivante pour la matrice de démixage :

$$W^{(q+1)} = W^{(q)} + \tau \left(\mathbf{I} - \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(\mathbf{z}_i^{(q)})\mathbf{z}_i^{(q)t} \right) W^{(q)}, \tag{2.66}$$

avec $\mathbf{z}_i^{(q)} = W^{(q)}\mathbf{x}_i$.

Remarque 2.6 (Décorrélation non linéaire) *Cette équation de mise à jour fait clairement apparaître une propriété intéressante. En effet, lorsque la convergence est atteinte, nous pouvons observer que :*

$$\begin{aligned} \left(\mathbf{I} - \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(\mathbf{z}_i^{(q)})\mathbf{z}_i^{(q)t}\right) &= 0 \\ \Rightarrow \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(\mathbf{z}_i^{(q)})\mathbf{z}_i^{(q)t} &= \mathbf{I}. \end{aligned} \tag{2.67}$$

décorrélation non linéaire

Ce qui signifie que les composantes de Z et de $\mathbf{g}(Z)$ sont décorrélées. Cette observation permet d'établir un lien entre les différents principes d'estimation déjà présentés et le principe de décorrélation non linéaires utilisé dans les premiers travaux sur l'analyse en composantes indépendantes, (Hérault et al. 1985). La figure (2.10) représente différentes formes paramétriques pouvant être postulées pour modéliser les sources et les fonctions g *correspondantes à titre d'illustration cette méthode pouvant être utilisée en postulant d'autre formes que des modèles de mélange. Le principe de décorrélation non linéaire peut d'ailleurs mener à d'autres solutions algorithmiques utilisant l'astuce du noyau (Bach et Jordan 2003).*

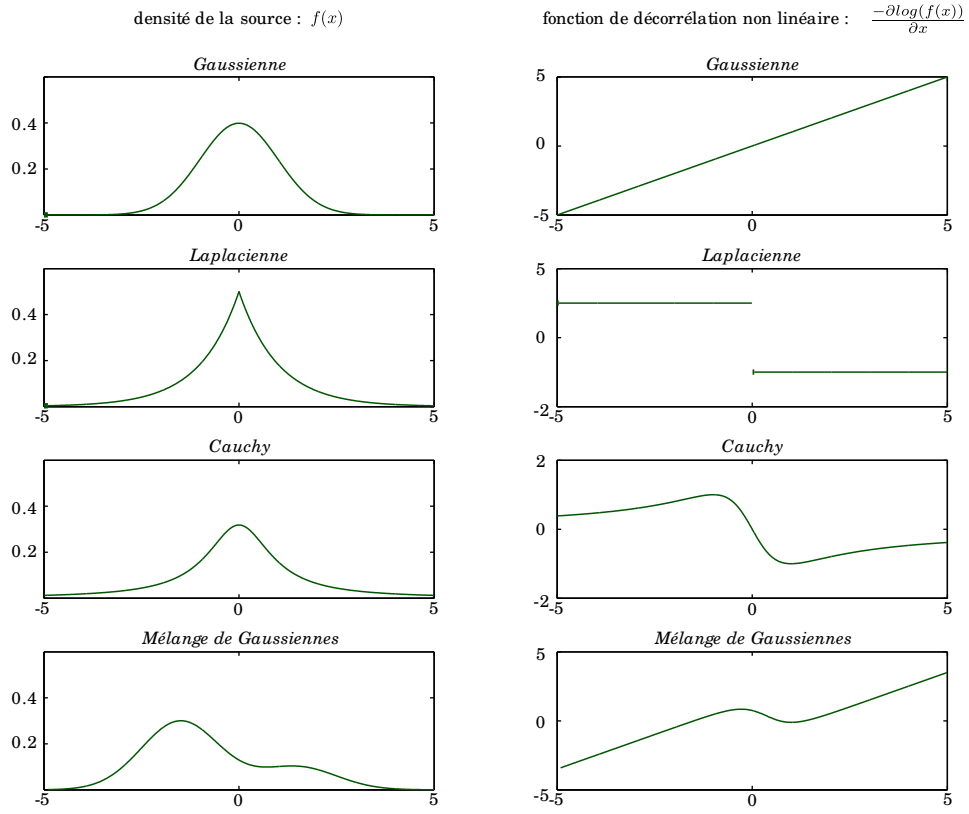


FIG. 2.10 – *Exemples de densités paramétriques pouvant être utilisées pour modéliser des sources et des fonctions de décorrélation non linéaire associées.*

Dans le cadre des modèles de mélange, les composantes de la fonction de décorrélation non linéaire g nécessaire pour maximiser la vraisemblance par rapport

à W sont données par :

$$\begin{aligned}\frac{-\partial \log(f^{z_s}(z_s))}{\partial z_s} &= \frac{-\partial \log\left(\sum_{k=1}^{K_s} \pi_k^s \varphi(z_s;\mu_k^s,\nu_k^s)\right)}{\partial z_s} \\ &= \sum_{k=1}^{K_s} t_k^s(z_s)\frac{(z_s-\mu_k^s)}{\nu_k^s}, \end{aligned} \tag{2.68}$$

avec $t_k^s(z_s)$ les probabilités a posteriori d'appartenance aux classes connaissant z_s, c'est-à-dire :

$$t_k^s(z_s) = \frac{\pi_k^s \varphi(z_s;\mu_k^s,\nu_k^s)}{\sum_{k'=1}^{K_s} \pi_{k'}^s \varphi(z_s;\mu_{k'}^s,\nu_{k'}^s)}. \tag{2.69}$$

Nous venons de voir comment faire croître la vraisemblance en modifiant W, en ce qui concerne la mise à jour des paramètres des sources, ceux-ci peuvent tout simplement être mis à jour en utilisant l'algorithme EM classique pour chacune des sources lorsque W est fixée. L'algorithme 4 récapitule sur ces différents éléments et montre comment ceux-ci peuvent être combinés pour estimer tous les paramètres intervenant dans le modèle.

Il est intéressant de noter que la méthode présentée ici peut être étendue pour traiter des modèles plus complexes : nombre de variables latentes différent du nombre de variables observées, modélisation du bruit... Nous proposerons dans le chapitre 4 de cette thèse une extension de cette méthode pour prendre en considération, une information partielle sur les variables latentes. Après cette présentation de l'analyse en composantes indépendantes et des travaux existants dans le domaine des modèles à variables latentes, la suite de ce chapitre est consacrée à l'introduction de la théorie des fonctions de croyance qui sera elle aussi mise à contribution dans les chapitres suivants de cette thèse.

Remarque 2.7 (IFA et modèle de mélange parcimonieux) *L'IFA peut être vue comme un modèle de mélange gaussien très particulier et extrêmement parcimonieux. En effet, les données observées suivent un modèle de mélange gaussien à $K_1 \times K_2 \times \cdots \times K_S$ composantes décrit seulement par $S \times S + (K_1 \times 3 - 2) + \cdots + (K_S \times 3 - 2)$ paramètres, ce qui est très faible comparé à un modèle de mélange complet intégrant le même nombre de composantes. Les variables latentes continues permettent de mettre en commun la structure de variance-covariance des données observées pour toutes les classes à l'aide de la matrice de mixage et l'hypothèse d'indépendance permet de limiter drastiquement le nombre de paramètres nécessaire à la description des données.*

Remarque 2.8 (ICA, IFA et minima locaux) *Tout comme dans le cadre des modèles de mélange, les algorithmes mis en place pour estimer les paramètres de l'ICA ou de l'IFA peuvent rencontrer des problème dûs à l'existence de minima/maxima locaux, en particulier lorsque les sources sont multi-modales (Vrins et al. 2005). Cette remarque doit être prise en compte lors de la mise en pratique de ce type d'algorithme afin d'éviter les solutions sous optimales associées à ces minima/maxima locaux.*

Algorithme 4: pseudo-code de l'analyse en facteurs indépendants sans bruit avec un algorithme GEM utilisant une montée de gradient naturel pour l'optimisation par rapport à la matrice de démixage.

Données : Matrice de données centrée : $\mathbf{X}$
 # Initialisation du vecteur de paramètres

$\boldsymbol{\psi}^{(0)} = (W^{(0)}, \boldsymbol{\pi}^{1(0)}, \ldots, \boldsymbol{\pi}^{S(0)}, \boldsymbol{\mu}^{1(0)}, \ldots, \boldsymbol{\mu}^{S(0)}, \boldsymbol{\nu}^{1(0)}, \ldots, \boldsymbol{\nu}^{S(0)}), q = 0$

tant que *test de convergence* **faire**
 # Mise à jour des sources
$\mathbf{Z} = \mathbf{X}.{W^{(q)}}^t$

 # Mise à jour des paramètres des sources / EM
pour tous les $s \in \{1, \ldots, S\}$ *et* $k \in \{1, \ldots, K_s\}$ **faire**
 # Etape E

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

pour tous les $s \in \{1, \ldots, S\}$ *et* $k \in \{1, \ldots, K_s\}$ **faire**
 # Etape M
 # Mise à jour des paramètres des sources

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

 # Mise à jour de G (2.68)
$\mathbf{G} = \mathbf{g}^{(q+1)}(\mathbf{Z})$

 # Calcul du gradient naturel (2.66)
$\Delta W = (\mathbf{I} - \frac{1}{N} \mathbf{G}^t.\mathbf{Z}).{W^{(q)}}^t$

 # Recherche linéaire sur τ
τ^* = RechercheLineaire($W^{(q)}$,ΔW)

 # Mise à jour de la matrice de démixage
$W^{(q+1)} = W^{(q)} + \tau^*.\Delta W$

 # Normalisation des sources
pour tous les $s \in \{1, \ldots, S\}$ **faire**
$\sigma_s^2 = \sum_{k=1}^{Ks} \pi_k^{s(q+1)} (\nu_k^{s(q+1)} + {\mu_k^{s(q+1)}}^2) - \left(\sum_{k=1}^{Ks} \pi_k^{s(q+1)} \mu_k^{s(q+1)} \right)^2$
pour tous les $k \in \{1, \ldots, K_s\}$ **faire**
$\mu_k^{s(q+1)} = \mu_k^{s(q+1)} / \sigma_s$
$\nu_k^{s(q+1)} = \nu_k^{s(q+1)} / \sigma_s^2$
$W_{s.}^{(q+1)} = W_{s.}^{(q+1)} / \sigma_s$
$q = q + 1$

Résultat : Paramètres estimés : $\widehat{\psi}^{ml}$, variables latentes estimées : $\widehat{\mathbf{Z}}^{ml}$

2.3 LA THÉORIE DES FONCTIONS DE CROYANCE

La théorie des probabilités a longtemps été la seule solution envisageable pour raisonner en présence d'informations incertaines. Cependant la dernière moitié du vingtième siècle a vu apparaître d'autres théories concurrentes s'appuyant sur une axiomatisation différente, dont font partie la théorie des possibilités Dubois et Prade (1985), la logique floue proposée par Zadeh (1965), la théorie des probabilités imprécises Walley (1991) et enfin la théorie des fonctions de croyance. Cette dernière a été initialement introduite par Dempster et Shafer, (Dempster 1967, Shafer 1976), puis reprise et étendue par Smets (1978; 1990a; 1993), Smets et Kennes (1994), Smets (1995; 2005b) dans une interprétation subjectiviste dénommée « Modèle des Croyances Transférables ». C'est ce cadre qui sera utilisée dans ce mémoire.

L'avantage de cette théorie est sans conteste sa flexibilité pour représenter l'information. En effet, comme nous allons le voir dans cette section, elle offre l'avantage de pouvoir représenter et manipuler aisément des informations aussi bien incertaines, qu'imprécises[2]. En séparant ces deux sources d'inexactitude dans la représentation d'une information, cette théorie permet de clarifier et de mieux formaliser certains problèmes. Pour présenter cette théorie nous allons tout d'abord effectuer un panorama des principales définitions nécessaires à sa construction dans le cas discret, c'est-à-dire lorsque les valeurs pouvant être prises par la variable d'intérêt sont en nombre fini.

2.3.1 Représentation de l'information

Concepts de base

Tout comme la théorie des probabilités, la théorie des fonctions de croyance suppose la construction d'un univers des possibles aussi appelé cadre de discernement qui sera noté Ω et sera supposé de cardinal fini $\Omega = \{\omega_1, \ldots, \omega_K\}$. Celui-ci contient les valeurs pouvant être prises par une variable d'intérêt et elles seront supposées distinctes. Dans le cadre de la théorie des fonctions de croyance, l'objet de base permettant de représenter l'information disponible sur la valeur prise par cette variable est la fonction de masse définie comme une fonction de l'ensemble de tous les ensembles de Ω vers $[0, 1]$ vérifiant une condition de normalité.

fonction de masse

Définition 2.6 (fonction de masse) *Une fonction de masse sur Ω est une fonction de l'ensemble de tous les ensembles de Ω dans $[0, 1]$, $2^\Omega \to [0, 1]$ telle que :*

$$\sum_{\omega \subseteq \Omega} m^\Omega(\omega) = 1. \tag{2.70}$$

Cette fonction diffère d'une distribution de probabilité classique en cela qu'elle est définie de l'ensemble de toutes les parties de Ω vers $[0, 1]$ et non pas de Ω vers $[0, 1]$. Une distribution de probabilité peut donc être vue comme un cas particulier de

[2] L'imprécision d'une information caractérise son aspect non précis, vague alors que l'incertitude concerne son caractère non certain, c'est-à-dire le fait qu'elle peut être vraie ou fausse

fonction de masse où seuls les éléments singletons ω_i sont affectés d'une masse non nulle. Les fonctions de masses de ce type seront appelées fonctions de masses bayésiennes. Les éléments ω de 2^Ω possédant une masse non nulle seront appelés ensemble focaux.

La masse attribuée à un ensemble focal modélise l'importance pouvant être accordée à l'hypothèse qu'elle supporte sans pour autant accréditer plus particulièrement aucun des sous-ensembles de l'ensemble focal en question.

Pour bien comprendre la différence entre une distribution de probabilité et une fonction de masse, nous pouvons observer la représentation de l'ignorance totale dans ces deux formalismes. Dans le cadre probabiliste classique, le principe de raison insuffisante ou d'entropie maximale conduit à représenter l'ignorance totale par une distribution uniforme. Chacun des singletons de Ω possède, dans ce cas, une masse égale à un sur le nombre d'éléments de Ω. Cette représentation a donc transformé une information imprécise (je ne sais rien) en une information incertaine (toutes les solutions sont équiprobables). A l'inverse, dans le cadre de la théorie des fonctions de croyance, l'ignorance totale est modélisée par l'affectation d'une masse de 1 à l'ensemble Ω lui-même, ce qui correspond bien à l'information initiale. Une telle fonction de masse est appelée fonction de masse vide.

Finalement, un dernier cas particulier de fonction de masse mérite d'être cité ; il concerne les fonctions de masses ne possédant qu'un seul élément focal de masse unité. De telles fonctions de masses seront appelées catégoriques dans la suite de ce mémoire.

hypothèse du monde ouvert

L'un des apports importants de P. Smets (Smets 1990b), a sans doute été la mise en évidence de l'intérêt de l'hypothèse du monde ouvert. Cette hypothèse relâche la contrainte classique $m(\emptyset) = 0$. La quantité $m(\emptyset)$ est alors interprétée comme la part de croyance accordée à l'hypothèse : la vérité n'est pas dans Ω, autrement dit le cadre de discernement n'est pas exhaustif.

A partir de la fonction de masse précédemment définie, il est possible de définir d'autres représentations de l'information contenue dans m. Deux d'entre elles ont une interprétation aisée, ce sont les fonctions de plausibilité et de crédibilité.

Définition 2.7 (fonction de plausibilité) *Une fonction de plausibilité est une fonction de l'ensemble de tous les ensembles de* Ω *dans* $[0,1]$*,* $2^\Omega \rightarrow [0,1]$ *définie à partir d'une fonction de masse* m^Ω *de la manière suivante :*

$$pl^\Omega(\omega) = \sum_{\alpha \cap \omega \neq \emptyset} m^\Omega(\alpha), \quad \forall \omega \subseteq \Omega. \tag{2.71}$$

Définition 2.8 (fonction de crédibilité) *Une fonction de crédibilité est une fonction de l'ensemble de tous les ensembles de* Ω *dans* $[0,1]$*,* $2^\Omega \rightarrow [0,1]$ *définie à partir d'une fonction de masse* m^Ω *de la manière suivante :*

$$bel^\Omega(\omega) = \sum_{\alpha \neq \emptyset \,,\, \alpha \subseteq \omega} m^\Omega(\alpha), \quad \forall \omega \subseteq \Omega. \tag{2.72}$$

fonction de crédibilité

Ces deux fonctions ont un sens concret dans le cadre de la théorie des fonctions de croyance. La fonction de crédibilité bel représente une mesure globale de la croyance attribuée à une hypothèse ; elle est égale à la somme des hypothèses

fonction de plausibilité

qui l'impliquent, qui la compose. La fonction de plausibilité pl représente quant à elle le degré de croyance maximal qui pourrait potentiellement être attribué à l'hypothèse, si de nouvelles informations étaient disponibles. Ces deux fonctions permettent d'encadrer la croyance en la véracité d'une hypothèse et sont donc une version optimiste et pessimiste de notre croyance.

Exemple 2.6 (fonction de masse, de plausibilité et de crédibilité) :

Supposons que dans le contexte d'un problème de diagnostic deux experts, l'un ingénieur mécanique, l'autre électricien sont interrogés sur la cause d'une panne d'un système industriel afin de savoir quelle équipe envoyer pour réparer la panne. Les causes de pannes possibles pour ce système sont supposées être $\Omega = \{a, b, c\}$:
- $\{a\}$: le système est victime d'un problème mécanique de type a ;
- $\{b\}$: le système est victime d'un problème mécanique de type b ;
- $\{c\}$: le système est victime d'un problème électrique.

Les deux experts fournissent les informations suivantes :
- Expert mécanique : « Le problème est certainement mécanique, et je pense qu'il est plus probable que celui-ci soit de type a. »
- Expert électrique : « Il est peu probable que le problème soit électrique. »

Après avoir interrogé ces deux experts pour quantifier ces différentes affirmations, ces informations peuvent être traduites de la manière suivante dans le cadre de la théorie des fonctions de croyance :

	Expert mécanique			Expert électrique		
	m_{me}	bel_{me}	pl_{me}	m_{el}	bel_{el}	pl_{el}
$\emptyset$	0	0	0	0	0	0
$\{a\}$	0.2	0	1	0	0	0.8
$\{b\}$	0	0	0.8	0	0	0.8
$\{c\}$	0	0	0.3	0.2	0.2	0.2
$\{a, b\}$	0.5	0.7	1	0.8	0.8	0.8
$\{a, c\}$	0	0.2	1	0	0.2	1
$\{b, c\}$	0	0	0.8	0	0.2	1
$\{a, b, c\}$	0.3	1	1	0	1	1

TAB. 2.1 – *Représentation de l'information dans le cadre de la théorie des fonctions de croyance. Fonctions de masse, de crédibilité et de plausibilité construites à partir de deux avis d'experts*

Travail sur un espace produit

En reprenant l'exemple 2.6, on conçoit l'intérêt de pouvoir fusionner les avis des deux experts pour produire un avis unique. En réalité, il est souvent nécessaire de travailler et de manipuler des fonctions de croyance définies sur des espaces produits. Différentes notions peuvent être particulièrement utiles dans ce contexte, telles que les notions de marginalisation et d'extension.

Lorsque l'information utile concerne uniquement l'un des éléments de l'espace

marginalisation produit, il est possible de marginaliser une fonction de masse définie sur $\Omega \times \Theta$ pour obtenir une fonction de masse sur Ω par exemple.

Définition 2.9 (Marginalisation) *Soit un espace produit $\Omega \times \Theta$. La marginalisation d'une fonction de masse $m^{\Omega\times\Theta}$ sur Ω noté $m^{(\Omega\times\Theta)\downarrow\Omega}$ est une fonction de masse sur Ω définie par :*

$$m^{(\Omega\times\Theta)\downarrow\Omega}(\omega) = \sum_{\alpha\subseteq\Omega\times\Theta\ ,\ \alpha^{\downarrow\Omega}=\omega} m^{\Omega\times\Theta}(\alpha), \quad \forall\omega \subseteq \Omega, \tag{2.73}$$

avec $\alpha^{\downarrow\Omega}$ la projection de α sur Ω : $\alpha^{\downarrow\Omega} = \{a : a \in \Omega, \alpha \cap (\{a\} \times \Theta) \neq \emptyset\}$

Cette notion recouvre la notion de marginalisation probabiliste. Elle est à mettre extension vide en parallèle avec la notion d'extension, qui permet d'effectuer l'opération inverse, c'est-à-dire construire une fonction de masse sur un espace produit $\Omega \times \Theta$ à partir d'une fonction de masse définie uniquement sur Ω de manière à ce que celle-ci soit la moins informative[3] possible tout en étant compatible avec m^{Ω}.

Définition 2.10 (Extension vide) *Soit un espace produit $\Omega \times \Theta$. L'extension d'une fonction de masse m^{Ω} sur $\Omega \times \Theta$ noté $m^{\Omega\uparrow(\Omega\times\Theta)}$ est une fonction de masse sur $\Omega \times \Theta$ définie par :*

$$m^{\Omega\uparrow(\Omega\times\Theta)}(\alpha) = \begin{cases} = m^{\Omega}(\omega) & \textit{si } \alpha = \omega \times \Omega, \\ = 0 & \textit{sinon}. \end{cases} \tag{2.74}$$

Ces deux opérations ne sont pas tout à fait opposées. En effet, il est aisée de démontrer que $m^{\Omega\uparrow(\Omega\times\Theta)\downarrow\Omega} = m^{\Omega}$, une extension suivie d'une marginalisation permet donc de retrouver la fonction de masse de départ. Inversement, une marginalisation suivie d'une extension ne permet généralement pas de rerouver la fonction de masse originelle. En effet dans ce cas de figure, la marginalisation entraîne une perte d'information, que l'extension ne peut reconstruire, sauf cas particuliers. Nous aurons donc : $m^{(\Omega\times\Theta)\downarrow\Omega\uparrow(\Omega\times\Theta)} \neq m^{\Omega\times\Theta}$.

Après avoir défini les opérations nécessaires a la manipulation de cadres de discernement construits en considérant l'espace produit de différents cadres, nous allons maintenant nous tourner vers le cœur de la théorie des fonctions de croyance qui concerne la prise en compte d'informations nouvelles.

2.3.2 Prise en compte de nouvelles informations

conditionnement Dans le cadre probabiliste la notion de conditionnement permet de prendre en compte de nouvelles informations lorsque celle-ci sont certaines. Ce concept, trouve une extension naturelle dans le cadre de la théorie des fonctions de croyance. :

Définition 2.11 (Conditionnement) *Soit m^{Ω} une fonction de masse définie sur Ω et ω un sous ensemble non vide de Ω. La fonction de masse conditionnelle à ω non normalisée $m^{\Omega}(.|\omega)$ est définie par :*

$$m^{\Omega}(\alpha|\omega) = \begin{cases} \sum_{\beta\cap\omega=\alpha} m^{\Omega}(\beta) & \textit{si } \alpha \subseteq \omega, \\ 0 & \textit{sinon}. \end{cases} \tag{2.75}$$

[3] Voir (Dubois et Prade 1986b, Dubois et al. 2001) pour une discussion sur les manières d'ordonner des fonctions de croyance suivant leur degré d'informativité.

Lorsque le cadre de discernement est supposé exhaustif, il est possible d'utiliser la notion de conditionnement normalisé.

Définition 2.12 (Conditionnement normalisé) *Soit m^Ω une fonction de masse définie sur Ω et ω un sous ensemble non vide de Ω. La fonction de masse conditionnelle à ω normalisée $m^\Omega(.||\omega)$ est définie par :*

$$m^\Omega(\alpha||\omega) = \begin{cases} \frac{m^\Omega(\alpha|\omega)}{1-m^\Omega(\emptyset|\omega)} & \textit{si } \alpha \subset \omega, \alpha \neq \emptyset \\ 0 & \textit{sinon}. \end{cases} \tag{2.76}$$

Lorsque la fonction de masse considérée est bayésienne, le conditionnement normalisé de celle-ci est équivalent au conditionnement probabiliste.

Nous allons maintenant voir comment la théorie des fonctions de croyance permet d'aller plus loin en permettant de prendre en considération de nouvelles informations incertaines. C'est ce que propose les opérateurs de combinaisons de la théorie des fonctions de croyance. Traditionnellement, des versions conjonctives et disjonctives sont utilisées (Smets 1993). La combinaison conjonctive présentée ci-dessous permet d'obtenir de bon résultats lorsque les sources sont réputées indépendantes et fiables.

combinaison conjonctive

Définition 2.13 (Combinaison conjonctive) *Soit m_1, m_2 deux fonctions de masses définies sur un cadre de discernement identique Ω, la combinaison conjonctive de m_1 et de m_2 notée $m^\Omega_{1\textcircled{\cap}2}$ est définie par :*

$$m^\Omega_{1\textcircled{\cap}2}(\omega) = \sum_{\alpha_1 \cap \alpha_2 = \omega} m^\Omega_1(\alpha_1) m^\Omega_2(\alpha_2), \quad \forall \omega \subseteq \Omega. \tag{2.77}$$

Cette opérateur correspond à la conjonction des deux fonctions de masse. La quantité $m^\Omega_{1\textcircled{\cap}2}(\emptyset)$ est appelée degré de conflit entre m_1 et m_2 et quantifie le désaccord entre les deux sources d'information. Cette quantité est égale à la somme des produits des masses allouées à des hypothèses incompatibles. Ce degré de conflit n'a de sens que lorsque l'hypothèse du monde ouvert est faite. Dans le cas contraire la règle de combinaison conjonctive normalisée (ou règle de Dempster) réaloue cette masse aux autres ensembles focaux.

Définition 2.14 (Combinaison conjonctive normalisée) *Soit m_1, m_2 deux fonctions de masses définies sur un cadre de discernement identique Ω. La combinaison conjonctive normalisée de m_1 et de m_2 notée $m^\Omega_{1\oplus 2}$ est définie par :*

$$m^\Omega_{1\oplus 2}(\omega) = \begin{cases} 0 & \textit{si } \omega = \emptyset, \\ \frac{m^\Omega_{1\textcircled{\cap}2}(\omega)}{1-m^\Omega_{1\textcircled{\cap}2}(\emptyset)} & \textit{si } \omega \subseteq \Omega, \omega \neq \emptyset. \end{cases} \tag{2.78}$$

Le résultat de cette combinaison est défini pourvu que $m^\Omega_{1\textcircled{\cap}2}(\emptyset) \neq 1$, c'est-à-dire pourvu que les deux fonctions de masse ne soient pas en totale contradiction. Ces deux règles de combinaison peuvent êtres dérivées de manière axiomatiques (Smets 1990b).

Il est intéressant de noter, que si m^Ω_1 (ou m^Ω_2) est une fonction de masse bayésienne, il en va de même pour le résultat de la combinaison : $m^\Omega_{1\oplus 2}(\omega)$. Ces opérateurs possèdent de nombreuses autres propriétés intéressantes, en particulier

la commutativité et l'associativité. Ils possèdent également un élément neutre naturel la fonction de masse vide. L'opérateur de combinaison conjonctive peut de plus être reformulé afin de mettre en évidence le lien entre la théorie des fonctions de croyance et la théorie des probabilités. Cette reformulation est donnée par le théorème des masses totales :

Théorème 2.2 (Théorème des masses totales, formulation générale) *Soit m_1, m_2 deux fonctions de masse définies sur un cadre de discernement identique Ω. La combinaison conjonctive de celles-ci $m^{\Omega}_{1\textcircled{\cap}2}$ est égale à :*

$$m^{\Omega}_{1\textcircled{\cap}2}(\omega) = \sum_{\alpha\subseteq\Omega} m^{\Omega}_1(\alpha) m^{\Omega}_2(\omega|\alpha), \quad \forall\omega \subseteq \Omega. \tag{2.79}$$

Cette formulation est en effet proche du théorème des probabilités totales rencontré dans le cadre probabiliste (Dubois et Prade 1986a, Smets 1993, Xu et Smets 1994). Ce théorème permet également d'effectuer des calculs de manière simplifiée lorsque l'espace considéré est un espace produit et que des fonctions de masses conditionnelles sont disponibles (Smets 1993). Si nous considérons deux cadres de discernement Ω et Θ, et un ensemble de fonction de masse conditionnelles $m^{\Theta|\Omega}(\,\cdot\,|\omega)$ pour tout $\omega \subseteq \Omega$. La combinaison de cet ensemble de fonctions de masse conditionnelles et d'une fonction de masse sur Ω, m^{Ω} donne le résultat suivant sur Θ :

Théorème 2.3 (Théorème des masses totales, cas d'un espace produit) *Soit $m^{\Theta|\Omega}(\,\cdot\,|\omega), \quad \forall\omega \subseteq \Omega$ un ensemble de fonctions de masse conditionnelles sur Θ et m^{Ω} une fonction de masse sur Ω, alors :*

$$m^{\Theta}(\theta) = \sum_{\omega\subseteq\Omega} m^{\Omega}(\omega) m^{\Theta|\Omega}(\theta|\omega), \quad \forall\theta \subseteq \Theta. \tag{2.80}$$

Une formule similaire peut être obtenue pour la fonction de plausibilité.

Théorème 2.4 (Théorème des plausibilités totales, cas d'un espace produit) *Soit $pl^{\Theta|\Omega}(\,\cdot\,|\omega), \quad \forall\omega \subseteq \Omega$ un ensemble de fonctions de plausibilité conditionnelles sur Θ et m^{Ω} une fonction de masse sur Ω, alors :*

$$pl^{\Theta}(\theta) = \sum_{\omega\subseteq\Omega} m^{\Omega}(\omega) pl^{\Theta|\Omega}(\theta|\omega), \quad \forall\theta \subseteq \Theta. \tag{2.81}$$

Nous n'avons, dans ce paragraphe, introduit qu'un seul type d'opérateur de combinaison qui est justifié lorsque les deux fonctions de masses sont supposées indépendantes et fiables. Il existe cependant, dans le cadre de la théorie des fonctions de croyance, d'autres opérateurs pouvant être utilisés lorsque ces hypothèses ne peuvent être posées. Il est en particulier possible de définir un opérateur de combinaison prudent lorsque les deux sources ne sont pas indépendantes (Denoeux 2008).

Exemple 2.7 (Combinaison conjonctive) :

Afin de prendre une décision quant à l'équipe de maintenance à envoyer (cf exemple 2.6), les deux fonctions de masse construites à partir des avis d'experts sont combinées. Cette combinaison peut être effectuée grâce à l'opérateur de combinaison conjonctive car les deux fonctions de masse peuvent être considérées comme indépendantes cognitivement (domaines d'expertises différents) et parce que les deux experts sont considérés comme fiables. Les résultats de la combinaison sont les suivants :

$$\begin{aligned}
m_{me\textcircled{\cap}el}(\emptyset) &= m_{me}(\{a\}) \times m_{el}(\{c\}) + m_{me}(\{a,b\}) \times m_{el}(\{c\}) \\
&= 0.2 \times 0.2 + 0.5 \times 0.2 = 0.14 \\
m_{me\textcircled{\cap}el}(\{a\}) &= m_{me}(\{a\}) \times m_{el}(\{a,b\}) \\
&= 0.2 \times 0.8 = 0.16 \\
m_{me\textcircled{\cap}el}(\{c\}) &= m_{me}(\{a,b,c\}) \times m_{el}(\{c\}) \\
&= 0.3 \times 0.2 = 0.06 \\
m_{me\textcircled{\cap}el}(\{a,b\}) &= m_{me}(\{a,b\}) \times m_{el}(\{a,b\}) + m_{me}(\{a,b,c\}) \times m_{el}(\{a,b\}) \\
&= 0.8 \times 0.5 + 0.3 \times 0.8 = 0.64.
\end{aligned}$$

Le résultat de la combinaison, possède un degré de conflit assez faible. Les deux sources ne sont donc pas en désaccord de manière importante. Nous pouvons également observer la spécialisation de la fonction de masse : plus aucune masse n'est allouée à l'ignorance, l'ensemble focal de plus grand cardinal est maintenant $\{a, b\}$ (c'est-à-dire probléme mécanique de type a ou b).

Nous allons voir maintenant comment une décision peut être prise sur la base d'information imprécises et incertaines représentées à l'aide de fonctions de croyance.

2.3.3 Prise de décision

L'extension de la théorie bayésienne de la décision dans le contexte de la théorie des fonctions de croyance ne conduit pas à une unique solution quant à la décision optimale. En effet, de manière générale, l'ordre induit entre les différentes hypothèses par la fonction de plausibilité est différent de l'ordre induit par la fonction de crédibilité.

Deux stratégies semblent a priori envisageables : sélectionner l'hypothèse ayant le plus grand degré de crédibilité ou sélectionner l'hypothèse ayant le degré de doute le plus faible, c'est-à-dire la plus plausible. Pour résoudre ce paradoxe et proposer une solution unique au problème de la décision, deux solutions sont en concurrence dans le cadre de la théorie des fonctions de croyance. Toutes deux s'appuient sur une transformation capable de convertir une fonction de masse en distribution de probabilité afin de pouvoir utiliser la théorie classique de la décision bayésienne. La première solution défendue par (Smets 1990a; 2005b), repose sur la transformation pignistique.

Définition 2.15 (Transformation pignistique) *Le résultat de la transformation pignistique d'une fonction de masse m^Ω est la distribution pignistique notée $betP^\Omega_m$. C'est*

une fonction de Ω *dans* $[0,1]$*, définie par :*

$$betP_m^{\Omega}(\{x\}) = \sum_{\alpha \neq \emptyset \ , \ \alpha \subseteq \Omega \ , \ x \in \alpha} \frac{m(\alpha)}{|\alpha|.(1 - m^{\Omega}(\emptyset))}, \quad \forall x \in \Omega. \tag{2.82}$$

transformation pignistique

Cette transformation alloue, à parts égales, la masse associée à un ensemble, à tous les éléments qui le compose. Cette transformation, possède une relation assez évidente avec le principe d'entropie maximale ou de raison insuffisante. Elle peut être justifiée par le fait quelle est la seule transformation invariante par rapport aux combinaisons linéaires de fonctions de masses. Si nous définissons une fonction de masse $m = \gamma.m_1 + (1-\gamma)m_2$ alors $betP_m$ est la seule transformation telle que $betP_m = \gamma.betP_{m_1} + (1-\gamma)betP_{m_2}$ pour tout scalaire $\gamma \in [0,1]$. Cette méthode de transformation possède une alternative intéressante qui s'appuie sur la fonction de plausibilité (Cobb et Shenoy 2006).

transformation plausibiliste

Définition 2.16 (Transformation plausibiliste) *Le résultat de la transformation plausibiliste d'une fonction de masse* m^{Ω} *est une fonction de* Ω *dans* $[0,1]$*, notée* plP_m^{Ω}*, définie par :*

$$plP_m^{\Omega}(\{x\}) = \frac{1}{\sum_{\alpha \in \Omega} pl^{\Omega}(\{\alpha\})} pl^{\Omega}(\{x\}), \quad \forall x \in \Omega \tag{2.83}$$

$$= \frac{1}{\sum_{\alpha \in \Omega} pl^{\Omega}(\{\alpha\})} \sum_{\alpha \cap \{x\} \neq \emptyset} m^{\Omega}(\alpha) \quad \forall x \in \Omega. \tag{2.84}$$

Cette transformation possède elle aussi une propriété très intéressante, celle d'invariance par rapport à la combinaison conjonctive normalisée. Si nous définissons $m = m_1 \oplus m_2$ alors plP_m est la seule transformation telle que $plP_m = plP_{m_1} \oplus plP_{m_2}$.

2.3.4 Concepts plus avancés

Indépendance

indépendance cognitive

Le concept classique d'indépendance dans la théorie des probabilités ne trouve pas d'extension immédiate dans le cadre de la théorie des fonctions de croyance. Différentes notions peuvent être définies, la forme la plus simple d'indépendance étant l'indépendance cognitive (Shafer 1976, p. 149).

Définition 2.17 (Indépendance cognitive) *Deux cadres de discernement* Ω *et* Θ *sont dits cognitivement indépendants par rapport à* $pl^{\Omega \times \Theta}$ *si et seulement si la propriété suivante est vérifiée :*

$$pl^{\Omega \times \Theta}(\omega \times \theta) = pl^{\Omega}(\omega) \, . \, pl^{\Theta}(\theta), \quad \forall \omega \subseteq \Omega, \forall \theta \subseteq \Theta. \tag{2.85}$$

Cette notion permet de retrouver la notion d'indépendance probabiliste lorsque $pl^{\Omega \times \Theta}$ est une mesure de probabilité. Cependant, ce concept ne bénéficie pas de toutes les propriétés de l'indépendance probabiliste et peut être vu comme une forme faible d'indépendance dans le contexte de la théorie des fonctions de croyance. Il est possible de se tourner vers les articles (Yaghlane et al. 2002a;b) pour une analyse en profondeur des différentes définitions possibles de l'indépendance dans le cadre de la théorie des fonctions de croyance.

Cadre de discernement continu

Les différentes définitions données dans les paragraphes précédents de cette section, supposaient toujours un cadre de discernement discret, c'est-à-dire composé d'un ensemble fini d'hypothèses. Elles peuvent être étendues à des cadres de discernement continus, en supposant que les ensembles focaux sont des intervalles (Smets 2005a). En effet, un intervalle étant décrit par deux nombres ses bornes (inférieure et supérieure), il est possible de mettre en relation l'ensemble des intervalles fermés de $\mathbb{R}$ et $\mathbb{R}^2$. La notion de fonction de masse est alors remplacée par la notion de densité de croyance, qui est définie comme une fonction $m^{\mathbb{R}}$ de l'ensemble des intervalles fermés de $\mathbb{R}$ vers $[0, +\infty)$ telle que

densité de croyance

$$\int_{-\infty}^{+\infty} \int_{x}^{+\infty} m^{\mathbb{R}}([x, y])dydx \leq 1. \tag{2.86}$$

Par convention, le complément à 1 de l'intégrale (2.86) est alloué à l'ensemble vide $\emptyset$. Comme dans le cas discret, la plausibilité $pl^{\mathbb{R}}([a, b])$ est définie comme la somme sur tous les intervalles qui ont une intersection non vide avec $[a, b]$:

$$pl^{\mathbb{R}}([a, b]) \triangleq \iint\limits_{[x,y]\cap[a,b]\neq\emptyset} m^{\mathbb{R}}([x, y])dydx. \tag{2.87}$$

Les différentes définitions données jusqu'à présent peuvent êtres étendues aux cas $\mathbb{R}^d, d > 1$ (Caron et al. 2008), en considérant des hypercubes et non plus des intervalles. Il est finalement possible de définir des fonctions de croyance sur des espaces intégrant à la fois des variables discrètes et des variables continues (Aregui et Denœux 2006).

Nous terminons cette introduction aux différents outils offerts par la théorie des fonctions de croyance par le théorème de Bayes généralisé (GBT), introduit par Smets (1978; 1993).

Théorème de Bayes Généralisé

Le théorème de Bayes de la théorie des probabilité est remplacé dans le contexte de la théorie des fonctions de croyance par le théorème de Bayes généralisé (Smets 1978; 1993, Delmotte et Smets 2004, Denœux et Smets 2006). Ce théorème fournit un moyen de renverser une fonction de croyance conditionnelle sans utiliser de connaissance a priori.

théorème de Bayes généralisé

Pour introduire ce théorème prenons comme exemple celui de deux espaces : $\mathcal{X}$ l'espace des observations et Θ l'espace des paramètres. Supposons que notre connaissance soit représentée par un ensemble de fonctions de masse conditionnelles $m^{\mathcal{X}|\Theta}(.|\theta_i), \theta_i \in \Theta$, représentant notre croyance en l'observation x si le véritable paramètre est θ_i. Nous observons $x \subseteq \mathcal{X}$. La question est alors la suivante : connaissant cette observation et l'ensemble de fonctions de masse conditionnelles, quel est notre croyance sur la véritable valeur de Θ ? La réponse à cette question est fournie par le GBT et stipule que la fonction de plausibilité sur Θ prend la forme suivante :

$$pl^{\Theta|\mathcal{X}}(\theta|x) = pl^{\mathcal{X}|\Theta}(x|\theta) = 1 - \prod_{\theta_i\in\theta}(1 - pl^{\mathcal{X}|\Theta}(x|\theta_i)). \tag{2.88}$$

Quand une information a priori sur la véritable valeur de Θ est disponible sous la forme d'une fonction de masse m_0^Θ, celle-ci peut être combinée avec le résultat obtenu grâce au GBT (2.88) pour définir notre croyance finale. Ce type de démarche permet de retrouver le théorème de Bayes classique lorsque les fonctions de masses conditionnelles $m^{\mathcal{X}|\Theta}(.|\theta_i)$ ainsi que la fonction de masse a priori m_0^Θ sont bayésiennes.

CONCLUSION DU CHAPITRE

Ce chapitre nous a permis d'introduire différents outils qui seront utilisés dans nos travaux. Nous avons ainsi pu évoquer la théorie des fonctions de croyance qui offre un cadre riche et flexible pour représenter l'information, combiner différentes sources en vue d'en tirer une conclusion (prendre une décision). Nous avons également présenté les modèles statistiques à variables latentes qui intègrent des variables inobservées, c'est-à-dire pour lesquelles aucune information n'est disponible. Nous avons en particulier présenté les modèles de mélange qui permettent de modéliser l'existence de différentes sous-populations aux propriétés différentes dans un jeu de données. Nous avons également mis en avant les modèles à variables latentes continues qui permettent de trouver des espaces de représentation intéressants.

Nous allons voir dans les chapitres à venir comment ces deux outils peuvent être utilisés conjointement pour fournir des solutions innovantes à différents problèmes concrets. Nous verrons en particulier comment il est possible de représenter une information partielle sur l'appartenance d'un individu à une classe à l'aide d'une fonction de croyance, et comment cette information peut être prise en considération lors de l'apprentissage des paramètres d'un modèle de mélange. La classe d'origine de chaque individu ne sera donc pas supposée être inconnue mais connue de manière partielle. Cette proposition fait l'objet du chapitre 3. Le chapitre 4, quant à lui propose d'utiliser une méthode similaire dans le cadre de l'analyse en facteurs indépendants.

3 Le problème de la labellisation incertaine / imprécise

Je me sens peu sûr de ma vérité, même si j'y crois.
Umberto Eco, **Le nom de la Rose (1980)**

Sommaire

Ce chapitre introduit une des principales contributions de cette thèse (Côme et al. 2009). Elle concerne la formulation du problème de la classification lorsque l'information disponible sur les étiquettes des exemples servant à l'apprentissage est imparfaite. Avant de formuler le problème et de présenter la solution que nous proposons pour le résoudre, nous effectuons un tour d'horizon des différentes variantes pouvant êtres imaginées entre les cadres supervisé (informations précises et certaines) et non supervisé (informations totalement imprécises) et nous détaillons les différentes contributions récentes ayant été proposées pour résoudre ceux-ci.

Enfin, ce chapitre sera également dédié à la présentation de différents résultats expérimentaux concernant notre méthode. Ceux-ci nous permettront d'analyser le comportement de celle-ci dans diverses situations et de fournir des éléments de réponse quant à son intérêt pratique. Nous verrons, en particulier comment il est possible de résoudre de manière efficace des problèmes de classification, même lorsque certaines des étiquettes fournies pour apprendre la fonction de classification sont erronées.

3.1 LES DIFFÉRENTS PROBLÈMES

Le cadre de l'apprentissage supervisé ne correspond pas toujours aux situations rencontrées dans la pratique. En effet, l'acquisition d'un volume important de données labellisées de manière précise et certaine peut dans certains domaines être très problématique, pour des raisons de coût ou bien de connaissances imparfaites sur le problème à résoudre. C'est souvent le cas lors de la résolution de problèmes de diagnostic, car dans ce contexte les données doivent le plus souvent être labellisées par des experts. Le travail d'étiquetage est donc souvent laborieux et coûteux ; de plus, celui-ci est soumis aux imprécisions et aux incertitudes que peut rencontrer l'expert lors de cette tâche. C'est pourquoi, le développement de solutions permettant de traiter des étiquettes imparfaites nous a particulièrement intéressé durant cette thèse.

La communauté de l'apprentissage statistique a tenté de répondre à ces besoins pratiques, en formalisant différents problèmes plus généraux que l'apprentissage supervisé tel que l'apprentissage semi-supervisé. où l'on dispose à la fois d'exemples parfaitement labellisés et d'exemples dont l'étiquette est inconnue. Les solutions développées dans ce cadre sont particulièrement adaptées lorsque l'étiquetage des données a un coût important. L'apprentissage semi-supervisé n'est pas le seul problème tentant de rapprocher les méthodes d'apprentissage statistique des besoins applicatifs, l'apprentissage partiellement supervisé, ou bien encore l'apprentissage en présence d'erreurs de labellisation vont dans ce sens.

Nous allons dans la première partie de ce chapitre présenter ces différents problèmes ainsi que les solutions envisagées dans la littérature pour y répondre, en commençant par le plus connu et le plus fréquemment rencontré : l'apprentissage semi-supervisé.

3.1.1 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est devenu un champ de recherche particulièrement actif dans le domaine de l'apprentissage statistique comme le montre l'édition très récente d'un ouvrage collectif sur cette problématique (Chapelle et al. 2006).

Problématique de l'apprentissage semi-supervisé

Le paradigme de l'apprentissage semi-supervisé est apparu assez tôt dans la littérature, en particulier dans le cadre des modèles génératifs (Hosmer 1973, McLachlan 1977). L'ensemble d'apprentissage est supposé être de la forme suivante :

$$\mathbf{X}^{ss} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M), \mathbf{x}_{M+1}, \ldots, \mathbf{x}_N\}, \tag{3.1}$$

M points de l'ensemble d'apprentissage sont donc associés à un label alors que les $N - M$ restants ne possèdent pas de labels. L'objectif de l'apprentissage semi-supervisé est tout simplement de construire une fonction de classification capable de prédire pour tout nouveau point x son étiquette à partir de ce jeu de données.

Les différentes classes possibles pour chaque individu seront notées comme précédemment $\mathcal{Y} = \{c_1, \ldots, c_K\}$

Remarque 3.1 (cadre inférentiel ou transductif) *Nous nous plaçons ici dans le cadre inférentiel qui doit bien être distingué du cadre transductif proposé par Vapnik (1999). Dans le cadre transductif, seules les étiquettes des points de l'ensemble d'apprentissage non labellisées doivent être estimées.*

Les solutions proposées au problème de l'apprentissage semi-supervisé dans la littérature s'appuient sur des hypothèses et sur des modèles différents. Nous retrouvons les méthodes génératives et discriminatives, ainsi que d'autre méthodes ne reposant pas sur un modèle probabiliste. Nous présentons rapidement ces différentes solutions dans les paragraphes qui suivent. Mais, avant cela, nous revenons sur la nature même du problème et nous tentons de décrire les situations pour lesquelles l'apprentissage semi-supervisé est possible, c'est-à-dire les situations pour lesquelles l'observation de points non labellisés peut être intéressante.

En effet, si aucune connaissance n'est introduite dans les relations entre les distributions de X et de $Y|X$, les points non labellisés ne sont d'aucun intérêt pour estimer la frontière de décision (Chapelle et al. 2006, chap. 2). Il est nécessaire de faire des hypothèses pour tirer parti de ces individus. L'hypothèse la plus couramment faite concerne le regroupement des classes dans les zones de fortes densités. Celle-ci peut être formulée de la manière suivante :

Les différentes classes forment des zones distinctes de forte densité ; celles-ci sont donc séparées par des zones de faible densité.

hypothèse du regroupement

La figure 3.1 représente un jeu de données satisfaisant cette hypothèse. Ce type de problème peut être résolu de différentes manières ; nous verrons en particulier comment le concept d'entropie conditionnelle peut être utilisé pour construire une régularisation favorisant les classifieurs respectant cette hypothèse.

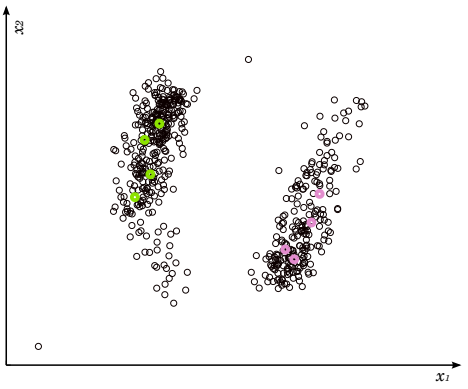


FIG. 3.1 – *Illustration de l'hypothèse de regroupement des classes : les différentes classes forment des zones de forte densité distinctes, séparées par des zones de faible densité. En vert les exemples de la classe c_1, en violet ceux de la classe c_2, en noir, les exemples non étiquetés.*

Il est également possible de prendre en compte l'information provenant des point

non labellisés, en spécifiant la forme de la loi jointe du couple (X, Y) ; c'est l'approche adoptée par les modèles génératifs. Dans ce contexte, c'est le modèle postulé qui définit l'influence des points non labellisés sur la solution obtenue. Lorsqu'un modèle de mélange gaussien est utilisé, ce type de solution se rapproche de l'hypothèse précédente. Chacune des classes (associée aux composantes du mélange) couvre une zone de forte densité. Cependant ces modèles ne supposent pas que celles-ci soient séparées par des zones de faible densité.

hypothèse de la sous-variété

Enfin, une autre hypothèse permet de prendre en compte les points non labellisés. Celle-ci se base sur une idée assez différente et tente de résoudre un autre problème rencontré par les méthodes de classification : le fléau de la dimension (Bellman 1957, Bouveyron 2006). c'est-à-dire la complexification du problème d'apprentissage lorsque la taille de l'espace des descripteurs augmente. Cette hypothèse peut être décrite de la manière suivante :

La population étudiée est distribuée à proximité d'une sous-variété de l'espace des descripteurs.

Les points non labellisés peuvent servir à estimer cette sous-variété, ce qui permet de réduire la dimension de l'espace et potentiellement d'améliorer les performances. La figure 3.2 illustre cette hypothèse où les individus sont tous à proximité d'une courbe, c'est-à-dire d'une sous-variété de dimension 1. Si nous parvenons à estimer celle-ci le problème de classification, au départ dans $\mathbb{R}^3$, peut être reformulé dans $\mathbb{R}$. Bien entendu cet exemple est illustratif, et la projection sur la sous-variété n'est pas indispensable à la résolution du problème de la figure 3.2. Cependant cette projection s'avère précieuse lorsque l'espace des descripteurs est de grande dimension.

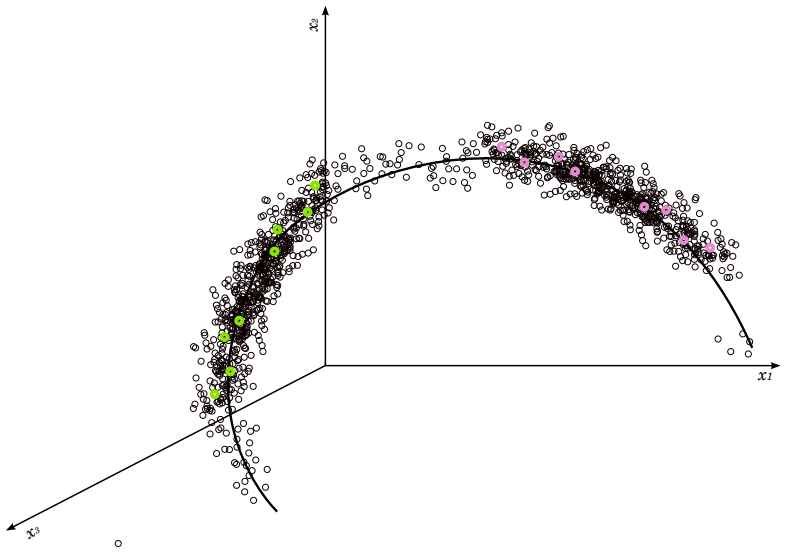


FIG. 3.2 – *Illustration de l'hypothèse de regroupement de tous les individus autour d'une sous-variété de l'espace des descripteurs. Ici la sous-variété est une courbe dans $\mathbb{R}^3$. En vert les exemples de la classe c_1, en violet ceux de la classe c_2, en noir, les exemples non étiquetés.*

Pour tirer partie de cette hypothèse, il est possible d'utiliser des techniques non supervisées de réduction de la dimension en tant que prétraitement. L'ACP qu'elle

soit linéaire ou à noyau, l'ACI, l'analyse en composantes curvilignes (Hérault et al. 1997) et bien d'autres méthodes de projection non linéaires peuvent être utilisées pour déterminer un espace de représentation intéressant à l'aide des points non labellisés.

Nous allons maintenant voir comment ces différentes hypothèses sont prises en compte par les méthodes semi-supervisés de type génératives puis de type discimionatives.

Approche générative

Les modèles génératifs ont été les premiers à être adaptés pour prendre en compte à la fois des données labellisées et non labellisées, (Hosmer 1973, McLachlan 1977). Il est en effet aisé de prendre en considération ce type de données car ces modèles postulent une loi jointe sur (X, Y).

D'un point de vue théorique, si l'hypothèse d'indépendance du processus générant la censure (le fait d'observer le label ou non) et du processus générant les données est faite, la prise en compte de données non labellisées ne pose pas de problème. Il suffit en effet, pour cela, de marginaliser le modèle sur les variables explicatives. La vraisemblance est alors décomposée en deux termes correspondants aux deux parties de l'ensemble d'apprentissage. Dans le cadre d'un modèle de mélange générique, la log-vraisemblance s'écrit :

$$\mathcal{L}(\boldsymbol{\psi}, \mathbf{X}^{ss}) = \sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k) \right) + \sum_{i=M+1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k) \right), \qquad (3.2)$$

où z_{ik} est un vecteur binaire indiquant l'appartenance de l'individus i à la classe k : $z_{ik} = 1$ si $y_i = c_k$, $z_{ik} = 0$ sinon.

Pour estimer les paramètres à partir d'une vraisemblance de cette forme il est possible d'utiliser un algorithme EM dans sa version classique, avec une unique modification qui concerne l'étape E. Lors de cette étape, les distributions a posteriori sur les différentes classes connaissant les variables observées et les paramètres courants ne sont en effet calculées que pour les individus non-labellisés. Lors de l'étape M, les véritables labels sont utilisés à la place des distributions a posteriori pour tous les individus labellisés. En utilisant des arguments classiques, il est aisé de démontrer qu'une telle approche conduit à un maximum local de la vraisemblance (McLachlan 1977).

Les modèles génératifs ont pour principal objectif l'estimation de la loi du couple (X, Y) ; la possibilité de classifier les données est une conséquence de cette estimation. Ils peuvent donc s'avérer sous optimaux en terme de classification lorsque le modèle n'est pas adapté aux données. Dans ce cas de figure, l'ajout de points non labellisés peut même dégrader les performances (Cozman et al. 2003). Ces approches offrent cependant l'opportunité de définir de manière simple des hypothèses sur la manière dont les points non labellisés doivent influencer la solution. Elles ont de plus déjà prouvé leur intérêt pour la résolution de différents problèmes concrets semi-supervisés, par exemple dans le cadre de la classification automatique de texte, (cf. exemple 3.4) (Nigam et al. 2000). Enfin, il est intéressant de

noter, comme les auteurs de ces travaux, la possibilité de pondérer les deux termes intervenant dans la vraisemblance afin de contrôler l'influence des points non labellisés sur la solution. Le critère à optimiser devient alors :

$$\mathcal{L}(\boldsymbol{\psi}, \mathbf{X}^{ss}) = \sum_{i=1}^{M}\sum_{k=1}^{K} z_{ik} \log\left(\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right) + \lambda \sum_{i=M+1}^{N} \log\left(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right), \tag{3.3}$$

où λ est un hyper-paramètre permettant de pondérer l'influence des points non labellisés sur la solution.

Une telle approche permet de donner plus ou moins d'importance aux points non-labellisés et peut ainsi conduire à une amélioration des performances lorsque le modèle est mal spécifié. Elle nécessite cependant de trouver la pondération optimale, laquelle est généralement obtenue à l'aide d'une procédure de validation croisée utilisant uniquement les données labellisées.

L'extension des modèles de mélange au cadre semi-supervisé ne pose pas de problème particulier, comme nous venons de le voir. En effet, ceux-ci postulent un modèle de la loi jointe sur les données et les classes, ce qui permet de prendre en considération l'information en provenance des points non labellisés. Le cadre discriminatif n'est a priori pas aussi bien adapté.

Approche discriminative

Dans un cadre discriminatif, le modèle porte sur la loi conditionnelle des classes connaissant les observations et il est dans ce contexte beaucoup plus difficile de tirer partie de l'information en provenance des points non labellisés. Cependant différentes solutions ont tout de même été proposées. Les méthodes discriminatives, utilisent une vraisemblance conditionnelle, ou un risque empirique pour estimer les paramètres définissant la frontière de décision. Les points non labellisés n'apportent pas d'information sur la localisation de cette frontière sauf si leur influence sur la solution peut être prise en compte par le biais d'un second terme ajouté au critère d'apprentissage.

Ce second terme peut être construit en utilisant l'hypothèse du regroupement et la notion d'entropie conditionnelle (Bengio et Grandvalet 2005), définie par :

Définition 3.1 (Entropie conditionnelle) *L'entropie conditionnelle $H(Y|X)$ de deux variables aléatoires Y définie sur $\mathcal{Y}$ et X définie sur $\mathcal{X}$ est donnée par :*

$$H(Y|X) = -\mathbb{E}[\log(p(Y|X))]\,. \tag{3.4}$$

Lorsque $\mathcal{X}$ est continue et $\mathcal{Y}$ discrète nous obtenons :

$$H(Y|X) = -\int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y,\mathbf{x}) \log\left(p(y|\mathbf{x})\right) .\mathrm{d}\mathbf{x}\,. \tag{3.5}$$

L'entropie conditionnelle mesure donc l'espérance de l'entropie de Y lorsque X est connue. Cette notion permet de quantifier le recouvrement entre les différentes classes, lorsqu'elle est appliquée dans le cadre d'un problème de classification. En

effet, lorsque les classes ont un degré de recouvrement important, les probabilités a posteriori ne sont pas proches de 1 dans les zones de recouvrement et l'entropie conditionnelle est importante, inversement, lorsque les classes sont bien séparées nous avons $p(y|\mathbf{x})$ proche de 1 pour toutes les valeurs de x et l'entropie conditionnelle est alors proche de zéro. En modélisant la densité de X par la distribution empirique il est alors possible de calculer un estimé de l'entropie conditionnelle associée à un modèle discriminant paramétré par $\boldsymbol{\psi}$:

$$H_{emp}(Y|X, \mathbf{X}^{ss}, \boldsymbol{\psi}) = -\frac{1}{N-M} \sum_{i=M+1}^{N} \sum_{k=1}^{K} p(c_k|\mathbf{x}_i; \boldsymbol{\psi}) \log \left(p(c_k|\mathbf{x}_i; \boldsymbol{\psi})\right). \tag{3.6}$$

Un critère discriminatif prenant en considération les points non labellisés peut ainsi être construit (Bengio et Grandvalet 2005) :

$$C(\boldsymbol{\psi}, \lambda, \mathbf{X}^{ss}) = \mathcal{L}_{cond}(\boldsymbol{\psi}; \mathbf{X}^{ss}) - \lambda . H_{emp}(Y|X, \mathbf{X}^{ss}, \boldsymbol{\psi}), \tag{3.7}$$

et donc

$$C(\boldsymbol{\psi}, \lambda, \mathbf{X}^{ss}) = \sum_{i=1}^{M} z_{ik} \log \left(p(c_k|\mathbf{x}_i; \boldsymbol{\psi})\right) + \frac{\lambda}{N-M} \sum_{i=M+1}^{N} \sum_{k=1}^{K} p(c_k|\mathbf{x}_i; \boldsymbol{\psi}) \log \left(p(c_k|\mathbf{x}_i; \boldsymbol{\psi})\right). \tag{3.8}$$

Une fois encore la valeur de λ qui contrôle l'influence des points non labellisés doit être déterminée en utilisant une procédure de validation croisée comme dans le cas des modèles génératifs.

Remarque 3.2 (a priori) *Cette solution au problème de l'apprentissage semi-supervisé peut être présentée d'un point de vue bayésien. L'entropie conditionnelle intervient alors pour la construction d'une densité a priori sur les paramètres de la forme* $p(\boldsymbol{\psi}) \propto exp(-\lambda . H_{emp}(\boldsymbol{\psi}))$.

L'hypothèse de localisation de la frontière dans des zones de faible densité peut aussi être traduite en des termes légèrement différents. En effet, supposer que la frontière de décision doit être placée dans des zones de faible densité est équivalent à supposer que la densité conditionnelle $p(y|\mathbf{x})$ doit varier faiblement dans les zones de forte densité. Une solution s'appuyant sur cette formalisation de l'hypothèse a été proposée par Corduneanu et Jaakkola (2003), les points non labellisés étant utilisés pour estimer la densité marginale sur $\mathcal{X}$: $p(\mathbf{x})$. Cette information permet ensuite de contraindre la densité conditionnelle $p(y|\mathbf{x}; \boldsymbol{\psi})$ à varier « doucement » dans les régions où la densité marginale $p(\mathbf{x})$ est importante.

Des principes géométriques permettent également de prendre en compte les points non labellisés dans une approche discriminative du problème de la classification. Les machines à vecteurs supports transductives (Joachims 1999) utilisent la notion de marge appliquée aux points non labellisés pour déplacer la frontière de décision vers des zones où peu de points de l'ensemble d'apprentissage peuvent être trouvés.

Remarque 3.3 (Convexité du problème d'optimisation) *Toutes ces solutions, qui dans un cadre supervisé mènent à des problèmes d'optimisation convexes (sans extremum*

locaux), conduisent à des problèmes d'optimisation non convexes lorsqu'une régularisation semi-supervisé leur est adjointe. Une exception à cette règle existe cependant et mérite d'être notée ; c'est l'algorithme proposé par De Bie (2005) pour résoudre le problème d'optimisation associé aux machines à vecteurs supports transductives. Cet algorithme relâche légèrement la formulation du problème et permet d'obtenir une solution unique.

Enfin, différentes méthodes d'apprentissage semi-supervisé inspirées de la théorie des graphes ont aussi été proposées récemment (Zhou et al. 2003, Zhu et Ghahramani 2002). Celles-ci se placent dans le cadre transductif et visent à propager l'information en provenance des points labellisés vers les points non labellisés en se servant d'un graphe construit à l'aide d'une mesure de similarité. Ces méthodes peuvent être vues comme implémentant l'hypothèse de regroupement des points autour d'une sous-variété, car la distance induite par de telles méthodes est locale.

Nous allons maintenant voir comment le problème de l'apprentissage partiellement supervisé est formalisé et comment celui-ci a été abordé dans le cadre discriminatif et dans le cadre génératif.

3.1.2 Apprentissage partiellement supervisé

L'apprentissage semi-supervisé n'est pas la formulation la plus générale du problème d'apprentissage. L'apprentissage partiellement supervisé généralise celui-ci en considérant un jeu de données de la forme :

$$\mathbf{X}^{ps} = \{(\mathbf{x}_1, C_1), \ldots, (\mathbf{x}_N, C_N)\}, \tag{3.9}$$

où C_i représente un ensemble de classes possibles pour l'individu i. Il est aisé de voir qu'en posant $C_i = \mathcal{Y}$ le label n'apporte aucune information sur la classe d'origine du point, l'individu est dans ce cas non-labellisé, de même en posant $C_i = c_k$ nous retrouvons une étiquette classique. L'apprentissage semi-supervisé est donc un cas particulier de l'apprentissage partiellement supervisé.

Les modèles de mélange ont été étendus pour pouvoir traiter des problèmes de ce type. La vraisemblance doit pour cela être calculée à partir de la loi jointe de $(\mathbf{x}_i, Y_i \in C_i)$:

$$\mathbb{P}(\mathbf{x}_i, Y_i \in C_i) = \sum_{\{k:c_k \in C_i\}} \mathbb{P}(\mathbf{x}_i, Y_i = c_k) = \sum_{k=1}^{K} \mathbb{P}(Y_i \in C_i \cap \{c_k\}) f(\mathbf{x}_i | Y_i = c_k), \tag{3.10}$$

avec

$$\mathbb{P}(Y_i \in C_i \cap \{c_k\}) = \begin{cases} 0 & \text{si } c_k \notin C_i \\ \pi_k & \text{si } c_k \in C_i. \end{cases} \tag{3.11}$$

Il est intéressant d'introduire ici la notation suivante $\mathbf{l}_i = (l_{i1}, \ldots, l_{iK}) \in \{0,1\}^K$ qui correspond simplement à une indicatrice des labels possibles tels que défini par le sous ensemble $C_i \subseteq \mathcal{Y}$: ($l_{ik} = 1$ si $c_k \in C_i$, $l_{ik} = 0$ sinon). En utilisant cette notation nous pouvons écrire :

$$\mathbb{P}(Y_i \in C_i \cap \{c_k\}) = l_{ik}\pi_k. \tag{3.12}$$

Ce qui permet de réécrire l'équation (3.10) plus simplement, en substituant $\mathbb{P}(Y_i \in \{C_i \cap c_k\})$ par $l_{ik}\pi_k$:

$$\mathbb{P}(\mathbf{x}_i, Y_i \in C_i) = \sum_{k=1}^{K} l_{ik}\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k). \tag{3.13}$$

Grâce à cela, la log-vraisemblance associée à un jeu de données $\mathbf{X}^{ps}$ s'écrit :

$$\mathcal{L}(\boldsymbol{\psi}, \mathbf{X}^{ps}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} l_{ik}\pi_k f(\mathbf{x}_i, \boldsymbol{\theta}_k) \right). \tag{3.14}$$

Une fois encore, l'algorithme EM (cf. algorithme 5) peut être utilisé pour maximiser cette vraisemblance, avec quelques modifications (Ambroise et Govaert 2000, Ambroise et al. 2001). Ces modifications concernent l'étape E de l'algorithme au sein de laquelle les probabilités a posteriori sont calculées en utilisant la relation suivante :

$$t_{ik}^{(q)} = \mathbb{P}\left(Y_i = c_k | \mathbf{x}_i, Y_i \in C_i; \boldsymbol{\psi}^{(q)}\right) = \frac{l_{ik}\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} l_{ik}\pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})}. \tag{3.15}$$

L'algorithme est ensuite strictement identique au cas non supervisé.

Algorithme 5: pseudo-code de l'algorithme EM pour les modèles de mélange dans le cadre partiellement supervisé

Données : Matrice des données : $\mathbf{X}$, labels partiels : $\{\mathbf{l}_i\}_{i=1\ldots N}$
Initialisation

$\boldsymbol{\psi}^{(0)}, q = 0$

tant que *test de convergence* **faire**
  # *Etape E*
  # *calcul des probabilités a posteriori*
  pour tous les $k \in \{1, \ldots, K\}$ **faire**
    $t_{ik}^{(q)} = \dfrac{l_{ik}\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} l_{ik}\pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})}, \quad \forall i \in \{1, \ldots, N\}$
  # *Etape M*
  # *maximisation de la fonction auxiliaire*

  $\boldsymbol{\psi}^{(q+1)} = \arg\max_{\boldsymbol{\psi} \in \Psi} Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik}^{(q)} \log\left(\mathbb{P}(\mathbf{x}_i, Y = c_k; \boldsymbol{\psi})\right)$

  $q = q + 1$

Résultat : Paramètres estimés : $\widehat{\boldsymbol{\psi}}^{ml}$

En ce qui concerne les méthodes discriminatives, différentes solutions on été proposées pour prendre en considération des points labellisés de manière imprécise. Une approche basée sur l'entropie conditionnelle étendant la solution proposée dans le cadre semi-supervisé a en particulier été proposée (Grandvallet 2002, Bengio et Grandvalet 2005).

Enfin, Hüllermeier et Beringer (2005) ont proposé d'étendre différentes méthodes classiques de classification supervisée tel que l'algorithme des k plus proches voi-

sins, les arbres de décision et les méthodes d'induction de règles afin de prendre en considération des points partiellement labellisés.

Nous allons maintenant aborder un autre problème : celui de l'apprentissage en présence d'erreur de labellisation.

3.1.3 Apprentissage en présence d'erreurs de labellisation

bruit d'étiquetage

Lors de la résolution d'un problème de classification, les labels utilisés lors de l'apprentissage sont classiquement supposés ne contenir aucune erreur. L'apprentissage en présence d'erreurs de labellisation lève cette hypothèse. Dans ce cadre, le jeu de données est supposé être :

$$\mathbf{X}^{bl} = \{(\mathbf{x}_1, \tilde{y}_1), \ldots, (\mathbf{x}_N, \tilde{y}_N)\}, \tag{3.16}$$

où $\tilde{y}_i$ est une étiquette pouvant être erronée.

D'un point de vue probabiliste, il est possible de définir un modèle génératif prenant en compte des étiquettes erronées. Celui-ci est présenté en figure 3.3.

Différents travaux ont proposé des solutions au problème de l'apprentissage avec de tels labels (Lawrence et Schölkopf 2001, Amini et Gallinari 2005, Li et al. 2007b). Ces méthodes supposent qu'une certaine proportion des étiquettes est erronée et tente d'estimer cette proportion et de corriger les labels. Lawrence et Schölkopf (2001), Li et al. (2007b) proposent une extension utilisant l'astuce noyau pour construire des classifieurs plus complexes. Amini et Gallinari (2005) proposent quant à eux d'adopter une démarche similaire mais en utilisant un critère discriminatif et en appliquant cette approche a l'apprentissage semi-supervisé.

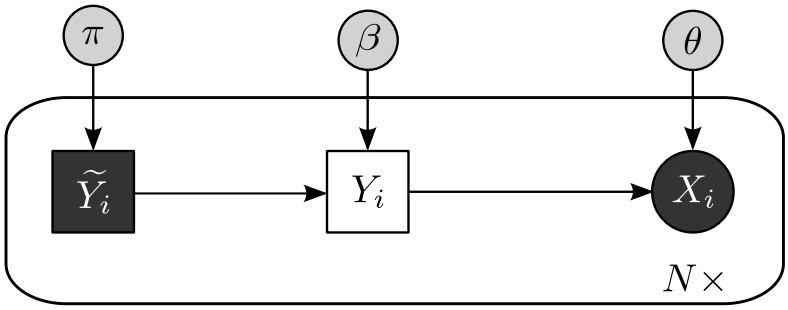


FIG. 3.3 – *Modèle graphique associé au bruit d'étiquetage.*

Ce modèle probabiliste est décrit à l'aide des éléments suivants :

$$\begin{aligned} p(\widetilde{Y} = c_k) &= \pi_k & \forall k \in \{1, \ldots, K\}, \\ p(Y = c_h | \widetilde{Y} = c_k) &= \beta_{kh} & \forall k, h \in \{1, \ldots, K\}, \\ p(\mathbf{x} | Y = c_h) &= f(\mathbf{x}; \boldsymbol{\theta}_h) & \forall h \in \{1, \ldots, K\}, \end{aligned}$$

avec $\sum_{k=1}^{K} \pi_k = 1$, $\sum_{h=1}^{K} \beta_{kh} = 1$ pour tout k. Pour pouvoir utiliser ce modèle, il est donc nécessaire d'estimer, en plus des paramètres classiques décrivant un modèle de mélange, une matrice stochastique contenant les probabilités β_{kh} qu'un individu appartienne véritablement à une classe h sachant qu'il a été étiqueté comme provenant de la classe k. Une approche de type maximum de vraisemblance peut être utilisée. La loi jointe des données observées est donnée par :

$$p(\mathbf{x}, \widetilde{Y} = c_k) = \pi_k \sum_{h=1}^{K} \beta_{kh} f(\mathbf{x}; \boldsymbol{\theta}_h). \tag{3.17}$$

La vraisemblance est donc égale à :

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}^{bl}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \tilde{z}_{ik} \log \left(\pi_k \sum_{h=1}^{K} \beta_{kh} f(\mathbf{x}_i; \boldsymbol{\theta}_h) \right), \tag{3.18}$$

où $\tilde{z}_{ik}$ est la variable binaire indiquant la valeur de $\widetilde{y}_i$, $\tilde{z}_{ik} = 1$ si $\widetilde{y}_i = c_k$; $\tilde{z}_{ik} = 0$ sinon.

Pour optimiser cette vraisemblance, qui une fois encore fait intervenir une marginalisation, il est possible de se tourner vers l'algorithme EM. Pour utiliser celui-ci, il est nécessaire de calculer la vraisemblance des données complétées : $(y_i, \tilde{y}_i, \mathbf{x}_i)$ qui est égale à :

$$\mathcal{L}_c(\boldsymbol{\psi}; \mathbf{y}, \mathbf{X}^{bl}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{h=1}^{K} \tilde{z}_{ik} z_{ih} \log \left(\pi_k \beta_{kh} f(\mathbf{x}_i; \boldsymbol{\theta}_h) \right). \tag{3.19}$$

Grâce à celle-ci, il est aisée de déterminer la fonction auxiliaire manipulée par l'algorithme :

$$\begin{aligned} Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \mathbb{E}[\mathcal{L}_c(\boldsymbol{\psi}; Y, \mathbf{X}^{bl}) | \mathbf{X}, \widetilde{\mathbf{y}}, \boldsymbol{\psi}^{(q)}] && (3.20) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{h=1}^{K} \tilde{z}_{ik}\, t_{ih}^{(q)} \log \left(\pi_k \beta_{kh}\, f(\mathbf{x}_i; \boldsymbol{\theta}_h) \right), && (3.21) \end{aligned}$$

avec :

$$t_{ih}^{(q)} = \mathbb{P}(Y = c_h | \widetilde{Y} = c_k, \mathbf{x}_i, \boldsymbol{\psi}^{(q)}) = \frac{\tilde{z}_{ik} \beta_{kh} . f(\mathbf{x}_i; \boldsymbol{\theta}_h)}{\sum_{h'=1}^{K} \tilde{z}_{ik} \beta_{kh'} . f(\mathbf{x}_i; \boldsymbol{\theta}_{h'})}. \tag{3.22}$$

L'étape E de l'algorithme nécessite donc de calculer ces quantités pour chacun des individus et chacune des classes. En ce qui concerne l'étape de maximisation de la fonction auxiliaire, nous retrouvons les formules usuelles pour les paramètres des lois conditionnelles aux classes.

Par contre pour la matrice stochastique $\boldsymbol{\beta}$, le problème de maximisation est nouveau. Pour prendre en compte les contraintes de normalité des différentes lignes de cette matrice nous devons former le lagrangien associé au problème d'optimisation sous contraintes. Celui-ci prend la forme suivante :

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \left(\sum_{h=1}^{K} \tilde{z}_{ik} t_{ih}^{(q)} \log \left(\beta_{kh} \right) + \lambda_k (1 - \sum_{h=1}^{K} \beta_{kh}) \right), \tag{3.23}$$

En dérivant celui-ci par rapport à chacun des β_{kh} nous obtenons :

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{kh}} = \sum_{i=1}^{N} \frac{\tilde{z}_{ik} t_{ih}^{(q)}}{\beta_{kh}} - \lambda_k. \tag{3.24}$$

En annulant ces dérivées nous obtenons différents systèmes d'équations faisant intervenir chacun un multiplicateur de Lagrange. En utilisant une démarche similaire à celle adoptée lors de la résolution du problème de la maximisation de la fonction Q par rapport aux proportions dans un modèle de mélange classique (cf. annexe .1), nous trouvons $\lambda_k = \sum_{i=1}^{N} \tilde{z}_{ik}$ pour toutes les valeurs possibles de k. Les différents éléments de la matrice doivent donc être mis à jour en utilisant :

$$\beta_{kh}^{(q+1)} = \frac{\sum_{i=1}^{N} \tilde{z}_{ik} t_{ih}^{(q)}}{\sum_{i=1}^{N} \tilde{z}_{ik}}. \tag{3.25}$$

Algorithme 6: pseudo-code de l'algorithme EM pour les modèles de mélange gaussien avec bruit d'étiquetage

Données : Matrice des données : $\mathbf{X}$, labels bruités : $\{\tilde{y}_i\}_{i=1\ldots N}$.
Initialisation

$$\boldsymbol{\psi}^{(0)} = \left(\boldsymbol{\beta}^{(0)}, \boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}, \Sigma_1^{(0)}, \ldots, \Sigma_K^{(0)}\right), q = 0$$

tant que *test de convergence* **faire**
- # *Etape E*
- # *calcul des probabilités a posteriori*
- **pour tous les** $h \in \{1, \ldots, K\}$ **faire**
  - $t_{ih}^{(q)} = \dfrac{\tilde{z}_{ik}\beta_{kh}.f(\mathbf{x}_i; \boldsymbol{\theta}_h)}{\sum_{h'=1}^{K} \tilde{z}_{ik}\beta_{kh'}.f(\mathbf{x}_i; \boldsymbol{\theta}_{h'})}, \quad \forall i \in \{1, \ldots, N\}$
- # *Etape M*
- # *maximisation de la fonction auxiliaire*
- **pour tous les** $h \in \{1, \ldots, K\}$ *et* $k \in \{1, \ldots, K\}$ **faire**
  - $\beta_{kh}^{(q+1)} = \frac{\sum_{i=1}^{N} \tilde{z}_{ik} t_{ih}^{(q)}}{\sum_{i=1}^{N} \tilde{z}_{ik}}$
  - $\boldsymbol{\mu}_h^{(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ih}^{(q)}} \sum_{i=1}^{N} t_{ih}^{(q)} \mathbf{x}_i$
  - $\boldsymbol{\Sigma}_h^{(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ih}^{(q)}} \sum_{i=1}^{N} t_{ih}^{(q)} (\mathbf{x}_i - \boldsymbol{\mu}_h^{(q+1)})^t (\mathbf{x}_i - \boldsymbol{\mu}_h^{(q+1)})$
- $q = q + 1$

Résultat : Paramètres estimés : $\widehat{\psi}^{ml}$

Nous venons de voir comment un algorithme EM pouvait être utilisé pour estimer les paramètres d'un modèle probabiliste intégrant la possibilité que certains labels soient erronées. Il est nécessaire pour cela d'estimer différents paramètres supplémentaires décrivant les probabilités d'inversion des classes.

Après avoir parcouru les différents problèmes pouvant être définis entre les cadres supervisé et non supervisé et en avoir présenté des solutions, nous nous tournons vers un problème encore plus général qui concerne l'apprentissage en présence d'étiquettes aussi bien imprécises qu'incertaines.

3.2 Modèle de mélange et labels doux

Cette section présente nos travaux sur l'extension des méthodes d'apprentissage statistique afin que celles-ci puissent prendre en compte des jeux de données de la forme suivante :

$$\mathbf{X}^{iu} = \{(\mathbf{x}_1, m_1^{\mathcal{Y}}), \ldots, (\mathbf{x}_N, m_N^{\mathcal{Y}})\}, \tag{3.26}$$

où chaque $m_i^{\mathcal{Y}}$ est une fonction de masse sur un ensemble $\mathcal{Y}$ de classes. Cette fonction de masse encode l'information disponible sur la véritable classe des individus. Dans notre proposition les descripteurs $\mathbf{x}_i$ seront supposés avoir été générés indépendamment par un modèle de mélange (cf. section 2.1.3). L'objectif des travaux présentés ici est d'étendre les approches d'estimation par maximum de vraisemblance pour pouvoir prendre en compte des jeux de données imprécis, incertains tel que $\mathbf{X}^{iu}$. L'objectif est donc de trouver un estimé du vecteur de paramètre :

$$\boldsymbol{\psi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K). \tag{3.27}$$

Pour cela, un critère étendant le critère de vraisemblance dans le cadre du modèle des croyances transférables, est tout d'abord dérivé.

Le concept de vraisemblance a des relations importantes avec les notions de possibilité et plus généralement de plausibilité, comme l'ont mis en évidence de nombreux auteurs (Smets 1982, Walley et Moral 1999, Shenoy et Giang 2005, Smets 1998, Monney 2003). De plus, sélectionner le vecteur de paramètres du modèle de mélange ayant la plausibilité maximale connaissant les observations est une stratégie de décision naturelle dans le cadre de la théorie des fonctions de croyance (Cobb et Shenoy 2006). Nous proposons donc comme principe d'estimation, de rechercher le vecteur de paramètres $\{\boldsymbol{\psi}\}$ ayant une plausibilité maximale conditionnellement aux données. Ce principe peut être formalisé par :

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} pl^{\boldsymbol{\Psi}}(\{\boldsymbol{\psi}\}|\mathbf{X}^{iu}). \tag{3.28}$$

Pour alléger les notations nous ne ferons pas de différence entre le singleton $\{\boldsymbol{\psi}\}$ et la valeur $\boldsymbol{\psi}$; la notation $pl^{\boldsymbol{\Psi}}(\{\boldsymbol{\psi}\}|\mathbf{X}^{iu})$ sera donc simplifiée en $pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})$.

Ce principe d'estimation (3.28) mène, dans le cas de figure qui nous intéresse, à une solution naturelle étendant les critères de type vraisemblance associés aux différentes formulations du problème d'apprentissage déjà présentées. Cette méthode conduit en effet au critère suivant :

Proposition 3.1 *Le logarithme de la plausibilité conditionnelle des paramètres $\boldsymbol{\psi}$ d'un modèle de mélange connaissant les données $\mathbf{X}^{iu}$ est donné par*

$$\log\left(pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})\right) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right) + \nu, \tag{3.29}$$

où les pl_{ik} sont les plausibilités de chaque classes k pour chaque exemple i en accord avec les labels m_i et ν est une constante indépendante de la valeur de $\boldsymbol{\psi}$.

Preuve. En utilisant le théorème de Bayes généralisé (2.88), la plausibilité des paramètres peut être exprimée en fonction de la plausibilité des observations :

$$pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}) = pl^{\mathcal{X}_1\times\ldots\times\mathcal{X}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N|\boldsymbol{\psi}). \tag{3.30}$$

En supposant que les fonctions de plausibilité de chacune des observations sont cognitivement indépendantes conditionnellement aux paramètres (2.85), cette plausibilité peut être décomposée en un produit sur l'ensemble des exemples d'apprentissage :

$$pl^{\mathbf{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}) = \prod_{i=1}^{N} pl^{\mathcal{X}_i}(\mathbf{x}_i|\boldsymbol{\psi}). \tag{3.31}$$

Par le théorème des plausibilités totales (2.81), la plausibilité d'une observation $pl^{\mathcal{X}_i}(\mathbf{x}_i|\boldsymbol{\psi})$ connaissant la valeur du vecteur de paramètres peut s'écrire :

$$pl^{\mathcal{X}_i}(\mathbf{x}_i|\boldsymbol{\psi}) = \sum_{C\subseteq\mathcal{Y}} m^{\mathcal{Y}_i}(C|\boldsymbol{\psi})\, pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|C,\boldsymbol{\psi}), \tag{3.32}$$

où $m^{\mathcal{Y}_i}(.|\boldsymbol{\psi})$ est une fonction de masse représentant notre croyance sur la classe de l'individu i. Cette fonction de masse provient de la combinaison de deux sources d'information : les labels « doux » $m_i^{\mathcal{Y}}$ et les proportions $\boldsymbol{\pi}$ des différentes classes. Ces proportions induisent une fonction de masse $m^{\mathcal{Y}}(\,\cdot\,|\boldsymbol{\pi})$ avec $m^{\mathcal{Y}}(\{c_k\}|\boldsymbol{\pi}) = \pi_k$ pour tout $c_k \in \mathcal{Y}$. En supposant que ces deux sources d'information sont distinctes, celles-ci peuvent être combinées conjonctivement (2.77) pour obtenir une seule fonction de masse sur $\mathcal{Y}_i$:

$$m^{\mathcal{Y}_i}(\,\cdot\,|\boldsymbol{\psi}) = m_i^{\mathcal{Y}} \textcircled{\cap} m^{\mathcal{Y}}(\,\cdot\,|\boldsymbol{\pi}).$$

Comme $m^{\mathcal{Y}}(\,\cdot\,|\boldsymbol{\pi})$ est une fonction de masse bayésienne, il en va de même pour le résultat de la combinaison $m^{\mathcal{Y}_i}(\,\cdot\,|\boldsymbol{\psi})$. Ce qui donne :

$$\begin{aligned} m^{\mathcal{Y}_i}(\{c_k\}|\boldsymbol{\psi}) &= \sum_{C\cap c_k\neq\emptyset} m_i^{\mathcal{Y}}(C) m^{\mathcal{Y}}(\{c_k\}|\boldsymbol{\pi}) = pl_{ik}\,\pi_k, \quad \forall k \in \{1,\ldots,K\} \\ m^{\mathcal{Y}_i}(C|\boldsymbol{\psi}) &= 0, \quad \forall C \subseteq \mathcal{Y} \text{ tel que } |C| > 1. \end{aligned} \tag{3.33}$$

De par cette propriété les seuls termes non nuls dans l'équations (3.32), correspondent aux sous ensembles C de $\mathcal{Y}$ qui ont un cardinal égal à 1. En conséquence il suffit de définir $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|c_k,\boldsymbol{\psi})$ pour $k \in \{1,...,K\}$.

Une difficulté apparaît cependant ici, en effet $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\,\cdot\,|c_k,\boldsymbol{\psi})$ est la fonction de plausibilité associée à la mesure de probabilité $f(\mathbf{x};\boldsymbol{\theta}_k)$ et la plausibilité d'une observation devient nulle si $\mathbf{x}_i$ est donné avec une précision infinie. Pour solutionner ce problème, il est nécessaire de considérer que $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|c_k,\boldsymbol{\psi})$ représente la plausibilité d'une région infinitésimale autour de $\mathbf{x}_i$ ayant pour volume $\mathrm{d}x_{i1}\ldots\mathrm{d}x_{iP}$ (où P est la dimension de l'espace des descripteurs). Dans ce cas de figure, la plausibilité d'une observation connaissant sa classe d'origine peut être approximée par le produit de la densité et du volume de la région entourant $\mathbf{x}_i$.

$$pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|c_k,\boldsymbol{\psi}) = f(\mathbf{x}_i;\boldsymbol{\theta}_k)\mathrm{d}x_{i1}\ldots\mathrm{d}x_{iP}. \tag{3.34}$$

En utilisant les différentes expressions dérivées jusqu'à maintenant c'est-à-dire (3.33,3.34), la plausibilité d'une observations (3.32) peut être écrite :

$$pl^{\mathcal{X}_i}(\mathbf{x}_i|\boldsymbol{\psi}) = \left(\sum_{k=1}^{K} pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right)\mathrm{d}x_{i1}\ldots\mathrm{d}x_{iP}. \tag{3.35}$$

En remplaçant dans (3.31) nous obtenons pour un jeu de données :

$$pl^{\mathbf{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}) = \prod_{i=1}^{N}\left[\left(\sum_{k=1}^{K} pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right)\mathrm{d}x_{i1}\ldots\mathrm{d}x_{iP}\right]. \tag{3.36}$$

Les termes $\mathrm{d}x_{ip}$ peuvent finalement être considérés comme des constantes multiplicatives qui n'affectent pas le problème d'optimisation. En prenant le logarithme de l'expression précédente (3.36), nous obtenons le résultat final (3.29). □

Remarque 3.4 *La fonction de masse m_i définissant le label de d'exemple i n'apparaît pas sous la forme d'une fonction de masse dans le critère que nous venons de définir (3.29), mais au travers des plausibilités des différentes classes pl_{ik}. Les fonctions de masses possédant le même profil de plausibilités (i.e., pour lesquelles les plausibilités des singletons sont identiques) sont donc traitées de la même manière par notre critère. Cette invariance provient de la nature probabiliste du modèle postulé. Le pouvoir expressif des fonctions de croyance n'est donc pas complètement utilisé ici. Cependant, chaque étiquette douce peut ainsi être décrite à l'aide de $|\mathcal{Y}|$ paramètres (à comparer aux $2^{\mathcal{Y}}$ valeurs nécessaires dans le cas d'une fonction de croyance quelconque), ce qui bien évidement allège la résolution.*

Remarque 3.5 *Dans la suite de ce mémoire nous noterons ce critère de la même manière qu'une fonction de vraisemblance :*

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{\psi}; \mathbf{X}^{iu}), \tag{3.37}$$

avec :

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}^{iu}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} pl_{ik}\, \pi_k\, f(\mathbf{x}_i; \boldsymbol{\theta}_k) \right). \tag{3.38}$$

Ce choix de notation est justifié par le fait que notre critère (3.38) étend le critère du maximum de vraisemblance dans le cas non supervisé (2.22), semi-supervisé (3.2) et partiellement supervisé (3.14) :

- *quand tous les labels « doux » m_i sont des fonctions de masse vide, nous avons $pl_{ik} = 1$, $\forall i, k$, et le critère non supervisé (2.22) est obtenu ;*
- *dans le cas semi-supervisé nous avons :*

$$pl_{ik} = \begin{cases} z_{ik}, & \forall i \in \{1, \ldots, M\},\ \forall k \\ 1, & \forall i \in \{M+1, \ldots, N\},\ \forall k, \end{cases}$$

  et notre critère (3.38) est équivalent à (3.2) ;
- *finalement, le critère partiellement supervisé (3.14) est retrouvé quand les étiquettes « douces » sont des fonctions de masses catégoriques ; dans ce cas, en reprenant les notations de la section précédente $pl_{ik} = l_{ik}$, $\forall i, k$.*

Nous avons ainsi dérivé à partir de principes généraux un critère permettant d'estimer les paramètres d'un modèle de mélange dans un contexte où l'information sur la classe d'origine des individus est imprécise et incertaine et prend la forme d'une fonction de masse. Nous allons voir comment un algorithme EM peut être mis au point pour optimiser ce critère. Il est intéressant de noter que grâce à cet algorithme, il est possible de traiter différentes situations, celle de l'apprentissage semi-supervisé, partiellement supervisé, non supervisé, supervisé et finalement la situation où les labels sont décrits par des fonctions de masse de croyance. Tous ces cas de figures seront traités identiquement par le même algorithme, la seule différence provenant des étiquettes fournies en entrée.

3.3 L'ALGORITHME EM POUR L'ESTIMATION DES PARAMÈTRES

Pour construire un algorithme EM capable d'optimiser $\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}^{iu})$, le chemin que nous allons emprunter suit la démarche classique rencontrée dans le cadre probabiliste (cf. section 2.1.1). L'étape E consiste essentiellement à utiliser l'information actuellement disponible c'est-à-dire les données observées mais aussi l'estimé courant des paramètres pour construire une distribution de probabilité sur les valeurs pouvant être prises par les variables latentes pour chacun des individus. Dans le cadre des labels « doux » ces deux sources d'informations sont complétées par une troisième source, les labels. A l'itération q, notre connaissance sur la classe d'origine d'un individu i provient donc de trois sources :

1. le label « doux » $m_i^{\mathcal{Y}}$ associé à l'exemple i ;
2. l'estimé courant des proportions $\boldsymbol{\pi}^{(q)}$, qui induit une fonction de masse bayésienne $m^{\mathcal{Y}}(\,\cdot\,|\boldsymbol{\pi}^{(q)})$ avec $m^{\mathcal{Y}}(\{c_k\}|\boldsymbol{\pi}^{(q)}) = \pi_k^{(q)}$;
3. le descripteur de l'individu $\mathbf{x}_i$ et l'estimé courant des paramètres $\boldsymbol{\psi}^{(q)}$ qui en utilisant (3.34) et le théorème de Bayes généralisé (2.88), fournissent une fonction de plausibilité sur $\mathcal{Y}_i$:

$$pl^{\mathcal{Y}_i|\mathcal{X}_i}(\{c_k\}|\mathbf{x}_i, \boldsymbol{\psi}) = pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|c_k, \boldsymbol{\psi}) = f(\mathbf{x}_i; \boldsymbol{\theta}_k)\mathrm{d}x_{i1} \ldots \mathrm{d}x_{iP}.$$

L'étape E consiste comme nous l'avons dit à construire une distribution de probabilité sur $\mathcal{Y}_i$ à partir de ces trois sources d'information. En supposant que toutes ces sources sont cognitivement indépendantes, elles peuvent être combinées par la règle de Dempster (2.78) pour obtenir une unique fonction de masse. Comme $m^{\mathcal{Y}}(\,\cdot\,|\boldsymbol{\pi}^{(q)})$ est une fonction de masse bayésienne, il en sera de même pour le résultat de la fusion. En notant $t_{ik}^{(q)}$ la masse affectée à $\{c_k\}$ après combinaison, le résultat de la fusion est donné par :

$$t_{ik}^{(q)} = \frac{pl_{ik}\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} pl_{ik'}\pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})}, \tag{3.39}$$

Cette expression est très proche de celle utilisée dans le cadre non supervisé (2.26) ainsi que dans le cadre partiellement supervisé (3.15). Nous allons voir qu'en utilisant cette expression il est possible de décomposer $\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}^{iu})$ en deux termes comme dans le cadre probabiliste (2.8) :

Proposition 3.2

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}^{iu}) = Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) - H(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}). \tag{3.40}$$

avec :

$$\begin{aligned} Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log\left(pl_{ik}\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)\right) \\ H(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log(t_{ik}). \end{aligned}$$

Preuve :

$$
\begin{aligned}
Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) - H(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log\left(pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right) - \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log(t_{ik}) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log\left(\frac{pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)}{pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)} \sum_{k'=1}^{K} pl_{ik'}\pi_{k'} f(\mathbf{x}_i;\boldsymbol{\theta}_{k'})\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log\left(\sum_{k'=1}^{K} pl_{ik'}\pi_{k'} f(\mathbf{x}_i;\boldsymbol{\theta}_{k'})\right) \\
&= \sum_{i=1}^{N} \log\left(\prod_{k=1}^{K}\left(\sum_{k'=1}^{K} pl_{ik'}\pi_{k'} f(\mathbf{x}_i;\boldsymbol{\theta}_{k'})\right)^{t_{ik}^{(q)}}\right) \\
&= \sum_{i=1}^{N} \log\left(\left(\sum_{k'=1}^{K} pl_{ik'}\pi_{k'} f(\mathbf{x}_i;\boldsymbol{\theta}_{k'})\right)^{\sum_{k=1}^{K} t_{ik}^{(q)}}\right) \\
&= \sum_{i=1}^{N} \log\left(\sum_{k'=1}^{K} pl_{ik'}\pi_{k'} f(\mathbf{x}_i;\boldsymbol{\theta}_{k'})\right) \\
&= \mathcal{L}(\boldsymbol{\psi}; \mathbf{X}^{iu}).
\end{aligned}
$$

□

La fonction H est exactement de la même forme que dans le cas classique et possède donc les mêmes propriétés. En utilisant exactement les mêmes arguments que dans le cadre probabiliste (cf. section 2.1.1), un algorithme alternant le calcul des quantités t_{ik} en utilisant (3.39) et la maximisation de la fonction Q fera croître la vraisemblance. L'algorithme EM optimisant le critère (3.37) est par conséquent l'algorithme classiquement rencontré dans le cadre des modèles de mélange excepté lors de l'étape E où les t_{ik} sont pondérées par les plausibilités des différentes classes $pl_{ik} = \sum_{C\cap c_k\neq\emptyset} m_i(C)$ tel que définies par les labels « doux ». L'impact des labels « doux » sur cet algorithme a donc une interprétation naturelle.

Discussion

L'algorithme théorique que nous venons de présenter peut aisément être implémenté, en prenant les précautions classiques pour éviter tout problème numérique. Ces précautions concernent en particulier le stockage et la manipulations des probabilités a posteriori t_{ik} sur une échelle logarithmique pour des considérations de précisions. Cet algorithme peut être adapté à différentes formes de densité conditionnelle. L'algorithme 7 fournit par exemple le pseudo-code dans le cas ou celles-ci sont gaussiennes.

D'autre part, d'un point de vue pratique, il est intéressant de noter que des labels « doux » permettent de résoudre l'un des points durs rencontrés lors de la mise en place d'un algorithme EM, à savoir son initialisation, qui a une influence non négligeable sur la qualité des résultats obtenus. Pour cela, il est possible d'utiliser la transformation pignistique (2.15) de chacun des labels m_i, pour initialiser les

Algorithme 7: pseudo-code de l'algorithme EM pour les modèles de mélange gaussien avec des labels « doux »

Données : Matrice des données : $\mathbf{X}$,
labels doux : $\{m_i\}_{i=1\dots N}$ ou $\{pl_i\}_{i=1\dots N}$.

 # Initialisation

$pl_{ik} = \sum_{C \cap c_k \neq \emptyset} m_i(C)$

$\boldsymbol{\psi}^{(0)} = \left(\pi_1^{(0)}, \dots, \pi_K^{(0)}, \boldsymbol{\mu}_1^{(0)}, \dots, \boldsymbol{\mu}_K^{(0)}, \Sigma_1^{(0)}, \dots, \Sigma_K^{(0)}\right), q = 0$

tant que *test de convergence* **faire**

 # Etape E

 # calcul des fonctions de masse courantes sur $\mathcal{Y}$

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

$$t_{ik}^{(q)} = \frac{pl_{ik}\pi_k^{(q)}\boldsymbol{\varphi}(\mathbf{x}_i; \boldsymbol{\mu}_k^{(q)}, \boldsymbol{\Sigma}_k^{(q)})}{\sum_{k'=1}^{K} pl_{ik'}\pi_{k'}^{(q)}\boldsymbol{\varphi}(\mathbf{x}_i; \boldsymbol{\mu}_{k'}^{(q)}, \boldsymbol{\Sigma}_{k'}^{(q)})}, \quad \forall i \in \{1, \dots, N\}$$

 # Etape M

 # maximisation de la fonction auxiliaire

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

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

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{(q)}} \sum_{i=1}^{N} t_{ik}^{(q)} \mathbf{x}_i$$

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{(q)}} \sum_{i=1}^{N} t_{ik}^{(q)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})^t (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})$$

$q = q + 1$

Résultat : Paramètres estimés : $\widehat{\boldsymbol{\psi}}^{ml}$

probabilités a posteriori t_{ik}. c'est-à-dire de poser :

$$t_{ik}^{(0)} = \sum_{C : c_k \in C} \frac{m_i^{\mathcal{Y}}(C)}{|C|}.$$

Il est aussi important de noter que la prise en compte de l'information en provenance des labels « doux » affecte l'étape E de l'algorithme et ne modifie aucunement l'étape M. Cette propriété permet d'envisager de nombreuses extensions à l'algorithme en particulier dans le cas gaussien. En effet, différentes améliorations (modèles parcimonieux, mélange d'ACP, HDDC) ont été évoquées (cf. section 2.1.3) pour traiter le problème des espaces de descripteurs de grande taille. Ces modifications affectent généralement l'étape M de l'algorithme et peuvent donc aisément être intégrées au sein de notre solution.

En ce qui concerne la complexité de cet algorithme, elle est identique à celle rencontrée dans le cadre non-supervisé. Des différences peuvent cependant être observées en ce qui concerne le nombre d'itérations nécessaire pour converger, en fonction de la qualité des labels. Des expérimentations présentées dans la section suivante permettront d'ailleurs d'éclairer ce point tout comme l'influence de la précision des labels sur le nombre de maximum locaux de la vraisemblance.

3.3.1 Liens avec des travaux précédents

L'idée d'adapter l'algorithme EM pour prendre en considération des labels « doux » a d'abord été introduite dans les travaux de Vannoorenberghe et Smets (2005), Vannoorenberghe (2007). Ils ont été repris pour traiter des données catégorielles par Jraidi et Elouedi (2007). Ces auteurs proposent une variante de l'algorithme EM, nommée CrEM (pour EM crédal), qui utilise une fonction auxiliaire $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)})$ modifiée pour prendre en compte des labels décrits par des fonctions de masse. L'objectif de ces travaux est donc similaire aux notres. Cependant notre solution diffère de ces travaux par plusieurs points.

Tout d'abord, l'algorithme CrEM n'est pas défini à partir d'un critère étendant le critère de maximum de vraisemblance tel que (3.38). En conséquence, l'interprétation des résultats obtenus n'est pas aussi aisée. Les liens entre ce type d'approche et les solutions existantes dans le cadre probabiliste (cf. remarque 3.5) n'a d'autre part pas pu être mis en évidence par ces travaux. Finalement, la convergence de l'algorithme vers un maximum local du critère n'a pas non plus pu être démontrée.

D'autre part, dans notre proposition, les labels « doux » $m_i^{\mathcal{Y}}$ apparaissent dans le critère (3.29) et dans la formule de mise à jour de l'étape E de l'algorithme EM (3.39) sous la forme des plausibilités des différentes classes pl_{ik}. Les plausibilités ne sont pas utilisées dans l'algorithme CrEM qui utilise des labels décrits à l'aide de $2^{\mathcal{Y}}$ valeurs, pour chaque $m_i^{\mathcal{Y}}$. Cette différence a une conséquence importante : les calculs nécessaires lors de l'étape E de l'algorithme CrEM sont de complexité $O(N.2^{\mathcal{Y}})$ alors que dans notre proposition, ces calculs ne nécessitent que $O(N.|\mathcal{Y}|)$ opérations. Lorsque le nombre de classes est important, cette différence a un impact non négligeable sur le temps de calcul.

3.4 Expérimentations

Pour pouvoir analyser le comportement de notre algorithme lorsque des labels « doux » sont utilisés, différentes expériences ont été effectuées. Nous avons tout d'abord essayé de mettre en évidence et de quantifier l'influence de la précision des labels sur la complexité du problème d'optimisation ainsi que sur la précision de l'estimation. Nous avons aussi voulu observer les bénéfices qui peuvent découler de l'usage de labels doux lorsque l'étiquetage est incertain, c'est-à-dire lorsque certains labels sont erronés. Ces expériences nous ont en particulier permis de mettre en évidence l'apport des labels « doux » dans ce contexte pour représenter une information sur la fiabilité des labels et améliorer ainsi de manière significative les performances.

3.4.1 Influence de la précision des labels

L'objectif de l'expérience décrite ici est de mettre en évidence l'influence de la précision des labels d'une part sur la difficulté du problème d'estimation et d'autre part sur la qualité des estimés obtenus. Pour ce faire, nous avons utilisé des données simulées pour lesquelles nous avons créé différents ensembles d'étiquettes

plus ou moins précises. A partir de ces différents éléments nous avons pu mesurer l'influence de la qualité des étiquettes sur le taux d'erreur de classification, sur le nombre de maximum locaux trouvés par l'algorithme EM, ainsi que sur le nombre d'itérations nécessaire à celui-ci pour converger. L'analyse des taux d'erreur de classification nous permettra de tirer des conclusions quant à l'influence de la qualité des labels sur la précision de l'estimation. L'analyse du nombre d'itération de l'algorithme EM et du nombre de minimum locaux, nous renseigneront sur l'impact des labels sur la complexité du problème d'optimisation. Avant de décrire en détails les conditions expérimentales utilisées, nous présentons la méthode utilisée pour créer différents ensembles d'étiquettes plus ou moins précises.

Génération d'étiquettes « douces »

Avant toute chose, nous devons déterminer comment quantifier la précision d'une étiquette, c'est-à-dire la quantité d'information apportée par cette étiquette sur la véritable classe de l'individu considéré. Différentes mesure d'incertitude existent dans le cadre de la théorie des fonctions de croyance (Klir et Wierman 1998). Parmi celles-ci la non spécificité permet de quantifier l'imprécision d'une fonction de masse et possèdent de nombreuses propriétés intéressantes (Ramer 1987). La définition de cette quantité est la suivante :

non spécificité

Définition 3.2 (Non spécificité) *La non spécificité NS d'une fonction de croyance m définie sur $\mathcal{Y}$ est donnée par :*

$$NS(m^{\mathcal{Y}}) = \sum_{C \subseteq \mathcal{Y}} m^{\mathcal{Y}}(C) \log_2(|C|). \tag{3.41}$$

Lorsque la fonction de masse est totalement précise, c'est-à-dire lorsque le label correspond une étiquette d'apprentissage supervisé, $NS(m)$ est égal à 0. Dans le cas où le label est vide, c'est-à-dire lorsque $\mathcal{Y}$ possède une masse de 1, (apprentissage non supervisé), la non spécificité est maximale et égale à $\log_2(|\mathcal{Y}|)$.

La non spécificité moyenne d'un ensemble d'étiquettes permet donc de quantifier la précision moyenne de celles-ci. En faisant varier cette non spécificité moyenne, il est possible de simuler des jeux de données avec des labels plus ou moins précis. C'est ce que nous avons fait dans cette expérience en faisant varier la non spécificité moyenne des labels que nous noterons ns entre 0.05 et 0.95 dans un problème de classification binaire.

Pour simuler des labels ayant une non spécificité moyenne fixée, nous avons tiré aléatoirement pour chaque exemple d'apprentissage, une non spécificité en utilisant une loi uniforme sur $[ns - 0.05, ns + 0.05]$. Une fois cette non spécificité fixée, nous avons pu déterminer la plausibilité de chacune des deux classes.

Imaginons par exemple qu'une valeur de non spécificité égale à 0.3 ait été tirée pour un individu et que la classe de cet individu soit k^*. Nous allons voir comment construire un profil de plausibilité ayant une non spécificité de 0.3. Comme seules les plausibilités des différentes classes nous intéressent, nous pouvons tout d'abord contraindre la fonction de masse à être consonante[1]. Comme le problème de classi-

[1] Se dit d'une fonction de masse présentant des ensembles focaux emboîtés.

fication est supposé binaire nous devons donc déterminer les masses associées à c_{k^*} et $\mathcal{Y}$. Nous supposons en effet, dans le cadre de cette expérience, que la véritable classe de l'individu possède la plausibilité la plus grande. Ce qui est équivalent à dire qu'il n'y a pas d'erreurs de labellisation. Cette hypothèse sera relâchée dans les expériences suivantes. Dans le contexte de cette expérience, comme le nombre de classes est égal à 2 nous obtenons en utilisant la définition de la non-spécificité donnée précédemment (3.41),

$$m^{\mathcal{Y}}(\{c_{k^*}\}) \times 0 + m^{\mathcal{Y}}(\mathcal{Y}) \times 1 = 0.3, \tag{3.42}$$

c'est-à-dire $m^{\mathcal{Y}}(\mathcal{Y}) = 0.3$. En utilisant le fait que la somme des masses des ensembles focaux d'une fonction de masse est égal à 1, nous trouvons pour la masse de la véritable classe : $m^{\mathcal{Y}}(\{c_{k^*}\}) = 1 - 0.3 = 0.7$. L'étiquette est donc définie dans cet exemple par $pl_{ik^*} = 1$ et $pl_{ik} = 0.3$ pour $k \neq k^*$, il est aisé de vérifier que cette étiquette possède bien une non spécificité de 0.3.

Influence de la précision des labels sur la qualité de l'estimation des paramètres

En utilisant cette démarche nous avons tout d'abord analysé l'influence de la qualité des labels sur la précision de l'estimation. Pour cela, les simulations suivantes ont été effectuées. Nous avons tout d'abord simulé N exemples d'apprentissage dans un espace de dimension 10 en utilisant un modèle de mélange gaussien à deux composantes. Deux tailles d'ensemble d'apprentissage ont été utilisées $N \in \{1000, 2000\}$. Les matrices de variance-covariance des deux composantes ont été considérées comme identiques et égales à la matrice identité. Les proportions de ces deux composantes ont elles aussi été considérées comme identiques et sont donc égales à 0.5. La distance entre les centres des deux composantes $\delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||$ a en revanche été modifiée entre les différentes expériences $\delta \in \{1, 2, 4\}$, pour pouvoir étudier différents cas, d'un modèle de mélange avec classes extrêmement mélangées, jusqu'à un modèle de mélange avec des classes bien séparées.

Pour chacun des exemples d'apprentissage ainsi générés, une non spécificité a ensuite été tirée comme expliqué précédemment, c'est-à-dire en utilisant une loi uniforme sur $[ns - 0.05, ns + 0.05]$, la non spécificité moyenne étant modifiée pour chacune des expériences $ns \in \{0.05, 0.1, 0.15, \ldots, 0.95\}$. Finalement, le label de l'individu considéré a été construit comme dans l'exemple précédent.

Les différents paramètres qui varient entre les expériences sont donc la taille du jeu de données $N \in \{1000, 2000\}$, la non spécificité moyenne des labels $ns \in \{0.05, 0.1, 0.15, \ldots, 0.95\}$ et la distance entre les centres des deux composantes $\delta \in \{1, 2, 4\}$. Le nombre de situations testées est donc égal à $2 \times 19 \times 3 = 114$.

Finalement pour éviter que les résultats ne dépendent de propriétés fortuites des jeux de données, toutes les expériences ont été effectuées 200 fois, et nous présentons les résultats sous la forme de boxplots. La mesure de qualité de l'estimation utilisée dans cette expérience est le taux d'erreur de classification estimé grâce à un ensemble de test de 5000 exemples. Dans cette expérience, la forme de modèle postulé est compatible avec celui ayant servi à simuler les données, les différences observées dans les taux d'erreur de classification sont donc uniquement dues à des différences de qualité d'estimation.

Finalement un dernier point mérite d'être détaillé pour décrire notre protocole expérimental ; celui-ci concerne la procédure d'initialisation de l'algorithme EM. En effet, pour éviter le problème des maximum locaux auxquels nous nous intéresserons dans l'expérience suivante, nous avons initialisé l'algorithme en utilisant les paramètres de la véritable distribution. De cette manière, l'algorithme EM convergera vers le maximum global et cet aspect du problème n'influencera pas les résultats obtenus.

La figure 3.4 présente les boxplots des résultats obtenus (taux d'erreur de classification) sur les 200 jeux de données simulées pour chacune des configurations expérimentales considérées (N, ns, δ). Cette figure montre clairement l'influence de la précision des étiquettes lorsque le problème est difficile, c'est-à-dire lorsque les classes sont mélangées ($\delta = 1, 2$). Ce résultat étant d'autant plus visible que la taille du jeu de données d'apprentissage est petite ($N = 1000$).

En revanche, lorsque le problème est simple (classes bien séparées et taille de l'ensemble d'apprentissage suffisante) la précision des labels ne semble pas affecter les résultats. Ces constatations sont en accord avec les résultats théoriques établis dans le contexte semi-supervisé (O'Neill 1978) qui ont montré que la quantité d'information apportée par des individus non labellisés dépendait de la distance entre les centres des différentes classes intervenant dans le problème.

Ces résultats montrent également que notre solution peut exploiter une information imprécise sur la classe des individus pourvus que celle-ci soit consistante avec la réalité. En effet, dans ces expériences, la véritable classe de chaque individu a toujours une plausibilité plus élevée que les autres. Cette hypothèse sera levée dans le dernier jeu d'expérimentations de ce chapitre.

Nous allons maintenant voir que la précision des labels a aussi une influence sur le paysage de vraisemblance, et donc sur la difficulté du problème d'optimisation qui est associée à l'estimation des paramètres.

Influence de la précision des labels sur le paysage de vraisemblance

L'expérience décrite ici, a pour but de montrer l'influence de la précision des labels sur le problème d'optimisation associé à l'estimation des paramètres. Nous avons étudié pour cela deux indicateurs de la complexité du problème d'optimisation : le nombre d'itérations nécessaire à l'algorithme EM pour converger vers un maximum local et le nombre de maximum locaux trouvés.

Le comportement de ces deux indicateurs a été analysé, comme dans l'expérience précédente, en fonction de la non spécificité moyenne des labels. Nous avons pour cela simulé $N = 1000$ exemples d'apprentissage en utilisant le même modèle que dans l'expérience précédente avec $\delta = 2$. La non spécificité moyenne des jeux de données simulés varie dans cette expérience entre 0.1 et 0.9. Pour tous les jeux de données ainsi générés, un algorithme EM a été lancé à partir de 200 initialisations aléatoires. Nous avons calculé le nombre moyen d'itérations avant convergence ainsi que le nombre de maxima locaux différents[2] trouvés pour ces 200 initialisa-

[2]Deux maxima locaux on été considérés comme différents lorsque la distance entre leurs vecteurs de paramètres était supérieure à 0.001.

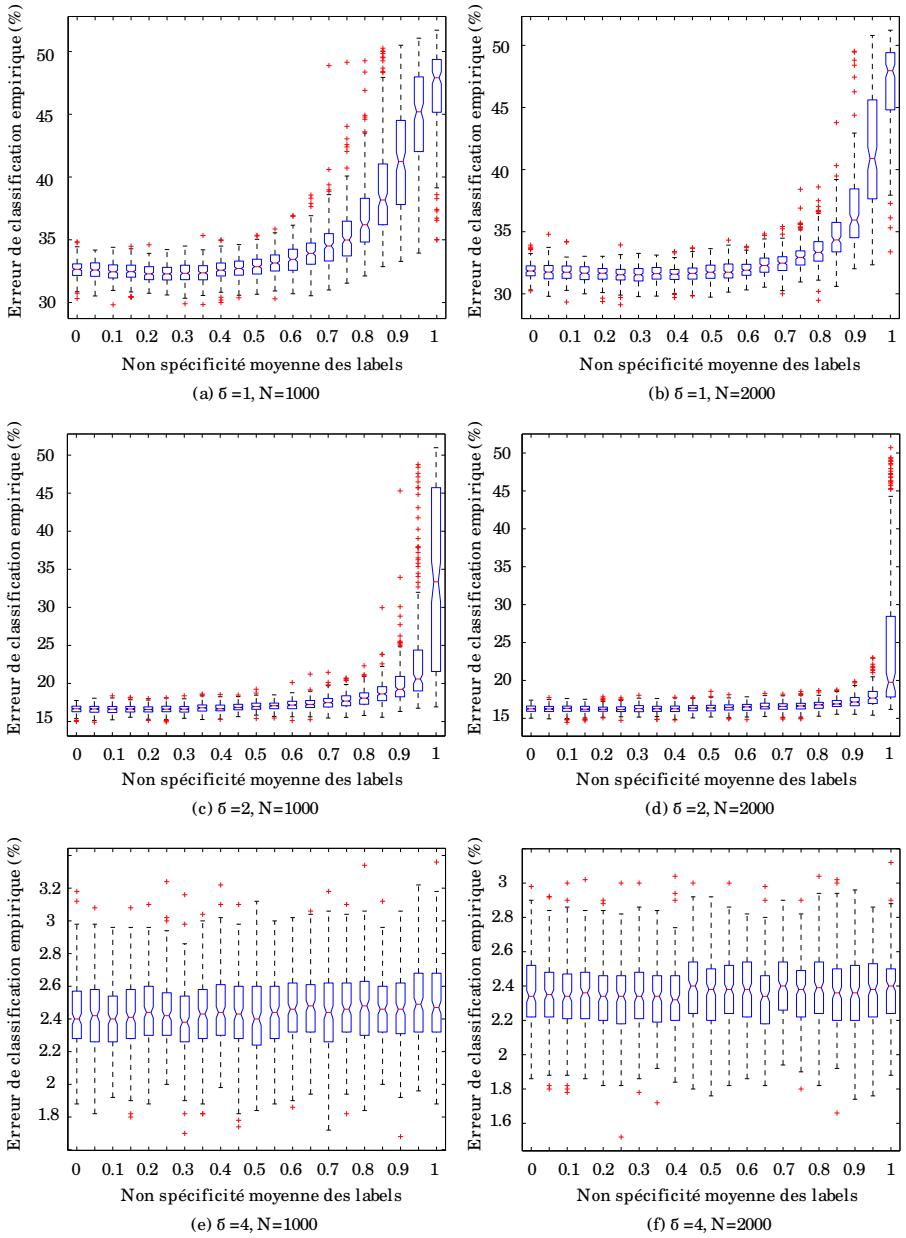


FIG. 3.4 – *Influence de la précision des labels sur la qualité de l'estimation des paramètres : boxplots de l'erreur de classification empirique (%) calculée à partir de 200 ensembles d'apprentissage indépendants, en fonction de la non spécificité moyenne des labels de l'ensemble d'apprentissage* $ns \in \{0.5, \ldots 0.95\}$*; les résultats de l'apprentissage supervisé* $ns = 0$ *et de l'apprentissage non supervisé* $ns = 1$ *sont également fournis à titre de comparaison. Finalement, ces résultats sont présentés pour deux tailles d'ensemble d'apprentissage (*$N = 1000$ *ou* $N = 2000$*) et pour différentes distances entre les centres des deux gausssiennes (*$\delta \in \{1, 2, 4\}$*).*

tions. Cette expérience a été répétée 10 fois avec des jeux de données différents et les résultats ont été moyennés. La figure 3.5 présente ces résultats.

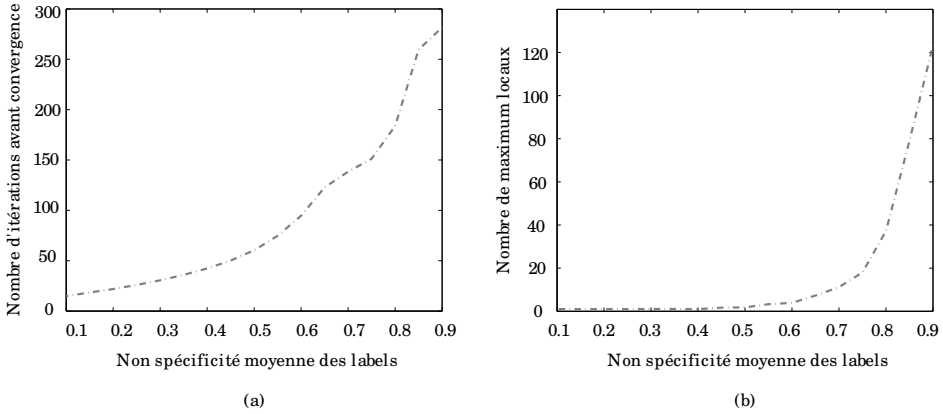


FIG. 3.5 – *Influence de la précision des labels sur la difficulté du problème d'optimisation : (a) nombre d'itérations moyen nécessaire pour que l'algorithme converge, (b) nombre de maximum locaux détectés sur* 200 *initialisations aléatoires de l'algorithme EM, en fonction de la non spécificité moyenne des labels de l'ensemble d'apprentissage.*

Comme attendu, l'apport d'information sur la véritable classe des individus a un impact important sur le problème d'optimisation. L'augmentation du nombre de maxima locaux est en effet quasiment exponentiel lorsque l'imprécision des labels augmente. De plus, lorsque la non spécificité moyenne est inférieure à 0.25, nous avons pu noter que quelque soit le point de départ aléatoire de l'algorithme, celui-ci convergeait vers une unique solution. Quand suffisamment d'information sur les classes d'origines des différents points est fournie, le problème d'optimisation devient convexe, comme dans le cas de l'apprentissage supervisé.

3.4.2 Simulations intégrant des erreurs de labellisation

Nous présentons dans cette section des résultats qui montrent l'intérêt des labels « doux » pour traiter des jeux de données où certains labels peuvent être erronés. Cette expérience vise en particulier à montrer comment les labels « doux » peuvent être profitables pour représenter une information fournie par un expert en prenant en compte toutes les dimensions d'une telle information. Nous avons pour cela imaginé la démarche suivante : pour chaque exemple i, l'expert fournit la classe qui lui semble la plus probable c_{k*} ainsi qu'une mesure de doute sur cette décision p_i. Ce doute est représenté par un nombre entre $[0, 1]$, qui peut être vu comme la probabilité que l'expert ne sache rien sur la véritable classe de l'individu.

Pour prendre en considération cette information dans le cadre de la théorie des fonctions de croyance, il est naturel d'affaiblir la fonction de masse catégorique associée à la classe la plus probable. Le coefficient d'affaiblissement est égal à p_i (Shafer 1976, page 251). Les labels construits à partir des informations fournies par l'expert sont donc des fonctions de masse simples tel que : $m_i^{\mathcal{Y}}(\{c_{k^*}\}) = 1 - p_i$, et $m_i^{\mathcal{Y}}(\mathcal{Y}) = p_i$. Les plausibilités des différentes classes sont données par $pl_{ik^*} = 1$

et $pl_{ik} = p_i$ pour tout $k \neq k^*$. Des labels de cette forme peuvent aisément être manipulés par notre méthode.

Simulation des labels

Des données réelles et simulées pour lesquelles la véritable classe de tous les individus est connue, ont été corrompues de la manière suivante pour simuler le bruit d'étiquetage : pour chaque exemple d'apprentissage i, un nombre p_i a été tiré pour définir le doute de l'expert sur la classe de cet individu. Cette probabilité a ensuite été utilisée pour modifier le label de cet individu. Un nombre aléatoire uniforme est pour cela tiré entre 0 et 1, si ce nombre est inférieur à p_i le label est modifié et devient erroné dans le cas contraire le véritbale label est conservé. Ce processus permet d'introduire dans l'étiquetage des erreurs de labellisation. La distribution de probabilité utilisée pour tirer les p_i définit le taux d'erreur de labellisation asymptotique de l'expert. Plus précisément, l'espérance des p_i est égale à ce taux d'erreur de labellisation asymptotique. Dans le cadre de cette expérience, nous avons utilisé des lois Beta pour tirer ces valeurs, et nous avons fait varier l'espérance de cette loi $\{0.1, 0.15, \ldots, 0.4\}$ en gardant la variance égale à 0.2. Les densités de ces différentes lois sont tracées sur la figure 3.6.

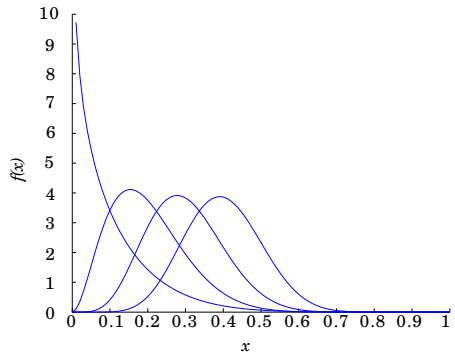


FIG. 3.6 – *Simulation de labels imprécis : densité de différentes lois Beta, d'espèrance* $\mu \in \{0.1, 0.15, \ldots, 0.4\}$ *et de variance égale à* 0.2.

Jeux de données simulées

Les premiers résultats de cette série d'expériences concernent des jeux de données simulées ; il n'y a donc pas de biais de modélisation dans cette expérience. Quatre jeux de données de taille $N \in \{500, 1000, 2000, 4000\}$ ont tout d'abord été simulés à partir d'un modèle de mélange identique à celui utilisé dans la section précédente, c'est-à-dire avec une distance $\delta = 2$ entre les centres des deux composantes gaussiennes en dimension 10 et de matrice de variance-covariance égale à la matrice identité. Les labels ont ensuite été simulés en utilisant le processus décrit dans la section précédente. Les résultats de notre approche ont été comparés à différentes méthodes :

1. Les modèles de mélange gaussien, dans un cadre supervisé, en utilisant les labels durs (potentiellement faux) fournis par l'expert.
2. Les modèles de mélange gaussien, dans un cadre non supervisé. Cette méthode n'utilise pas les informations données par l'expert.

3. Une stratégie utilisant un modèle de mélange gaussien dans un contexte semi-supervisé. Cette stratégie prend en compte l'information sur la fiabilité des labels de la manière suivante : tous les points pour lesquels le doute de l'expert est raisonnable ($p_i \leq 0.5$) sont considérés comme labellisés, dans le cas contraire ($p_i > 0.5$) ceux-ci sont considérés comme non étiquetés. Cette stratégie sera nommée « apprentissage semi-supervisé adaptatif ».
4. Le modèle génératif de bruit d'étiquetage introduit par Lawrence et Schölkopf (2001), qui apprend les probabilités d'inversion des labels. Cette solution vise à résoudre elle aussi le problème du bruit d'étiquetage mais ne prend pas en compte l'information sur la fiabilité des labels. L'implémentation utilisée est celle proposée dans Li et al. (2007b).

Il est intéressant de noter que les méthodes (1, 2, 3) peuvent être mises en place à l'aide de l'algorithme générique proposé dans ce chapitre, la seule différence provenant des labels utilisés en entrée. La méthode (4) est basée sur un modèle génératif légèrement plus complexe puisqu'il intègre la possibilité d'inversion des labels, les paramètres de ce modèle sont cependant eux aussi estimés par un algorithme EM (cf. section 3.1.3)

Le tableau 3.1 et la figure 3.7 présentent les performances obtenues par toutes ces méthodes. Le taux d'erreur de classification a pour cela été estimé en utilisant un ensemble de test de 5000 observations. Ce taux correspond au pourcentage d'erreurs entre les véritables classes de ces 5000 individus et les classes prédites par chacun des modèles. Les résultats ont été moyennés sur 100 jeux de données simulés pour éviter tout artefact dû aux différents tirages aléatoires entrant dans le processus de génération des données. Pour toutes les méthodes, l'algorithme EM a été initialisé avec les véritables valeurs des paramètres.

Comme prévu, quand le taux d'erreur de labellisation augmente, le taux d'erreur de l'apprentissage supervisé augmente aussi. Notre solution ne souffre pas autant de cette augmentation du taux d'erreur de labellisation. Elle obtient de meilleures performances que l'apprentissage supervisé et semi-supervisé, quelque soit le taux d'erreur de labellisation. L'information sur la fiabilité des labels est donc exploitée efficacement par notre méthode pour obtenir des résultats quasiment stables, même lorsque le taux d'erreur de labellisation devient important. La comparaison entre notre méthode et l'apprentissage non supervisé est aussi largement à l'avantage de celle-ci, excepté lorsque le nombre d'exemples d'apprentissage est très important. Dans ce cas de figure, l'apprentissage non supervisé obtient de bons résultats. Finalement, le modèle génératif de bruit d'étiquetage obtient lui aussi de bons résultats, en particulier lorsque le taille de l'ensemble d'apprentissage est importante $N \in \{2000, 4000\}$. Les labels « doux » permettent cependant d'obtenir de meilleurs résultats pour des faibles tailles de jeux de données.

Enfin, nous avons intégré aux résultats, les résultats d'un ensemble de tests de significativité (test de signe bilatéral) qui ont été effectués entre la meilleure méthode (au sens du taux d'erreur de classification) et toutes les autres méthodes. Si l'hypothèse nulle (médiane de la différence des deux distributions égale à 0) est rejetée pour tous les tests, avec un risque de première espèce de 5%, cette méthode est considérée comme significativement plus performante que toutes les autres. Dans ce cas, le taux d'erreur de cette méthode est présenté en gras dans le tableau 3.1.

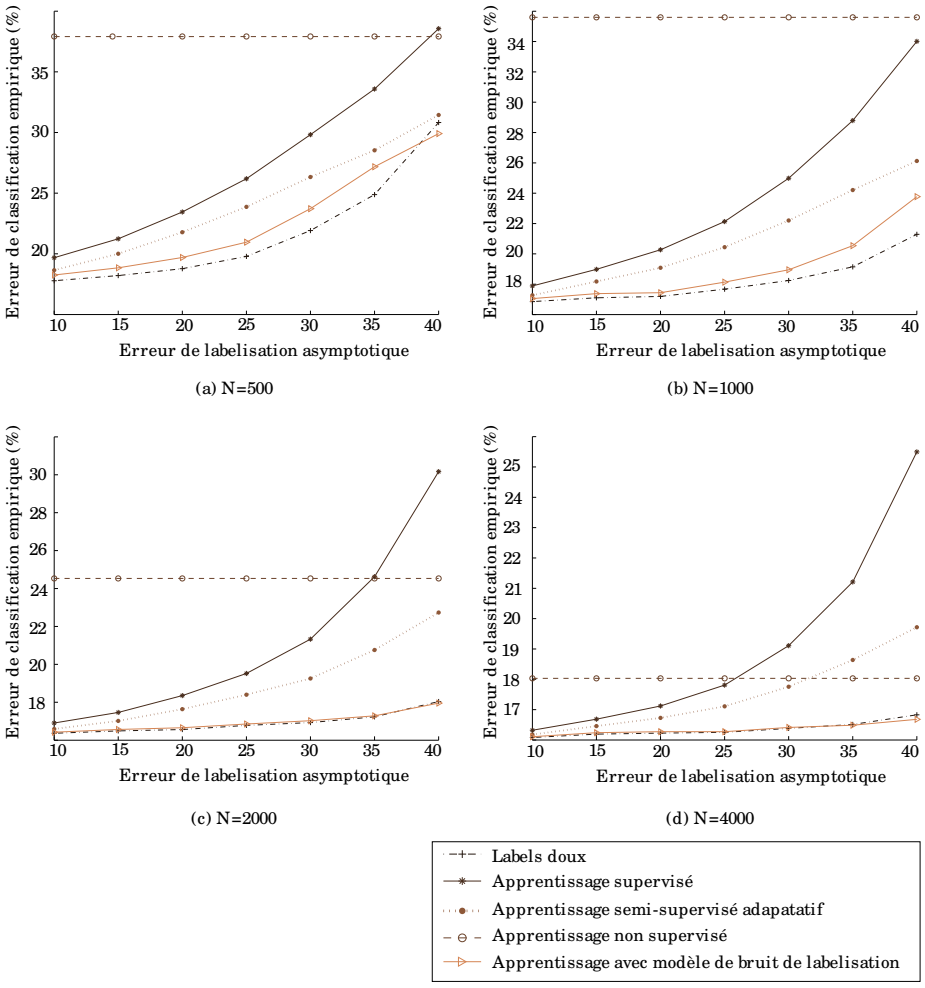


FIG. 3.7 – *Expérience sur le bruit d'étiquetage sur données simulées : erreur de classification empirique (%) moyennée sur 100 ensembles de labels indépendants, en fonction de l'erreur de labellisation asymptotique de ces labels, pour différentes tailles de l'ensemble d'apprentissage et pour chacune des méthodes testées.*

		Taux d'erreur d'étiquetage asymptotique (%)						
		10	*15*	*20*	*25*	*30*	*35*	*40*
$N = 500$	Labels « doux »	**17.8**	**18.2**	**18.8**	**19.8**	**21.9**	**24.9**	30.8
	Apprentissage Supervisé	19.7	21.2	23.4	26.2	29.8	33.6	38.6
	Apprentissage Non supervisé	37.9	37.9	38.0	37.9	37.9	37.9	37.9
	Apprentissage Semi supervisé	18.7	20.0	21.8	23.9	26.3	28.5	31.4
	Modèle de bruit d'étiquetage	18.3	18.8	19.7	21.0	23.7	27.2	29.9
$N = 1000$	Labels « doux »	**16.9**	**17.1**	**17.2**	**17.7**	**18.2**	**19.1**	**21.3**
	Apprentissage Supervisé	17.9	19.0	20.3	22.1	25.0	28.8	34.0
	Apprentissage Non supervisé	35.6	35.6	35.6	35.6	35.6	35.7	35.5
	Apprentissage Semi supervisé	17.3	18.2	19.1	20.4	22.2	24.2	26.1
	Modèle de bruit d'étiquetage	17.1	17.4	17.4	18.1	19.0	20.5	23.8
$N = 2000$	Labels « doux »	**16.4**	**16.5**	**16.6**	16.8	**16.9**	17.2	18.0
	Apprentissage Supervisé	16.9	17.5	18.4	19.5	21.3	24.6	30.2
	Apprentissage Non supervisé	24.5	24.5	24.5	24.6	24.5	24.5	24.5
	Apprentissage Semi supervisé	16.6	17.0	17.6	18.4	19.3	20.8	22.7
	Modèle de bruit d'étiquetage	16.4	16.6	16.7	16.9	17.0	17.3	18.0
$N = 4000$	Labels « doux »	16.1	**16.2**	16.2	16.3	16.4	16.5	16.8
	Apprentissage Supervisé	16.3	16.7	17.1	17.8	19.1	21.2	25.5
	Apprentissage Non supervisé	18.0	18.1	18.1	18.1	18.0	18.0	18.0
	Apprentissage Semi supervisé	16.2	16.5	16.7	17.1	17.8	18.6	19.7
	Modèle de bruit d'étiquetage	16.1	16.2	16.3	16.3	16.4	16.5	**16.7**

TAB. 3.1 – *Taux d'erreur de classification empirique (%) moyenné sur 100 jeux de données indépendants, pour différents taux d'erreur de labellisation asymptotique (%) et différentes tailles de jeu de données. Les taux d'erreurs en gras sont significativement plus faibles d'après un test de signe bilatéral avec une erreur de première espèce de 5 %.*

Pour conclure sur cette expérience, notre solution obtient de bien meilleurs résultats que l'apprentissage supervisé lorsque les labels sont contaminés par des erreurs. L'apprentissage supervisé est en effet dans ce cas de figure une méthode hasardeuse puisque certains labels sont faux. Notre méthode permet aussi d'obtenir des résultats de meilleur qualité que l'apprentissage non supervisé qui est une méthode trop conservative dans le cadre de cette expérience puisqu'aucune information fournie par l'expert n'est utilisée. Finalement pour les jeux de données de faible taille, notre méthode est également plus efficace que le modèle génératif de contamination des labels. Cette expérience permet donc de conclure sur la capacité de notre méthode à prendre en compte une information complexe sur la véritable classe d'un individu, et sa capacité à traiter des problèmes où certains individus sont mal labellisés.

Jeux de données réels

Pour confirmer les résultats précédemment obtenus sur données simulées nous avons aussi mené des expériences similaires sur des jeux de données réels. Un biais de modélisation peut donc être présent dans les expérimentations qui suivent puisque les données n'ont pas été simulées en utilisant un modèle de mélange gaussien. Nous avons pour cette série d'expériences utilisé des données librement

disponibles[3] (Asuncion et Newman 2007), dont les caractéristiques sont résumées dans le tableau 3.2.

Le premier jeu de données *Iris* est bien connu de la communauté statistique puisqu'il s'agit de données déjà étudiées par Fisher. Ce jeu de données contient des mesures effectuées sur trois espèces différentes d'iris. Le jeu de données *Crabes* concerne lui reconnaissance de différentes espèces de crabes cette fois, le sexe de chacun des individus devant également être déterminé à partir de mesures morphologiques. Le jeu de données *Vins* contient les résultats d'analyses chimiques effectuées sur différents cépages de vins italiens. Finalement, le jeu de données *Cancer du sein (Wisconsin)* traite du problème de la reconnaissance de tumeurs à partir de 30 variables qui décrivent les caractéristiques du noyau des cellules présentes dans une image sous microscope obtenue suite à une biopsie. La tâche consiste à déterminer si la tumeur est maligne ou bénigne.

nom	# dimensions	# exemples	# classes
Iris	4	150	3
Crabes	5	200	4
Vins	13	178	3
Cancer du sein (Wisconsin)	30	569	2

TAB. 3.2 – *Caractéristiques des jeux de données réels étudiés.*

Aucun pré-traitement n'a été effectué sur ces données, excepté un centrage et une réduction. En ce qui concerne les étiquettes, le même procédé que précédemment a été mis en place pour introduire des erreurs de labellisation. Enfin, l'erreur de classification a été estimée grâce à une validation croisée (10 blocs), 9/10 du jeu de données étant utilisé pour l'apprentissage, le reste servant à l'estimation du taux d'erreur. Finalement, 30 jeux de labels différents ont été simulés et les résultats donnent la moyenne des performances sur ces 30 jeux de labels.

En ce qui concerne l'initialisation de l'algorithme EM, nous avons utilisé une stratégie différente pour chacune des méthodes afin de prendre en considération la nature de l'information dont chacune d'entre elles dispose. Dans le cadre non supervisé nous avons utilisé 100 initialisations aléatoires [4]. Pour résoudre le problème du « label switching », engendré par ces initialisations aléatoires, l'erreur de classification a été calculée en accord avec la meilleure permutation des classes. En ce qui concerne les labels « doux », l'initialisation utilise la transformation pignistique des labels, comme expliqué précédemment. De cette manière, seul un point de départ a été utilisé pour cette méthode. La même démarche a été utilisée pour l'apprentissage semi-supervisé adaptatif.

Dans cette expérience sur données réelles nous avons également comparé les résultats de notre méthode avec une approche non paramétrique qui peut aussi travailler avec des labels doux. Cette approche, Denœux (1995), Denoeux et Zouhal (2001) basée sur le principe de l'algorithme des k plus proches voisins a été testée

[3] http ://mlearn.ics.uci.edu/MLRepository.html
et http ://rweb.stat.umn.edu/R/library/MASS/html/crabs.html.

[4] Les centres sont tirés en utilisant une distribution gaussienne dont les paramètres sont estimés sur toute la population, les matrices de variance-covariance sont quant à elle initialisées avec celles de la population entière et sont égales.

dans deux versions, l'une prend en compte le doute de l'expert en affaiblissant les labels, alors que la seconde utilise les labels durs. Le nombre de voisins pris en compte dans l'élaboration de la règle de décision a été fixé a priori à 10.

		Taux d'erreur d'étiquetage asymptotique (%)						
		10	*15*	*20*	*25*	*30*	*35*	*40*
Iris	Labels « doux »	2.9	3.0	3.0	3.6	**4.2**	**4.2**	**6.2**
	Apprentissage Supervisé	7.0	9.9	11.7	14.2	16.6	19.4	23.6
	Apprentissage Non supervisé	12.4	12.4	12.4	12.4	12.4	12.4	12.4
	Apprentissage Semi supervisé	4.9	8.3	9.0	12.4	14.5	16.2	21.3
	Modèle de bruit d'étiquetage	2.9	3.1	3.3	4.0	6.2	8.2	15.3
	k-ppv TBM	5.1	5.2	6.4	7.0	10.4	13.1	18.1
	k-ppv TBM avec affaiblissement	4.7	5.0	5.4	6.0	8.0	8.0	12.0
Vins	Labels « doux »	1.1	**1.2**	**1.9**	**2.8**	**4.4**	**6.4**	**8.2**
	Apprentissage Supervisé	6.2	9.6	12.8	15.8	20.1	23.9	28.6
	Apprentissage Non supervisé	31.6	31.6	31.6	31.6	31.6	31.6	31.6
	Apprentissage Semi supervisé	3.4	6.2	9.5	11.7	14.0	17.0	18.1
	Modèle de bruit d'étiquetage	1.6	1.7	2.5	3.9	6.1	8.6	12.4
	k-ppv TBM	2.5	3.3	4.1	5.7	9.2	10.7	17.2
	k-ppv TBM avec affaiblissement	2.4	3.0	3.4	4.6	6.1	7.6	11.0
Crabes	Labels « doux »	6.0	5.9	6.1	6.2	6.3	**6.4**	**6.8**
	Apprentissage Supervisé	8.3	9.8	10.8	12.8	15.0	17.2	21.0
	Apprentissage Non Supervisé	7.6	7.6	7.6	7.6	7.6	7.6	7.6
	Apprentissage Semi supervisé	7.2	8.8	9.7	11.1	12.6	13.5	16.7
	Modèle de bruit d'étiquetage	6.0	5.9	6.0	6.3	6.3	8.0	10.0
	k-ppv TBM	23.5	25.0	27.4	27.9	31.0	33.5	37.3
	k-ppv TBM avec affaiblissement	23.3	25.8	27.0	28.3	30.2	32.7	35.8
Cancer du sein	Labels « doux »	5.1	5.5	6.3	6.5	**7.3**	**8.5**	**8.5**
	Apprentissage Supervisé	7.7	9.1	10.5	12.2	15.0	20.2	24.9
	Apprentissage Non Supervisé	11.2	11.2	11.2	11.2	11.2	11.2	11.2
	Apprentissage Semi supervisé	6.2	7.6	9.5	10.3	10.9	14.7	13.9
	Modèle de bruit d'étiquetage	5.9	6.1	6.9	7.3	8.2	10.9	13.3
	k-ppv TBM	4.0	4.9	7.3	10.3	15.1	22.6	31.1
	k-ppv TBM avec affaiblissement	**3.5**	**4.0**	**4.6**	**5.9**	7.8	12.1	17.0

TAB. 3.3 – *Taux d'erreur de classification empirique (%) estimé par validation croisée (10 blocs) moyenné sur 30 jeux de labels indépendants, pour différents taux d'erreur de labellisation asymptotique (%) et différentes tailles de jeu de données. Les taux d'erreurs en gras sont significativement plus faibles d'après un test de signe bilatéral avec une erreur de première espèce de 5 %.*

Comme pour les expériences sur données simulées, la capacité de notre méthode à prendre en compte, de manière efficace, l'information sur la fiabilité des labels est clairement visible, comme le montre le tableau de résultats 3.3 et la figure 3.8.

Pour tous les problèmes réels étudiés, les résultats sont assez stables même quand l'erreur de labellisation est importante. Nous pouvons aussi noter que notre méthode obtient des performances largement meilleures que l'apprentissage supervisé et semi-supervisé quand le taux d'erreur de labellisation est assez important. Cette méthode permet aussi d'obtenir des résultats bien meilleurs que l'apprentissage non supervisé qui, en plus du peu d'information utilisé, semble en partie

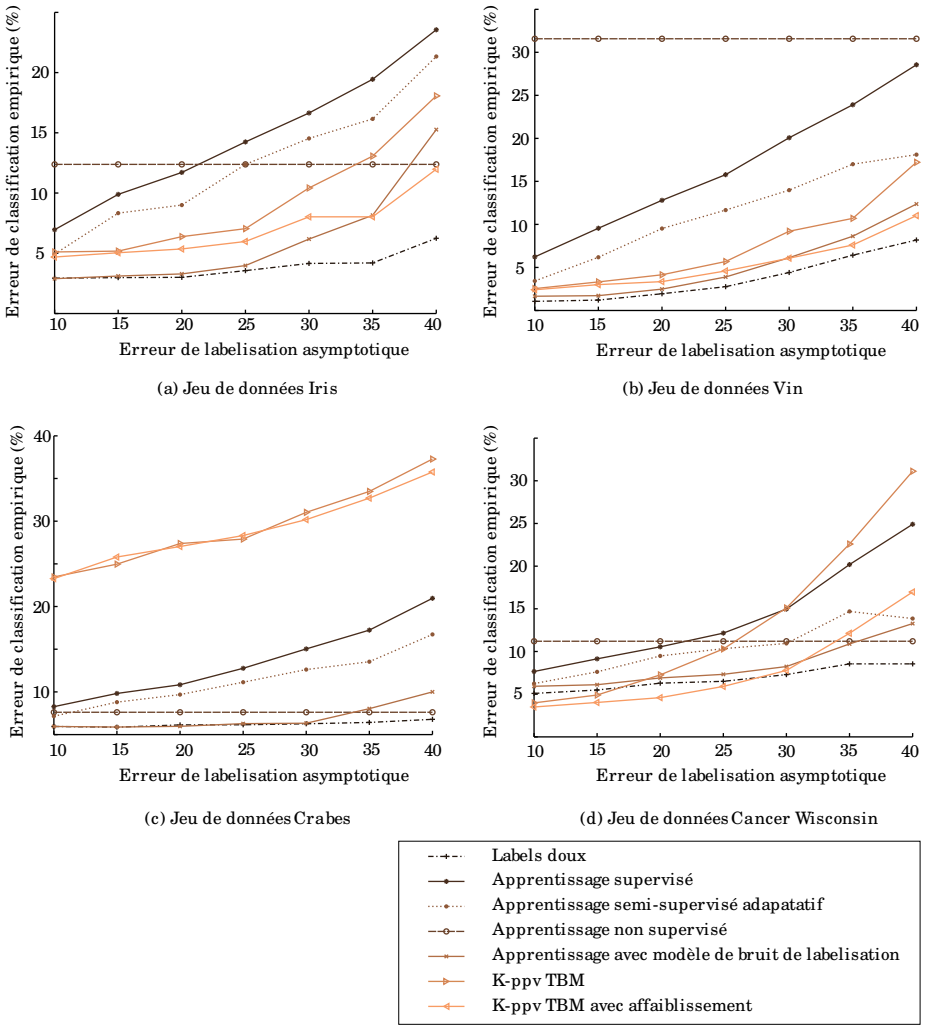


FIG. 3.8 – *Expérience sur le bruit d'étiquetage sur données réelles : erreur de classification empirique (%), estimée par validation croisée (10 blocs), moyennée sur 30 ensembles de labels indépendants, en fonction de l'erreur de labellisation asymptotique, pour différents jeux de données réels : Iris (a), Vin (b), Crabes (c) et Cancer Wisconsin (d).*

affecté par le problème des minima locaux. En ce qui concerne la méthode basée sur le modèle génératif de bruit d'étiquetage (4), celle-ci obtient des résultats assez similaires à la notre pour des taux d'erreur de labellisation faibles. Cependant, l'information sur la fiabilité des labels permet, lorsque le taux d'erreur est plus important, d'obtenir de bien meilleurs résultats.

L'algorithme des k plus proches voisins basé sur la théorie des fonctions de croyance obtient lui aussi de bons résultats lorsque l'information sur le doute de l'expert est prise en compte. Les mauvais résultats obtenus sur le jeu de données *Crabes* constitue une exception, en ce qui concerne cette méthode. Ces résultats peuvent s'expliquer par le fait que toutes les variables de ce jeu de données sont très fortement corrélées entre elles et que cette forte corrélation n'est pas reliée au problème de classification, mais s'explique par l'existence d'une variable latente influençant l'ensemble des variables étudiées : l'âge des différents individus.

Pour conclure sur cette série d'expériences, nous pouvons dire que les labels doux ont prouvé leur capacité à prendre en compte une information complexe sur la fiabilité des labels pour obtenir de bons résultats même lorsque certains labels sont faux.

Conclusion du chapitre

Ce chapitre nous a permis de présenter une solution innovante au problème de l'apprentissage lorsque l'information sur les classes d'origine des différents individus servant à l'apprentissage de la règle de classification, est partielle. Cette solution repose sur l'extension du critère de maximum de vraisemblance afin de prendre en compte des étiquettes imprécises, incertaines modélisées à l'aide des fonctions de croyance. Nous avons mis en avant un critère simple et implémentable, basé sur la maximisation de la plausibilité des paramètres d'un modèle génératif. Les relations, entre ce critère et les solutions déjà existantes dans le cadre probabiliste, ont également pu être mises à jour. Enfin, des expériences sur jeux de données réelles et simulées nous ont permis de valider l'apport de cette solution pour traiter des situations où les labels peuvent être erronés mais où une information sur la fiabilité des dits labels est disponible. En conclusion, notre méthode offre de nombreuses possibilités pour représenter l'information disponible sur les labels des différents individus et permet de tirer parti de celle-ci quelque soit sa forme.

4 Analyse en composantes indépendantes et informations a priori

Le ciel à Paris a ses propres lois qui opèrent indépendamment de la ville en dessous.
Paul Auster , **La chambre dérobée (1993)**

Sommaire

Ce chapitre est consacré à la présentation de deux extensions de l'analyse en composantes indépendantes (ACI) permettant de prendre en compte des informations supplémentaires sur les données et sur le problème à traiter. Ces deux extensions qui ont été motivées par notre application seront présentées ici de manière générale. Leur utilisation dans un contexte applicatif concret, celui du diagnostic des circuits de voie ferroviaire, fera l'objet du dernier chapitre de cette thèse.

Après avoir rapidement présenté différentes extensions de l'ACI déjà proposées dans la littérature, nous verrons dans ce chapitre comment des connaissances sur la structure du processus de mixage, formalisées grâce à des hypothèses d'indépendance entre variables observées et variables latentes, peuvent être prises en compte dans la procédure d'estimation des paramètres de l'ACI. Nous détaillerons

une solution pour intégrer au modèle de l'ACI de telles connaissances dans le cadre d'une approche de type vraisemblance et nous présenterons des résultats de simulation illustrant l'intérêt de cette approche.

Enfin, nous verrons comment nos travaux sur la labellisation « douce » qui avaient pour cadre les modèles de mélange, peuvent être utilisés dans le contexte de l'analyse en facteurs indépendants (IFA) pour tirer parti d'informations partielles sur les différents individus servant à l'apprentissage. Des expériences nous permettront d'analyser le comportement de cette méthode dans un contexte de labellisation partielle et de constater son intérêt par rapport au contexte d'utilisation classique de l'ACI et de l'IFA, à savoir le contexte non-supervisé.

4.1 Travaux existants

L'analyse en facteurs indépendants (IFA) et plus généralement l'analyse en composantes indépendantes (ACI) présentées en section 2.2 ont fait l'objet de nombreuses extensions afin de pouvoir prendre en compte des connaissances supplémentaires sur le problème étudié. Ces informations a priori peuvent concerner les deux éléments du modèle, à savoir :

1. La matrice de mixage
2. Les sources

Des connaissances, réelles ou supposées, peuvent être introduites dans le modèle de l'analyse en composantes indépendantes au travers d'hypothèses supplémentaires sur la forme de la matrice de mixage ou sur la forme des variables latentes recherchées. Les principales hypothèses étudiées concernent la positivité des variables latentes ou de la matrice de mixage ou bien encore des deux (Moussaoui 2005) et l'hypothèse de parcimonie des sources ou de la matrice de mélange (Jutten et Comon 2007b, chap. 10).

L'hypothèse de positivité se justifie dans de nombreuses applications par la nature même des variables observées et des variables latentes recherchées. Cette hypothèse a par exemple été utilisée avec succès en analyse d'images (Bakir et al. 2006), ou bien encore, pour analyser des données spectroscopiques issues de la chimie (Li et al. 2007a, Moussaoui 2005). Dans ces deux cas de figure ces contraintes découlent directement de connaissances physiques sur le problème étudié.

contraintes de positivité

L'analyse en composantes indépendantes lorsqu'elle utilise ce type de contraintes se rapproche des méthodes de décomposition matricielle, en particulier de la factorisation matricielle non négative (Non Negative Matrix Factorization NNMF en anglais), (Lee et Seung 1999). Les contraintes de positivité peuvent être introduites directement dans le problème d'optimisation de l'ACI ou au travers de la forme des densités postulées pour les sources. Il est intéressant de noter l'intérêt des approches de type vraisemblance dans ce cadre car celles-ci permettent d'introduire aisément ce type de connaissance dans le modèle en utilisant des formes spécifiques pour les densités des sources, par exemple des densités à support positif.

parcimonie

Les approches bayésiennes (Jutten et Comon 2007b, chap. 12) peuvent également se montrer pertinentes pour introduire des connaissances a priori sur la matrice de mixage. En effet, l'utilisation d'une loi a priori sur les coefficients de cette matrice permet de prendre en considération des connaissances supplémentaires sur celle-ci telle que la positivité ou la parcimonie (Hyvärinen et Karthikesh 2002, Zhang et Chan 2006). Dans ce dernier cas de figure, il est possible d'utiliser une loi de Laplace comme a priori sur les coefficients de la matrice, cette approche étant équivalente à l'ajout d'une pénalisation de type L_1 sur ces mêmes coefficients. Une telle approche peut également permettre d'intégrer au modèle des informations plus spécifiques à l'application considérée aux travers de lois a priori (Knuth 1999).

hypothèse markovienne

D'autres extensions ont été proposées afin de prendre en considération une structure temporelle ou spatiale (Attias 2000), (Jutten et Comon 2007b, p. 500-514). Ce type de modèle se rapproche des modèles de type Markov caché en introdui-

sant une dépendance entre les états des différentes sources entre deux instants consécutifs.

Dans ce chapitre nous présentons deux autres extensions. La première de celles-ci concerne la prise en compte d'informations sur la physique du processus de mélange. Cette extension peut être employée dans toutes les applications de l'analyse en composantes indépendantes où des hypothèses d'indépendance entre variables observées et variables latentes peuvent être faites. La deuxième qui étendra les travaux présentés au chapitre précédent concerne la prise en compte d'informations partielles sur les individus servant à l'apprentissage ; cette extension ne concernera que l'analyse en facteurs indépendants.

4.2 Prise en compte d'information sur le processus de mixage des sources

Notre première proposition pour prendre en considération des informations supplémentaires sur le processus de mixage est assez simple. Elle concerne la prise en compte explicite d'hypothèses d'indépendance entre variables observées et variables latentes issues la plupart du temps de connaissance physique du processus de mélange. Ce type d'approche, qui n'a pas été utilisée à notre connaissance dans le contexte de l'analyse en composantes indépendantes, est par contre largement utilisée dans le contexte de l'analyse factorielle (Bartholomew et Martin 1999, pages 43-44, 175-176) et plus particulièrement dans le domaine de la modélisation par équation structurelle (Structural Equation Modeling SEM en anglais) (Bollen 1989)).

4.2.1 Principe

L'ACI sans bruit repose sur le modèle suivant : les données observées X sont supposées dépendre de manière déterministe et linéaire d'un ensemble de causes indépendantes et inobservées Z. Le modèle est donc de la forme :

$$\mathbf{x} = A\,\mathbf{z}, \tag{4.1}$$

avec A la matrice de mixage de taille $S \times S$. Les variables latentes ou sources $Z_1, \ldots, Z_S$ sont supposées suivre une loi ayant pour densité $f^{\mathcal{Z}_1}, \ldots, f^{\mathcal{Z}_S}$ et celles-ci sont indépendantes. La densité sur l'ensemble des variables latentes s'écrit donc :

$$f^{\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_S}(z_1, \ldots, z_S) = \prod_{s=1}^{S} f^{\mathcal{Z}_s}(z_s). \tag{4.2}$$

Enfin comme la relation entre variables observées et latentes est supposée être déterministe (4.1) la densité conditionnelle d'une variable observée connaissant les variables latentes est donnée par :

$$f^{\mathcal{X}_h|\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_S}(x_h|\mathbf{z}) = \delta(x_h - A_{h.}\mathbf{z}), \tag{4.3}$$

avec $A_{h.}$ la h^{e} ligne de la matrice A et δ la fonction de Dirac.

L'objectif de la présente section est d'analyser l'influence des hypothèses de la forme,

$$X_h \perp\!\!\!\perp Z_g, \tag{4.4}$$

sur ce modèle et sur ses performances. Ce type d'hypothèse supplémentaire peut être interprétée de la manière suivante :

La variable latente Z_g n'influence pas la variable observée X_h.

Les hypothèses de cette forme peuvent être utiles dans certaines applications de l'ACI ou de l'IFA en améliorant les performances de la méthode lorsque les hypothèses sont justifiées. Ce type d'hypothèse a en effet un impact direct sur le modèle, comme le montre la proposition suivante :

Proposition 4.1 *Dans le cadre du modèle de l'ACI sans bruit on a :*

$$X_h \perp\!\!\!\perp Z_g \Leftrightarrow A_{hg} = 0. \tag{4.5}$$

Preuve. La notion d'indépendance est définie par :

$$X_h \perp\!\!\!\perp Z_g \Leftrightarrow f^{\mathcal{X}_h \times \mathcal{Z}_g}(x_h, z_g) = f^{\mathcal{X}_h}(x_h) \times f^{\mathcal{Z}_g}(z_g). \tag{4.6}$$

Or, la densité jointe sur $\mathcal{X}_h \times \mathcal{Z}_g$ est donnée dans le cadre de l'ACI sans bruit par :

$$\begin{aligned}
f^{\mathcal{X}_h \times \mathcal{Z}_g}(x_h, z_g) &= \int_{\mathbb{R}^{S-1}} f^{\mathcal{X}_h \times \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_S}(x_h, z_1, \ldots, z_S) \prod_{l=1, l\neq g}^{S} \mathrm{d}z_l \qquad (4.7)\\
&= \int_{\mathbb{R}^{S-1}} \left(\prod_{s=1}^{s} f^{\mathcal{Z}_s}(z_s) \right) \times f^{\mathcal{X}_h | \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_S}(x_h | z_1, \ldots, z_S) \prod_{l=1, l\neq g}^{S} \mathrm{d}z_l \\
&= \int_{\mathbb{R}^{S-1}} \prod_{s=1}^{S} f^{\mathcal{Z}_s}(z_s) \times \delta(x_h - A_{h.}\mathbf{z}) \prod_{l=1, l\neq g}^{S} \mathrm{d}z_l \\
&= f^{\mathcal{Z}_g}(z_g) \times \left(\int_{\mathbb{R}^{S-1}} \prod_{l=1, l\neq g}^{S} f^{\mathcal{Z}_l}(z_l) \times \delta(x_h - A_{h.}\mathbf{z}) \mathrm{d}z_l \right). \qquad (4.8)
\end{aligned}$$

En utilisant (4.8), on identifie donc :

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

l'intégrale ne doit pas dépendre de z_g (la g^{e} ligne de $\mathbf{z}$) ce qui n'est possible que si et seulement si $A_{hg} = 0$. □

contraintes sur la matrice de mixage

Faire une telle hypothèse supprime donc un degré de liberté dans le modèle et élimine un paramètre en contraignant la matrice de mixage. En utilisant une hypothèse différente, il est possible d'obtenir une contrainte sur la matrice de démixage, comme le montre la proposition suivante.

Proposition 4.2 *Dans le cadre de l'ACI sans bruit, on a :*

$$(Z_g \perp\!\!\!\perp X_h) | X_1, \ldots, X_{h-1}, X_{h+1}, \ldots, X_S \Leftrightarrow W_{gh} = 0. \tag{4.9}$$

Preuve. La définition de l'indépendance conditionnelle donne :

$$(Z_g \perp\!\!\!\perp X_h)|X_1,\ldots,X_{h-1},X_{h+1},\ldots,X_S$$
$$\Leftrightarrow$$
$$f(z_g|x_1,\ldots,x_S) = f(z_g|x_1,\ldots,x_{h-1},x_{h+1},\ldots,x_S), \quad (4.10)$$

or la densité conditionnelle sur $\mathcal{Z}_g$ connaissant les variables observées $X_1,\ldots,X_S$ est donnée par :

$$f(z_g|x_1,\ldots,x_S) = \delta(z_g - W_{g.}\mathbf{x}),$$

avec $W_{g.}$ la g^{e} ligne de la matrice de démixage $W = A^{-1}$. On obtient donc, $\delta(z_g - W_{k.}\mathbf{x})$ ne doit pas dépendre de x_h (la h^{e} ligne de $\mathbf{x}$), ce qui n'est possible que si et seulement si $W_{gh} = 0$. □

contraintes sur la matrice de démixage

L'hypothèse d'indépendance conditionnelle $(Z_g \perp\!\!\!\perp X_h)|X_1,\ldots,X_{h-1},X_{h+1},\ldots,X_S$ induit donc quant à elle une contrainte sur la matrice de démixage. Ces deux hypothèses ne sont pas équivalentes. En effet, l'inverse d'une matrice contenant des zéros ne contient pas forcément des zéros aux mêmes places.

D'un point de vue génératif, il semble plus naturel d'utiliser des hypothèses de la forme (4.4), car elles portent directement sur le processus de génération des données et sont de plus, plus faciles à interpréter. Pour toutes ces raisons, ce sont donc des contraintes sur la matrice de mixage que nous considérerons dans la suite de cette section.

Ce type de contrainte supprime un paramètre du modèle ; le premier bénéfice attendu concerne l'erreur d'estimation. Celle-ci doit logiquement diminuer lorsque le nombre de paramètres à estimer diminue. Faire des hypothèses de la forme que nous venons de décrire peut permettre d'obtenir de meilleurs résultats lorsque celles-ci concordent avec la réalité.

Nous allons maintenant voir comment de telles contraintes peuvent être implémentées dans un algorithme d'analyse en composantes indépendantes lorsque les densités des variables latentes sont fixées a priori.

4.2.2 L'intégration des contraintes au problème d'optimisation

La principale difficulté pour l'introduction de contraintes sur l'indépendance statistique de certaines variables latentes vis a vis de certaines variables observées réside dans la reformulation du problème d'estimation de l'ACI en fonction de la matrice de mixage et non plus en fonction de la matrice de démixage comme c'est le cas classiquement (Hyvärinen 2001).

Lorsque l'hypothèse d'une relation déterministe entre variables latentes et observées est faite, la densité sur $\mathcal{X}$ peut être construite à partir des densités sur $\mathcal{Z}_1,\ldots,\mathcal{Z}_S$ grâce à la relation suivante (cf. annexe .4) :

$$p^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(A)|}\prod_{s=1}^{S} f^{\mathcal{Z}_s}\left((A^{-1}\mathbf{x})_s\right). \quad (4.11)$$

La vraisemblance de la matrice de mixage A est alors donnée par :

$$L(A;\mathbf{X}) = \prod_{i=1}^{N} \frac{1}{|\det(A)|} \left(\prod_{s=1}^{S} f^{\mathcal{Z}_s}((A^{-1}\mathbf{x}_i)_s) \right), \tag{4.12}$$

pour un jeu de données i.i.d. $\mathbf{X}$. Il est alors possible d'écrire la log-vraisemblance d'une matrice de mixage A quelconque aussi bien que d'une matrice de démixage W quelconque :

$$\mathcal{L}(A;\mathbf{X}) = -N\log(|\det(A)|) + \sum_{i=1}^{N}\sum_{s=1}^{S} \log\left(f^{\mathcal{Z}_s}((A^{-1}\mathbf{x}_i)_s)\right), \tag{4.13}$$

$$\mathcal{L}(W;\mathbf{X}) = N\log(|\det(W)|) + \sum_{i=1}^{N}\sum_{s=1}^{S} \log\left(f^{\mathcal{Z}_s}((W\mathbf{x}_i)_s)\right). \tag{4.14}$$

En supposant que les densités des sources $f^{\mathcal{Z}_1}, \ldots, f^{\mathcal{Z}_S}$ sont connues, il est possible de maximiser, l'une ou l'autre de ces fonctions par rapport à A ou W, en utilisant des méthodes de type gradient simple ou gradient naturel, comme expliqué au chapitre 2 section 2.2.3.

Comme les contraintes considérées portent sur la matrice de mixage, il est plus aisé de travailler sur la log-vraisemblance de A (4.13) pour prendre en considération celles-ci. Le calcul de la dérivée de la log-vraisemblance par rapport aux différents coefficients de la matrice de mixage permet d'obtenir la formule de mise à jour suivante (cf. annexe .3) :

$$\Delta A^{(q)} = ({A^{(q)}}^{-1})^t \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i^{(q)}\right) {\mathbf{z}_i^{(q)}}^t - \mathbf{I} \right), \tag{4.15}$$

$$A^{(q+1)} = A^{(q)} + \tau.\Delta A^{(q)}, \tag{4.16}$$

avec $\mathbf{z}_i^{(q)} = {A^{(q)}}^{-1}\mathbf{x}_i$ et $\mathbf{g}$ une fonction de $\mathbb{R}^S \to \mathbb{R}^S$ définie par :

$$\mathbf{g}(\mathbf{z}) = \left[\frac{-\partial \log(f^{\mathcal{Z}_1}(z_1))}{\partial z_1}, \ldots, \frac{-\partial \log(f^{\mathcal{Z}_S}(z_S))}{\partial z_S} \right]^t. \tag{4.17}$$

Cette formule de mise à jour correspond à une simple montée de gradient. Il est également possible d'utiliser la formule de mise à jour correspondant à l'utilisation d'un gradient naturel qui est donnée par :

$$\Delta_{nat} A^{(q)} = A\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i^t - \mathbf{I} \right), \tag{4.18}$$

$$A^{(q+1)} = A^{(q)} + \tau \Delta_{nat} A^{(q)}, \tag{4.19}$$

(cf. annexe .3).

gradient avec contraintes

Pour maximiser la log-vraisemblance sous la contrainte de nullité de certains coefficients de A, il suffit de considérer le gradient des seuls coefficients non nuls, d'utiliser une montée de gradient sur ces seuls coefficients et d'initialiser l'algorithme avec une matrice respectant les contraintes. Plus formellement, l'algorithme s'écrit :

$$\begin{aligned} A^{(0)} &= M \bullet A^{(0)} \\ A^{(q+1)} &= A^{(q)} + \tau\, M \bullet \Delta A^{(q)}, \end{aligned} \tag{4.20}$$

avec $\bullet$ le produit d'Hadamard entre deux matrices (produit éléments par éléments) et $\Delta A^{(q)}$ une direction de montée obtenue en utilisant un gradient simple (4.15) ou un gradient naturel (4.18) et M une matrice binaire représentant le masque des contraintes :

$$M_{hk} = \begin{cases} 0 & \text{si} \quad Z_k \perp\!\!\!\perp X_h, \\ 1 & \text{sinon.} \end{cases} \tag{4.21}$$

L'utilisation d'un gradient naturel a été testée durant nos travaux, mais elle s'est avérée parfois instable numériquement (problème de convergence lorsque l'initialisation est éloignée de la solution).

Remarques pratiques

pré-traitements

Lors de la mise en place d'un tel algorithme, il est important de noter qu'aucun pré-traitement de type blanchiment, analyse en composantes principales, ... classiquement utilisé pour améliorer le conditionnement du problème et séparer l'information utile du bruit, ne peut être effectué sur les données. En effet, celui-ci aurait pour effet immédiat de détruire les hypothèses d'indépendance faites précédemment, les nouvelles variables obtenues après blanchiment étant des combinaisons linéaires de toutes les variables initiales. L'ACI avec contraintes sur la matrice de mixage doit donc être effectuée sur les données initiales, les seuls prétraitements autorisés et même encouragés étant le centrage et la réduction de celles-ci. En conséquence, pour utiliser cette méthode lorsque le bruit n'est pas intégré au modèle, il est obligatoire d'extraire autant de composantes indépendantes que de variables observées, certaines composantes pouvant ensuite être identifiées comme des composantes de bruit.

modélisation des sources

algorithme GEM

D'autre part, nous avons considéré que les densités des sources étaient fixées a priori, mais rien n'empêche d'introduire plus de flexibilité dans le modèle en considérant des formes paramétriques pour celles-ci. L'analyse en facteurs indépendants peut en particulier être étendue pour prendre en compte de telles contraintes, il suffit pour cela d'utiliser l'algorithme GEM (algorithme 4) du chapitre 2, et de remplacer dans celui-ci l'équation de mise à jour de la matrice de démixage par l'équation de mise à jour de la matrice de mixage prenant en compte les contraintes (4.20). Le pseudo code de cet algorithme est donné ci-après, (algorithme 8).

Nous allons maintenant présenter quelques résultats expérimentaux visant à illustrer l'intérêt d'une telle approche.

4.2.3 Expérimentation

Pour quantifier l'apport de la prise en compte d'informations a priori sur le processus de mélange, le protocole expérimental suivant a été élaboré. Nous avons tout d'abord utilisé un modèle spécifique pour simuler différents jeux de données. Ce modèle correspond à celui de l'IFA avec $S = 6$ sources et un nombre identique de variables observées. Les densités des sources utilisées pour les simulations sont représentées sur la figure 4.1, celles-ci couvrent différentes situations, sources uni-

Algorithme 8: pseudo-code de l'analyse en facteurs indépendants sans bruit, avec contraintes sur la matrice de mixage en utilisant un gradient naturel.

Données : Matrice de données centrée $\mathbf{X}$, matrice des contraintes M
 # Initialisation du vecteur de paramètres

$\boldsymbol{\psi}^{(0)} = (A^{(0)}, \boldsymbol{\pi}^{1(0)}, \ldots, \boldsymbol{\pi}^{S(0)}, \boldsymbol{\mu}^{1(0)}, \ldots, \boldsymbol{\mu}^{S(0)}, \boldsymbol{\nu}^{1(0)}, \ldots, \boldsymbol{\nu}^{S(0)})$
$A^{(0)} = M \bullet A^{(0)}$, $q = 0$

tant que *test de convergence* **faire**
 # Mise à jour des sources
$\mathbf{Z} = \mathbf{X}\left(A^{(q)^{-1}}\right)^t$

 # Mise à jour des paramètres des sources / EM
pour tous les $s \in \{1, \ldots, S\}$ *et* $k \in \{1, \ldots, K_s\}$ **faire**
 # Etape E
$t_{ik}^{s(q)} = \dfrac{\pi_k^{s(q)} \varphi(z_{is}; \mu_k^{s(q)}, \nu_k^{s(q)})}{\sum_{k'=1}^{K_s} \pi_{k'}^{s(q)} \varphi(z_{is}; \mu_{k'}^{s(q)}, \nu_{k'}^{s(q)})}, \quad \forall i \in \{1, \ldots, N\}$

pour tous les $s \in \{1, \ldots, S\}$ *et* $k \in \{1, \ldots, K_s\}$ **faire**
 # Etape M
 # Mise à jour des paramètres des sources

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

 # Mise à jour de G (4.34)
$\mathbf{G} = \mathbf{g}^{(q+1)}(\mathbf{Z})$

 # Calcul du gradient naturel (4.18)
$\Delta A = \left(A^{(q)^{-1}}\right)^t \left(\frac{1}{N} \sum_{i=1}^N \mathbf{g}\left(\mathbf{z}_i^{(q)}\right) \mathbf{z}_i^{(q)^t} - \mathbf{I}\right)$

 # Recherche linéaire sur τ
$\tau^* =$ RechercheLineaire($A^{(q)}$,$M \bullet \Delta A$)

 # Mise à jour de la matrice de mixage (cf. annexe .5)
$A^{(q+1)} = A^{(q)} + \tau^*.M \bullet \Delta A$

 # Normalisation des sources
pour tous les $s \in \{1, \ldots, S\}$ **faire**
$\sigma_s^2 = \sum_{k=1}^{Ks} \pi_k^{s(q+1)} (\nu_k^{s(q+1)} + \mu_k^{s(q+1)^2}) - \left(\sum_{k=1}^{Ks} \pi_k^{s(q+1)} \mu_k^{s(q+1)}\right)^2$
pour tous les $k \in \{1, \ldots, K_s\}$ **faire**
$\mu_k^{s(q+1)} = \mu_k^{s(q+1)} / \sigma_s$
$\nu_k^{s(q+1)} = \nu_k^{s(q+1)} / \sigma_s^2$
$A_{s.}^{(q+1)} = \sigma_s A_{s.}^{(q+1)}$
$q \leftarrow q + 1$

Résultat : Paramètres estimés : $\widehat{\psi}^{ml}$, variables latentes estimées : $\widehat{\mathbf{Z}}^{ml}$

modale ou bi-modale, source leptokurtique[1] ou platikurtique[2]. Finalement la matrice de mixage utilisée a été simulée en utilisant une loi normale centrée réduite pour chacun des coefficients, des zéros étant introduits au sein de celle-ci de manière aléatoire (chaque coefficient ayant une probabilité 0.3 d'être nul), lors de cette expérience 12 coefficients de la matrice étaient nuls sur 36.

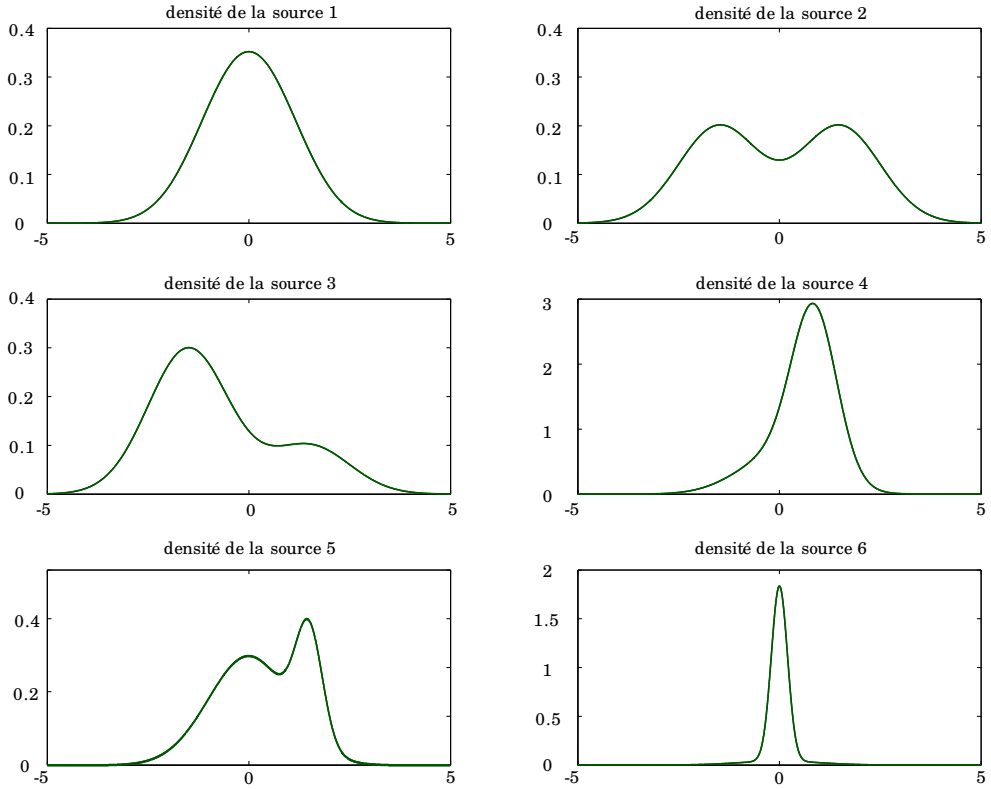


FIG. 4.1 – *Densités des sources utilisées pour les simulations, ces densités correspondent à des mélanges de loi normale monodimensionnelle à 2 composantes, les paramètres des différentes sources ont été choisis de manière à couvrir différentes formes de sources unimodale, bi-modale, leptokurtique, platikurtique...*

Grâce à des jeux de données simulées, nous avons pu comparer les résultats de l'IFA classique, c'est-à-dire n'utilisant pas la connaissance quant à la localisation des zéros dans la matrice de mixage, et ceux de l'IFA avec prise en compte des contraintes sur la matrice de mixage (notée IFAc). La comparaison a été menée sur des jeux de données de tailles différentes $N \in \{100, \ldots, 700\}$, en utilisant une mesure de performance bien connue dans le domaine de l'analyse en composantes indépendantes, l'indice de performance d'Amari (Cichocki et Amari 2002, p. 308). Cette mesure de performance est basée sur l'analyse de la matrice Q définie par :

$$Q = H\widehat{W}, \tag{4.22}$$

indice de performance d'Amari

où H est la matrice de mixage exacte ayant servie aux simulations et $\widehat{W}$ est la matrice de démixage estimée. Si le système de mixage a été correctement identifié

[1] Adjectif decrivant une variable aléatoire dont le kurtosis ou coefficient d'applatissement (rapport du moment centré d'ordre 4 et du carré du moment centré d'ordre 2) est > 3, 3 étant le kurtosis de la loi normale centrée réduite.

[2] Adjectif decrivant une variable aléatoire dont le kurtosis est < 3.

cette matrice Q doit être égale à une matrice de permutation P quelconque multipliée par une matrice diagonale de rang plein D elle aussi quelconque. En effet, dans ce cas de figure, les sources ont été estimées aux deux indéterminations du modèle près (permutation, facteur d'échelle). En partant de ce constat S. Amari a proposé l'indicateur $A_p(Q)$ suivant pour mesurer les performances d'un algorithme d'ACI :

$$A_p(Q) = \frac{1}{S}\sum_{i=1}^{S}\left(\left(\sum_{j=1}^{S}\frac{|Q_{ij}|}{max(|Q_{i.}|)} - 1\right) + \left(\sum_{j=1}^{S}\frac{|Q_{ji}|}{max(|Q_{.i}|)} - 1\right)\right). \tag{4.23}$$

Cet indice mesure l'écart entre la matrice Q obtenue et la matrice de la forme $P\,D$ la plus semblable au sens où : il est nul lorsque la matrice de mixage a été parfaitement estimée (aux indéterminations du modèle près) et qu'il croît lorsque l'erreur d'estimation augmente.

Toutes les expériences ont été réalisées avec 30 jeux de données d'apprentissage différents et nous présentons les résultats moyens des deux méthodes (IFA avec et sans contraintes) sur ces 30 jeux de données. Finalement, le problème des maximums locaux, rencontré par les algorithmes GEM servant à l'estimation, a été traité en utilisant une procédure d'initialisation aléatoire des paramètres (25 initialisations différentes parmi lesquelles est conservé celle qui conduit à l'estimé ayant la plus grande vraisemblance). Les résultats sont présentés en figure 4.2. Nous pouvons observer l'apport de l'information a priori qui fait chuter l'indice de 0.5 points en moyenne. Quelque soit la taille de l'ensemble d'apprentissage utilisé, l'indice de performance d'Amari est meilleur lorsque les connaissances a priori sur la forme de la matrice de mixage sont prises en considération.

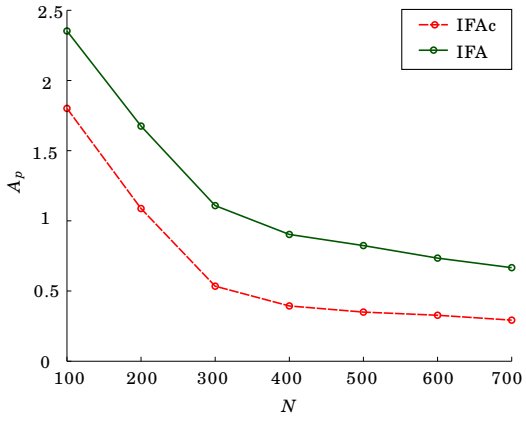


FIG. 4.2 – *Indice de performance d'Amari moyenné sur 30 jeux de données, en fonction du nombre d'individus ayant servi à l'estimation, pour l'IFA et pour l'IFA avec prise en compte des contraintes sur la matrice de mixage (IFAc).*

Nous allons maintenant présenter notre seconde contribution qui concerne la prise en compte d'informations partielles sur les individus utilisés lors de l'apprentissage de l'analyse en facteurs indépendants.

4.3 EXTENSION AVEC LABELS DOUX

L'objectif des travaux présentés ici est similaire à celui du chapitre précédent, c'est-à-dire proposer une extension d'un modèle génératif intégrant des variables latentes pour pouvoir prendre en compte des informations partielles sur les individus de l'ensemble d'apprentissage.

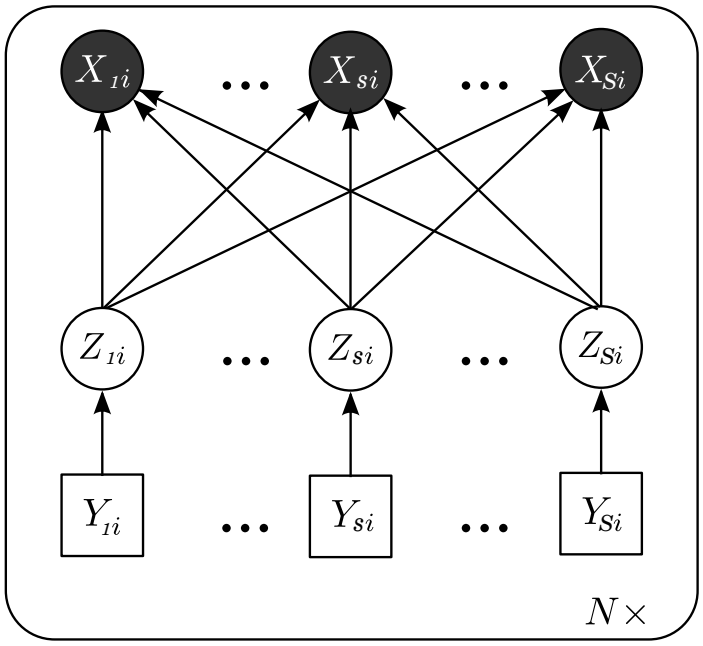


FIG. 4.3 – *Modèle graphique de l'analyse en facteurs indépendants (les paramètres du modèle n'ont pas été représentés pour ne pas alourdir le schéma).*

Nous allons voir dans cette section, plus précisément, comment une information imprécise incertaine sur la composante d'origine d'un individu i (c'est-à-dire une information sur la valeur de la variable latente Y_{si} du modèle de l'IFA, cf. figure 4.3), peut être prise en compte lors de l'apprentissage. La solution proposée s'appuie sur une démarche similaire à celle adoptée au chapitre précédent. La théorie des fonctions de croyance nous permettra tout d'abord de construire un critère d'estimation des paramètres prenant en compte les informations partielles disponibles sur les variables latentes. Nous proposerons ensuite un algorithme capable de maximiser ce critère par rapport aux paramètres du modèle. Enfin, nous étudierons le comportement de cette méthode sur des données simulées dans un contexte semi-supervisé.

4.3.1 Fonction de vraisemblance généralisée

Comme dans le cadre des modèles de mélange, le jeu de données est supposé être constitué de mesures dans $\mathbb{R}^S$ pour différents individus et d'un ensemble de fonctions de masses pour chacun des individus. Cependant, à l'inverse du chapitre précédent, chaque individu n'est plus décrit à l'aide d'une unique fonction de masse spécifiant notre connaissance sur la composante d'origine de l'individu mais d'un ensemble de fonction de masses $m^{\mathcal{Y}_1}, \ldots, m^{\mathcal{Y}_S}$ spécifiant notre connaissance sur la composante d'origine de l'individu pour chacune des S sources. Plus formellement

le jeu de données est défini par :

$$\mathbf{X}^{iu} = \{(\mathbf{x}_1, m_1^{\mathcal{Y}_1}, \ldots, m_1^{\mathcal{Y}_S}), \ldots, (\mathbf{x}_N, m_N^{\mathcal{Y}_1}, \ldots, m_N^{\mathcal{Y}_S})\}, \tag{4.24}$$

avec $m_i^{\mathcal{Y}_s}$ une fonction de masse sur le cadre de discernement $\mathcal{Y}_s$ correspondant aux différentes composantes de la densité de la source s.

Le modèle postulé dans cette section correspond au modèle de l'IFA sans bruit défini à l'aide des équations (4.1,4.2,4.3). Nous supposons seulement en plus que les S sources suivent un modèle de mélange gaussien et que leurs densités sont de la forme :

$$f^{\mathcal{Z}_s}(z_s) = \sum_{k=1}^{K_s} \pi_k^s \varphi\left(z_s, \mu_k^s, \nu_k^s\right), \tag{4.25}$$

avec $\varphi(., \mu, \nu)$ la densité de la loi normale monodimensionnelle de moyenne μ et de variance ν. Le modèle comprend, outre la matrice de démixage W, les paramètres des densités des sources, c'est-à-dire les proportions, les moyennes et les variances des différentes composantes de chacune des sources. L'ensemble de ces paramètres sera noté $\boldsymbol{\psi}$:

$$\boldsymbol{\psi} = (W, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^S, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^S, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^S). \tag{4.26}$$

Pour tirer parti de ces informations nous proposons d'utiliser une démarche similaire à celle adoptée aux chapitre précédent. Nous recherchons donc le vecteur de paramètre $\hat{\boldsymbol{\psi}}$ tel que :

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}). \tag{4.27}$$

Nous allons voir que ce principe mène ici aussi à un critère naturel étendant les critères probabilistes classiques.

Proposition 4.3 *Si nous supposons que les différents labels sont indépendants entre eux et indépendants des observations* $\mathbf{X}$ *générées de manière i.i.d. suivant le modèle génératif associé à l'IFA, le logarithme de la plausibilité conditionnelle des paramètres du modèle* $\boldsymbol{\psi}$ *sachant le jeu de données* $\mathbf{X}^{iu}$*, est donné par :*

$$\log\left(pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})\right) = N\log(|\det(W)|) + \sum_{i=1}^{N}\sum_{s=1}^{S}\log\left(\sum_{k=1}^{K_s} pl_{ik}^s \pi_k^s \varphi\left((W\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right) + cst. \tag{4.28}$$

avec pl_{ik}^s *la plausibilité que l'individu* i *provienne de la composante* k *de la variable latente* s*, ces plausibilités devant être calculées à partir des labels doux* $m_i^{\mathcal{Y}_s}$*, et* cst *étant une constante indépendante de* $\boldsymbol{\psi}$*.*

Preuve. En utilisant le théorème de Bayes généralisé (2.88), la plausibilité des paramètres peut être exprimée en fonction de la plausibilité des observations :

$$pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}) = pl^{\mathcal{X}_1\times\ldots\times\mathcal{X}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N|\boldsymbol{\psi}). \tag{4.29}$$

En supposant que les observations ainsi que les labels sont cognitivement indépendants conditionnellement aux paramètres (2.85), cette plausibilité peut être décomposée en un produit sur l'ensemble des exemples d'apprentissage :

$$pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}) = \prod_{i=1}^{N} pl^{\mathcal{X}_i}(\mathbf{x}_i|\boldsymbol{\psi}). \tag{4.30}$$

En utilisant la relation linéaire et déterministe entre $\mathbf{x}_i$ et $\mathbf{z}_i$, l'hypothèse d'indépendance des composantes de $\mathbf{z}$ et l'hypothèse d'indépendance des différents labels d'un même individu, nous obtenons :

$$pl^{\mathbf{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}) = \prod_{i=1}^{N} |\det(W)| \prod_{s=1}^{S} pl^{\mathcal{Z}_{is}}((W\mathbf{x}_i)_s|\boldsymbol{\psi}). \tag{4.31}$$

Or en utilisant (3.35) nous pouvons écrire :

$$pl^{\mathcal{Z}_{is}}((W.\mathbf{x}_i)_s|\boldsymbol{\psi}) \propto \sum_{k=1}^{Ks} pl_{ik}^{s}.\pi_k^s \varphi((W\mathbf{x}_i)_s; \mu_k^s, \nu_k^s) \tag{4.32}$$

et donc :

$$pl^{\mathbf{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}) \propto \prod_{i=1}^{N} |\det(W)| \prod_{s=1}^{S} \sum_{k=1}^{Ks} pl_{ik}^{s}\, \pi_k^s \varphi((W\mathbf{x}_i)_s; \mu_k^s, \nu_s^s) \tag{4.33}$$

En prenant le logarithme de l'expression précédente (4.33), nous obtenons le résultat final (4.28). □

Remarque 4.1 *Lien avec les critères probabilistes : comme dans le contexte des modèles de mélange, notre critère (4.28) permet de retrouver les critères probabilistes comme cas particulier. En effet, si tous les labels sont vides, nous avons $pl_{ik}^{s} = 1$ pour toutes les composantes, tous les individus et toutes les sources et nous retrouvons alors le critère non supervisé. En utilisant des labels catégoriques, nous retrouvons facilement, à partir de notre critère, les critères semi-supervisé et partiellement supervisé.*

Le critère proposé pour intégrer des connaissances partielles au modèle de l'IFA a donc une forme simple et facile à interpréter, proche de celle associée au critère non supervisé. Une fois encore les informations sur les composantes d'origine d'un individu interviennent aux travers des plausibilités des différentes composantes pl_{ik}^{s}, celles-ci venant pondérer les contributions des différentes composantes.

Nous allons voir maintenant comment un algorithme de type GEM peut être utilisé pour estimer les différents paramètres du modèle.

4.3.2 Considérations pratiques

La mise en place pratique d'un algorithme capable d'optimiser le critère (4.28), ne pose pas de problème particulier. En effet, si l'on fixe la matrice de mixage ou de démixage, nous retrouvons S problèmes d'estimation de paramètres de modèles de mélange avec des étiquettes imprécises, incertaines. Dans ce cas, l'algorithme EM présenté au chapitre précédent (cf. algorithme 7), peut être utilisé pour faire croître la vraisemblance par rapport aux paramètres de chacune des sources. De la même manière, si les paramètres des sources sont fixés, un gradient ou un gradient naturel peut être utilisé pour faire croître la vraisemblance par rapport à la matrice de mixage ou de démixage. Il est donc possible d'utiliser un algorithme d'optimisation alternée de type GEM pour maximiser le critère (4.28). Cet algorithme est identique à l'algorithme classique pour l'IFA sans bruit (cf. algorithme 4), les seules modifications concernent les étapes E d'estimation des probabilités

a posteriori d'appartenance aux composantes des différentes sources et le calcul de la fonction g qui dépend des paramètres des sources mais aussi des étiquettes. Celle-ci est donnée par :

$$\begin{aligned} g_s(z_{is}) &= \frac{-\partial \log\left(\sum_{k=1}^{K_s} pl_{ik}^s \pi_k^s \varphi(z_{is}; \mu_k^s, \nu_k^s)\right)}{\partial z_{is}} \\ &= \sum_{k=1}^{K_s} t_{ik}^s \frac{(z_{is} - \mu_k^s)}{\nu_k^s}, \end{aligned} \tag{4.34}$$

avec t_{ik}^s les probabilités a posteriori d'appartenance aux composantes de la source s connaissant $z_{is} = (W\mathbf{x}_i)_s$ et les étiquettes, c'est-à-dire :

$$t_{ik}^s = \frac{pl_{ik}^s \pi_k^s \varphi(z_{is}; \mu_k^s, \nu_k^s)}{\sum_{k'=1}^{K_s} pl_{ik'}^s \pi_{k'}^s \varphi(z_{is}; \mu_{k'}^s, \nu_{k'}^s)}. \tag{4.35}$$

L'algorithme 9 présente en détail les différentes étapes de l'optimisation. Il est intéressant de remarquer que les labels « doux » ont un impact à la fois sur l'estimation des paramètres des sources et sur le calcul du gradient permettant d'optimiser la vraisemblance par rapport à la matrice de mixage ou de démixage.

Un autre point mérite d'être souligné, celui de l'indétermination ou non du modèle par rapport aux permutations des sources. La partie expérimentale de cette section montrera en particulier que l'apport d'information aux travers de labels « doux » peut lever ce problème. En effet, si les labels sont informatifs et différents d'une source à l'autre, le critère n'est plus invariant par rapport aux permutations.

Remarque 4.2 *La complexité de l'algorithme proposé pour estimer les paramètres de l'IFA lorsque des données labellisées de manière douce sont disponibles est strictement identique à la complexité de l'algorithme dans le cadre non supervisé. Les labels peuvent cependant avoir un impact non négligeable sur le nombre d'itérations nécessaires pour parvenir à la convergence.*

4.3.3 Expérimentations

Afin d'étudier l'intérêt de l'approche proposée, différentes expérimentations ont été effectuées. Nous avons tout d'abord chercher à savoir quelle était l'influence de la labellisation dure de certaine sources sur l'indétermination du critère par rapport aux permutations des sources. Nous avons également chercher à mesurer l'influence de l'apport d'information aux travers de labels sur la difficulté du problème d'optimisation en calculant des indicateurs permettant de « quantifier » celle-ci tel que le nombre de maximums locaux différents trouvés par l'algorithme d'estimation à partir d'un nombre constant d'initialisations aléatoires. Enfin, nous avons cherché à observer l'impact de la labellisation sur la qualité de l'estimation. Les expériences présentées ici concernent essentiellement l'influence d'un étiquetage partiel mais correct des données sur les performances de l'IFA. Un travail complémentaire sur la prise en compte du bruit d'étiquetage grâce à des labels doux fait partie de nos perspectives.

Algorithme 9: pseudo-code de analyse en facteurs indépendants sans bruit avec labellisation douce en utilisant une montée de gradient naturel pour l'optimisation par rapport à la matrice de démixage.

Données : Matrice de données centrée $\mathbf{X}$, Labels doux $\{pl_i^s\}_{i=1\ldots N, s=1\ldots S}$

 # Initialisation du vecteur de paramètres

$$\boldsymbol{\psi}^{(0)} = (W^{(0)}, \boldsymbol{\pi}^{1(0)}, \ldots, \boldsymbol{\pi}^{S(0)}, \boldsymbol{\mu}^{1(0)}, \ldots, \boldsymbol{\mu}^{S(0)}, \boldsymbol{\nu}^{1(0)}, \ldots, \boldsymbol{\nu}^{S(0)}), \quad q = 0$$

tant que *test de convergence* **faire**

 # Mise à jour des sources

$\mathbf{Z} = \mathbf{X}.{W^{(q)}}^t$

 # Mise à jour des paramètres des sources / EM

pour tous les $s \in \{1, \ldots, S\}$ *et* $k \in \{1, \ldots, K_s\}$ **faire**

 # Etape E

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

pour tous les $s \in \{1, \ldots, S\}$ *et* $k \in \{1, \ldots, K_s\}$ **faire**

 # Etape M

 # Mise à jour des paramètres des sources

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

 # Mise à jour de G (4.34)

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

 # Calcul du gradient naturel (4.18)

$\Delta W = (\mathbf{I} - \frac{1}{N} \mathbf{G}^t.\mathbf{Z}).{W^{(q)}}^t$

 # Recherche linéaire sur τ

$\tau^* =$ RechercheLineaire($W^{(q)}$,ΔW)

 # Mise à jour de la matrice de démixage

$W^{(q+1)} = W^{(q)} + \tau^*.\Delta W$

 # Normalisation des sources

pour tous les $s \in \{1, \ldots, S\}$ **faire**

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

pour tous les $k \in \{1, \ldots, K_s\}$ **faire**

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

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

$q \leftarrow q + 1$

Résultat : Paramètres estimés : $\widehat{\psi}^{ml}$, variables latentes estimées : $\widehat{\mathbf{Z}}^{ml}$

Utilité de l'étiquetage partiel pour le problème de permutation des sources

Afin de mettre en évidence l'influence de la labellisation sur le problème de permutations des sources nous avons utilisé notre algorithme sur le même jeu de données qu'en section 4.2.3 dans trois configurations différentes :
- en utilisant des labels non informatifs (non-supervisée) ;
- en utilisant 25% d'individus labellisés pour les trois premières sources ;
- en utilisant 25% d'individus labellisés pour toutes les sources.

Nous présentons en figure 4.4 des exemples de résultats obtenus par notre algorithme dans ces trois situations sur des jeux de données de 500 individus simulés suivant le modèle de l'IFA avec les sources présentées en figure 4.1. Ces résultats sont fournis sous la forme de matrices contenant les valeurs absolues des coefficients de corrélations entre les sources estimées et les sources réelles, lesquelles sont évaluées à l'aide d'un jeu de données indépendant de 5000 individus.

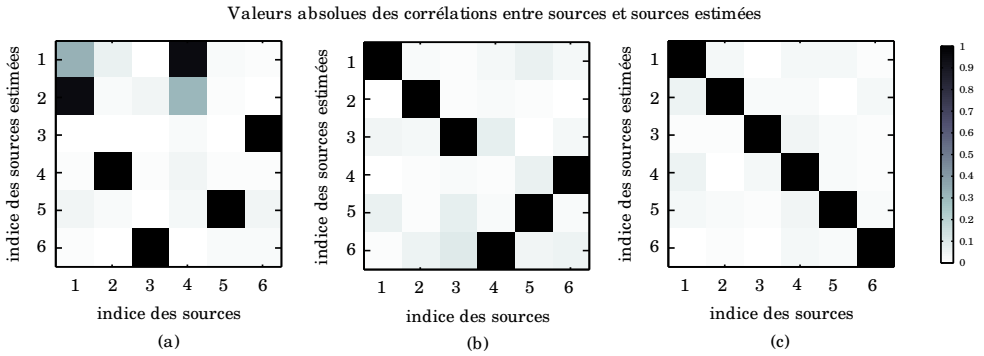


FIG. 4.4 – *Expérience IFA non supervisée et semi-supervisée et indétermination du modèle par rapport aux permutations des sources : matrice des valeurs absolues des corrélations entre les sources estimées et les sources ayant servi à la simulation pour l'IFA classique (a), pour l'IFA semi-supervisée avec* 25% *d'individus labellisés sur les trois premières sources (b) et pour l'IFA semi-supervisée avec* 25% *d'individus labellisés sur toutes les sources (c).*

Nous pouvons observer sur cette figure, l'intérêt de la labellisation pour résoudre le problème des permutations. Les sources pour lesquelles des individus labellisés ont été fournis à l'algorithme ont en effet clairement été retrouvées sans permutations (valeurs absolues du coefficient de corrélation très proche de 1 sur la diagonale et très faibles ailleurs). Il est important de souligner que ces résultats dépendent certainement de la forme des densités des sources. Ils sont donc donnés à titre illustratif, la quantité d'individus labellisés nécessaire pour retrouver les sources sans permutation dépend sans aucun doute de nombreux paramètres, à savoir la forme de la densité de la source considérée, la forme des densités des autres sources et la quantité d'individus disponibles pour l'apprentissage...

D'autre part, d'un point de vue qualitatif et hormis les problèmes de permutations, il semble en observant ces matrices que les résultats obtenus soient de meilleure qualité lorsque des données partiellement labellisées sont utilisées (meilleur contraste dans les matrices). L'expérience présentée en section 4.3.3 nous permettra de mieux quantifier cette amélioration. Mais avant cela, nous allons étudier l'influence de la labellisation sur le problème d'optimisation.

Influence de l'étiquetage sur le problème d'optimisation

Nous avons également voulu quantifier l'influence des labels durs sur la difficulté du problème d'optimisation. Pour cela, nous avons calculé le nombre de maximums locaux différents trouvés par l'algorithme GEM à partir de 100 initialisations aléatoires en fonction du pourcentage d'individus labellisés $\{10\%, 20\%, \ldots, 90\%\}$. Dans cette expérience la labellisation concerne l'ensemble des sources. L'expérience a été répétée avec 10 jeux de données différents simulés suivant le modèle de l'IFA avec les 6 sources présentées en figure 4.1, contenant chacun 500 individus. L'algorithme utilisé correspond à l'algorithme GEM avec une montée de gradient de type gradient naturel sur la matrice de mixage. Les résultats sont donnés en figure 4.5. Pour savoir si deux maximums de la fonction de vraisemblance étaient identiques, nous avons comparé les matrices de démixage estimées. En prenant garde à l'indétermination d'échelle, le critère

$$||\widehat{W}_1 - \widehat{W}_2||^2/S^2, \tag{4.36}$$

avec S le nombre de sources, nous a permis de tester l'égalité de deux solutions $\widehat{W}_1$ et $\widehat{W}_2$. Pour ne pas prendre en compte l'indétermination d'échelle dans le décompte des maximums, nous avons à la fois contraint les variances des sources estimées à être égales à 1 et les coefficients de la diagonale de la matrice de démixage à être positifs pour éviter les problèmes de signe. Finalement, deux maximums locaux de la vraisemblance ont été considérés comme identiques quand :

$$\frac{||\widehat{W}_1 - \widehat{W}_2||^2}{S^2} < 0.01. \tag{4.37}$$

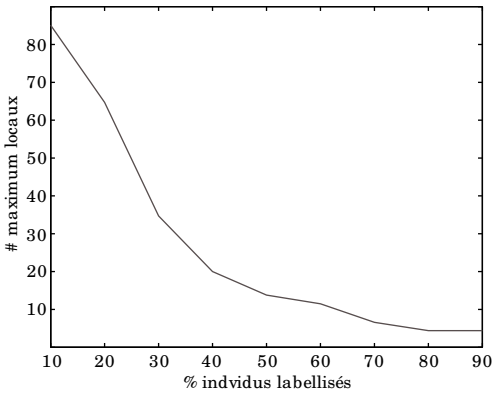


FIG. 4.5 – *Influence de l'étiquetage sur le problème d'optimisation : nombre de maximums locaux différents trouvés par l'algorithme GEM à partir de* 100 *initialisations aléatoires. Les résultats sont des moyennes sur 10 jeux de données différents, et sont présentés en fonction du pourcentage d'individus labellisés. L'algorithme d'optimisation correspond à l'algorithme GEM avec une montée de gradient de type gradient naturel sur la matrice de mixage.*

Nous pouvons observer sur cette figure l'influence importante de la labellisation sur le nombre de maximums locaux. Plus le nombre d'individus labellisés est important, plus l'algorithme a un nombre réduit de maximums locaux à gérer. Une fois encore les résultats présentés ici sont illustratifs et dépendent de nombreux paramètres, même si de manière générale il semble logique que l'apport d'information améliore le conditionnement du problème.

Nous allons maintenant essayer de quantifier l'impact de la labellisation sur la qualité des estimations obtenues.

Influence de l'étiquetage sur la qualité de l'estimation

En utilisant toujours le même modèle de simulation, nous avons analysé l'influence de la quantité d'individus labellisés sur les performances de l'algorithme. L'indice de performance d'Amari (4.23) a pour cela été calculé sur 30 jeux de données d'apprentissage indépendants de 500 individus, contenant chacun, un pourcentage variable d'individus labellisés : $\{5\%, 10\%, 15\%, \ldots, 50\%\}$. Dans cette expérience la labellisation concerne l'ensemble des sources. Comme à l'accoutumé, le problème des maximums locaux a été résolu en utilisant différentes initialisations aléatoires pour l'algorithme (10 initialisations différentes par jeu de données). Nous avons également calculé un indice permettant de mesurer la qualité de l'estimation différent de l'indice d'Amari car ce dernier est sensible aux permutations des sources, cet indicateur noté $r2$ est défini par :

$$r2(\widehat{W}) = \frac{1}{S}\sum_{s=1}^{S} r^2_{\widehat{z}_s, z_s}, \tag{4.38}$$

avec $\widehat{z}_s = (\widehat{W}\mathbf{x})_s$ et $r_{x,y}$ l'estimateur de Pearson de la corrélation. Celui-ci correspond à la moyenne sur l'ensemble des sources du coefficient de détermination (coefficient de corrélation au carré, taux de variance expliquée) entre sources réelles et sources estimées. Cette corrélation étant elle-même estimée à l'aide d'un ensemble de données indépendant de 5000 individus simulés suivant le modèle. Ce critère est égal à 1 lorsque toutes les sources ont été correctement estimées sans aucune permutation et égal à 0 si les sources ont été correctement estimées mais permutées. La figure 4.6 présente les résultats obtenus sous la forme de boxplots sur les 30 jeux de données.

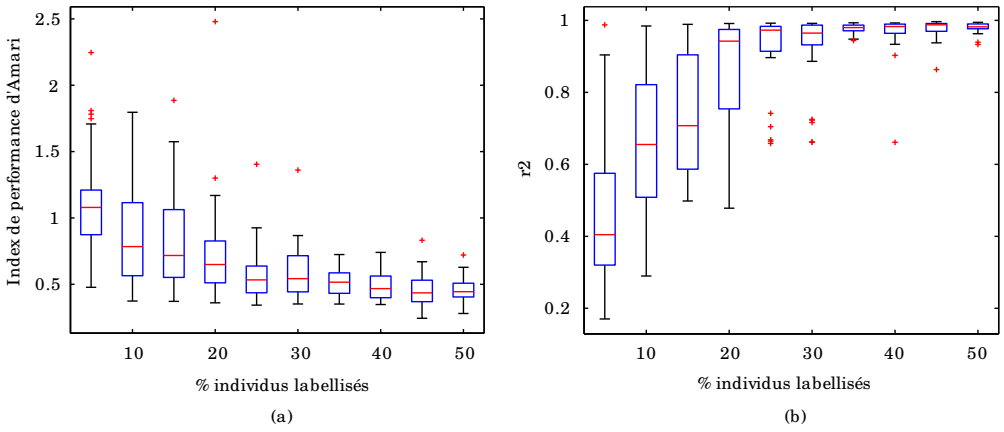


FIG. 4.6 – *Influence de l'étiquetage sur la qualité de l'estimation d'une IFA semi-supervisée : boxplot de l'indice de performance d'Amari (a) et de l'indicateur $r2$ (b) en fonction du pourcentage d'individus labellisés $\{5\%, 10\%, 15\%, \ldots, 50\%\}$ de l'ensemble d'apprentissage pour 30 jeux de données indépendants.*

Nous pouvons clairement observer sur ces deux figures l'influence bénéfique de l'information apportée par les labels. L'indice de performance d'Amari décroît de manière importante lorsque le pourcentage de données labellisées augmente. De même les résultats du critère $r2$ sont intéressants, lorsque peu d'individus sont labellisés $< 20\%$ nous pouvons observer une grande variabilité de cet indicateur,

ce qui s'explique par le fait que l'algorithme converge vers des maximums locaux correspondant à des permutations des sources. Par contre, lorsque le pourcentage d'individus labellisés croît $> 20\%$, cet indicateur est extrêmement proche de 1 pour la majorité des jeux de données. L'algorithme a donc quasiment toujours convergé vers un maximum correspondant à une matrice de mixage non permutée. Lorsque le pourcentage d'individu labellisés est de l'ordre de 20 à 30%, la plupart des jeux de données ont permis de retrouver les sources sans permutation mais quelques individus atypiques correspondent encore à des maximums locaux sous optimaux.

En conclusion, les différentes expériences menées jusqu'à présent sur données simulées, nous ont permis de mettre en avant la possibilité d'améliorer la qualité de l'estimation des paramètres de l'IFA lorsque des données labellisées sont fournies. Celles-ci nous ont également permis de mettre en lumière l'influence importante de la labellisation sur le problème de l'indétermination du modèle de l'IFA par rapport aux permutations des sources et sur la difficulté du problème d'optimisation associé à l'IFA. Plus la quantité d'information sur les individus servant à l'apprentissage est importante plus le nombre de maximums locaux rencontrés par l'algorithme est faible et plus il est aisé pour celui-ci de converger vers un maximum intéressant.

Conclusion du chapitre

Ce chapitre a été l'occasion de présenter nos travaux sur deux extensions de l'analyse en facteurs indépendants (IFA). La première concernait la prise en considération d'hypothèses a priori d'indépendance entre certaines variables latentes et certaines variables observées. Nous avons montré que ce type d'hypothèse se traduisait par des contraintes de nullité de certains coefficients de la matrice de mixage et qu'il était aisé d'en tirer parti dans le cadre d'une approche de type vraisemblance. Une expérience sur données simulées nous a permis de montrer l'influence positive de ce type de contraintes sur les résultats de l'IFA, lorsque ces hypothèses sont étayées évidement.

La seconde contribution proposée visait à étendre les travaux du chapitre précédent au cadre de l'analyse en facteurs indépendants. Nous avons pour cela défini un critère permettant de prendre en compte des informations partielles sur les composantes d'origine des individus servant à l'estimation des paramètres. Nous avons également proposé un algorithme simple de type GEM pour optimiser ce critère et nous avons finalement étudié le comportement de notre méthode sur données simulées. Ces expériences nous ont permis de démontrer l'intérêt de la supervision même partielle de l'IFA, à la fois pour la simplification du problème d'estimation et pour la qualité du résultat du démixage.

5 Application au diagnostic des circuits de voie ferroviaire

Qui se résigne à chercher des preuves d'une chose
à laquelle il ne croit pas ou dont la prédication ne l'intéresse pas ?
Jorge Luis Borges , **Trois versions de Judas, dans Fictions (1944)**

Sommaire

DANS ce chapitre, nous présentons l'application pratique à l'origine de ces travaux de thèse. Celle-ci concerne le diagnostic d'un élément essentiel de la chaîne de contrôle commande des trains sur le réseau français, les circuits de voie (CdV). Après avoir décrit l'application, ces enjeux et ces particularités, nous présenterons l'approche adoptée pour mettre au point un système de diagnostic des « condensateurs d'accord », composant essentiel au bon fonctionnement des circuits de voie. Ce système de diagnostic sera bâti autour de l'analyse d'un signal enregistré grâce à un véhicule d'inspection spécifique.

L'objectif final de nos propositions s'inscrit dans le contexte du développement d'une maintenance préventive efficace, limitant au maximum le temps d'indisponibilité du système grâce à une localisation précise des défauts couplée à l'estimation de leur gravité.

Nous verrons en particulier, comment les différentes propositions théoriques détaillées dans les chapitres 3 et 4 de cette thèse peuvent être mises à profit pour l'utilisation d'informations imparfaites lors de l'apprentissage et pour la prise en compte d'informations a priori dans le cadre de l'IFA au travers de contraintes sur la matrice de mixage. Ainsi, le diagnostic des « condensateurs d'accord » prendra en compte toute l'information experte disponible et tirera parti des connaissances physiques existantes sur le système.

5.1 Description du système complexe

5.1.1 Généralités

circuit de voie

Les circuits de voie (CdV) ont un rôle essentiel en signalisation ferroviaire, puisqu'ils permettent de détecter de façon automatique et continue la présence d'un véhicule sur une portion de voie donnée et commandent grâce à cette information l'activation de panneaux de signalisation placés en bord de voie, ceux-ci indiquant aux conducteurs la vitesse maximale autorisée.

Le système de signalisation est conçu de façon à assurer la sécurité des circulations. Il permet notamment d'éviter les collisions entre deux trains circulant dans le même sens sur la même voie. Pour cela, la voie est divisée en zones appelées cantons, dont la longueur peut varier entre quelques centaines de mètres et quelques kilomètres suivant la vitesse maximale autorisée sur la ligne. Un espacement d'au moins un canton entre deux trains est imposé. Un feu de signalisation à l'entrée de chaque canton indique si un véhicule peut ou non circuler sur la portion de voie correspondant et sont ces feux sont commandés grâce aux CdV.

Le fonctionnement d'un CdV classique est assez simple (cf. figure 5.1), il est principalement constitué de 3 éléments :

- un émetteur, branché à l'une des extrémités de la zone. Il délivre un courant qui peut être, selon les types de CdV, continu, alternatif sinusoïdal, alternatif modulé ou impulsionnel ;
- une ligne de transmission, constituée par les 2 files de rails ;
- un récepteur, branché à l'autre extrémité de la zone. Il assure le filtrage, l'amplification et la transformation du signal reçu via les rails et agit sur un relais appelé relais de voie. Les contacts de ce relais sont utilisés pour établir ou couper les circuits électriques associés à la signalisation de la ligne.

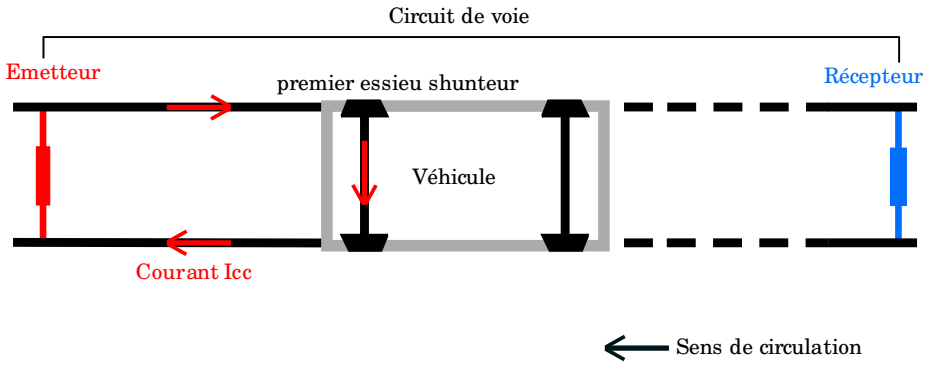


FIG. 5.1 – *Schéma de principe d'un circuit de voie non compensé*

Remarque 5.1 (Joints électriques de séparation) *Dans les versions initiales de CdV les émetteurs et récepteurs étaient associés en voie à la réalisation de joints mécaniques qui isolent électriquement les rails des cantons adjacents. Dans les versions actuelles, la coupure n'existe plus (on parle de « long rails soudés ») et l'isolation s'effectue à l'aide de bouchons électriques qui « bloquent » les courants en provenance*

des CdV voisins et qui font partie intégrante de l'émetteur et du récepteur. On parle de joints électriques de séparation (JES).

En général, un véhicule roule du récepteur vers l'émetteur. Lorsqu'aucun véhicule n'est présent sur la zone délimitant le CdV (voie libre), le signal délivré par l'émetteur parvient au récepteur à travers la ligne de transmission, et le relais de voie est excité. Le feu d'entrée du canton est vert, ce qui indique que la circulation est autorisée. En revanche, lorsqu'un véhicule est présent (voie occupée), son premier essieu agit comme une faible résistance, appelée shunt, qui court-circuite la transmission. Dans ce cas, le niveau du signal parvenant à la réception n'est plus suffisant et le relais de voie se désexcite. Le feu passe au rouge, ce qui signifie qu'aucun autre véhicule n'est autorisé à circuler sur ce canton ; un avertissement est également présenté en entrée du canton précédent.

Les CdV sont donc un élément essentiel à la chaîne de contrôle commande des trains. Nous allons voir que leur rôle est encore plus important dans le contexte français, en particulier en ce qui concerne les lignes à grande vitesse (LGV).

5.1.2 Les CdV et la grande vitesse

Le principe de signalisation décrit dans le paragraphe précédent est valable pour les lignes où la vitesse n'excède pas 220 km/h. Mais sur les lignes à grande vitesse (LGV) où les trains peuvent circuler jusqu'à 300 km/h, la signalisation est différente car l'observation des panneaux latéraux par le conducteur est trop difficile. Les informations de signalisation sont transmises directement en cabine au conducteur, au moyen d'afficheurs spécifiques indiquant les vitesses limites autorisées, les annonces d'arrêt,... Si le conducteur ne les respecte pas, une procédure automatique d'arrêt d'urgence est déclenchée.

transmission voie-machine

Ce mode de signalisation nécessite un système de transmission continue d'information entre la voie et les véhicules pour mettre à jour les afficheurs. Plusieurs technologies ont été développées par différents pays. En France, la technologie mise au point pour résoudre ce problème utilise directement les CdV comme système de transmission. Cette technologie appelée TVM (Transmission Voie-Machine), est également utilisée en Corée, et sur les lignes Eurostar et Thalys. Pour cela, le signal délivré par l'émetteur du CdV est modulé en fréquence, la modulation contenant les informations à transmettre. La transmission proprement dite s'effectue par couplage électromagnétique entre les rails et deux bobines embarquées, montées en différentiel, qui prélèvent une image du courant modulé, un mètre environ en amont du premier essieu. Ce signal est transmis à un processeur qui le filtre et le décode. L'information décodée donne des indications sur la vitesse maximale autorisée, la pente moyenne de la voie sur le canton considéré, la longueur du canton, l'occupation des cantons précédents[1]... Elle est ensuite envoyée à un ordinateur de bord qui génère le profil idéal de vitesse qui est affiché en cabine.

La TVM utilise un type particulier de CdV. Il a la structure d'un CdV classique

[1] Un canton est lui même composé d'une ou plusieurs zones de CdV selon la configuration de la portion de voie constituant le canton.

(émetteur, ligne de transmission, récepteur), et mesure entre 800 m et 2500 m. Pour éviter d'éventuelles interférences dues aux CdV encadrants, quatre fréquences différentes de porteuses sont utilisées de façon alternée sur les deux voies parallèles (voie 1 et voie 2). Sur la voie 1, les deux fréquences de CdV utilisées sont $f_1 = 1700$ Hz et $f_2 = 2300$ Hz. Sur la voie 2, il s'agit de $f_1 = 2000$ Hz et $f_2 = 2600$ Hz. Nous allons décrire le CdV-TVM plus en détails dans le paragraphe suivant, car c'est sur le diagnostic de ce type de CdV que nos travaux ont porté.

5.1.3 Description des CdV utilisés sur les lignes à grande vitesse

Un schéma type de de CdV-TVM est présenté en figure 5.2.

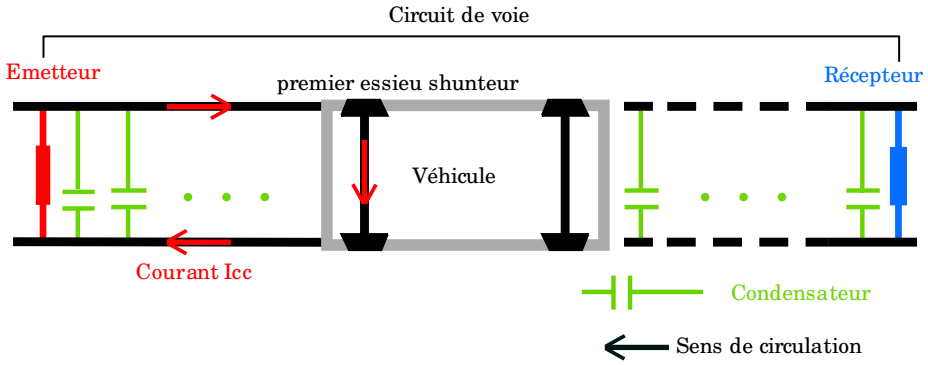


FIG. 5.2 – *Schéma de principe d'un circuit de voie compensé de type TVM*

Par rapport à un CdV classique, il présente une particularité importante. Pour que la transmission entre la voie et la machine soit efficace, il est nécessaire que le signal émis par l'émetteur se maintienne tout au long du CdV à un niveau « suffisant ». Or la voie a un comportement resistif et inductif. Pour limiter l'affaiblissement du signal résultant de la nature selfique, des condensateurs appelés « condensateurs de compensation » sont placés entre les rails à intervalles réguliers le long du CdV, selon une règle de pose prédéfinie (60, 80 ou 100 m). L'affaiblissement du courant le long du CdV a alors l'allure présenté sur la figure 5.3.

5.1.4 Les défauts possibles des CdV

Les CdV peuvent être sujets à différents types de défauts affectant la TVM, et donc la sécurité des circulations. On ne parlera ici que des défauts propres aux éléments constitutifs du CdV (JES, émetteur, récepteur et condensateurs). D'autres problèmes, liés à l'isolation de la voie ou aux rails cassés, peuvent être rencontrés.

Une première famille de défauts concerne les interférences entre différents CdV. Ce phénomène, appelé « diaphonie » peut être :

- une diaphonie transversale, due à un phénomène d'induction entre voies parallèles : un CdV situé sur une voie vient perturber un CdV situé en vis-à-vis sur la voie voisine ;

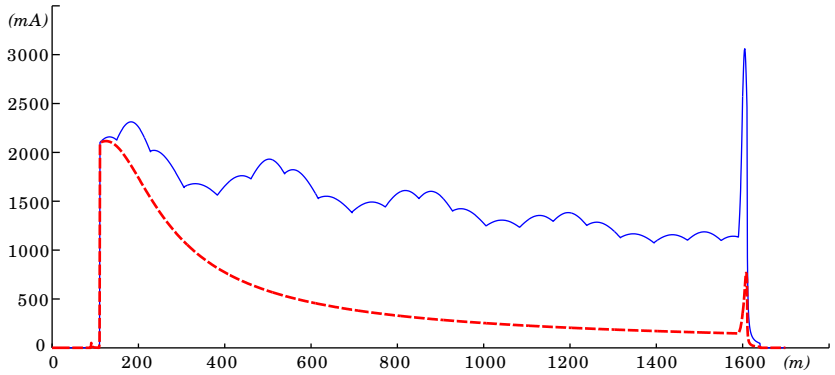


FIG. 5.3 – *Exemple de signal Icc sur circuit de voie compensé (– bleue) et non compensé (- - rouge).*

- une diaphonie longitudinale, due à un JES défectueux côté réception : un CdV vient perturber le CdV voisin situé sur la même voie.

Des défauts peuvent aussi survenir au niveau de l'émetteur et du récepteur (problèmes de connectique, vieillissement des composants), entraînant soit un niveau d'émission trop faible, soit une mauvaise réception du signal. Le risque est que l'information de signalisation ne parvienne pas en cabine, auquel cas le TGV est arrêté.

Une autre catégorie de défauts, beaucoup plus fréquente, concerne les condensateurs de compensation. Les défauts les plus classiques sont :

- un ou plusieurs condensateurs arrachés, par exemple suite à des travaux de maintenance sur les voies ;
- des problèmes de connectique, entre les condensateurs et les rails (composants mal fixés,oxydation...) ;
- une augmentation des pertes du condensateur lorsque le composant vieilli ou est soumis à des conditions climatiques extrêmes. Contrairement aux défauts évoqués ci-dessus, ce type de défauts est associé à un mode de dégradation interne du composant.

Ces défauts peuvent être particulièrement gênants, car les condensateurs servent à limiter l'affaiblissement linéique su signal support de la TVM. S'ils n'assurent plus leur fonction, le niveau de signal TVM devient trop faible, les informations de signalisation ne parviennent plus en cabine, et le TGV est automatiquement arrêté. Ceci peut perturber une grande partie du trafic ferroviaire. A titre d'information, environ 220 incidents de ce type ont été recensés en 2005 sur le réseau national.

Etant donné les conséquences des défauts de CdV sur le trafic des lignes à grande vitesse, et les répercussions que cela peut avoir sur l'ensemble du réseau ferroviaire, il est primordial, non seulement d'assurer une maintenance efficace du parc de CdV de type TVM, mais aussi de disposer d'outils performants de diagnostic afin de pouvoir agir avant l'apparition de pannes. Nous nous intéresserons plus

particulièrement à la détection des défauts de condensateurs, car ceux-ci sont plus fréquents.

5.1.5 Inspection des CdV sur le réseau français

A la SNCF, plusieurs mesures sont utilisées pour le diagnostic de l'UM71C-TVM :

- des mesures manuelles à voie libre, réalisées par des agents de maintenance. La tension à l'émetteur est relevée tous les 6 mois, ainsi que la tension aux bornes du récepteur, pour détecter d'éventuelles défaillances. La connectique est également inspectée, une fois par an, pour vérifier l'état des liaisons entre les émetteurs et les récepteurs et la voie ;
- des mesures réalisées par un véhicule d'inspection spécifique, appelé voiture *IRIS*. Cet engin parcourt les LGV toutes les 4 à 5 semaines depuis Septembre 2007, et effectue une série de mesures à 300 km/h de façon à détecter les variations de caractéristique de certains constituants. Nous allons décrire plus précisément ces enregistrements, car ils constituent la principale source d'information pour la mise au point d'un système de diagnostic automatique des CdV de type TVM.

Pour acquérir les mesures d'intérêt pour le diagnostic des CdV, la voiture *IRIS* est équipée des capteurs TVM (bobines) utilisés également sur les TGV commerciaux et d'un second capteur spécifiquement sous la forme d'une boucle inductive placée sous la voiture de mesure et conçue pour la surveillance des « condensateurs d'accord ». La voiture *IRIS* permet donc de relever deux mesures :

- une mesure de l'amplitude du courant efficace détecté par les bobines en fonction de la position du véhicule, indépendamment de la vitesse de circulation. Ce courant est appelé courant de court-circuit, noté Icc (cf. figure 5.4) ; il est relevé pour les 4 fréquences de fonctionnement des CdV ;
- une mesure d'impédance transversale, qui traduit l'accord à 25 kHz entre le dispositif de mesure et les composants reliant les rails et présentant impédance capacitive, comme par exemple les condensateurs de compensation ; cette mesure sera notée Zt.

Nos travaux ont porté sur l'analyse des signaux Icc pour le diagnotic des « condensateurs d'accord ». Le signal Zt n'a pas été utilisé dans cette étude et nous ne reviendrons pas en détails sur celui-ci. Ce choix a été motivé par la nature plus riche du signal Icc permettant d'envisager une maintenance préventive, grâce à l'extraction d'indicateurs de l'état de dégradation d'un condensateur donné. Le signal Zt a été conçu comme un détecteur binaire permettant uniquement de détecter l'absence d'un condensateur mais n'apportant pas d'information sur son état de dégradation.

5.1.6 Logiciel de traitement et de stockage des relevés d'inspections

La voiture *IRIS* dispose d'un enregistreur qui permet de stocker les résultats de mesure sous forme numérique. Ces mesures sont de plus couplées à des informations de localisation permettant de retrouver la position de la voiture *IRIS* à un

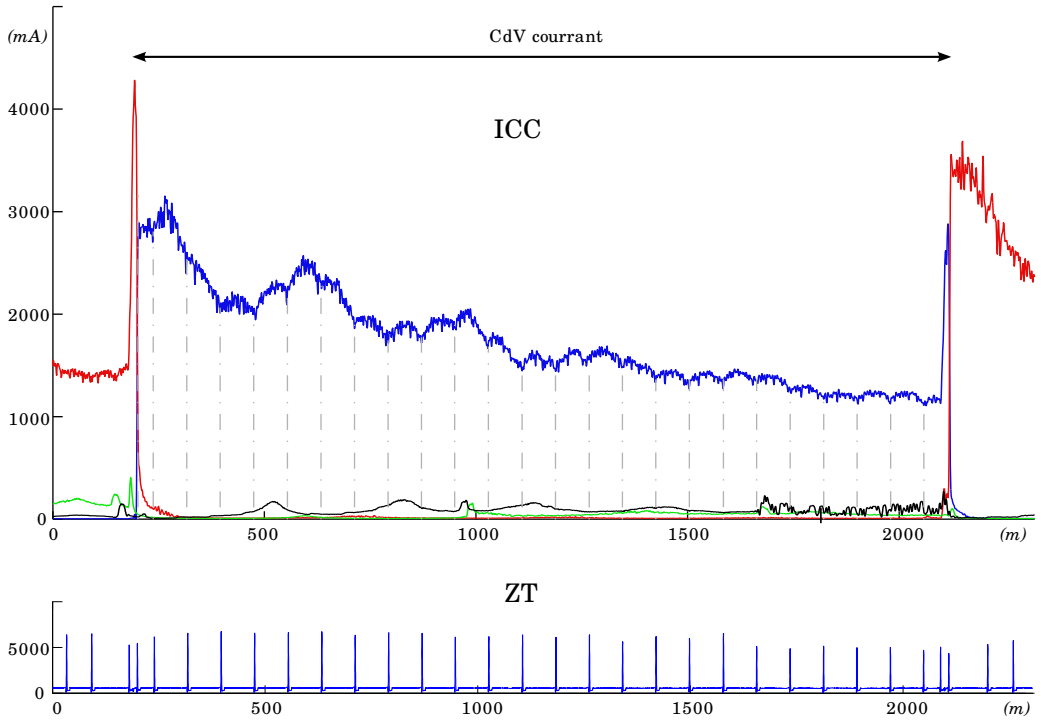


FIG. 5.4 – *Exemple de signal d'inspection réel (amplitude du courant porteur Icc), d'un circuit de voie TVM de fréquence 2000 Hz (en bleu), relevé à l'aide du véhicule IRIS. La fréquence 2600 Hz (correspondant aux CdV encadrant) est représentée en rouge.*

instant de l'enregistrement donné sur le réseau LGV français. A partir des relevés d'inspection effectués entre le 01/09/2007 et le 31/12/2007 par le véhicule d'inspection *IRIS*, une base de données d'étude a été mise en place à l'INRETS. Nous avons en particulier mis au point un programme segmentant les enregistrements sur chacun des CdV avec une localisation connue. Ces enregistrements ont ensuite été insérés dans une base de données MYSQL contenant des informations sur l'infrastructure du réseau ferroviaire français. Grâce à cette base de données, nous avons pu développer un programme Matlab interfacé avec la base de données permettant d'analyser et de traiter les différents enregistrements de manière simple. Une copie d'écran de ce programme est présentée en figure 5.5.

Cet outil permet en particulier de rechercher des enregistrements correspondant à différents critères portant sur la date de la mesure, la ligne, la fréquence utilisée par le CdV et le type de CdV recherché. La structure de données permet en particulier de retrouver tous les enregistrements d'un CdV particulier à différentes date ou bien encore tous les enregistrements concernant les CdV de la ligne LN1 fonctionnant à une fréquence de 2300 Hz par exemple. A partir de ces critères le programme cherche en base les enregistrements correspondant et récupère ceux-ci. Il est ensuite possible de naviguer dans les résultats de la requête en passant d'un enregistrement à l'autre et d'effectuer des traitements simples sur les signaux.

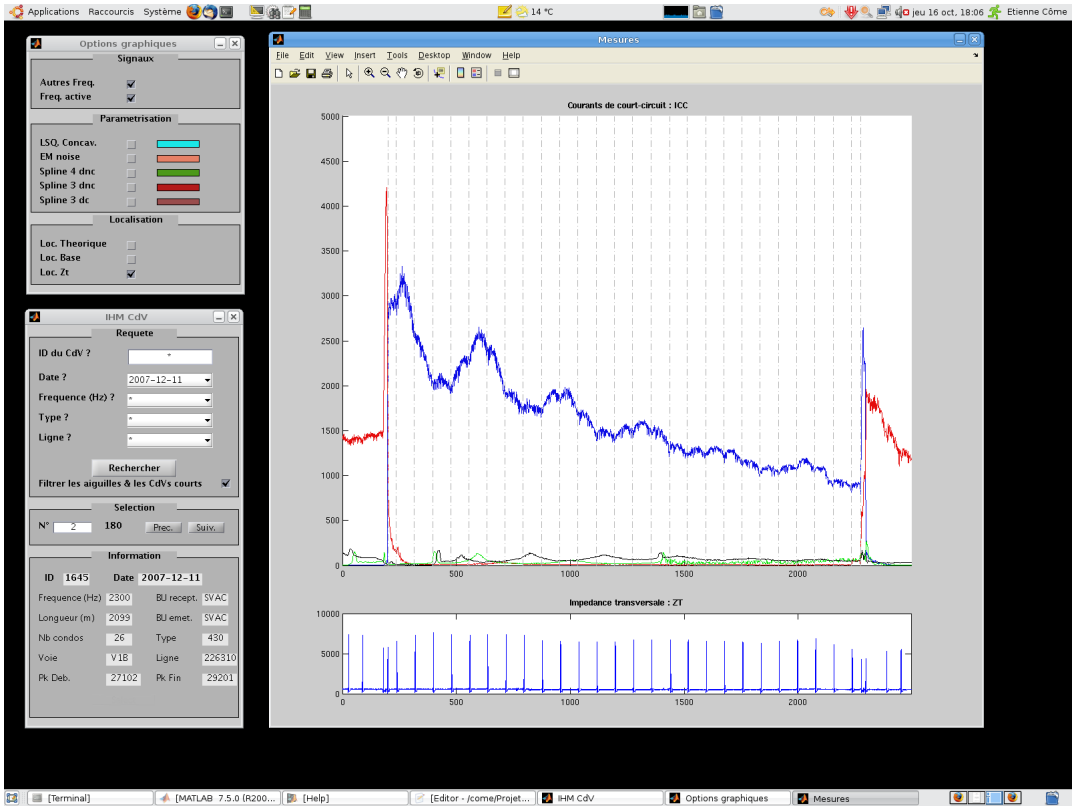


FIG. 5.5 – *Exemple d'utilisation du logiciel IHMCdVbase.*

5.2 LES INFORMATIONS À DISPOSITION POUR LE DIAGNOSTIC

5.2.1 Spécificité de l'application

multi-composants

multi-défauts

La mise au point d'un système de diagnostic permettant de décider si un système multi-composant est dans un mode de fonctionnement dégradé ou non est une tâche complexe. En effet, lorsque le système étudié présente différents composants qui peuvent tous être défectueux, et que de plus le système de mesure sur lequel repose le diagnostic mélange l'influence de l'état de chacun des composants, le nombre de mode de fonctionnement possible devient extrêmement important. L'association d'une mesure à un état de fonctionnement de chacun des composants est clairement une tâche complexe.

Les travaux qui font l'objet de ce chapitre ont pour objectif la mise au point d'un système de diagnostic automatique des « condensateurs d'accord » des CdV à partir du signal Icc. Les propositions faites visent à construire un système capable de fournir à partir d'un signal Icc, un indicateur de l'état de dégradation de chacun des condensateurs constitutifs du circuit de voie sur lequel le signal Icc a été prélevé. Cet indicateur peut être envisagé sous deux formes, continue : il est alors une image des caractéristiques électriques du composant ; discrète : il définit dans ce cas l'appartenance du composant à l'une des classes de fonctionnement (bon fonctionnement, défaut naissant, défaut grave). Nous verrons en particulier comment

l'analyse en facteurs indépendants peut être utilisée pour fournir ces deux types d'indicateurs.

Une première remarque importante nous a guidé dans la poursuite de nos travaux, celle-ci concerne les données disponibles. En effet, les procédures de maintenances actuelles de la SNCF vont conduire au stockage d'un nombre important de relevés d'inspection. Il sera donc possible dans un futur très proche de disposer de base de données de signaux d'inspection réels de taille importante. Cependant ces données ne seront pas labellisées, c'est-à-dire qu'aucune information sur le mode de fonctionnement du CdV considéré ne sera a priori disponible. La labellisation d'une partie d'entre elles représente de plus une tâche complexe, longue et coûteuse. Cette remarque a orienté nos travaux vers des solutions algorithmiques pouvant prendre en compte des données labellisées de manière imprécise/incertaine qui ont fait l'objet des chapitres 3 et 4 de cette thèse. Nous revenons en détail sur ces différents aspects en relation avec l'application visée dans la section 5.2.3.

La seconde remarque porte sur la structure du problème. Nous allons voir que des connaissances physiques sur le système étudié permettent de formuler des hypothèses intéressantes qui nous ont semblées pertinentes à intégrer au modèle utilisé pour le diagnostic. Le paragraphe qui suit fait le point sur ces informations a priori et sur le modèle génératif mis au point pour prendre en compte celles-ci.

5.2.2 Les connaissances physiques comme a priori pour la modélisation

Structure du signal Icc

Les signaux de mesure Icc sont fortement structurés. Comme le montre la figure 5.4, ceux-ci présentent une succession d'arches dont les jonctions correspondent aux emplacements des condensateurs. Nous pouvons également observer sur ces signaux une décroissance exponentielle, un comportement oscillant de grande longueur d'onde ($\approx$ 400 m) ainsi qu'un bruit de mesure. Toutes ces caractéristiques s'expliquent aisément à l'aide d'une approche physique phénoménologique (Oukhellou et al. 2006).

Le signal Icc étant structuré en portions correspondant aux différents morceaux de voie comprises entre deux condensateurs, nous avons décidé d'utiliser cette information pour paramétrer les signaux. Des polynômes de 2^{nd} degré ont été utilisés pour approximer le signal sur chacune de ces portions ; en introduisant la contrainte de continuité des différentes portions de signal et en postulant un bruit gaussien, la paramétrisation des signaux de mesure s'est vue formulée comme un problème d'approximation par polynômes par morceaux (splines) coïncidant avec le découpage naturel induit par la disposition des condensateurs à la voie. La figure 5.6 présente le résultat de cette paramétrisation sur un signal réel de fréquence 2600 Hz. Nous pouvons observer une bonne adéquation du modèle au signal sur cette figure. Les coefficients des polynômes correspondant aux différentes portions de signal permettent de résumer de manière drastique l'information apportée par les signaux d'inspection. Ce sont ces coefficients qui ont été utilisés comme descripteurs, c'est-à-dire comme variables observées. Chacun des polynômes décrivant une arche dépend de 3 coefficients ; les contraintes de continuité éliminent l'un

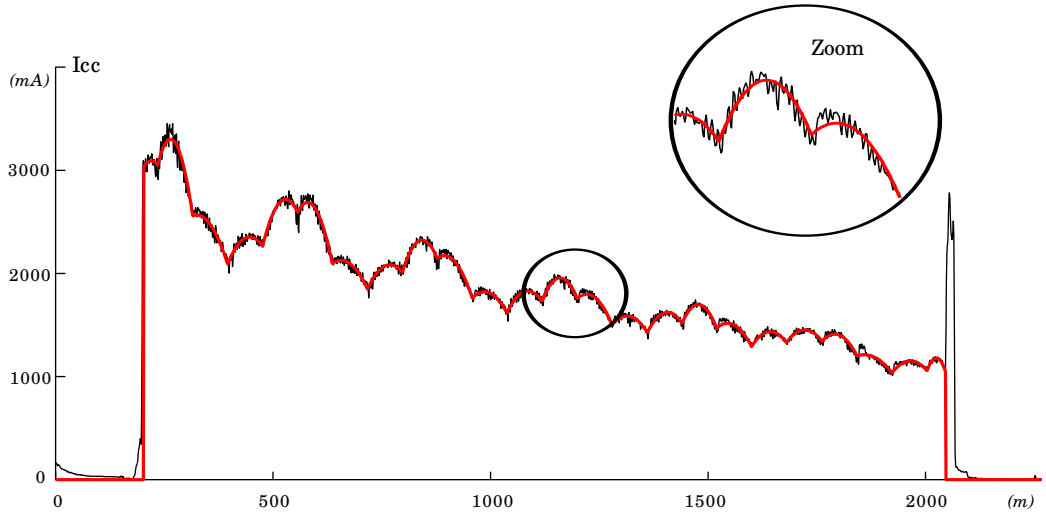


FIG. 5.6 – *Exemple de paramétrisation de signal Icc réel de fréquence 2600 Hz par splines.*

d'entre eux et nous disposons à l'issue de cette phase de paramétrisation d'un vecteur de $2 \times N_c + 1$ paramètres, avec N_c le nombre de condensateurs du CdV. Ce sont ces paramètres qui ont été utilisés comme variables observées dans la méthode de diagnostic que nous allons décrire.

Structure du bruit

L'hypothèse d'un bruit gaussien centré posé préalablement à l'approximation par spline du signal Icc peut être affinée à l'aide de connaissances sur la physique du système de mesure lui-même. En effet, la génération d'un signal Icc repose sur le court-circuit effectif des rails par les essieux du véhicule de mesure, en particulier celui situé en tête. Selon la qualité des surfaces en contact au niveau de l'interface roue/rail, de la stabilité transversale des essieux de l'humidité,... ce shunt s'avère plus ou moins bon. Cela se traduit par une variation locale du signal Icc toujours négative lorsque l'impédance de court-circuit augmente. Ce phénomène s'observe sur les enregistrements cf. figure 5.7, en particulier sur ceux effectués par l'ancien véhicule d'inspection (voiture *Hélène*).

Le bruit peut alors être approximé par un mélange gaussien à 2 composantes, la 1re composante correspondant aux réalisations négatives dûes aux variation de shunt et la 2de à un bruit électronique centré. L'utilisation d'un algorithme GEM pour la régression suivant ce modèle de mélange gaussien a fait l'objet 2 publications (Samé et al. 2006; 2007). La figure 5.7 présente le résultat de ce traitement sur une portion de signal et l'on constate un résultat d'approximation situé au dessus de la simple régression aux moindres carrés.

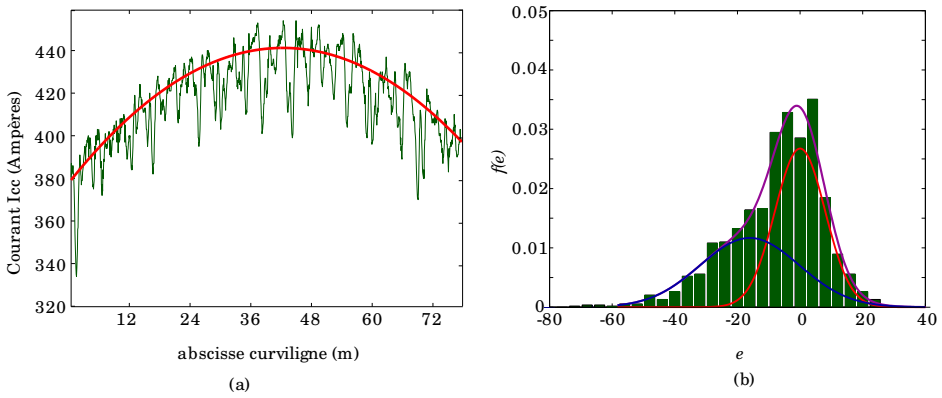


FIG. 5.7 – *Exemple de débruitage par morçeau sur des signaux enregistrés par le véhicule d'inspection Hélène à l'aide d'un modèle paramètrique du bruit utilisant un modèle de mélange gaussien : (a) signal réel et signal débruité, (b) densité des résidus et modèle de mélange estimé*

Structure amont-aval

Le CdV est composé de multiples éléments (émetteur, condensateurs, récepteur) organisés spatialement le long d'un axe. L'émetteur est le seul élément actif qui génère un courant et cet axe se trouve donc orienté.

Lorsque le véhicule d'inspection circule sur cet axe orienté il prélève le signal Icc fonction de l'abscisse curviligne x. On pose $x = 0$ la position de l'émetteur et $x = L$ celle du récepteur. A l'abscisse x_v, le véhicule court circuite le courant en provenance de l'émetteur et ce dernier ne peut donc atteindre les éléments du CdV situés aux abscisses x tel que $x_v < x < L$. Les défauts éventuels situé en aval ($x > x_v$) ne peuvent pas influencer la mesure Icc(x_v), ce qui peut se réumer par la propriété suivante :

Le signal Icc se décompose en autant d'arches qu'il y a de condensateurs d'accord. Chaque arche observée est influencée par les condensateurs situées en amont.

Cette propriété peut facilement être illustrée grâce au modèle électrique des CdV développé par l'INRETS (Aknin et al. 2003, Oukhellou et al. 2006). La figure 5.8 présente des signaux Icc simulés grâce à ce modèle électrique. Le même circuit de voie a été simulé : sans aucun défaut, avec 1 défaut sur le condensateur 5, et deux défauts respectivement sur les condensateurs 5 et 9. Nous pouvons facilement observer sur cette figure que seules les portions de signal (et donc les coefficients correspondants servant au diagnostic) situées à droite du condensateur défectueux sont influencées par ce défaut.

Le diagnostic des CdV doit donc prendre en considération les différentes remarques que nous venons de faire. Celles-ci peuvent être résumées de la manière suivante :

- les sous-systèmes sont organisés spatialement sur un axe orienté (de l'émetteur vers le récepteur) ;
- les signatures (arches) des sous-systèmes (condensateurs) sont liées spatialement de façon unidirectionnelle : l'allure de la signature d'un sous-système dé-

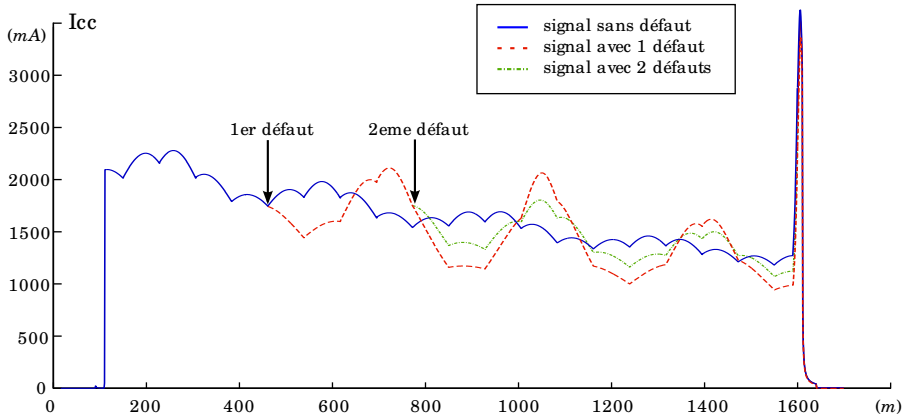


FIG. 5.8 – *Exemples de signaux d'inspections simulés à l'aide d'un modèle électrique du système, sans aucun condensateur défectueux (– bleu), avec un défaut sur le condensateur 5 (- - rouge) et avec 2 condensateurs défectueux respectivement sur les condensateurs 5 et 9 (-.-vert).*

pend de son état mais aussi de l'état des sous-systèmes situés en amont. En revanche, elle ne dépend pas de l'état des sous-systèmes situés en aval.

Ces différentes remarques nous ont conduit à l'utilisation du modèle graphique présenté sur la figure 5.9 avec X_{si} les coefficients extraits de l'arche s du signal, Z_{si} une variable latente continue correspondant à l'état de dégradation du condensateur s et Y_{si} une variable latente discrète représentant l'appartenance du condensateur à l'un des différents modes de fonctionnement (pas de défaut, défaut léger, défaut grave).

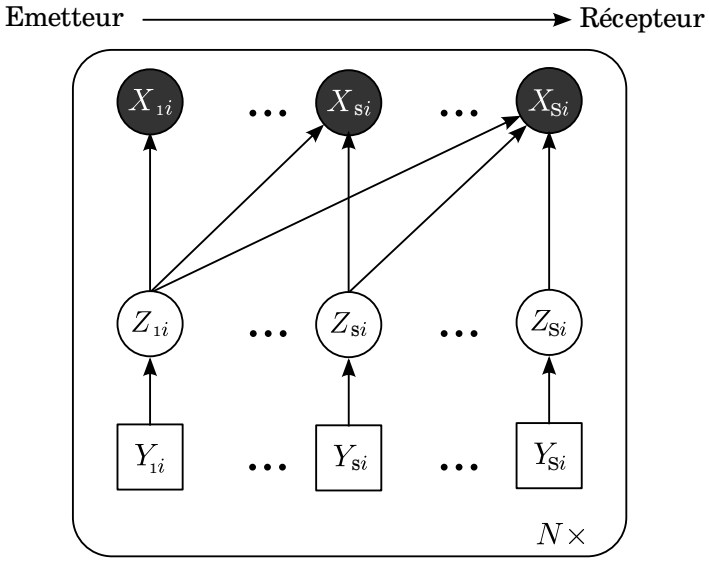


FIG. 5.9 – *Modèle génératif triangulaire pour le diagnostic des CdV représenté à l'aide d'un modèle graphique, intégrant des variables latentes continues et discrètes.*

Nous pouvons clairement observer le lien entre ce modèle génératif et le modèle de l'analyse en facteurs indépendants. La différence provenant des hypothèses supplémentaires d'indépendance amont-aval pouvant être introduites dans le cadre

du diagnostic des CdV. Une solution assez simple pour leur prise en compte a été présentée dans la première partie du chapitre 4 de cette thèse et elle sera directement reprise dans les évaluations à venir.

Nous abordons maintenant le problème de la labellisation dans le cadre du diagnostic des CdV.

5.2.3 L'expertise imparfaite pour la labellisation

Les données réelles disponibles pour le diagnostic des CdV, correspondent à des enregistrements effectués par le véhicule d'inspection *IRIS*. Après segmentation nous disposons d'un ensemble de signaux tels que celui présenté en figure 5.4. Plusieurs points quant à la possible labellisation de ces signaux réels doivent être mis en exergue. En effet, la solution envisagée pour « nourrir » le système de diagnostic en informations labellisées, repose sur la collaboration d'experts qui ont l'habitude d'observer les signaux de contrôle Icc. Les experts SNCF sont en effet capables de déterminer si un condensateur est défectueux et de fournir par exemple un étiquetage du type : (bon fonctionnement, défaut naissant, défaut grave) pour tous les condensateurs d'un CdV donné à partir de l'analyse des enregistrements. Cependant cette tâche est longue et coûteuse et soumise de plus aux imprécisions et incertitudes que peut générer l'expert. Il est par exemple possible que celui-ci rencontre des difficultés à déterminer si un condensateur présente un défaut naissant ou un défaut grave. L'information fournie par l'expert sur le mode de fonctionnement d'un condensateur peut être imprécise. Il est également possible que celui-ci soit incertain quant à la labellisation qu'il ait fournie. L'utilisation de label doux semble donc particulièrement adaptée dans cette application pour représenter l'information experte disponible.

Les différentes particularités de l'application ayant maintenant été présentées, nous donnons des résultats permettant de juger de l'intérêt de la méthode proposée, en commençant par décrire les premiers travaux effectués durant cette thèse sur le diagnostic des CdV.

5.3 Résultats selon l'approche supervisée

Une première solution s'appuyant sur le modèle génératif triangulaire (cf. figure 5.9) associé aux signaux Icc a été développée durant cette thèse dans un cadre purement supervisé (Côme et al. 2007, Côme 2007). Ces travaux reprenaient l'idée du modèle génératif contraint dans sa structure par des informations sur la physique du problème. L'originalité des travaux présentés dans ces deux publications concernait l'utilisation d'un prétraitement basé sur l'analyse canonique des corrélations en utilisant l'astuce noyau pour transformer les données observées afin que celles-ci dépendent de manière linéaire des variables d'intérêt au sein d'un modèle génératif. Ces travaux ont permis de montrer qu'un tel prétraitement associé à un modèle linéaire conduisait à des résultats encourageants.

Nous avons, à la suite de ces travaux, poursuivi dans cette voie en l'étendant hors

du cadre supervisé, et nous avons proposé des solutions demandant un travail moindre de labellisation.

5.4 RÉSULTATS SELON L'APPROCHE LABELLISATION PARTIELLE

Pour valider la pertinence de l'approche labellisation partielle, une base de données a dû être constituée ; les données réelles n'étant pas labellisés elles ne permettent pas de quantifier les performances de la méthode de diagnostic, nous avons créé une base de données simulées. Cette base de données a été constituée à l'aide du modèle électrique du CdV (Aknin et al. 2003, Oukhellou et al. 2006) permettant de générer des signaux d'inspection en faisant varier les caractéristiques électriques des différents composants. Grâce à cette base de donnée nous disposons d'information sur les variables d'intérêts du problème et nous avons ainsi pu comparer les résultats obtenus par notre méthode à ces véritables valeurs. Nous présentons dans la section suivante la démarche utilisée pour constituer cette base de données.

5.4.1 Constitution d'une base de données simulées

Les paramètres structurels des CdV ayant servi aux simulations ont été gardés constants et correspondent à ceux d'un CdV de type TVM de longueur 1500 m composé de 18 condensateurs et de fréquence 2000 Hz. Les caractéristiques suivantes ont par contre été tirées de manière aléatoire :

- les capacités des condensateurs $C_1, C_2, \ldots, C_{18}$;
- le coefficient de désadaptation λ de la voie (paramètre de nuisance) qui ajuste l'amplitude de l'onde sinusoïdale de grande longueur d'onde ($\approx$ 400 m) présente dans le signal Icc.

Les lois utilisées pour tirer ces caractéristiques ont été déterminées de manière à obtenir une base de signaux simulés réaliste. Le coefficient de désadaptation a été simulé en utilisant une loi uniforme sur $[1, 1.1]$ et les capacités en utilisant un modèle de mélange reflétant les différentes sous populations de condensateurs :

- une classe c_1 correspondant aux condensateurs en parfait état répondant aux spécifications données par le constructeurs. Cette classe représente dans nos simulations 95% des condensateurs ($\pi_1 = 0.95$). Elle suit une loi normale de moyenne égale à la valeur nominale des condensateurs posés en voie $\mu_1 = 22$ µF et de variance telle que 95% de cette sous population réponde aux tolérances fournies par le constructeur (10%), c'est-à-dire $\nu_1 = (2.2/1.96)^2 \approx 1.26$;
- une classe c_2 correspondant aux condensateurs défectueux. Cette classe représente dans la base 3% des condensateurs ($\pi_2 = 0.03$). Cette classe suit elle aussi une loi normale, la moyenne de celle-ci a été fixée à $\mu_2 = 10$ µF et sa variance à $\nu_2 = 6$;
- une classe c_3 correspondant aux condensateurs arrachés. Cette classe représente 2% des condensateurs ($\pi_3 = 0.02$). Tous les individus générés dans cette classe se sont vus attribuer une capacité égale à 0.

En utilisant cette démarche, deux bases de données ont été constituées ; la première contient 500 CdV et servira lors de l'apprentissage des paramètres de la méthode (base d'apprentissage) ; la seconde contient 2000 CdV et servira à quantifier les performances (base de test). On notera que ces bases de données nous placent d'emblée dans la perspective de la résolution d'un problème multi-défauts, contexte qui avait été peu abordé lors des travaux précédents cette thèse (Debiolles 2007).

5.4.2 Description du modèle et du protocole expérimental

Le modèle proposé pour le diagnostic est celui de l'IFA partiellement labellisée avec contraintes sur la matrice de mixage ; ce modèle a été présenté en détail dans le chapitre 4 de cette thèse. Cependant quelques points méritent d'être éclaircis quant à son utilisation dans le contexte du diagnostic des CdV.

En effet, comme nous l'avons fait remarquer lors de la description de la méthode de paramétrisation des signaux d'inspection, le nombre de variables observées extraites du signal de contrôle est égal à $2 \times N_c$, avec N_c le nombre de condensateurs du CdV, c'est-à-dire le nombre de variables latentes d'intérêt pour le diagnostic. Comme dans le cadre du modèle de l'IFA sans bruit le nombre de variables latentes doit correspondre au nombre de variables observées nous avons extrait $2 \times N_c$ variables latentes, la moitié d'entre elles correspondant aux états de dégradation des condensateurs et l'autre moitié à des variables de bruit. Les densités des variables latentes relatives aux condensateurs ont été supposées correspondre à celles de modèles de mélanges à trois composantes (correspondant aux trois modes de fonctionnement) alors que les variables de bruit ont tout simplement été modélisées à l'aide de variables aléatoires gaussiennes. Finalement, les contraintes sur la matrice de mixage ont été imposées de manière à ce que le modèle respecte les hypothèses issues de la physique du processus de mesure (indépendance amont-aval).

En ce qui concerne le protocole expérimental, nous avons essentiellement voulu mettre en évidence l'apport des hypothèses d'indépendance entre variables latentes et variables observées ainsi que l'influence de la labellisation sur les résultats. Nous avons, pour étudier ce dernier point, décidé d'étudier tout d'abord les performances de la méthode dans un contexte proche du contexte réel pouvant être envisagé, c'est-à-dire 500 CdV pour l'apprentissage dont 50% sont labellisés. Nous nous sommes donc placés dans un cadre partiellement supervisé, la labellisation ne concernant bien entendu que les variables d'intérêts.

Enfin, pour étudier le comportement de la méthode suivant la quantité d'observations labellisées nous avons réalisé différentes expériences en modifiant le pourcentage de CdV labellisés et en introduisant ou non les contraintes sur la matrice de mixage.

Le modèle présentant deux niveaux d'interprétation et d'analyse des résultats, nous étudierons tout d'abord les résultats correspondant aux variables latentes continues puis ceux concernant les variables latentes discrètes.

5.4.3 Exploitation des variables latentes continues

Résultats détaillés avec 50% de CdV labellisés

Le modèle de l'IFA permet d'obtenir pour chaque CdV de la base de test une valeur pour chaque variable latente correspondant aux états de dégradations des condensateurs. L'échelle de ces variables est arbitraire, mais elles peuvent tout de même être exploitées avantageusement pour restituer les résultats du diagnostic à l'utilisateur. Nous montrons grâce à la figure 5.10 que celles-ci permettent de visualiser de manière intéressante les résultats ; en effet, après normalisation, il est par exemple possible d'associer à ces variables une échelle de couleur permettant une lecture rapide des résultats du diagnostic, ou bien encore d'utiliser celles-ci pour représenter les résultats grâce à un « diagramme en étoile » (Chambers et al. 1983). La figure 5.10 présente ces deux solutions pour 4 CdV de la base de test. Le tableau 5.1 permet quant à lui de vérifier que les résultats de la méthode concorde avec les valeurs des capacités des différents condensateurs.

N° du CdV	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
1	**0.00**	22.30	22.05	23.20	**8.36**	22.19	21.80	22.08	22.21
2	21.38	21.89	21.53	**8.43**	21.77	19.59	23.15	20.70	22.12
3	23.28	22.63	21.55	23.60	**10.15**	21.95	21.76	22.12	20.73
4	22.29	22.19	21.05	23.05	22.47	22.00	20.40	21.97	20.77
	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}
1	22.75	23.46	20.77	21.65	23.94	23.72	21.59	21.55	21.66
2	21.92	22.04	21.48	22.33	22.80	20.31	21.61	21.66	19.62
3	22.53	22.90	22.28	22.27	21.32	**0.00**	21.96	19.81	22.02
4	19.17	23.44	20.58	22.06	22.18	21.67	**4.99**	21.19	20.98

TAB. 5.1 – *Valeurs des capacités (en µF) des 4 CdV extraient de la base de test servant à illustrer les résultats de la méthode de diagnostic.*

D'un point de vue plus global, les performances peuvent être quantifiées grâce aux corrélations et aux coefficients de détermination entre les sources estimées et les capacités des condensateurs. Le tableau 5.2 donne ces résultats pour les 18 condensateurs constitutifs du CdV. Nous pouvons observer grâce à celui-ci la bonne adéquation entre les variables latentes et les capacités des condensateurs, les résultats se dégradant légèrement le long du CdV (passage d'un coefficient de corrélation de 0.93 à $0.67s$), les deux derniers condensateurs semblent plus difficile à diagnostiquer. Ceci s'explique par le fait qu'en extrémité de CdV le signal Icc est influencé par la quasi totalité des variables latentes.

Evolution des performances en fonction du nombre de CdV labellisés et apport des contraintes.

Afin d'étudier les performances de notre méthode dans différents contextes, nous avons fait varier le pourcentage de CdV labellisés de 0% à 100% en utilisant toujours la base d'apprentissage constitué de 500 CdV. Nous présentons pour juger de la qualité des résultats obtenus la moyenne du coefficient de corrélation (en valeur absolue) entre les sources estimées et les capacités des condensateurs, cette

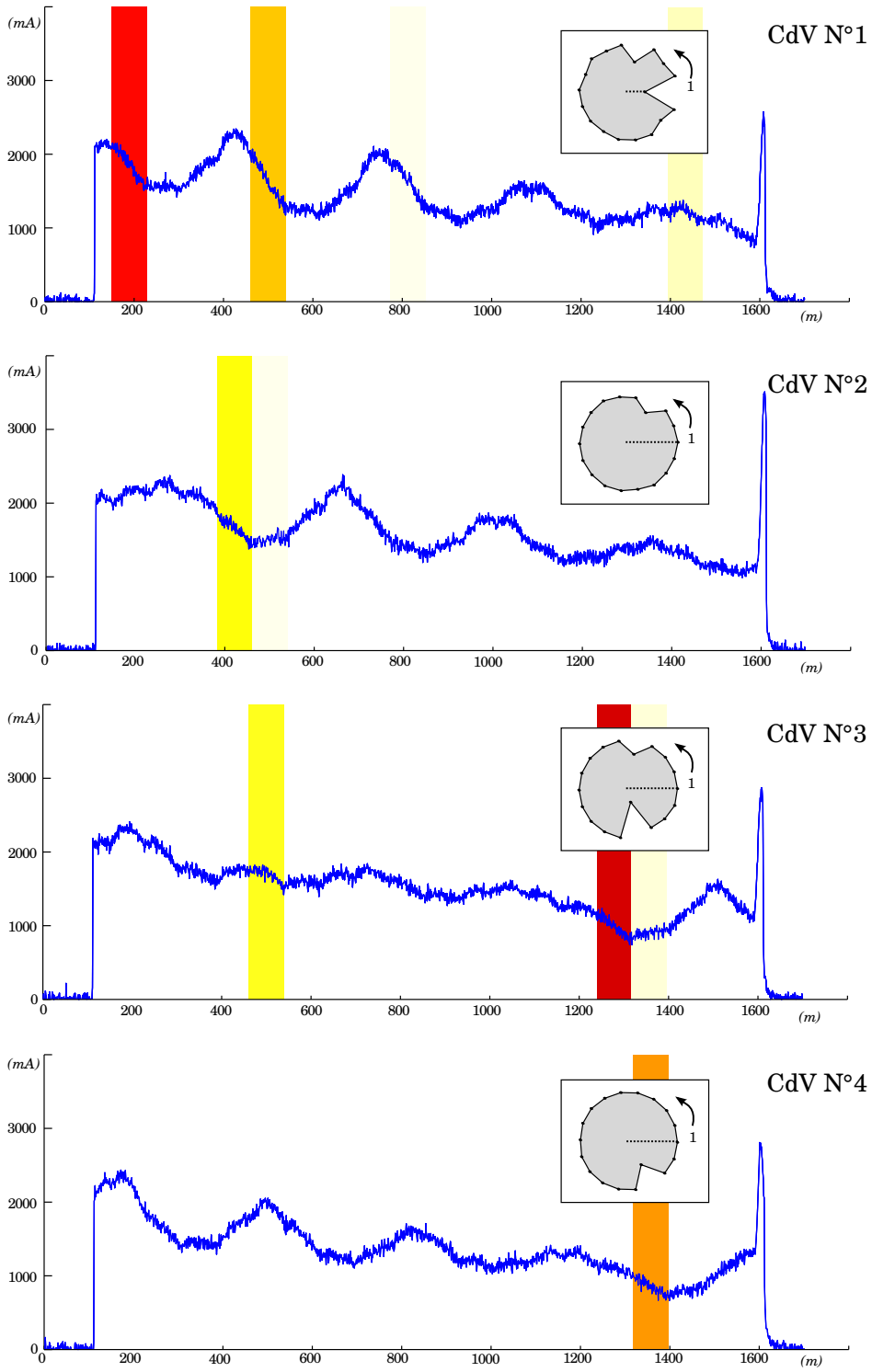


FIG. 5.10 – *Exemples de résultats de diagnostic sur la base de données de test. Les valeurs des variables latentes estimées sont représentées à l'aide de couleurs. Un « diagramme en étoile » permet également de visualiser celles-ci pour l'ensemble du CdV. L'apprentissage a été réalisé en utilisant l'IFA partiellement supervisée avec contraintes sur la matrice de mixage et avec 250 observations de labellisés sur les 500 observations de la base d'apprentissage. Les variables latentes ont été normalisées et un échelle de couleurs commune à toutes les variables a été utilisée. Celle-ci est uniforme entre O et la valeur maximale des variables latentes et correspond à un dégradé du blanc au rouge. Les diagrammes en étoile utilisent l'inverse de cette variable translatée de la valeur maximale, les condensateurs défectueux se trouve donc proche du centre du diagramme.*

N° condensateur	1	2	3	4	5	6	7	8	9
$r_{\widehat{z}_i,z_i}$	0.93	0.92	0.92	0.91	0.89	0.88	0.84	0.90	0.86
$r^2_{\widehat{z}_i,z_i}$	0.87	0.84	0.85	0.82	0.79	0.77	0.71	0.80	0.74
N° condensateur	10	11	12	13	14	15	16	17	18
$r_{\widehat{z}_i,z_i}$	0.85	0.80	0.79	0.82	0.84	0.79	0.82	0.67	0.69
$r^2_{\widehat{z}_i,z_i}$	0.72	0.63	0.63	0.67	0.70	0.63	0.68	0.45	0.47

TAB. 5.2 – *Résulats de l'IFA partiellement supervisé avec contraintes sur la matrice de mixage (250 CdV labellisés, 250 CdV non labellisés). Valeurs absolues des coefficients de corrélation et coefficients de détermination entre les sources estimées et les capacités des condensateurs, estimées sur l'ensemble de test (2000 CdV).*

moyenne étant prise sur l'ensemble des condensateurs. La figure 5.11 présente l'évolution de ce critère pour l'IFA avec labellisation partielle et contraintes sur la matrice de mixage (- - rouge) et pour l'IFA avec labellisation partielle sans contraintes sur la matrice de mixage (– vert).

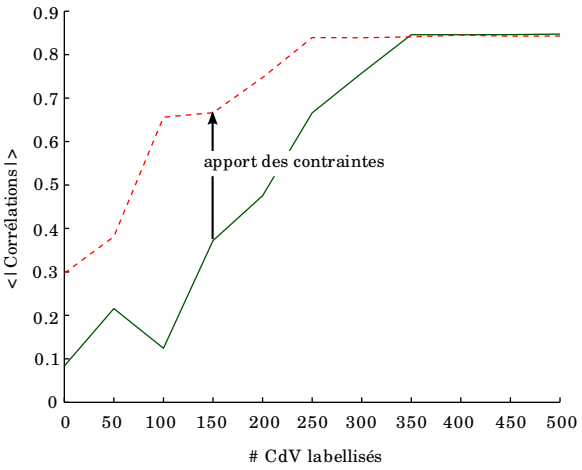


FIG. 5.11 – *Résultats de l'IFA partiellement labellisée avec (- - rouge) et sans contraintes (– vert). Evolution de la moyenne des coefficients de corrélations entre les sources estimées et les capacités des condensateurs, en fonction du nombre de CdV labellisés. Les différentes solutions ont été obtenues à l'aide d'un algorithme GEM avec montée de gradient naturel, en utilisant 20 initialisations aléatoires et en conservant la meilleure d'entres elles.*

Cette figure permet de mettre clairement en évidence l'intérêt des contraintes lorsque le pourcentage de CdV labellisés est faible. Les deux courbes présentent une forme voisine : forte croissance puis stagnation. Cependant, les hypothèses d'indépendance permettent une croissance plus rapide et une arrivée plus précoce sur le plateau. A l'opposé, lorsque le nombre de CdV labellisés est important (>350), les contraintes n'améliorent pas significativement les performances.

La labellisation permet d'améliorer de manière très importante les performances, en particulier en supprimant le problème de l'indétermination par rapport aux permutations des sources (le critère utilisé étant sensible à celles-ci) ; par contre il ne semble pas y avoir de différences significatives lorsque le nombre de CdV labellisés est plus important. Cette constatation a une influence directe sur l'application pratique de notre méthode : il ne semble en effet pas nécessaire de labelliser l'ensemble des CdV réels pour obtenir des performances intéressantes, se contenter de labelliser la moitié d'entre eux devrait permettre d'obtenir des résultats intéressants. Cette constatation étant encore plus remarquable lorsque les contraintes sont prises en considération, en effet, dans ce cas de figure, le plateau du critère est atteint pour seulement 250 CdV labellisés. Grâce aux contraintes, il paraît donc possible de ne labelliser que 250 CdV et d'obtenir tout de même de bons résultats.

A l'extrême, le contexte non supervisé (aucun CdV de labellisé) obtient un résultat intéressant d'une corrélation de 0.3 obtenu grâce aux seules contraintes.

Enfin, le plateau observé sur cette figure à 0.84 s'explique par la nature légèrement non linéaire existant entre variables latentes et variables observées dans cette application (Côme et al. 2007). Cette valeur correspond au meilleure niveau pouvant être atteint à l'aide d'un modèle linéaire, une amélioration des performances est envisageable si un modèle non linéaire est mis en place.

En conclusion, les variables latentes continues extraites grâce à l'IFA avec labellisation partielle et contraintes sur la matrice de mixage sont très encourageants sur cette base de données simulées, et permettent d'espérer obtenir de bons résultats sur données réelles. Les contraintes permettent de diminuer le nombre de CdV labellisés nécessaires à l'obtention de bon résultats.

Comparaison partiellement supervisé / supervisé

Nous étudions ici l'apport des individus non labellisés sur les performances en comparant les résultats obtenus par l'IFA partiellement supervisée, et ceux de l'IFA supervisée où seuls les individus labellisés sont utilisés, les individus supplémentaires non labellisés sont ignorés. Les résultats sur notre application sont présentés en figure 5.12, en prenant en compte (a) ou non (b) les contraintes sur la matrice de mixage.

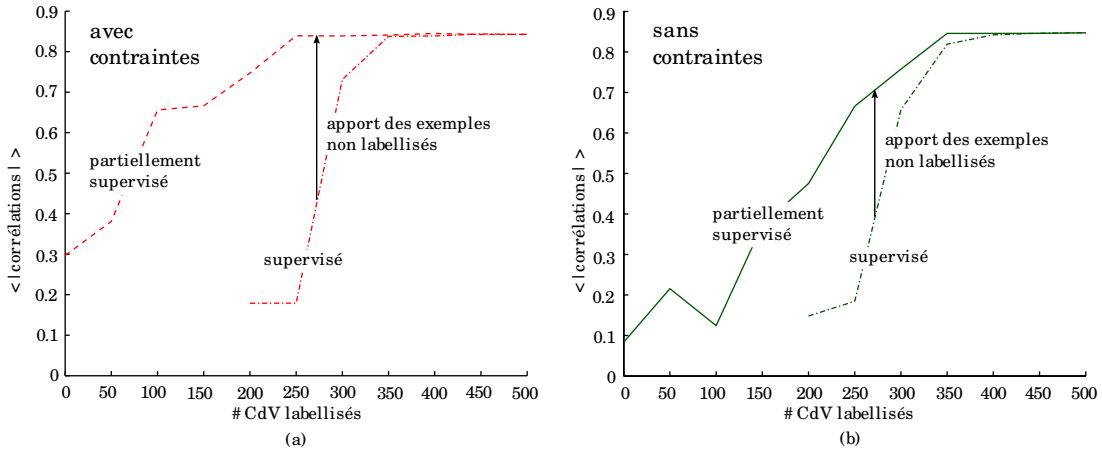


FIG. 5.12 – *Résultats de l'IFA partiellement labellisée avec (a) et sans contraintes (b). Evolution de la moyenne des valeurs absolues des coefficients de corrélations entre les sources estimées et les capacités des condensateurs, en fonction du nombre de CdV labellisés en utilisant ou non des individus supplémentaires non labellisés. Les différentes solutions ont été obtenues à l'aide d'un algorithme GEM avec montée de gradient naturel, en utilisant 20 initialisations aléatoires et en conservant la meilleure d'entres elles.*

Nous pouvons observer sur cette figure que les individus non labellisés permettent d'améliorer les performances, en particulier lorsque le nombre d'individus labellisés est faible. De plus, lorsque les contraintes de structure sont prises en considération, l'apport des individus non labellisés est plus clairement visible. Grâce aux contraintes et aux individus non labellisés il suffit de labelliser 250 CdV pour ob-

tenir des résultats équivalents à ceux obtenus par l'apprentissage supervisé avec 400 CdV (sans contraintes) ; soit un gain important en ce qui concerne la tâche d'étiquetage. L'apport des individus non labellisés observé ici est sans aucun doute extrêmement dépendant du problème, et de l'adéquation du modèle aux données. Ces résultats sont donc à contextualiser relativement à l'apllication traitée, des gains plus ou moins important peuvent être obtenus suivant l'application.

Enfin, un point mérite d'être noté : l'impact des points non labellisés sur l'existence ou non de minimum locaux. En effet, nous avons vu au chapitre précédent que le pourcentage d'individus labellisés avait une influence directe sur cet aspect. Augmenter le nombre d'individus non labellisés entraine donc une augmentation du nombre de minimum locaux et du même coup de la difficulté du problème d'optimisation. Cette constatation doit être prise en compte lors de la mise en pratique de méthodes mêlant individus non labellisés et labellisés. Les bons résultats présentés en figure 5.12 peuvent s'en trouver remis en cause.

5.4.4 Exploitation des variables latentes discrètes

Après cette étude des résultats obtenus au niveau des variables latentes continues, nous allons analyser les résultats de notre méthode vis à vis des variables latentes discrètes. Comme précédemment, nous étudierons tout d'abord les résultats de celle-ci dans un contexte « type » (30% de CdV labellisés et prise en compte des contraintes sur la matrice de mixage), puis nous examinerons l'évolution des performances suivant que les contraintes sont prises en compte ou pas et suivant le pourcentage de CdV labellisés.

Résultats détaillés avec 250 CdV labellisés

Pour analyser les résultats concernant les variables latentes discrètes, nous pouvons tout d'abord donner la matrice de confusion entre classes réelles et classes estimées par notre méthode ; cette matrice de confusion est donnée par le tableau 5.3 pour l'ensemble des condensateurs du CdV. Les conventions suivantes ont été utilisées pour représenter les différentes classes réelles possibles et les différentes décisions possibles :

- R_0 : le condensateur ne présentait pas de défaut ;
- R_1 : le condensateur présentait un défaut léger ;
- R_2 : le condensateur présentait un défaut grave ;
- D_0 : le condensateur a été diagnostiqué comme non défectueux ;
- D_1 : le condensateur a été diagnostiqué comme légèrement défectueux ;
- D_2 : le condensateur a été diagnostiqué comme gravement défectueux.

	R_0	R_1	R_2
D_0	33671	124	12
D_1	455	830	175
D_2	76	149	508

TAB. 5.3 – *Matrice de confusion, classes réelles / classes estimées, sur l'ensemble de test de 2000 CdV soit 2000 × 18 = 36000 condensateurs. Les classes réelles sont en colonne et les décisions en ligne.*

Nous pouvons observer de bons résultats de décision sur cette matrice de confusion avec cependant quelques confusions entre les classes voisines c'est-à-dire entre les classes R_0 et R_1 et entre les classes R_1 et R_2. En regroupant les deux classes de défauts (R_1, R_2), il est possible d'obtenir la matrice de détection présentée sur le tableau 5.4. Cette matrice simplifiée permet de calculer les indicateurs usuels en détection : BD le taux de bonnes détections, FA le taux de fausses alarmes, FD le taux de fausses détections.

TAB. 5.4 – *Matrice de confusion simplifiée classes réelles / classes estimées, sur l'ensemble de test de 2000 CdV soit 2000 × 18 condensateurs.*

	R_0	R_1
D_0	33671	136
D_1	531	1662

Ces indicateurs sont les suivants :

$$BD = nb_{(R_1,D_1)}/nb_{R_1} = 92.43\% \tag{5.1}$$

$$FA = nb_{(R_0,D_1)}/nb_{R_0} = 1.55\% \tag{5.2}$$

$$FD = nb_{(R_0,D_1)}/nb_{D_1} = 24.22\% \tag{5.3}$$

Les taux de bonnes détections et de fausses alarmes sont satisfaisants, et permettent d'envisager une mise en exploitation de la méthode si des résultats équivalents sont obtenus sur données réelles. Au vue des proportions très différentes des deux classes, le taux de fausses détections est lui aussi satisfaisant.

Evolution des performances en fonction du nombre de CdV labellisés et apport des contraintes.

En ce qui concerne l'évolution des performances en fonction de la labellisation et de l'utilisation ou non de contraintes sur la matrice de mixage, nous donnons sur la figure 5.13 l'évolution du taux de fausses alarmes (a) et de bonnes détections (b) en fonction du nombre de CdV labellisés.

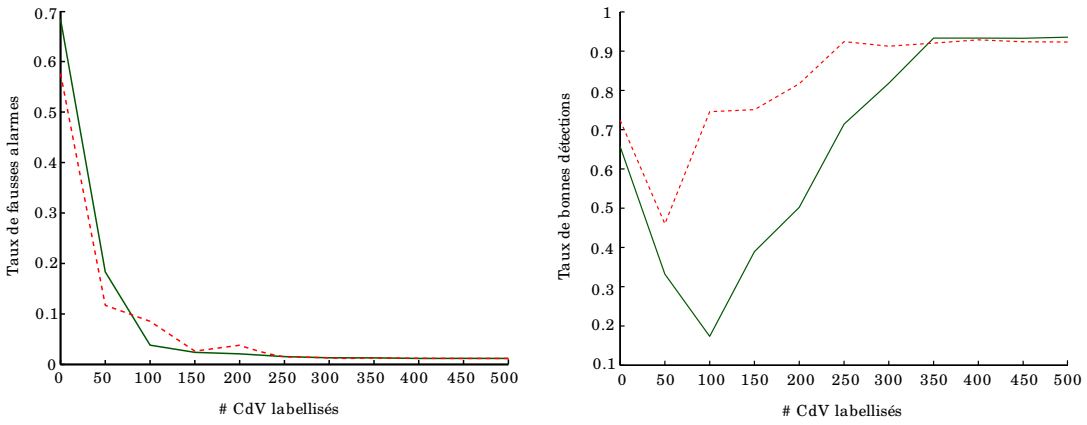


FIG. 5.13 – *Evolution du taux de fausses alarmes (a) et de bonnes détections (b) en fonction du pourcentage de CdV labellisés pour l'IFA avec contraintes (- - rouge) et l'IFA sans contraintes (– vert). Ces taux sont évalués à l'aide de la base de test, l'algorithme GEM a été initialisé à partir de 20 initialisations aléatoires et la meilleure d'entre elles a été conservée.*

Nous pouvons faire quelques observations sur ces résultats. Lorsque le nombre d'individus labellisés est trop faible (<100) le taux de fausses alarmes est très important et les résultats inexploitables. Au dessus de cette valeur, les fausses alarmes se stabilisent autour de 1.5%. En ce qui concerne le taux de bonnes détections, nous pouvons également observer deux régimes ; le premier correspond à un nombre trop faible d'individus labellisés (<100) et le taux de bonnes détections est alors intéressant, mais comme nous venons de le dire, les fausses alarmes sont extrêmement nombreuses. Le taux de bonnes détections diminue au départ, mais ceci s'explique par la diminution extrêmement importante en parallèle des fausses alarmes. Finalement, lorsque le nombre d'individus labellisés est suffisant pour que le taux de fausses alarmes soit stable, nous pouvons observer l'intérêt des contraintes et de la labellisation qui permettent d'atteindre avec seulement 250 individus labellisés un taux de bonnes détections intéressant de 92%.

Conclusion du chapitre

Ce chapitre nous a permis d'éclairer l'utilisation des propositions théoriques présentées dans les chapitres 3 et 4 de cette thèse sur une application réelle. Le problème du diagnostic des CdV au travers de l'analyse du courant de court circuit (Icc), s'est vu résolu grâce au modèle de l'IFA partiellement supervisée avec contraintes et de bons résultats ont été obtenus sur données simulées.

Nous avons également pu observer dans ce chapitre l'intérêt de disposer de deux niveaux d'interprétation des résultats, comme c'est le cas dans le cadre de l'IFA, pour restituer les résultats à l'utilisateur. Enfin, l'apport d'une labellisation partielle et de la prise en considération d'hypothèses réalistes d'indépendance entre certaines variables observées et certaines variables latentes a pu être mis en évidence dans le cadre de cette application validant ainsi l'intérêt pratique des propositions théoriques de cette thèse.

CONCLUSION GÉNÉRALE

SYNTHÈSE DES TRAVAUX

Labellisation douce pour la classification

Les travaux effectués durant les trois années de thèse ont été inspirés par des besoins applicatifs réels. Ceux-ci nous ont conduit à travailler sur le problème de la labellisation douce permettant de sortir du cadre trop restrictif de l'apprentissage supervisé. La théorie des fonctions de croyance et une approche générative du problème de l'apprentissage se sont révélées pertinentes dans ce contexte pour utiliser des données imparfaites en classification.

Les méthodes discriminatives ont longtemps été préférées dans la communauté des chercheurs travaillant sur l'apprentissage statistique car elles ont prouvé leur supériorité lorsque les données sont abondantes et labellisées de manière précise et certaine. Cette thèse souhaitait démontrer que d'autres approches pouvaient être proposées dans un contexte plus réaliste où les données labellisées de manière précise et certaine sont rares. Dans ce contexte, l'approche générative présente des avantages certains. Elle nécessite certes un travail important de modélisation de la densité jointe, mais une fois ses paramètres estimés la loi conditionnelle suffit pour prédire sur les données de test. Ce travail de modélisation supplémentaire est de plus intéressant pour introduire des hypothèses sur la structure des données, ce qui permet de trouver des solutions pertinentes même lorsque la qualité de la labellisation peut être remise en cause ou lorsque les labels sont imprécis. L'approche générative présente un grand intérêt dans de nombreux cas de figure, comme l'ont montrés les expériences du chapitre 3 de cette thèse.

Nos travaux effectués dans le cadre de la théorie des fonctions de croyance sur l'apprentissage des paramètres des modèles de mélange ont abouti à la définition d'un critère génératif permettant d'utiliser des labels de différentes natures, ce qui rend la solution proposée très flexible et à même de traiter nombre de problèmes concrets. Cette solution possède les avantages et les inconvénients inhérents aux méthodes génératives, en particulier une forte dépendance de l'adéquation du modèle aux données. Enfin d'un point de vue pratique le problème d'estimation des paramètres lorsque les labels sont imprécis/ incertains a été résolu grâce à la mise au point d'un algorithme EM, facilement implémentable.

Labellisation douce dans le cadre de l'IFA

L'utilisation de labels doux qui s'est montrée pertinente dans le cadre des modèles de mélange, a également été étendue au contexte de l'analyse en facteurs indépendants. Ce modèle génératif extrêmement particulier et parcimonieux permet de traiter élégamment le cas des problèmes à nombre de classes important. L'extension de la méthode précédente à ce modèle n'a pas posé de problème particulier, laissant espérer qu'il en serait de même pour d'autres modèles à variables latentes. L'intégration des labels doux, dans un algorithme GEM pour l'estimation des paramètres, s'est elle aussi révélée similaire dans l'esprit, au développement précédent. Enfin, des expériences sur jeux de données artificielles ont permis de mettre en évidence l'intérêt de l'approche semi-supervisée dans le cadre de l'IFA.

Prise en compte d'hypothèses d'indépendance entre variables observées et variables latentes dans le contexte de l'ACI

L'incorporation d'informations additionnelles au jeu de données est une voie naturel d'amélioration des performances lors de la résolution d'un problème d'apprentissage. En effet, de telles informations permettent de restreindre l'espace de recherche des solutions, ce qui peut mener à une amélioration des performances. Ces informations peuvent prendre la forme d'hypothèses sur le processus de génération des données et les modèles génératifs peuvent aisément les prendre en compte.

Nous avons proposé dans cet esprit de prendre en considération des hypothèses supplémentaires sur l'indépendance entre certaines variables latentes et certaines variables observées. Ces hypothèses se sont vues retranscrites dans le contexte de l'ACI en contraintes de nullité de coefficients de la matrice de mixage. Leur prise en compte pratique lors de l'estimation des paramètres a été résolue grâce à l'utilisation d'un algorithme GEM. Enfin, des expériences ont permis de montrer que de telles hypothèses pouvaient être pertinentes dans un cadre non supervisé tout comme dans un contexte de labellisation « douce » des sources, en particulier dans le contexte applicatif à l'origine de cette thèse. En effet, l'analyse des résultats de l'IFA partiellement supervisée sur ce problème, a permis de mettre en évidence que de telles hypothèses permettaient de réduire le nombre d'individus labellisés nécessaires pour obtenir des performances intéressantes.

Perspectives

De nombreux éléments introduits dans cette thèse mériteraient d'être approfondis. Nous en dressons une liste non exhaustive en distinguant ceux qui relèvent d'une démarche de recherche à court, moyen ou long terme.

Court terme

Travaux sur d'autres modèles

A court terme, il semble pertinent d'envisager l'extension des solutions proposées dans cette thèse à d'autres modèles. Les modèles de mélange de loi multinomiales mériteraient en particulier d'être étudiés dans le cadre d'une approche « labellisation douce ». En effet, dans le cadre de ceux-ci la labellisation peut concerner aussi bien les variables observées (celles-ci étant discrètes), que les variables latentes. Dans ce cadre leur distinction deviendrait somme toute artificielle et les labels doux permettraient de proposer une solution extrêmement flexible pour représenter les informations disponibles sur les observations de la base d'apprentissage.

Finalisation des travaux sur l'IFA avec labels doux pour le diagnostic des circuits de voie

L'application nécessite elle aussi d'autres développements. En effet, les travaux présentés dans cette thèse concernent des données issues de simulation. Une base de données réelle étiquetée de manière imprécise, incertaine par un expert SNCF est en cours de constitution. La méthode devra donc être testée sur celles-ci. D'autres part, les données en cours d'acquisition par la SNCF possèdent une nature intrinsèquement historisée, le réseau étant inspecté périodiquement. La méthode proposée ne tient pas en compte de cette particularité. Il semble envisageable d'intégrer au modèle un aspect temporel qui pourrait se baser par exemple sur les travaux évoqués lors du chapitre 4 pour prendre en considération une hypothèse markovienne quant à la structure des données (Attias 2000), (Jutten et Comon 2007b, p. 500-514).

Moyen terme

Poursuite des expérimentations autour de l'IFA et des labels « doux »

Le travail effectué sur l'apprentissage avec des étiquettes douces dans le contexte de l'IFA mériterait d'être étendu au problème du bruit d'étiquetage.

La modélisation du bruit proposée dans le cadre de nos travaux sur l'IFA au travers de l'extraction de composantes indépendantes supplémentaires représentant le bruit, devrait également être comparée à la solution alternative consistant à l'intégrer directement au modèle. Dans ce cas de figure, la relation

$$\mathbf{x} = A\,\mathbf{z}, \tag{5.4}$$

avec A une matrice carrée, est remplacée par un modèle de la forme

$$\mathbf{x} = A\,\mathbf{z} + \xi, \tag{5.5}$$

avec A une matrice rectangulaire. Ce type de modèle nécessiterait des solutions algorithmiques plus lourdes que l'algorithme GEM proposé dans cette thèse (Attias

1999) ; les approches variationnelles peuvent être envisagées comme solution à ce problème (Lawrence et Bishop 2000).

Enfin, la prise en compte d'une relation non linéaire entre variable observées et variables latentes peut également être envisagée, les travaux tels que (Valpola et al. 2003) peuvent constituer un point de départ intéressant pour aborder cette problématique.

Long terme

Travail sur des solutions algorithmiques permettant de traiter le problème des minimums locaux

Nous avons vu dans cette thèse que l'utilisation d'individus non labellisés ou labellisés de manière imprécise/incertaine complexifiait le problème d'optimisation par rapport au cadre supervisé en cela qu'il était alors nécessaire de prendre en considération l'existence de minimums locaux dans le critère à optimiser (tout comme c'est le cas classiquement en apprentissage non supervisé). Cet aspect doit bien évidement être solutionné et le développement de solutions algorithmiques pour le traiter mériterait d'être étudié en profondeur. Les stratégies d'optimisation globale telles que les métaheuristiques (Siarry et al. 2003) ou les méthodes de recuit déterministe (Ueda et Nakano 1995, Rose 1998) pourraient en particulier être explorées. L'utilisation de stratégies d'initialisations « intelligentes », c'est à dire tirant parti des individus pour lesquels une information précise et fiable est disponible peut également être envisagée et étudiée pour aider à résoudre ce problème.

Etude théorique de l'apport de la labellisation

A plus long terme, une étude sur l'apport de la labellisation douce sur les performances asymptotiques peut être envisagée. Dans ce cadre, des questions théoriques portant sur les conditions qu'un label imprécis/incertain doit remplir pour améliorer l'estimation des paramètres devront être abordées. De même, des questions sur la quantification de l'apport d'un label suivant sa nature devrait être envisagées, dans l'esprit des travaux de O'Neill (1978) comparant l'apport des individus labellisés et non labellisés sur l'estimation d'un modèle de mélange gaussien.

Labels doux et sélection de modèle

Le problème de la sélection de modèle dans un cadre « labels doux » mériterait elle aussi des investigations poussées. Les critères de sélections de modèle couramment utilisés tels que les critères BIC, AIC,... peuvent-ils être étendus dans ce cadre ? Sous quelle forme ? Quelles sont alors les propriétés de tels critères ?

Enfin, le problème de la validation des performances est un problème en soi dans le cadre « label doux » ; en effet les labels n'étant pas considérés comme parfaits,

le taux d'erreur en généralisation ne peut être évalué et il n'existe donc pas de solution triviale à la validation des performances qui reste aujourd'hui un problème ouvert.

Annexes

.1 Mise à jour des proportions lors de l'étape M de l'algorithme EM pour les modèles de mélange

La mise à jour des proportions effectuée lors de l'étape M de l'algorithme EM, correspond à la maximisation de la fonction auxiliaire Q par rapport à $\boldsymbol{\pi}$. La fonction Q est de plus définie dans le cadre des modèles de mélange par :

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log(\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)), \tag{6}$$

avec :

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} \pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})} \tag{7}$$

Cette maximisation à une solution analytique de la forme suivante :

$$\pi_k^{(q+1)} = \sum_{i=1}^{N} t_{ik}^{(q)}/N. \tag{8}$$

Preuve : Nous avons :

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log(\pi_k) + cst, \tag{9}$$

où cst est une constante indépendante de $\boldsymbol{\pi}$. Afin de maximiser Q par rapport à $\boldsymbol{\pi}$ en prenant en considération la contrainte $\sum_{k=1}^{K} \pi_k = 1$, nous formons le lagrangien :

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log(\pi_k) + \lambda.(1 - \sum_{k=1}^{K} \pi_k), \tag{10}$$

où λ est le multiplicateur de lagrange associé à la contrainte. En dérivant le lagrangien par rapport aux proportions nous obtenons :

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_k} = \frac{\sum_{i=1}^{N} t_{ik}^{(q)}}{\pi_k} - \lambda, \quad \forall k \in \{1, \ldots, K\}, \tag{11}$$

pour maximiser Q par rapport à $\boldsymbol{\pi}$ nous devons trouver les valeurs des proportions telles que ces dérivées s'annules, c'est à dire telles que :

$$\begin{aligned} \frac{\sum_{i=1}^{N} t_{i1}^{(q)}}{\pi_1} &= \lambda \\ \vdots &= \vdots \\ \frac{\sum_{i=1}^{N} t_{iK}^{(q)}}{\pi_K} &= \lambda \end{aligned}$$

En multipliant chacune de ces équation par la proportion correspondante et en les sommant toutes nous obtenons :

$$\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} = \lambda.(\pi_1 + \ldots + \pi_K) \tag{12}$$

et donc $\lambda = N$, ce qui permet d'obtenir la formule de mise à jour (8) en remplaçant λ par N dans (12). □

.2 Gradient de la log-vraisemblance de l'ACI par rapport à la matrice de démixage

La log vraisemblance d'une matrice de démixage W dans le cadre de l'ACI sans bruit est donnée par :

$$\mathcal{L}(W;\mathbf{X}) = \sum_{i=1}^{N}\sum_{s=1}^{S} \log\left(f^{\mathcal{Z}_s}\left((W\mathbf{x}_i)_s\right)\right) + N\log(|\det(W)|), \tag{13}$$

en supposant que les densités des sources $f^{\mathcal{Z}_1}, \ldots, f^{\mathcal{Z}_S}$ sont connues.

Pour calculer le gradient de $\mathcal{L}(W;\mathbf{X})$ par rapport à W nous devons calculer la dérivée du logarithme de la valeur absolue du déterminant d'une matrice par rapport à l'un de ces éléments, cette dernière est donnée par (voir MacKay (1996), Petersen et Pedersen (2008, p. 8)) :

$$\frac{\partial \log(|\det(X)|)}{\partial X_{lk}} = (X^{-1})_{kl}, \tag{14}$$

Nous obtenons, en utilisant cette propriété, la dérivée de la log-vraisemblance précédente par rapport à un élément l,k de la matrice de démixage :

$$\begin{aligned}\frac{\partial \mathcal{L}(W;\mathbf{X})}{\partial W_{lk}} &= N(W^{-1})_{kl} + \sum_{i=1}^{N} \frac{\partial \log\left(f^{\mathcal{Z}_l}\left((W\mathbf{x}_i)_l\right)\right)}{\partial W_{lk}} \\ &= N(W^{-1})_{kl} - \sum_{i=1}^{N} \mathbf{x}_{ik} g_l\left((W\mathbf{x}_i)_l\right),\end{aligned} \tag{15}$$

avec $g_l(z)$ l'opposé de la dérivée du logarithme de la densité de la source l :

$$g_l(z) = \frac{-\partial \log(f^{\mathcal{Z}_l}(z))}{\partial z} \tag{16}$$

En prenant des notations matricielles, nous pouvons définir la fonction g :

$$\begin{aligned}\mathbf{g} &: \mathbb{R}^S \to \mathbb{R}^S \\ \mathbf{g}(\mathbf{z}) &= \left[\frac{-\partial \log(f^{\mathcal{Z}_1}(z_1))}{\partial z_1}, \ldots, \frac{-\partial \log(f^{\mathcal{Z}_S}(z_S))}{\partial z_S}\right]^t.\end{aligned} \tag{17}$$

Ce qui nous permet d'obtenir la matrice contenant la dérivée de la log-vraisemblance par rapport à chaque coefficient de W :

$$\begin{aligned}\frac{\partial \mathcal{L}(W;\mathbf{X})}{\partial W} &= N(W^{-1})^t - \sum_{i=1}^{N} \mathbf{g}\left(W\mathbf{x}_i\right)\mathbf{x}_i^t \\ &\propto (W^{-1})^t - \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}\left(W\mathbf{x}_i\right)\mathbf{x}_i^t.\end{aligned} \tag{18}$$

.3 Gradient de la log-vraisemblance de l'ACI par rapport à la matrice de mixage

La log-vraisemblance d'une matrice de mixage A dans le cadre de l'ACI sans bruit est donnée par :

$$\mathcal{L}(A;\mathbf{X}) = \sum_{i=1}^{N}\sum_{s=1}^{S} \log\left(f^{\mathcal{Z}_s}\left((A^{-1}\mathbf{x}_i)_s\right)\right) - N\log(|\det(A)|), \tag{19}$$

en supposant que les densités des sources $f^{\mathcal{Z}_1},\ldots,f^{\mathcal{Z}_S}$ sont connues.

Pour calculer le gradient de $\mathcal{L}(A;\mathbf{X})$ par rapport à A, il est nécessaire de calculer la dérivée d'un élément de l'inverse d'une matrice par rapport à un second élément de la même matrice, cette dérivée est donnée par :

$$\frac{\partial (X^{-1})_{sp}}{\partial X_{lk}} = -(X^{-1})_{sl}(X^{-1})_{kp}, \tag{20}$$

voir (MacKay 1996), (Petersen et Pedersen 2008, p. 8). En utilisant (14) et (20) nous obtenons donc la dérivée de la log-vraisemblance par rapport à un élément de l,k de la matrice de mixage :

$$\begin{aligned}
\frac{\partial \mathcal{L}(A;\mathbf{X})}{\partial A_{lk}} &= -N(A^{-1})_{kl} + \sum_{i=1}^{N}\sum_{s=1}^{S} \frac{\partial \log\left(f^{\mathcal{Z}_s}\left((A^{-1}\mathbf{x}_i)_s\right)\right)}{\partial A_{lk}} \\
&= -N(A^{-1})_{kl} + \sum_{i=1}^{N}\sum_{s=1}^{S} \frac{\partial \log\left(f^{\mathcal{Z}_s}\left(\sum_{p=1}^{S}(A^{-1})_{sp}(\mathbf{x}_i)_p\right)\right)}{\partial A_{lk}} \\
&= -N(A^{-1})_{kl} + \sum_{i=1}^{N}\sum_{s=1}^{S} \left(\sum_{p=1}^{S}(A^{-1})_{sl}(A^{-1})_{kp}(\mathbf{x}_i)_p\right) g_s\left((A^{-1}\mathbf{x}_i)_s\right) \\
&= -N(A^{-1})_{kl} + \sum_{i=1}^{N}(A^{-1}\mathbf{x}_i)_k \sum_{s=1}^{S}(A^{-1})_{sl} g_s\left((A^{-1}\mathbf{x}_i)_s\right) \\
&= -N(A^{-1})_{kl} + \sum_{i=1}^{N}(A^{-1}\mathbf{x}_i)_k \sum_{s=1}^{S}(A^{-1})_{sl} (\mathbf{g}(A^{-1}\mathbf{x}_i))_s \\
&= -N(A^{-1})_{kl} + \sum_{i=1}^{N}(A^{-1}\mathbf{x}_i)_k \left((A^{-1})^t \mathbf{g}(A^{-1}\mathbf{x}_i)\right)_l,
\end{aligned} \tag{21}$$

avec g_s l'opposé de la fonction score de la source s (16) et g le vecteur contenant les opposées des fonctions scores de toutes les sources (17). La dérivée matricielle de la log-vraisemblance de l'ACI par rapport à la matrice de mixage est donc donnée par :

$$\begin{aligned}
\Delta A &\propto \frac{\partial \mathcal{L}(A;\mathbf{X})}{\partial A} \propto -N.(A^{-1})^t + \sum_{i=1}^{N}(A^{-1})^t \mathbf{g}\left(A^{-1}.\mathbf{x}_i\right)(A^{-1}\mathbf{x}_i)^t \\
&\propto -(A^{-1})^t + \frac{1}{N}\sum_{i=1}^{N}(A^{-1})^t \mathbf{g}\left(\mathbf{z}_i\right)\mathbf{z}_i^t \\
&\propto (A^{-1})^t\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}\left(\mathbf{z}_i\right)\mathbf{z}_i^t - \mathbf{I}\right),
\end{aligned} \tag{22}$$

avec $\mathbf{z}_i = A^{-1}\mathbf{x}_i$. Le gradient naturel $\Delta_{nat}A$ correspondant est quant à lui donné par (Lewicki et al. 1997, Amari et al. 1996) :

$$\Delta_{nat}A = AA^t\Delta A = A\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}\left(\mathbf{z}_i\right)\mathbf{z}_i^t - \mathbf{I}\right). \tag{23}$$

.4 Densité d'une transformation

Ces résultats peuvent être trouvés dans (Hyvärinen 2001, p. 35-36). Soit $\mathbf{x}, \mathbf{z}$ deux vecteurs de dimension d reliés par la relation déterministe suivante :

$$\mathbf{x} = \mathbf{h}(\mathbf{z}). \tag{24}$$

Si la transformation inverse de $\mathbf{h}$, $\mathbf{h}^{-1}$ existe et est unique alors la densité de X peut être obtenue à partir de la densité de Z comme suit :

$$f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(J_h(\mathbf{h}^{-1}(\mathbf{x})))|} f^{\mathcal{Z}}\left(\mathbf{h}^{-1}(\mathbf{x})\right), \tag{25}$$

où J_h est la matrice Jacobienne de $\mathbf{h}$, c'est à dire la matrice des dérivées partielles du premier ordre de $\mathbf{h}$. Dans le cas d'une transformation linéaire $\mathbf{x} = A\mathbf{z}$ inversible, nous obtenons donc :

$$f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(A)|} f^{\mathcal{Z}}\left(A^{-1}\mathbf{x}\right). \tag{26}$$

.5 ALGORITHME DE RECHERCHE LINÉAIRE

Lors de l'utilisation d'une méthode de type montée de gradient pour maximiser un critère du type

$$\arg\max_{\mathbf{w}} L(\mathbf{w}), \tag{27}$$

il est nécessaire de fixer le pas τ^* utilisé dans la règle de mise à jour des paramètres w, cette règle étant de la forme :

$$\mathbf{w}^{(q+1)} = \mathbf{w}^{(q)} + \tau^* \Delta\mathbf{w}^{(q)}, \tag{28}$$

avec $\Delta\mathbf{w}^{(q)}$ une direction de montée acceptable. Pour déterminer ce pas, il est possible de ce tourner vers des méthodes de recherche linéaire (Nocedal et Wright 1999, pages 30-63), qui comme leur nom l'indique recherche sur la direction définie par la direction de montée $\Delta\mathbf{w}^{(q)}$, une valeur du pas τ permettant d'améliorer la solution courante. Différentes solutions sont possibles pour cela. Nous avons utilisé une méthode très simple, partant d'un pas trop grand et diminuant celui-ci jusqu'à ce qu'une amélioration suffisante du critère soit obtenue (backtracking en anglais). Le pseudo-code de ce type de solution est le suivant :

Algorithme 10: pseudo-code de la recherche linéaire par backtracking

Données : pas de la recherche linéaire $\rho \in [0, 1]$, pas initial τ
tant que $L(\mathbf{w}^{(q)} + \tau\Delta\mathbf{w}^{(q)}) \leq L(\mathbf{w}^{(q)}) + \tau||\Delta\mathbf{w}^{(q)}||$ **faire**
    # *Diminution du pas*
    $\tau = \tau \times \rho$
$\tau^* = \tau$

Résultat : τ^*

.6 Exemples de signaux Icc aux quatre fréquences de fonctionnement

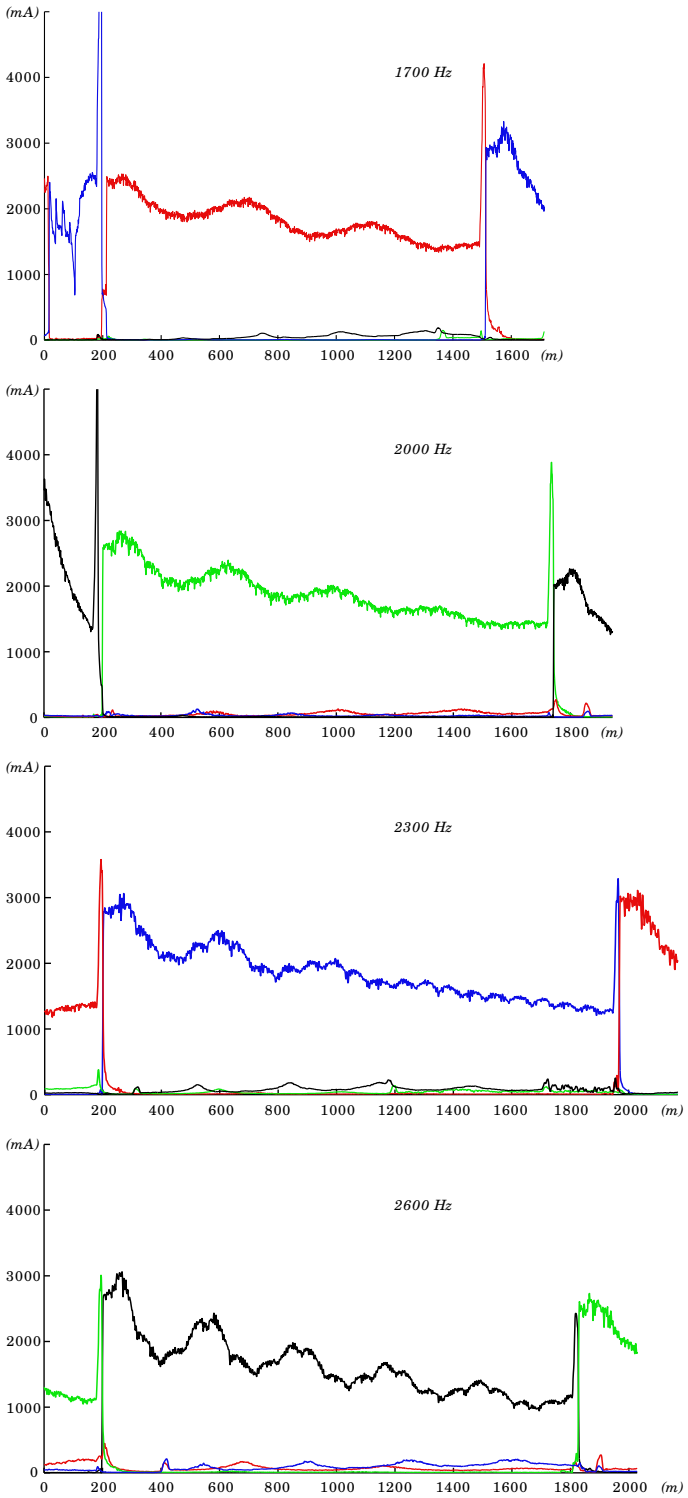


Fig. 14 – *Exemples de signaux d'inspection réels (amplitude du courant porteur Icc) sur circuit de voie UM71C-TVM, aux 4 fréquences de fonctionnement.*

BIBLIOGRAPHIE

M. Aizerman, E. Braverman et L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

H. Akaike. Information theory and an extension of the maximum liekelihood principle. Dans *Proceedings of the Second International Symposium on Information Theory*, 1973.

P. Aknin et H. Cygan. Improving the detection of rail cracks by using recursive radon transform. Dans *Proceedings of Railway Engineering Conference*, 2004.

P. Aknin, L. Oukhellou et F. Vilette. Track circuit diagnosis by automatic analysis of inspection car measurements. *WCRR*, 2003.

S. Amari, A. Cichocki et H. H. Yang. A new learning algorithm for blind signal separation. Dans *Proceedings of the 8th Conference on Advances in Neural Information Processing Systems (NIPS)*, volume 8, pages 757–763. MIT Press, 1996.

C. Ambroise, T. Denoeux, G. Govaert et P. Smets. Learning from an imprecise teacher : probabilistic and evidential approaches. Dans *Proceedings of the 10th International symposium on applied stochastic models and data analysis (ASMDA)*, volume 1, pages 100–105, 2001.

C. Ambroise et G. Govaert. EM algorithm for partially known labels. Dans *Proceedings of the 7th Conference of the International Federation of Classication Societies (IFCS)*, pages 161–166. Springer, 2000.

R. Amini et P. Gallinari. Semi-supervised learning with an imperfect supervisor. *Knowledge Information Systems*, 8(4):385–413, 2005. ISSN 0219-1377.

C. Archambeau, N. Delannay et M. Verleysen. Mixtures of robust probabilistic principal component analyzers. *Neurocomputing*, 71(7-9):1274–1282, 2008.

C. Archambeau, J. Lee et M. Verleysen. On convergence problems of the em algorithm for finite gaussian mixtures. Dans *European Symposium on Artificial Neural Networks (ESANN)*, pages 99–106, 2003.

A. Aregui et T. Denœux. Novelty detection in the belief functions framework. Dans *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-based systems (IPMU)*, volume 1, pages 412–419, 2006.

A. Asuncion et D. J. Newman. UCI machine learning repository, 2007.

H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

H. Attias. Independent factor analysis with temporally structured factors. Dans *Proceedings of the 12th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 386–392. MIT Press, 2000.

F. R. Bach et M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003. ISSN 1533-7928.

A. D. Back et A. S. Weigend. A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8(5), 1997.

T. Bakir, A. Peter, R. Riley et J. Hackett. Non-negative maximum likelihood ICA for blind source separation of images and signals with application to hyperspectral image subpixel demixing. Dans *Proceedings of the IEEE International Conference on Image Processing*, pages 3237–3240, 2006.

J. D. Banfield et A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

D. J. Bartholomew et K. Martin. *Latent variable models and factor analysis*. Arnold, London, 1999. Seconde édition.

L. E. Baum, T. Petrie, G. Soules et N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

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

R. Bellman. *Dynamic Programming*. Princeton university Press, 1957.

Y. Bengio et Y. Grandvalet. Semi-supervised learning by entropy minimization. Dans *Proceedings of the 17th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 529–536. MIT Press, 2005.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 35:192–236, 1974.

C. Biernacki. Testing for a global maximum of the likelihood. *Journal of Computational and Graphical Statistics*, 14(3):657–674, 2005.

C. Biernacki, G. Celeux et G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

C. Biernacki, G. Celeux et G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41:561–575, 2003.

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

C. Bishop et J. Lasserre. Generative or discriminative ? Getting the best of both worlds. *Bayesian Statistics*, 8:3–23, 2007.

B. B. Biswal et J. L. Ulmer. Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. *Journal of Computer Assisted Tomography*, 23(2):265–271, 1999.

K. A. Bollen. *Structural Equations with Latent Variables*. Wiley, 1989.

M. Borga, T. Landelius et H. Knutsson. A unified approach to PCA, PLS, MLR and CCA. Rapport technique, Computer Vision Laboratory, Linkõping University, 1995.

B. E. Boser, I. M. Guyon et V. N. Vapnik. A training algorithm for optimal margin classifiers. Dans *Proceedings of the 5th ACM Workshop on Computational Learning Theory (COLT)*, pages 144–152, 1992.

G. Bouchard. *Generative models in supervised statistical learning with applications to digital image categorization and structural reliability.* Thèse de doctorat, Université Joseph Fourier - Grenoble 1, 2005.

C. Bouveyron. *Modélisation et classification des données de grande dimension, Application à l'annalyse d'images*. Thèse de doctorat, Université Joseph Fourrier - Grenoble 1, 2006.

C. Bouveyron, S. Girard et C. Schmid. High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52(1):502–519, 2007.

L. Breiman, J. H. Friedman, R. A. Olshen et C. J. Stone. *Classification and Regression Trees*. CRC Press, 1984.

O. Cappé et E. Moulines. Online EM algorithm for latent data models, 2007. URL `http://arxiv.org/abs/0712.4273`.

J. F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4(4):112–114, 1997.

J. F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

J. F. Cardoso et B. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, 1996.

F. Caron, B. Ristic, E. Duflos et F. Vanheeghe. Least committed basic belief density induced by a multivariate gaussian : Formulation with applications. *International Journal of Approximate Reasoning*, 48(3):417–438, 2008.

G. Celeux et J. Diebolt. A random inputation principle : The stochastic EM algorithm. Rapport Technique 901, INRIA, 1988.

G. Celeux et G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computation Statistics and Data Analysis*, 14:315–332, 1992.

G. Celeux et G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

J. Chambers, W. Cleveland, B. Kleiner et P. Tukey. *Graphical Methods for Data Analysis*. Wadsworth, 1983.

O. Chapelle, B. Schölkopf et A. Zien, éditeurs. *Semi-Supervised Learning*. MIT Press, 2006.

W. Chu et Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2004.

A. Cichocki et S. Amari. *Adaptive Blind Signal and Image Processing*. Wiley, 2002.

B. R. Cobb et P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.

E. Côme. Diagnostic de systèmes spatialement répartis, par modèle génératif et méthode à noyau. application au diagnostic des circuits de voie ferroviaires. Dans J. Marais et J. Berbineau, éditeurs, *Actes INRETS, Communiquer, naviguer, surveiller. Innovations pour des transports plus sûrs, plus efficaces et plus attractifs*, volume 112, pages 97–106, 2007.

E. Côme, L. Oukhellou, P. Aknin et T. Denœux. Diagnostic de système spatialement répartis, modèle génératif et méthode à noyau. Dans *Actes du 13ème colloque GRETSI*, pages 84–84, 2007.

E. Côme, L. Oukhellou, T. Denœux et P. Aknin. Learning form partially supervised data using mixture models and belief functions. *Pattern recognition*, 42:334–348, 2009.

P. Comon. Independent Component Analysis, a new concept ? *Signal Processing*, 36(3):287–314, 1994. Special issue on Higher-Order Statistics.

A. Corduneanu et T. Jaakkola. On information regularization. Dans *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pages 151–158, 2003.

T. M. Cover et J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

F. G. Cozman, I. Cohen et M. C. Cirelo. Semi-supervised learning of mixture models. Dans *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 99–106, 2003.

T. De Bie. *Semi-supervised learning based on kernel methods and graph cut algorithms*. Phd thesis, K.U.Leuven (Leuven, Belgium), 2005.

A. Debiolles. *Diagnostic de systèmes complexes à base de modèle interne, reconnaissance des formes et fusion d'informations. Application au diagnostic des Cricuits de Voie ferroviaires*. Thèse de doctorat, Université de Technologie de Compiègne, 2007.

F. Delmotte et P. Smets. Target identification based on the Transferable Belief Model interpretation of Dempster-Shafer model. *IEEE Transactions on Systems, Man and Cybernetics A*, 34(4):457–471, 2004.

A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(5):804–813, 1995.

T. Denoeux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.

T. Denœux et P. Smets. Classification using belief functions : the relationship between the case-based and model-based approaches. *IEEE Transactions on Systems, Man and Cybernetics part B*, 36(6):1395–1406, 2006.

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

G. Druck, C. Pal, A. McCallum et X. Zhu. Semi-supervised classification with hybrid generative/discriminative methods. Dans *Proceedings of the 13th International Conference on Knowledge Discovery and Data mining (KDD)*, pages 280–289, 2007.

D. Dubois et H. Prade. *Théorie des Possibilités. Applications à la Représentation des Connaissances en Informatique*. Masson, Paris, 1985.

D. Dubois et H. Prade. On the unicity of dempster's rule of combination. *International Journal of Intelligent Systems*, 1:133–142, 1986a.

D. Dubois et H. Prade. A set-theoretic view of belief functions : logical operations and approximations by fuzzy sets. *International Journal of General Systems*, 12 (3):193–226, 1986b.

D. Dubois, H. Prade et P. Smets. New semantics for quantitative possibility theory. Dans *Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU)*, pages 410–421. Springer, 2001.

B. Dubuisson. *Diagnostic et Reconnaissance des formes*. Hermès, 1990.

B. Dubuisson. *Automatique et Statistique pour le Diagnostic*. Hermès, 2001a.

B. Dubuisson. *Diagnostic, intelligence artificielle et reconnaissance des formes*. Hermès, 2001b.

R. O. Duda, P. E. Hart et D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley, 2000.

B. Efron, T. Hastie, I. Johnstone et R. J. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

B. Efron et R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1994.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, pages 309–368, 1922.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

E. Fix et J. L. Hodges. Discriminatory analysis, nonparametric discrimination : Consistency properties. Rapport Technique 4, USAF School of Aviation Medicine, 1951.

C. Fraley et A. Raftery. Model-based clustering, discriminant analysis, and density estimation. *American Statistical Association*, 97:611–631, 2002.

C. Fraley et A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, 2007.

J. Friedman et J. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23:881–889, 1974.

G. Govaert, éditeur. *Analyse des données*. Hermès, 2003.

Y. Grandvallet. Logistic regression for partial labels. Dans *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, volume III, pages 1935–1941, 2002.

D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–15, 2006.

T. Hastie, T. Tibshirani et J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. Statistics. Springer, 2006.

S. Haykin et Z. Chen. The cocktail party problem. *Neural Computation*, 17(9): 1875–1902, 2005.

J. Hérault, C. Jutten et B. Ans. Détection de gandeur primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non-supervisé. Dans *Actes du 10ème colloque GRETSI*, pages 1017–1022, 1985.

J. Hérault, A. Oliva et A. Guerin-Dugue. Scene categorisation by curvilinear component analysis of low frequency spectra. Dans *5th European Symposium on Artificial Neural Network (ESANN)*, pages 91–96, 1997.

A. E. Hoerl et R. W. Kennard. Ridge Regression : biaised estimation for non-orthogonal problems. *Technometrics*, 12:55–67, 70.

D. W. Hosmer. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 29:761–770, 1973.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441,498–520, 1933.

E. Hüllermeier et J. Beringer. Learning from ambiguously labeled examples. Dans *Proceedings of the 6th International Symposium on Intelligent Data Analysis (IDA)*, pages 168–179, 2005.

A. Hyvärinen. Fast and robust fixed point algorithms for independant component analysis. *IEEE, Transaction on Neural Networks*, 10(3), 1999.

A. Hyvärinen. *Independant Component Analysis*. Wiley, 2001.

A. Hyvärinen et R. Karthikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*, 49(1):151–162, 2002.

S. Ikeda. ICA on noisy data : a factor analysis approach. Dans *Advances in independant component analysis*, pages 201–215. Springer, 2000.

V. V. Ivanov. On linear problem which are not well-posed. *Soviet Math Docl*, 3(4): 981–983, 1962.

T. Jaakkola et M. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

T. Jebara. *Discriminative, Generative and Imitative learning*. Phd thesis, Media Laboratory, MIT, 2001.

T. Joachims. Transductive inference for text classification using support vector machine. Dans *Proceedings of the 6th International Conference on Machine Learning (ICML)*, pages 202–209. Morgan Kaufmann, 1999.

M. Jordan. *Learning in Graphical Models (Adaptive Computation and Machine Learning)*. MIT Press, 1998.

M. Jordan. *An introduction to graphical models*. Berkeley, U. C., 2006.

M. Jordan, Z. Ghahramani, T. Jaakkola et L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

I. Jraidi et Z. Elouedi. Belief classification approach based on generalized credal EM. Dans *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, pages 524–535. Springer, 2007.

C. Jutten et P. Comon, éditeurs. *Séparation de source 1, concepts de base et analyse en composantes indépendantes*. Hermès, 2007a.

C. Jutten et P. Comon, éditeurs. *Séparation de source 2, au-delà de l'aveugle et application*. Hermès, 2007b.

H. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23:187–200, 1958.

A. Khlaifi, A. Ionescu et Y. Candau. Identification des sources et quantification de leur contribution aux niveaux de pm10. etude d'un site industriel en italie. *Journal Européen des Systèmes Automatisés*, 39:437–453, 2005.

G. Kimeldorf et G. Wahba. Some results on tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

K. Kiviluoto et E. Oja. Independent component analysis for parallel financial time series. Dans *International Conference on Neural Information Processing (ICONIP)*, pages 895–898, 1998.

G. J. Klir et M. J. Wierman. *Uncertainty-Based Information. Elements of Generalized Information Theory*. Springer, 1998.

K. H. Knuth. A bayesian approach to sources separation. Dans *Proceedings of the International Conference on Independent Component Analysis (ICA)*, pages 283–288, 1999.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

J. Lasserre, C. Bishop et T. Minka. Principled hybrids of generative and discriminative models. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

F. Lauer et G. Bloch. Incorporating prior knowledge in support vector machines for classification : A review. *Neurocomputing*, 71(7-9):1578–1594, 2008.

S. Lauritzen et D. Spieger. Local computations with probabilities on graphical structures and their applications to experts systems. *Journal of the Royal Statistical Society, Series B*, 50(2):175–224, 1988.

N. Lawrence et C. Bishop. Variational bayesian independent component analysis. Rapport technique, University of Manchester, 2000.

N. D. Lawrence et B. Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. Dans *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 306–313. Morgan Kaufmann, San Francisco, CA, 2001.

D. D. Lee et H. S. Seung. Learning the parts of objects by non negative matrix factorization. *Nature*, 401(6755), 1999.

M. S. Lewicki, T. J. Sejnowski et H. Hughes. Learning nonlinear overcomplete representations for efficient coding. Dans *Proceedings of the 10th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 815–821. MIT Press, 1997.

H. Li, T. Adal, W. Wang, D. Emge et A. Cichocki. Non-negative matrix factorization with orthogonality constraints and its application to raman spectroscopy. *The Journal of VLSI Signal Processing*, 48(1-2):83–97, 2007a.

Y. Li, L. Wessels, D. De Ridder et M. Reinders. Classification in the presence of class noise using a probabilistic kernel fisher method. *Pattern Recognition*, 40: 3349–3357, 2007b.

H. Lodhi, J. S. Taylor, N. Cristianini et C. Watkins. Text classification using string kernels. Dans *Proceedings of the 13th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 563–569. MIT Press, 2000.

D. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Non publié, 1996.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. Dans *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

S. Makeig, T. P. Jung, A. J. Bell, D. Gharamani et T. Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences (USA)*, 94:10979–10984, 1997.

J. M. Marín et C. Robert. *Bayesian Core, A practical approach to computational Bayesian statistics*. Springer, 2007.

J. M. Marín, M. T. Rodríguez et M. P. Wiper. Using weibull mixture distributions to model heterogeneous survival data. *Communications in statistics. Simulation and computation*, 34(3):673–684, 2005.

G. J. McLachlan. Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *Journal of the American Statistical Association*, 72(358):403–406, 1977.

G. J. Mclachlan et T. Krishnan. *The EM algorithm and Extension*. Wiley, 1996.

G. J. Mclachlan et D. Peel. *Finite Mixture Models*. Wiley, 2000.

X. L. Meng et D. B. Rubin. Maximum likelihood estimation via the ECM algorithm : A general framework. *Biometrika*, 80(2):267–278, 1993.

P. A. Monney. *A Mathematical Theory of Arguments for Statistical Evidence*. Contributions to Statistics. Physica-Verlag, 2003.

E. Moulines, J. Cardoso et E. Cassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. Dans *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 3617–3620, 1997.

S. Moussaoui. *Séparation de sources non-négatives. Application au traitement des signaux de spectroscopie*. Thèse de doctorat, Université Henri Poincaré - Nancy I, 2005.

R. Neal et G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. Dans M. I. Jordan, éditeur, *Learning in Graphical Models*. Kluwer, 1998.

A. Ng et M. Jordan. On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. Dans *Proceedings of the 14th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 841–848. MIT Press, 2001.

A. Ng, M. Jordan et Y. Weiss. On spectral clustering : Analysis and an algorithm. Dans *Proceedings of the 15th Conference on Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2002.

K. Nigam, A. McCallum, S. Thrun et T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.

J. Nocedal et S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, 1999.

T. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73(364):821–826, 1978.

L. Oukhellou, P. Aknin et E. Delechelle. Infrastructure system diagnosis using empirical mode decomposition and hilbert transform. Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, 2006.

E. Parzen. On estimation of probability function and mode. *Annals of Mathematical Statistics*, 33(3), 1962.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

K. B. Petersen et M. S. Pedersen. The matrix cookbook, 2008.

M. Pontil, S. Mukherjee et F. Girosi. On the noise model of support vector machine regression. *A.I. Memo, MIT Artificial Intelligence Laboratory*, 1651:1500–1999, 1998.

J. Qiu, M. Hue, A. Ben-Hur, J. P. Vert et W. S. Noble. An alignment kernel for protein structures. *Bioinformatics*, 23(9):1090–1098, 2007.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

L. Rabiner et B. H. Juang. *Fundamentals of Speech Recognition*. Signal Processing. Prentice Hall, 1993.

A. Ramer. Uniqueness of information measure in the theory of evidence. *Fuzzy Sets and Systems*, 24:183–196, 1987.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

S. Roberts et R. Everson, éditeurs. *Indepedent Component Analysis, Principles and Practices*. Cambridge Univeristy Press, 2001.

K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Dans *Proceedings of the IEEE*, volume 86, pages 2210–2239, 1998.

F. Rosenbalt. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

S. Roweis et Z. Ghahramani. A unifying review of linear gaussian models. Rapport technique, University of Toronto, 1997.

D. Rubin et D. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47 (1):69–76, 1982.

D. E. Rumelhart, G. E. Hinton et R. J. Williams. Learning representations by backpropagating errors. *Nature*, 323:533–536, 1986.

A. Samé. *Modèles de mélange et classification de données acoustiques en temps réel*. Thèse de doctorat, Université de Technologie de Compiègne, 2004.

A. Samé, E. Côme, L. Oukhellou et P. Aknin. Un algorithme GEM pour le débruitage de signaux. Dans *Actes de la 13ème Rencontres de la Societée Française de Classification*, pages 195–199, 2006.

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

G. Saporta. *Probabilités analyse des données et statistique*. Technip, 1990.

N. Saravanan, V. N. S. Siddabattuni et K. I. Ramachandran. A comparative study on classification of features by SVM and PSVM extracted using morlet wavelet for fault diagnosis of spur bevel gear box. *Expert Systems with Applications*, 35 (3):1351–1366, 2008.

B. Schölkopf, C. Platt, J. Shawe-Taylor, A. J. Smola et R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 2001.

B. Schölkopf et A. J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.

G. Schwarz. Estimating the number of components in a fnite mixture model. *Annals of Statistics*, 6:461–464, 1978.

A. K. Seghouane et A. Cichocki. Bayesian estimation of the number of principal components. *Signal Processing*, 87(3):562–568, 2007.

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

C. E. Shannon. A mathematical theory of communication. *Bell Systems Techinal Journal*, 27:379–423,623–656, 1948.

R. Shapire. Strength of weak learnability. *Journal of Machine Learning*, 5:197–227, 1990.

J. Shawe-Taylor et N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

P. P. Shenoy. Conditional independence in uncertainty theories. *Uncertainty in Artificial Intelligence*, 8:284–291, 1992.

P. P. Shenoy. Representing conditional independence relations by valuation networks. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2):143–165, 1994.

P. P. Shenoy et P. H. Giang. Decision making on the sole basis of statistical likelihood. *Artificial Intelligence*, 165(2):137–163, 2005.

P. P. Shenoy et J. Kohlas. Computation in valuation algebras. *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, 5:5–39, 2000.

P. Siarry, J. Dréo, A. Pétrowski et E. Taillard. *Métaheuristiques pour l'optimisation difficile*. Eyrolles, 2003.

P. Smets. *Un modèle mathématico-statistique simulant le processus du diagnostic médical*. Thèse de doctorat, Université Libre de Bruxelles, 1978.

P. Smets. Possibilistic inference from statistical data. Dans A. Bellester, D. Cardus et E. Trillas, éditeurs, *Proceedings of the 2nd World Conference on Mathematics at the service of Man*, pages 611–613, 1982.

P. Smets. Constructing the pignistic probability function in a context of uncertainty. Dans M. Henrion, R. D. Schachter, L. N. Kanal et J. F. Lemmer, éditeurs, *Proceedings of the 5th Conference on Uncertainty in Artificial Intelligence*, pages 29–40. North-Holland, 1990a.

P. Smets. Belief functions : The disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9(1):1–35, 1993.

P. Smets. The axiomatic justification of the transferable belief model. Rapport technique, Université Libre de Bruxelles., 1995.

P. Smets. Numerical representation of uncertainty. Dans D. M. Gabbay et P. Smets, éditeurs, *Handbook of Defeasible reasoning and uncertainty management systems*, volume 3, pages 265–309. Kluwer Academic Publishers, 1998.

P. Smets. Belief functions on real numbers. *International Journal of Approximate Reasoning*, 40(3):181–223, 2005a.

P. Smets. Decision making in the tbm : the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005b.

P. Smets et R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66: 191–243, 1994.

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

C. Spearman. General intelligence, objectively determined and measured. *American Journal of psychology*, 15:201–293, 1904.

N. A. Thikonov. On solving ill posed problems and method of regularization. *Doklady Akademii Nauk*, 153:501–504, 1963.

L. L. Thurstone. *Multiple Factor Analysis*. University of Chicago Press, 1947.

M. E. Tipping et C. Bishop. Mixtures of principal component analyzers. *Neural Computation*, 11(2):443–482, 1997a.

M. E. Tipping et C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1997b.

D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society*, 46(2):257–267, 1984.

N. Ueda et R. Nakano. Deterministic annealing variant of the EM algorithm. Dans *Proceedings of the 7th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 545–552. MIT Press, 1995.

N. Ueda, R. Nakano, Z. Ghahramani et G. E. Hinton. Split and merge EM algorithm for improving gaussian mixture density estimates. *The Journal of VLSI Signal Processing*, pages 133–140, 2000.

N. Valentin et T. Denœux. A neural network-based software sensor for coagulation control in a water treatment plant. *Intelligent Data Analysis*, 5:23–39, 2001.

H. Valpola, T. Östman et J. Karhunen. Nonlinear independent factor analysis by hierarchical models. Dans *Proceedings of the International Conference on Independent Component Analysis (ICA*, pages 257–262, 2003.

P. Vannoorenberghe. Estimation de modèles de mélanges finis par un algorithm EM crédibiliste. *Traitement du Signal*, 24(2):103–113, 2007.

P. Vannoorenberghe et P. Smets. Partially supervised learning by a Credal EM approach. Dans *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ESQUARU)*, pages 956–967, 2005.

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

V. N. Vapnik et V. Chervonenkis. Teoriya raspoznavaniya obrazov : Statisticheskie problemy obucheniya (theory of pattern recognition : Statistical problems of learning), 1974.

J. P. Vert, R. Thurman et W. S. Noble. Kernels for gene regulatory regions. Dans *Proceedings of the 18th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 1401–1408. MIT Press, 2006.

R. Vigàrio. Extraction of ocular artifacts from EEG using independant component analysis. *Electroencephalography and clinical neurophysiology*, 103(3):395–404, 1997.

F. Vrins, J. Lee et Verleysen M. Can we always trust entropy minima in the ica context ? pages 1107–1111, 2005.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics (SIAM), 1990.

P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

P. Walley et S. Moral. Upper probabilities based on the likelihood function. *Journal of the Royal Statistical Society B*, 161:831–847, 1999.

J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, 58(301):236–244, 1963.

C. F. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

H. Xu et P. Smets. Evidential reasoning with conditional belief functions. Dans *Proccedings of the Conference on Uncertainty in Artificial Intelligence*, pages 98–605, 1994.

L. Xu et M. Jordan. On convergence properties of the EM algorithm for Gaussian Mixtures. *Neural Computation*, 8(1):129–151, 1996.

B. B. Yaghlane, Ph. Smets et K. Mellouli. Belief function independence i. the marginal case. *International Journal of Approximate Reasoning*, 2002a.

B. B. Yaghlane, Ph. Smets et K. Mellouli. Belief function independence ii. the conditional case. *International Journal of Approximate Reasoning*, 2002b.

L. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

K. Zhang et L. W. Chan. ICA with sparse connections. Dans *Proceedings of Intelligent Data Engineering and Automated Learning Conference (IDEAL)*, pages 530–537. Springer, 2006.

D. Zhou, O. Bousquet, T. Lal et B. Schölkpf. Learning with local and global consistency. Dans *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 321–328. MIT Press, 2003.

J. Zhu et T. Hastie. Kernel logistic regression and the import vector machine. Dans *Proceedings of the 14th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 1081–1088. MIT Press, 2001.

X. Zhu et Z. Ghahramani. Learning from labelled and unlabelled data with label propagation. Rapport technique, Carnegie Mellon University, 2002.

G. Zwingelstein. *Diagnostic des défaillances*. Hermès, 2002.

Index

NOTATIONS

De manière générale, les lettres capitales $X, Y, Z, \ldots$ représentent des variables aléatoires, les lettres calligraphiées $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ représentent les domaines de définition de ces variables aléatoires.

Les vecteurs et donc les réalisations des variables aléatoires multidimensionnelles sont dénotés par des lettres minuscules grasses $\mathbf{x}, \mathbf{y}, \ldots$ (vecteurs colonness), les matrices contenant des données par des lettres grasses majuscules $\mathbf{X}$, une application linéaire par une lettre majuscule $A, B, W, \ldots$

Enfin, les paramètres sont représentés par des lettres grecs minuscules $\theta, \pi, \ldots$; grasse lorsqu'il s'agit de vecteur $\boldsymbol{\psi}, \boldsymbol{\theta}, \ldots$; leurs domaines de définitions étant quant à eux représentés grâce à la lettre grecs majuscule correspondante $\boldsymbol{\Psi}, \boldsymbol{\Theta}, \ldots$

D'autres notations utiles à la compréhension de cette thèse suivent :

Indices	
i, j	indice sur un ensemble d'observations
k	indice sur un ensemble de classes ou groupes
s	indice sur un ensemble de variables latentes continues
p	indice sur un ensemble de variables observées

Constantes du problème	
P	dimension du vecteur forme
N	nombre d'observations
M	nombre d'observations labellisées
K	nombre de classes
S	nombre de variables latentes continues

Variables	
$\mathbf{x}$	observations à valeurs continues $\in \mathbb{R}^P$
$\mathbf{z}$	variables latentes continues $\in \mathbb{R}^S$
y	variable à valeurs discrètes $\in \{1, \ldots, K\}$

Jeu de données	
$\mathbf{X}$	Jeu de données non supervisé

$\mathbf{y}$	Vecteurs d'étiquettes
$\mathbf{X}^{ss}$	Jeu de données d'apprentissage semi supervisé
$\mathbf{X}^{ps}$	Jeu de données d'apprentissage partiellement supervisé
$\mathbf{X}^{bl}$	Jeu de données d'apprentissage avec bruit de label
$\mathbf{X}^{iu}$	Jeu de données d'apprentissage avec des labels doux

Probabilité

$p(\,.\,)$	notation générique de la densité associée à une mesure de probabilité discrète ou continue
$f(\,.\,)$	notation générique d'une densité
$f(\,.\,;\boldsymbol{\psi})$	notation générique d'une densité paramétrée par $\boldsymbol{\psi}$
$\varphi(\,.\,;\mu,\nu)$	densité d'une loi normale monodimensionnelle de moyenne μ et de variance ν
$\boldsymbol{\varphi}(\,.\,;\boldsymbol{\mu},\boldsymbol{\Sigma})$	densité d'une loi normale multidimensionnelle de moyenne $\boldsymbol{\mu}$ et de matrice de variance $\boldsymbol{\Sigma}$
$\mathbb{P}(.)$	fonction de probabilité
$\mathbb{E}[X]$	espérance de X
$\mathbb{E}[Y\|X=x]$	espérance de Y conditionnellement à $X=x$
$X \perp\!\!\!\perp Y$	X indépendante de Y
$(X \perp\!\!\!\perp Y)\|Z$	X indépendante de Y conditionnellement à Z

Lois de probabilités

$\mathcal{N}(\,.\,;\boldsymbol{\mu},\boldsymbol{\Sigma})$	loi normale multidimensionnelle de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$
$\mathcal{M}(\,.\,;\boldsymbol{\pi})$	loi multinomiale de paramètre $\boldsymbol{\pi}$

Estimation

$L(\boldsymbol{\psi};\mathbf{X})$	fonction de vraisemblance associée au paramètre $\boldsymbol{\psi}$ par rapport au jeu de données $\mathbf{X}$
$\mathcal{L}(\boldsymbol{\psi};\mathbf{X})$	fonction de log-vraisemblance associée au paramètre $\boldsymbol{\psi}$ par rapport au jeu de données $\mathbf{X}$
$\widehat{\boldsymbol{\psi}}$	estimateur de $\boldsymbol{\psi}$

Théorie des fonctions de croyances

2^{Ω}	l'ensemble des parties de l'ensemble Ω
$\|\Omega\|$	cardinal de Ω
$m^{\Omega}(.)$	fonction de masse de croyance sur l'ensemble Ω
$pl^{\Omega}(.)$	fonction de plausibilité sur l'ensemble Ω

$bel^{\Omega}(.)$	fonction de croyance sur l'ensemble Ω
$q^{\Omega}(.)$	fonction de comodalité sur l'ensemble Ω
$\textcircled{\cup}$	combinaison conjonctive
$\textcircled{\cap}$	combinaison disjonctve
$\oplus$	combinaison conjonctive normalisée
$\uparrow$	extension vide
$\downarrow$	marginalisation

Algèbre linéaire

$\mathbf{x}$	vecteur colonne
A, B	matrice
A_{ij}	intersection de la i^{e} ligne et de la j^{e} colonne de A
$A_{i.}$	i^{e} ligne de A
$A_{.j}$	j^{e} colonne de A
$(\mathbf{x})_i$	i^{e} ligne de $\mathbf{x}$
$\mathbf{x}^t$	transposé de $\mathbf{x}$
A^{-1}	inverse de A
$\mathrm{tr}(A)$	trace de A
$\det(A)$	déterminant de A
$(\mathbf{v}_l, \lambda_l)$	l^{e} plus grand couple de valeur propre vecteur propre

LISTE DE PUBLICATIONS

REVUES INTERNATIONALES AVEC COMITÉ DE LECTURE

- E. Côme, L. Oukhellou, T. Denœux et P. Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42:334–348, 2009.

- L. Oukhellou, E. Côme, L. Bouillaut et P. Aknin. Combined use of sensor data and structural data processed by bayesian network. application to a railway diagnosis aid scheme. *Transportation Research Part C*, 16:755–767, 2008.

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

CONFÉRENCES INTERNATIONALES AVEC ACTES

- E. Côme, Z.L Cherfi, L. Oukhellou, T. Denœux et P. Aknin. Semi-supervised IFA with prior knowledge on the mixing process. An application to a railway device diagnosis. Dans *Proceedings of the 8^{th} International Conference on Machine Learning and Applications (ICMLA), San-Diego*, Décembre 2008.

- E. Côme, L. Oukhellou, T. Denœux et P. Aknin. Mixture model estimation with soft labels. Dans *Proceedings of the 4^{st} Conference on Soft Methods in Probability and Statistics (SMPS), Toulouse*, pages 165–174, 2008.

- E. Côme, L. Bouillaut, P. Aknin et L. Oukhellou. Hidden markov random field, an application to railway infrastructure diagnosis. Dans *Proceedings of the 1^{st} IFAC Workshop on dependable control of discret systems (DCDS), Paris*, pages 155–160, 2007.

- E. Côme, L. Bouillaut, P. Aknin et A. Samé. Bayesian network for railway infrastructure diagnosis. Dans *Proceedings of the 11^{th} International Conference on Information Processing and Management of Uncertainty (IPMU), Paris*, pages 1436–1442, 2006.

- A. Debiolles, L. Oukhellou, P. Aknin, T. Denœux et E. Côme. Linear and non linear regression using PLS feature selection and NN on a defect diagnosis application. Dans *Proceedings of the International Conference On Machine Intelligence (ICMI), Tozeur, Tunisie*, CD-ROM, 2005.

CONFÉRENCES FRANCOPHONES AVEC ACTES

- E. Côme, L. Oukhellou, P. Aknin et T. Denœux. Diagnostic de systèmes spatialement répartis, modèle génératif et méthode à noyau. Dans *Actes du 11es Colloque du groupe de recherche et d'étude en traitement du signal (GRETSI), Troyes*, pages 633–636, 2007.

- E. Côme. Diagnostic de systèmes spatialement répartis, par modèle génératif et méthode à noyau. Application au diagnostic des circuits de voie ferroviaires. Dans J. Marais and M. Berbineau, editors, *Actes INRETS, Communiquer, naviguer, surveiller. Innovations pour des transports plus sûrs, plus efficaces et plus attractifs*, volume 112, pages 97–106, 2007.

- A. Samé, E. Côme, L. Oukhellou et P. Aknin. Un algorithme GEM pour le débruitage de signaux. Dans *Actes des 13es Rencontres de la Société Francophone de Classification (SFC)*, pages 195–199, 2006.

COMMUNICATIONS SANS ACTES

- L. Oukhellou, E. Côme, P. Aknin et T. Denœux. Diagnostic de systèmes répartis à l'aide de modèles génératifs en contexte supervisé ou partiellement supervisé. *Workshop Surveillance, Sûreté et Sécurité des grands systèmes, Troyes*, 2008.

- E. Côme, L. Oukhellou, et P. Aknin. Diagnosis of complex system by combined use of RKCCA and graphical model. *Workshop on Current Challenge in Kernel Methods, Bruxelles*, 2006.

- E. Côme, A. Samé, L. Oukhellou, et P. Aknin. Régression et modèle de mélange pour le débruitage de signaux. *2es Rencontres Inter-Associations : La classification et ses applications, Lyon*, 2006.

Ce document a été préparé à l'aide de l'éditeur de texte Gedit, du logiciel de composition typographique LaTeX 2_ε et du logiciel de dessin vectoriel Inkscape.

Résumé Le thème principal de cette thèse concerne la formalisation et la résolution du problème de l'apprentissage statistique lorsque les informations disponibles sur une ou plusieurs variables d'intérêt discrètes sont imprécises, incertaines. La solution proposée s'appuie sur une approche générative et sur l'utilisation de la théorie des fonctions de croyance afin de représenter l'information disponible sur ces variables. Nous montrons tout d'abord, comment des labels « doux », prenant la forme de fonctions de masse de croyance, peuvent être utilisés pour estimer les paramètres d'un modèle de mélange grâce à un critère étendant les critères rencontrés dans le cadre probabiliste. Le problème d'optimisation associé est quant à lui résolu grâce à une extension de l'algorithme EM. Une démarche similaire, dans le cadre de l'analyse en facteurs indépendants, modèle génératif extrêmement parcimonieux faisant intervenir un ensemble de variables d'intérêts discrètes, est également présentée et étudiée. D'autres part, une solution pour tirer parti d'informations sur le processus de génération des données dans le cadre de ce modèle est proposée. Enfin, des résultats concernant un problème réel de diagnostic permettent de juger de l'intérêt de ces propositions. Ce problème de diagnostic concerne un élément essentiel de la chaîne de contrôle-commande des trains sur le réseau français : le circuit de voie.

Mots-clés : Apprentissage statistique, Labellisation douce, Modèles génératifs, Fonctions de croyance, Algorithme EM, Analyse en facteurs indépendants, Diagnostic

Title Statistical learning of generative models for complex system fault diagnosis with soft labels and spatial constraints

Abstract The main topic of this thesis concerns the formalization and the resolution of statistical learning problem involving imperfect information on one or several discrete variables of interest. The solution advocates is build on top of the Dempster-Shaffer theory of evidence and a generative approach. We show first, how « soft » labels defined as a Dempster-Shafer basic belief assignments can be employed to define a criterion generalizing the likelihood function which can be used to compute estimates of mixture model parameters. A variant of the EM algorithm dedicated to the optimization of this criterion is furthermore proposed. A similar approach is also studied in the context of independent factor analysis, a parsimonious generative model dealing with several discrete variables. A solution to leverage prior knowledge on the generative process underlying this model is also supplied. Finally, results from a real diagnosis application demonstrates the interest of these proposals. This diagnosis application concerns an essential component of the French railway infrastructure : the track circuit.

Keywords : Statistical learning, Soft labels, Generative models, Demspter-Shaffer theory of evidence, EM algorithm, Independant Factor Analysis, Diagnosis