UNIVERSITÉ DE TECHNOLOGIE DE COMPIÈGNE
HEUDIASYC (UTC) - GRETTIA (IFSTTAR)

T H È S E

présentée en vue d'obtenir le titre de
Docteur de l'Université de Technologie de Compiègne

Spécialité : Technologie de l'Information et des Systèmes

par

Zohra Leila Cherfi

Diagnostic de systèmes complexes en contextes non supervisé et partiellement supervisé. Application au circuit de voie ferroviaire

Thèse soutenue le 10/10/2011 devant le jury composé de :

Président :	Walter Schön	-	HEUDIASYC (UTC)
Rapporteurs :	Pierre Beauseroy	-	LM2S (UTT)
	Arnaud Martin	-	IRISA (Université de Rennes 1)
Directeurs :	Patrice Aknin	-	GRETTIA (IFSTTAR)
	Thierry Denœux	-	HEUDIASYC (UTC)
Encadrante :	Latifa Oukhellou	-	GRETTIA (IFSTTAR)
Examinateur :	Emmanuel Ramasso	-	FEMTO-ST (ENSMM Besançon)

Remerciements

Je tiens tout d'abord à exprimer mes plus profonds remerciements à mes deux directeurs de thèse Patrice Aknin et Thierry Denœux pour leurs conseils avisés et leurs encouragements. Je tiens également à remercier mon encadrante Latifa Oukhellou qui m'a orientée et conseillée tout au long de cette thèse.

Je voudrais aussi exprimer mes sincères remerciements à Mr. Pierre Beauseroy et Mr. Arnaud Martin, rapporteurs de ce travail, pour leur lecture attentive de la thèse et les remarques constructives qu'ils m'ont faites. Merci également à Mr. Walter Schön et Mr. Emmanuel Ramasso d'avoir accepté de faire parti du jury et de m'avoir fait l'honneur d'évaluer mon travail.

Je remercie bien sûr l'ensemble du personnel du GRETTIA que j'ai pu côtoyer tout au long de cette thèse, avec une pensé particulière pour Allou, Etienne, Laurent, Olivier et Sébastien qui ont toujours su se montrer disponibles. Mes collègues thésards Inès, Raissa, Nicolas, Wissam, Rony et Johanna qui m'ont encouragée et avec qui j'ai eu le plaisir d'apprendre et de me détendre. Enfin, merci à mes anciens voisins de bureau Faicel et Cyril pour leur soutient et leurs conseils, ainsi qu'à Roland pour m'avoir initiée aux sorties *running*.

Je tiens aussi à adresser mes remerciements aux membres du personnel de la SNCF qui ont permis et qui ont participé à l'aboutissement de ces travaux, Mr. Santos, Mr. Ducloux, Mr. Havet, Mr. Parisot et Mr. Servais.

Enfin, bien évidemment, je remercie de tout cœur mes parents et mes proches qui m'ont supportée, encouragée et permis d'en arriver là.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction Générale

1.1 Contexte général

Ces dernières années, les développements liés à la sûreté de fonctionnement et à la disponibilité des systèmes industriels se sont vus accorder un intérêt croissant en accord avec l'enjeu économique qu'ils représentent. Dans le but de maintenir ces systèmes à un niveau de fonctionnement satisfaisant, la mise en œuvre de stratégies définissant une politique de maintenance intelligente est nécessaire.

Dans le cas des systèmes critiques, l'efficacité d'une telle stratégie a longtemps été considérée relativement au coût sécuritaire que peut entraîner une panne imprévue. En conséquence, des politiques de maintenance systématique étaient souvent privilégiées sur ce type de systèmes. Toutefois, ce type d'approche entraînant une fréquence d'intervention élevée et des remplacement couteux et parfois superflus, des approches prévisionnelles sont à présent favorisées avec pour objectif de réduire les coûts liés à l'entretien tout en préservant une qualité de service élevée.

Dans le cadre d'une maintenance prévisionnelle, la décision d'une action de maintenance doit reposer sur une analyse suffisamment pertinente de l'état du système. Un diagnostic est alors nécessaire : il s'agit tout d'abord de récupérer des informations les plus fines possibles sur l'état de fonctionnement courant du système et de fournir ensuite des décisions robustes quant à la présence éventuelle de défauts, leur localisation et leur gravité.

Le diagnostic est un champ disciplinaire à part entière qui s'est initialement développé dans la communauté automatique. Plus récemment, les communautés reconnaissance des formes, statistique et traitement du signal s'en sont emparés avec un certain succès. Ces nouvelles approches tirent en effet partie de l'existence de bases de données d'inspection (ou surveillance) de grande taille qui sont presque systématiquement constituées dans les grands domaines industriels. Ces bases de données proposent une lecture quasi exhaustive des états de fonctionnement des systèmes auxquelles elles sont dédiées. Les observations qu'elles comprennent sont de dimensions souvent trop importantes (variables redondantes, corrélées, non pertinentes ...) mais les méthodes issues de la statistique sont bien armées pour les analyser.

Deux grandes évolutions sont à noter dans les méthodes de diagnostic par reconnaissance des formes. La première concerne la volonté d'une lecture dynamique (ou temporelle) des données d'inspection. Il s'agit alors de replacer le résultat du

diagnostic dans une série de décisions successives prises « en ligne » au fur et à mesure de la fourniture des observations.

Si le champ de l'analyse de données propose une panoplie large de méthodes efficientes quelque soit le type de données à disposition, la nécessité de prendre en compte certaines spécificités liées au système à diagnostiquer est souvent gage de meilleure efficacité. Ce point est la deuxième catégorie d'évolution qu'il semble important de citer ici : la prise en compte d'a priori dans les modèles. A priori sur la structure des données, a priori sur la correspondance de certains états de fonctionnement avec certaines observations ... etc. Cette volonté d'intégrer des a priori trouve tout son sens dans les domaines où préexiste une forte expertise qu'il est utile de prendre en compte dans le développement de la méthode de diagnostic automatique. Elle permet de plus de diffuser et de faire accepter plus aisément les nouvelles méthodes de diagnostic au sein de la communauté d'expert concernée.

1.2 Contexte applicatif

Ces travaux de thèse s'inscrivent dans le cadre de la maintenance des infrastructures ferroviaires, et plus particulièrement des circuits de voie. Ce système intervient dans le contrôle-commande des trains et a pour fonction principale de détecter de façon automatique et continue la présence d'un véhicule sur une portion de voie donnée.

Sur les lignes à grande vitesse, le circuit de voie sert également de support à la transmission d'informations entre la voie et le train comme par exemple la vitesse limite transmise ainsi au conducteur. Ce système constitue un organe de sécurité important, dont la défaillance peut provoquer d'importants problèmes de signalisation et entraîner l'arrêt de l'exploitation. Une maintenance efficace de ce système est donc primordiale.

La maintenance des circuits de voie repose sur l'analyse de signaux d'inspection qui fournissent une signature des éléments qui composent ce système. Cette analyse présente certaines difficultés car les signaux en question ne permettent pas une observation directe de l'état des sous composants.

L'objectif des présents travaux est de mettre au point des outils d'analyse automatique des signaux d'inspection des circuits de voie, d'une part pour pour accélérer le processus de détection et de localisation des défauts avérés, et d'autre part pour anticiper l'apparition de défauts pénalisants.

1.3 Organisation de la thèse

Ce mémoire débute par le chapitre 2 qui présente l'application pratique ayant motivé ces travaux et met en avant la problématique du diagnostic. Après avoir décrit le système « circuit de voie », son fonctionnement et ses spécificités, la suite

du chapitre introduit la notion de diagnostic automatique et passe en revue ses principales approches. Une attention particulière a été portée aux méthodes de diagnostic par reconnaissance des formes qui ont déjà été investies lors de travaux précédents sur l'application circuit de voie [Debiolles 2007, Côme 2009].

Le chapitre 3 est consacré à la description des différents modèles et outils utilisés dans cette thèse. Il fait principalement référence aux modèles à variables latentes. Leur formulation est notamment intéressante pour l'introduction de connaissances supplémentaires en lien avec les spécificités de l'application ou avec la nature séquentielle des données.

Le chapitre 4 est principalement consacré à la mise en œuvre de méthodes non supervisées pour le diagnostic des circuits de voie et détaille deux extensions de modèles à variables latentes continues que constituent l'Analyse en Composantes Indépendants (ICA) et l'Analyse en Facteurs Indépendants (IFA). Si l'objet de la première extension est la prise en compte de connaissances structurelles, celui de la seconde concerne l'intégration de dépendances temporelles dans le modèle pour le traitement de données séquentielles. Cette méthode intègre des modèles de Markov cachés afin de prendre en compte l'aspect dynamique des données. L'intérêt de l'ensemble de ces propositions sera illustré à l'aide de résultats obtenus sur des signaux synthétiques équivalents aux signaux de surveillance prélevés sur les circuits de voie.

Le chapitre 5 explore l'intérêt d'intégrer des informations partiellement fiables, obtenues grâce à différents experts du domaine applicatif. L'approche proposée s'appuie sur l'Analyse en Facteurs Indépendants (IFA) et sur l'utilisation de la théorie des fonctions de croyance pour représenter l'information imparfaite et pour fusionner les avis d'expert à l'aide de différents schémas de combinaison. Ce chapitre met en avant une formulation de l'IFA faisant intervenir une information partielle sur les classes d'origine des individus et permet de poser un cadre général pour l'apprentissage en situation intermédiaire entre supervisée et non supervisée. L'approche est finalement illustrée par le diagnostic des circuits de voie à partir de signaux réels. Les signaux en question ont été fournis par la SNCF dans le cadre du projet ANR DIAGHIST, dont l'objectif a été de mettre en place une politique de maintenance prévisionnelle des circuits de voie.

Chapitre 2

Contexte applicatif et problématique de diagnostic

Sommaire

2.1 Introduction

Ce chapitre présente l'application pratique ayant motivé ces travaux de thèse et évoque les différentes approches pouvant être envisagées pour mettre en place une procédure de diagnostic automatique. La première partie de ce chapitre est consacrée à la description du système considéré et détaille la problématique du diagnostic de celui-ci en tant que système complexe. Les points abordés ont pour but de répondre aux questions suivantes :

- Qu'est-ce qu'un circuit de voie ?
- Quel rôle tient-il dans l'exploitation ferroviaire ?
- Quelles sont les défaillances possibles de ce système ?
- Quelles données sont utilisées pour sa surveillance ?

La seconde partie du chapitre introduit la notion de diagnostic automatique et passe en revue les différentes approches de diagnostic possibles. L'objectif de cette section est de retracer le cheminement des travaux précédemment menés dans le cadre du diagnostic des circuits de voie en vue de mieux positionner les démarches proposées.

2.2 Contexte ferroviaire

2.2.1 Généralités sur les Circuits de Voie (CdV)

Les circuits de voie (CdV) tiennent un rôle essentiel en signalisation ferroviaire car ils permettent de détecter de façon automatique et continue la présence d'un véhicule sur une portion de voie donnée. Ils commandent ainsi l'activation des feux de signalisation placés en bord de voie afin d'indiquer aux agents de conduite les autorisations de franchissement et les arrêts obligatoires.

Le système de signalisation est conçu pour assurer la sécurité des circulations et éviter le phénomène de rattrapage, qui désigne la collision entre deux trains circulant dans le même sens sur la même voie. Afin d'empêcher ce type d'accident, la voie est divisée en zones appelées cantons, dont la longueur peut varier de quelques centaines de mètres à quelques kilomètres suivant les caractéristiques de la ligne, sa vitesse de circulation en particulier. Un espacement d'au moins un canton entre deux trains est imposé, afin de disposer du temps et de la distance de freinage nécessaires en cas de besoin. A l'entrée de chaque canton, un feu de signalisation indique si un véhicule peut circuler sur la portion de voie en question. Ces feux sont régis par les CdV.

Le CdV est principalement constitué des éléments suivants (figure 2.1) :

- 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 deux 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 des signaux d'entrée des cantons.

Aux deux extrémités du CdV, des circuits bouchons électriques isolent électriquement les rails des cantons adjacents et empêchent la propagation du signal émis par un CdV sur les CdV voisins. Ces circuits bouchons 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. Lorsque la voie est libre et qu'aucun véhicule n'est présent sur la zone délimitant le CdV, 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, si la voie est occupée, autrement dit lorsqu'un véhicule est présent, son premier essieu agit comme une faible résistance, appelée shunt, et court-circuite la transmission. Dans ce cas, le niveau du signal parvenant au récepteur 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.

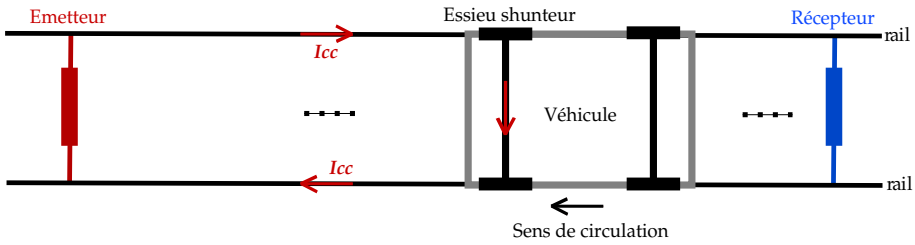


FIGURE 2.1 – *Schéma d'un circuit de voie non compensé.*

Les CdV sont donc des éléments essentiels à la chaîne de contrôle-commande des trains. Nous allons voir que leur rôle est encore plus important sur le réseau français des lignes à grande vitesse (LGV).

2.2.2 La signalisation sur les lignes à 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. Sur les lignes à grande vitesse, où les trains peuvent circuler jusqu'à 320 km/h, la signalisation est différente car les panneaux latéraux sont difficiles à observer par le conducteur. Les informations de signalisation sont donc transmises directement en cabine, au moyen d'afficheurs spécifiques indiquant les vitesses limites autorisées ou les annonces d'arrêt. Si le conducteur ne les respecte pas, une procédure automatique d'arrêt d'urgence se déclenche automatiquement.

Un tel mode de signalisation nécessite un système de transmission continue d'information entre la voie et les véhicules pour mettre à jour les informations affichées en cabine. Plusieurs technologies ont été développées par différents pays. Dans certains pays, ce système consiste à utiliser un câble rayonnant posé en voie pour transmettre au train la distance avant le prochain point d'arrêt, la vitesse maximale autorisée et les paramètres de freinage. Ce système est connu sous le nom de LZB (Linienzugbeeinflussung) et est utilisé en Allemagne, en Autriche et en Espagne [Uebel 1989][Canarail 2003].

En France, la technologie mise au point pour résoudre ce problème utilise directement les CdV comme système de transmission. Cette technologie appelée TVM pour Transmission Voie-Machine, est également utilisée en Corée du Sud, ainsi que 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, deux mètres 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 ... etc. Elle est ensuite envoyée à un ordinateur de bord qui génère le profil idéal de vitesse qui est affiché en cabine.

Hormis cette particularité liée à la modulation, la TVM utilise un CdV classique (émetteur, ligne de transmission, récepteur). Pour éviter d'éventuelles interférences dues aux CdV encadrants, quatre fréquences différentes de porteuses sont utilisées de façon alternée sur les deux voies parallèles (figure 2.2).

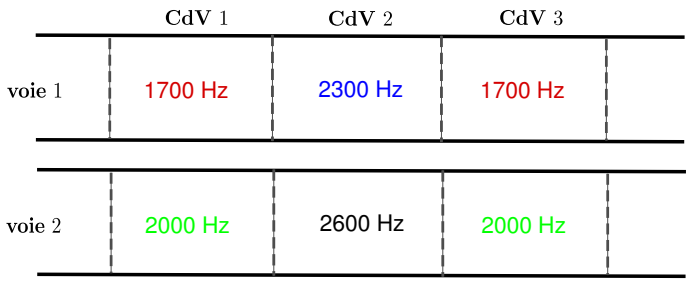


FIGURE 2.2 – *Fréquences de porteuses alternées sur les deux voies parallèles. Sur la voie 1, les deux fréquences de CdV utilisées sont* $f_1 = 1700\,\text{Hz}$ *et* $f_2 = 2300\,\text{Hz}$. *Sur la voie 2, il s'agit de* $f_1 = 2000\,\text{Hz}$ *et* $f_2 = 2600\,\text{Hz}$.

Condensateurs de compensation

La seconde particularité du CdV utilisé pour la TVM (CdV-TVM) est qu'il est plus long que le CdV classique. Ceci est directement lié à la distance de freinage qui augmente avec la vitesse de circulation et conduit à des distances entre émetteurs et récepteurs pouvant atteindre 2500 m. Pour que la transmission entre la voie et la machine soit efficace sur de telles longueurs, il est nécessaire que l'amplitude du signal émis se maintienne à un niveau suffisant malgré l'affaiblissement inéluctable du signal lié au comportement dissipatif de la voie.

Le rail présente en effet une isolation imparfaite vis à vis du ballast et une impédance répartie de type selfique. Pour améliorer la transmission, une compensation de cette inductance linéique est réalisée par l'installation ponctuelle de condensateurs reliant les rails (figure 2.3). Ces *condensateurs de compensation* sont la seconde particularité des CdV-TVM.

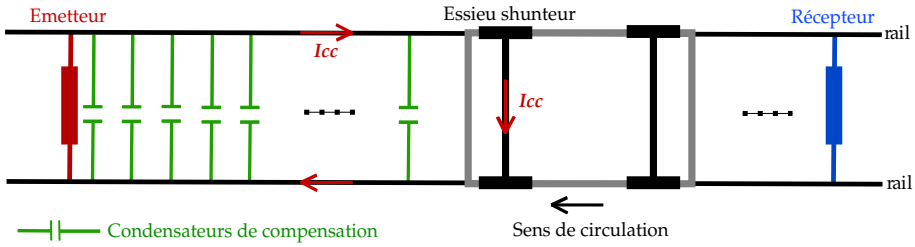


FIGURE 2.3 – *Schéma d'un circuit de voie compensé de type TVM.*

Les condensateurs de compensation utilisés sur les CdV sont des condensateurs au polypropylène, de capacité 22 μF($\pm$10 %). Leur diélectrique est électrolytique, ce qui assure un bon fonctionnement pour des fréquences inférieures au MHz, comme dans le cas des CdV et leur fiabilité est bonne.

Physiquement, ces composants sont placés entre les rails, sous la surface du ballast, selon des règles de pose qui fixent un intervalle constant de 60 m entre chaque condensateur pour les fréquences 1700 Hz et 2000 Hz, et 80 m pour les deux autres fréquences.

2.2.3 Les défauts possibles des CdV

Les CdV peuvent être sujet à différents types de défauts affectant la TVM, et donc la sécurité des circulations. Nous ne présenterons 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 mais ne seront pas abordés.

Une première famille de défauts concerne les interférences entre différents CdV, ce phénomène est appelé *diaphonie*. Deux types de diaphonie sont possibles :

1. 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 ;
2. diaphonie longitudinale : due à un JES défectueux, 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.

Une autre catégorie de défauts, beaucoup plus fréquents, concerne les condensateurs de compensation. Une attention particulière leur est portée compte tenu de leur nombre important, environ 100000 pour l'ensemble du réseau français. 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, corrosion ...) ;
- une augmentation des pertes du condensateur lorsque le composant vieillit 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 du 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 grandement le trafic ferroviaire.

Étant 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 dans le cadre de ces travaux uniquement à la détection des défauts sur les *condensateurs de compensation*, car ceux-ci sont plus fréquents.

2.2.4 Inspection des CdV

A la SNCF, plusieurs mesures sont utilisées pour le diagnostic des CdV destinés à la 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, 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 2008, 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 même capteurs TVM (bobines) que ceux utilisés sur les TGV commerciaux et d'un second capteur spécifique qui se présente sous la forme d'une boucle inductive placée sous la voiture de mesure et conçue pour détection de présence des condensateurs de compensation. La voiture *IRIS* permet de relever deux mesures :

- une mesure de l'amplitude du courant efficace détecté par les bobines en fonction de la position du véhicule. Ce courant est appelé courant de court-circuit, noté *Icc* (figure 2.4). Il est relevé pour les 4 fréquences de fonctionnement des CdV ;
- une mesure d'impédance transversale, qui traduit l'accord entre le dispositif de mesure par boucle et les composants reliant les rails (comme par exemple les condensateurs de compensation) et présentant une impédance faible à 25 kHz ; cette mesure est notée *Zt* (figure 2.5).

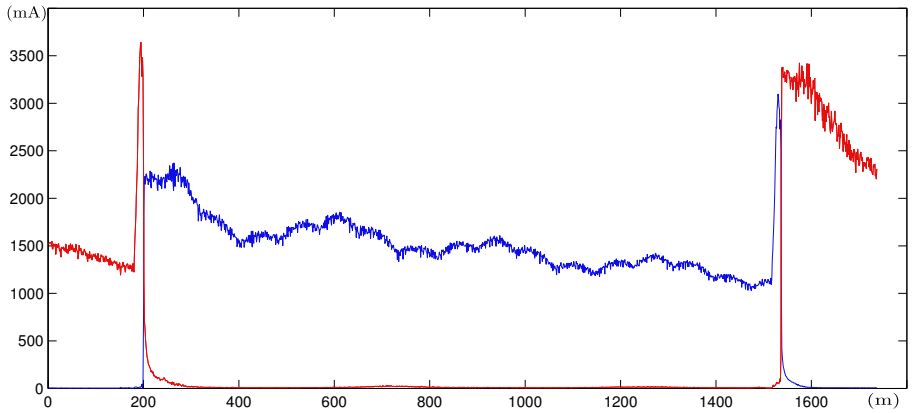


FIGURE 2.4 – *Exemple de signal Icc réel d'un circuit de voie TVM de fréquence* 2300 Hz *(en bleu), relevé à l'aide du véhicule IRIS. La fréquence* 1700 Hz *(en rouge) correspond aux CdV encadrants.*

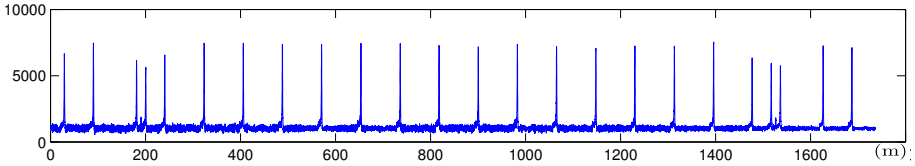


FIGURE 2.5 – *Signal Zt réel correspondant au même circuit de voie, relevé à l'aide du véhicule IRIS.*

C'est à partir de l'analyse de ces signaux qu'est réalisée la surveillance des condensateurs présents sur les CdV. L'observation de certaines variations sur ces mesures permet de signaler un comportement défectueux. Dans le cadre de cette thèse, nos travaux ont porté uniquement sur l'analyse des signaux *Icc* pour le diagnostic des condensateurs. Le signal *Zt* n'a pas été pris en compte pour l'accomplissement de cette tâche, hormis pour aider à la segmentation du signal *Icc*. En effet, le signal *Zt* a été conçu comme un détecteur binaire permettant uniquement

de détecter l'absence d'un condensateur mais n'apportant aucune information sur son état de dégradation. De plus par sa nature plus riche, le signal *Icc* permet d'envisager une maintenance préventive, grâce à l'extraction d'indicateurs de l'état de dégradation pour un condensateur donné.

2.2.5 Vers un diagnostic automatique des CdV

En complément de l'expertise existante sur l'analyse de ces signaux réalisée par le service de maintenance, les travaux présentés ici visent à automatiser leur dépouillement, au delà du repérage des situations de simple franchissement d'un seuil en amplitude. Le système CdV est un système complexe réparti constitué de plusieurs condensateurs et le diagnostic de ces sous-systèmes à partir d'un signal d'inspection prélevé sur le système global n'est pas chose aisée. En effet, ce signal ne signe pas le comportement des condensateurs individuellement mais mélange l'influence de l'état de chacun d'entre eux. Dans le cas d'un système présentant plusieurs composants défectueux, le nombre de modes de fonctionnement multi-défauts possibles devient extrêmement important et l'association d'une mesure à un état de fonctionnement pour chacun des composants est une tâche complexe. Toutefois, la forme même du signal *Icc* est porteuse d'un certain nombre d'informations (figure 2.6) qu'il est central d'intégrer dans un outil de diagnostic automatique. Les paragraphes suivants décrivent les liens entre certaines particularités physiques de l'application et l'allure du signal de mesure *Icc*.

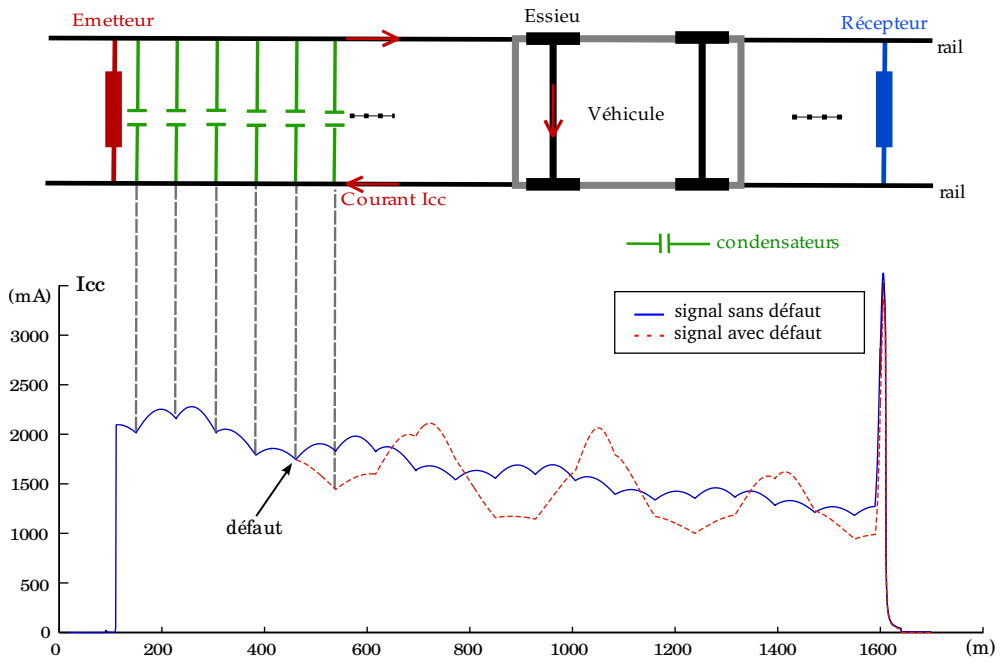


FIGURE 2.6 – *CdV et signal Icc.*

Structure du signal Icc et génération des descripteurs

Les signaux de mesure Icc sont fortement structurés. Comme le montre la figure 2.4, ceux-ci présentent une succession d'arches dont les jonctions correspondent aux emplacements des condensateurs (figure 2.6). Nous pouvons également observer sur ces signaux une décroissance exponentielle, un comportement oscillant de grande longueur d'onde ($\approx 400\,\mathrm{m}$) ainsi qu'un bruit de mesure. L'ensemble de ces caractéristiques s'expliquent aisément à l'aide d'une approche physique phénoménologique [Oukhellou *et al.* 2006].

Le signal Icc étant structuré en segments correspondant aux différentes zones de voie comprises entre deux condensateurs, nous avons choisi d'utiliser cette information pour paramétrer les signaux et réduire ainsi la dimension de l'espace où ils sont représentés. Des polynômes de 2^{nd} degré ont été utilisés pour approximer le signal sur chacune de ces portions, en introduisant des contraintes de continuité entre les différentes portions du signal. La paramétrisation des signaux de mesure s'est vue formulée comme un problème de régression polynomiale avec contraintes afin de coïncider avec le découpage naturel induit par la disposition des condensateurs à la voie. La figure 2.7 présente le résultat de cette paramétrisation sur un signal réel de fréquence 2300 Hz.

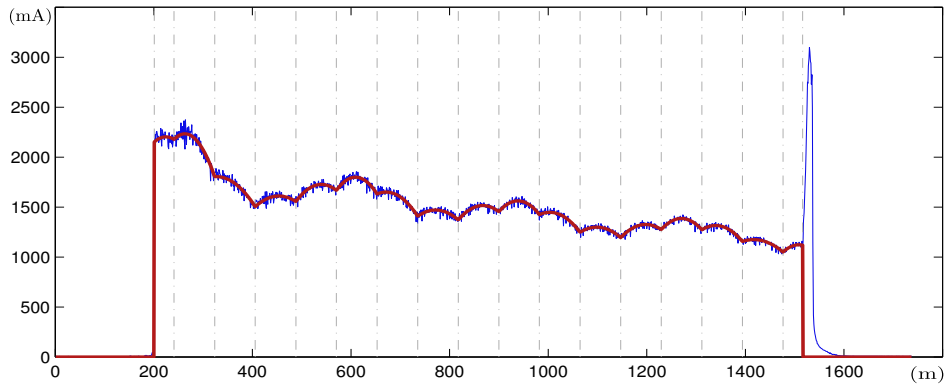


FIGURE 2.7 – *Paramétrisation d'un signal réel de fréquence* 2300 Hz*, relevé à l'aide du véhicule IRIS.*

Nous pouvons observer une bonne adéquation du modèle au signal sur la figure précédente (figure 2.7). L'ensemble des polynômes correspondants aux différentes portions de signal permettent de résumer l'information apportée par un signal d'inspection. Chacun des polynômes décrivant une arche dépend de 3 coefficients mais les contraintes de continuité entre polynômes éliminent l'un d'entre eux. A l'issue de cette phase de paramétrisation, on dispose ainsi d'un vecteur de $2 \times N_c + 1$ paramètres, où N_c est le nombre de condensateurs du CdV. Ces paramètres définissent l'espace de représentation des méthodes de diagnostic présentées dans cette thèse.

Dépendance amont-aval

Le CdV est composé de multiples éléments (émetteur, condensateurs, récepteur) organisés spatialement le long d'un axe (figure 2.3) . 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 en 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 situés en aval ($x > x_v$) ne peuvent donc pas influencer la mesure $\mathrm{Icc}(x_v)$, ce qui peut se résumer par la propriété suivante :

- Le signal Icc se décompose en autant d'arches qu'il y a de condensateurs de compensation et l'allure de chaque arche est influencée par l'état des condensateurs situées en amont.

Cette propriété peut facilement être illustrée grâce au modèle électrique du CdV développé par l'INRETS [Aknin *et al.* 2003, Oukhellou *et al.* 2006]. La figure 2.8 présente des signaux Icc simulés grâce à ce modèle électrique. Le même circuit de voie a été simulé : sans aucun défaut, avec un défaut sur le condensateur 5, et avec deux défauts respectivement sur les condensateurs 5 et 9.

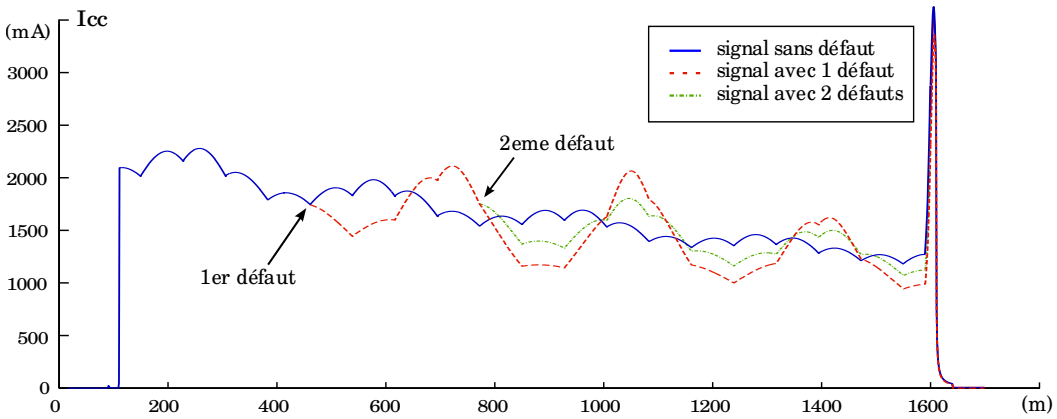


FIGURE 2.8 – *Allure d'un signal Icc sans défaut (bleu), avec un défaut (rouge), avec deux défauts (vert).*

Nous pouvons facilement observer à travers 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.

Dans le cadre du diagnostic d'un système complexe, les informations disponibles sur sa structure et sur son fonctionnement constituent des connaissances a priori qu'il est important d'exploiter pour une bonne modélisation de celui-ci. Dans le cas des CdV, ces spécificités peuvent être résumées de la manière suivante :

- le CdV peut être considéré comme un système global composé de plusieurs sous-systèmes (les condensateurs de compensation) ;
- les sous-systèmes sont organisés spatialement sur un axe orienté (de l'émetteur vers le récepteur) ;
- les signatures (arches) des sous-systèmes sont liées spatialement de façon unidirectionnelle : l'allure de la signature d'un sous-système dépend 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.

La problématique du diagnostic revient à déterminer l'état des sous-composants à partir de données de surveillance prélevées sur le système global. Différentes méthodes peuvent être envisagées pour résoudre la question. Celles-ci peuvent être regroupées en familles d'approches : approches par système expert [Zwingelstein 2002], approches à base de modèle physique [Isermann 1997] et approches à base de reconnaissance des formes [Dubuisson 2001]. La suite de ce chapitre passe en revue certaines de ces méthodes, le but de ce qui suit n'est pas de faire une présentation exhaustive des méthodes existantes mais de résumer les travaux menés précédemment en fonction des approches adoptées pour le diagnostic des CdV.

2.3 Diagnostic à base de modèles

Principalement issues des travaux menés par les automaticiens, les méthodes de diagnostic à base de modèles sont utilisées lorsque l'on dispose d'un modèle analytique du système lorsqu'il est en bon fonctionnement. Une phase de validation expérimentale du modèle en question est nécessaire avant de l'exploiter. On distingue deux grandes familles de modèles [Isermann 2006] ; celles qui utilisent des modèles de connaissance dites « boîte blanche » reposant entièrement sur les lois physiques régissant le système et celles dites de représentation ou « boîte noire » qui décrivent le comportement global du système à l'aide de relations entrées-sorties. Des modèles mixtes combinant ces deux types d'approches peuvent être employés dans le cas de systèmes complexes.

2.3.1 Principe

L'idée clé d'une procédure de diagnostic à base de modèles est la notion de redondance analytique qui compare, en temps réel, le comportement attendu du système prédit par un modèle à celui observé par des capteurs. Toutes différences entre les observations et ces prédictions sont interprétées comme la présence d'un ou de plusieurs défauts. Pour qu'une telle procédure de diagnostic puisse être adoptée, il est nécessaire de mettre au point un modèle du système et de le valider. Une fois le modèle validé, la procédure de diagnostic en elle-même se fait en plusieurs étapes (figure 2.9) :

- la génération de résidus [Weber 1999, Basseville 1988], cette étape est nécessaire en vue de déterminer des grandeurs sensibles aux défauts,
- l'analyse des résidus, ces derniers sont évalués dans une phase de décision afin de décider de la présence ou non de défauts,
- la localisation et l'identification du type de défaillance.

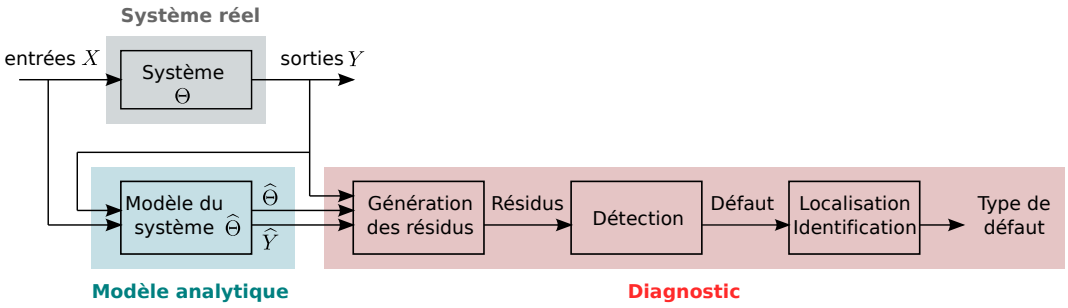


FIGURE 2.9 – *Principe du diagnostic à base de modèle*

Dans le cas où un défaut est détecté, il est ensuite localisé et identifié. Cette dernière étape requiert nécessairement une connaissance a priori sur les types de dysfonctionnements pouvant affecter le système. La suite donne plus de détails sur la réalisation de ces différentes étapes.

Génération des résidus

La difficulté principale de cette approche est la génération de résidus robustes. Cette robustesse est à juger au regard de la sensibilité aux perturbations, aux bruits et aux erreurs de modélisation qui ne doivent pas être confondues avec les défauts. Plusieurs méthodes de génération de résidus existent, parmi lesquelles on peut citer :

- Les approches par espace de parité [Chen & Patton 1999] dont le principe consiste à vérifier la consistance entre des relations issues des mesures ou celles entre les entrées du système et les mesures.
- Les approches par estimation d'état ou à base d'observateurs où le résidu peut être simplement constitué de l'erreur entre la mesure et son estimation ou d'une combinaison linéaire des composantes de cette erreur. Les premiers travaux ont été proposés par [Beard 1971, Frank 1990] dans un cas déterministe et par [Willsky 1976, Basseville 1988] à l'aide de filtres de Kalman dans un cas stochastique.
- Les approches par identification paramétrique qui utilisent l'hypothèse que les défauts modifient les paramètres du modèle du système [Isermann 1984, Isermann 1997]. Les résidus sont ici les paramètres du modèle qu'il faut identifier. Si on tient compte des propriétés statistiques des bruits sur le système et sur les mesures, on obtient, en plus des valeurs des paramètres, une estimation de la précision avec laquelle ils sont identifiés. Pour cela, on peut

utiliser des techniques statistiques, comme les estimateurs des moindres carrés généralisés, le maximum de vraisemblance, ou encore l'estimateur de la variable instrumentale [Zwingelstein 2002]. Dans le cas où on ne dispose pas d'un modèle du bruit, il faut faire appel à des méthodes déterministes, comme la méthode du modèle de référence. Les différents paramètres sont alors estimés grâce à la minimisation d'une fonction de coût. Cette fonction prend en compte l'écart entre les caractéristiques déduites du modèle (pour un jeu de paramètres donné), et ces mêmes caractéristiques mesurées sur le système.

Détection et localisation de défaut

Une fois les résidus générés, la deuxième étape consiste à les exploiter dans le cadre d'un système de diagnostic pour décider de la présence ou non d'un défaut dans le système. Les méthodes de détection sont regroupées sous le terme de test de cohérence [Adrot 2000]. Leur objectif est de vérifier l'adéquation entre les grandeurs observées, et les sorties issus du modèle, décrivant le comportement attendu du système. Ces tests de cohérence peuvent varier d'un simple procédé qui consiste à comparer les résidus à des seuils fixes à des tests statistiques plus élaborés [Basseville & Benveniste 1986, Basseville & Nikiforov 1993, Zhang *et al.* 1994]. La localisation du défaut peut être réalisée en utilisant des bancs d'observateurs pour lesquels chaque résidu peut être sensible à un défaut particulier (bancs DOS pour Dedicated Observer Scheme) ou inversement les résidus sont sensibles à tous les défauts sauf un en particulier (GOS pour Generalized Observer Scheme) [Alcorta Garcia & Frank 1996, Frank 1990].

2.3.2 Approche à base de modèles pour le diagnostic des CdV

L'approche de diagnostic des CdV à base de modèles a essentiellement été développée au cours des travaux de thèse de A. Debiolles [Debiolles 2007]. La méthode développée a montré des résultats prometteurs en matière de détection de défaut sur signaux simulés et sur signaux réels. Toutefois, les parties localisation et quantification de défaut se sont heurtées à une difficulté majeure. En effet, cette dernière étape s'est montrée particulièrement sensible à la bonne estimation de certains paramètres physiques du CdV liés à la voie et utilisés dans le modèle mathématique permettant de simuler les courants *Icc*. D'autres techniques de diagnostic, ne nécessitant pas de modèle physique du système, ont donc été envisagées. Les sections suivantes présentent ce type d'approche.

2.4 Diagnostic à base de reconnaissance des formes

L'avantage des méthodes de diagnostic par *reconnaissance des formes* est qu'elles ne présupposent pas l'existence d'un modèle physique du système à analyser, mais uniquement la présence d'observations des différents modes de fonctionnement regroupées dans une base de données. Cette base qualifiée de *base d'apprentissage*

est constituée d'observations décrites par un ensemble de variables explicatives (extraites des signaux de mesure et constituant le vecteur forme), et éventuellement complétée des modes de fonctionnement associés à chacune des observations. Une telle approche vise ainsi à apprendre, à partir d'un ensemble d'apprentissage, une règle de décision permettant d'affecter à toute nouvelle observation une des classes de fonctionnement [Bishop 2006, Dubuisson 1990].

2.4.1 Principe

La construction d'un système de diagnostic par reconnaissance de formes peut être décomposé en plusieurs étapes (figure 2.10) où la qualité de réalisation de chaque étape du processus dépend souvent de celle des étapes précédentes [Duda *et al.* 2000]. Pour une meilleure compréhension, la suite résume ce processus selon deux phases principales :

- Acquisition des données et sélection de variables.
- Classification/régression et évaluation des performances.

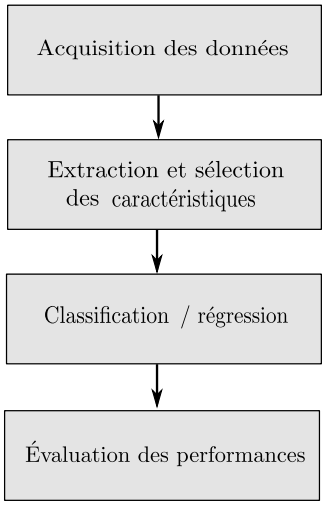


FIGURE 2.10 – *Principe du diagnostic à base de reconnaissance de formes*

Acquisition des données et sélection de variables

Cette première phase a pour but de déterminer l'espace de représentation final dans lequel le modèle de diagnostic opérera en définissant et sélectionnant les caractéristiques qui décrivent au mieux l'état du système à diagnostiquer. Pour effectuer cette tâche correctement il est souvent nécessaire de s'appuyer sur des connaissances expertes du système afin d'identifier quels indicateurs peuvent être pertinents pour discriminer un système en bon état de fonctionnement d'un système défectueux. Les données en question sont des vecteurs réalisations de variables qui correspondent à

des mesures réalisées sur un système physique ou à des informations collectées lors de l'observation d'un phénomène. Elles sont rarement utilisables directement et bien souvent il est nécessaire de leur appliquer certains prétraitements ou d'en extraire une paramétrisation. La nature des traitements pouvant être utilisés sur les données est étroitement liée au contexte pratique de l'application et il est difficile d'en faire une énumération exhaustive. On peut toutefois citer les prétraitements suivants [Theodoridis & Koutroumbas 2006] :

- La segmentation et la localisation en vue de détecter sur des données acquises de façon continue, une séquence discrète de segments qui correspond à des réalisations du phénomène à analyser.
- Le débruitage dont le but est de limiter l'influence du bruit pouvant être présent sur les signaux ou les images. Celui-ci entraîne des variations dans la représentation d'un même objet et son influence varie selon le contexte.
- La normalisation des données pour que celles-ci varient dans une même plage. Ce type de traitement améliore souvent la stabilité numérique des étapes de traitement ultérieures.
- L'invariance des descripteurs à un certain nombre de transformations qui ne doivent pas affecter le résultat du diagnostic. On peut par exemple éliminer ainsi des paramètres de nuisances relatifs aux conditions de mesure qui rajouteraient un biais.

A l'issue de l'étape d'acquisition intervient la sélection de variables. Celle-ci consiste à traiter les mesures recueillies pour extraire une représentation stable et concise caractéristique des fonctionnements normaux ou anormaux du système. Cette étape revient généralement à réduire la dimension de l'espace de représentation notamment afin d'éviter la redondance d'information dans les variables utilisées et de limiter la dégradation des performances à cause d'une dimension trop élevée de l'espace de représentation [Bellman 1957, Blum & Langley 1997].

Les techniques de réduction de dimension sont traditionnellement scindées en deux catégories selon qu'elles cherchent à sélectionner, parmi les variables initiales, un sous-ensemble de variables plus pertinentes ou plus informative (*feature selection*), ou qu'elles utilisent toute l'information des variables initiales en les projetant le mieux possible dans un espace de dimension réduite (*feature extraction*) [Theodoridis & Koutroumbas 2006].

Dans la première catégorie se rangent les méthodes de recherche optimale exhaustive lorsque la dimension initiale n'est pas trop grande, les méthodes séquentielles de type branch and bound [Duda *et al.* 2000]. Dans la seconde catégorie, on retiendra les approches statistiques exploratoires comme l'Analyse en Composantes Principales (ACP) [Jackson 1991, Jolliffe 2002], la Décomposition en Valeurs Singulières (SVD) [Golub & Kahan 1965], l'Analyse en Composantes Indépendantes (ACI) [Hyvärinen *et al.* 2001], ou encore les méthodes prenant en compte l'objectif de modélisation à réaliser comme l'analyse factorielle discriminante [Fukunaga 1990],

la régression aux moindres carrés partiels [Tenenhaus 1998]. On retiendra également les méthodes non linéaires comme l'analyse en composantes principales à noyau (KPCA) et l'analyse en composante curviligne (CCA) [Demartines & Hérault 1997].

Classification/régression et évaluation des performances

Le but de cette phase est de construire des frontières de décision entre les classes afin de partitionner l'espace de représentation en autant de régions que de modes de fonctionnement. La méthode utilisée permettra de définir une règle de décision, qui sera utilisée par la suite pour classer automatiquement toute nouvelle observation.

L'évaluation des performances d'un modèle de classification se quantifie au travers de sa capacité de généralisation [Hastie *et al.* 2006], laquelle est mesurée à l'aide d'une estimation du taux d'erreur de prédiction sur une base de données, appelée *base de test*, n'ayant pas servi à l'apprentissage. On cherche alors à déterminer un modèle dont la complexité est adaptée au problème à traiter en tentant de satisfaire le compromis biais-variance : un modèle trop complexe risque d'apprendre par cœur l'ensemble d'apprentissage (*surapprentissage*), mais généralisera très mal sur de nouvelles données. En revanche un modèle trop simple, obtiendra peut-être de moins bons résultats sur l'ensemble d'apprentissage, mais de part sa simplicité sera moins affecté par les variations des données d'apprentissage. Celui-ci généralisera mieux sur de nouvelles données mais risque de mal modéliser le problème. La complexité optimale est celle qui conduit à une erreur de généralisation minimale. Chaque méthode de classification dispose d'un ou plusieurs paramètres appelés *hyperparamètres* et permettant de contrôler le compromis biais-variance.

Pour tenter de résoudre ce dilemme, deux volets complémentaires doivent être considérés : l'un concerne la *sélection de modèle* où il s'agit de comparer plusieurs modèles et d'en sélectionner le meilleur au sens de sa capacité de prédiction, l'autre volet concerne l'*évaluation des performances* où il s'agit d'estimer le pouvoir prédictif du modèle choisi sur de nouvelles données. Idéalement, la base de données est à diviser en trois parties : une base d'apprentissage pour l'ajustement des différents modèles, lesquels sont ensuite comparés en terme d'erreur de prédiction sur la base de validation et enfin la base de test qui sert à estimer les capacités de généralisation du modèle ainsi choisi. Lorsqu'on ne dispose pas de bases de données suffisamment riches pour être scindées de la sorte, on peut se tourner vers les techniques classiques de ré-échantillonnage telles que la validation croisée qui réalise un découpage du jeu de données en blocs ou le bootstrap qui effectue un tirage aléatoire avec remise des échantillons [Efron & Tibshirani 1994]. Bien que largement employées, ces solutions peuvent être très coûteuses en temps de calcul vu le nombre important d'apprentissage qu'elles impliquent et ce, particulièrement lorsque plusieurs hyper-paramètres du modèle doivent être ajustés. Une première validation croisée est nécessaire pour ajuster les paramètres du modèle qui une fois fixés, seront utilisés dans une deuxième étape de validation croisée pour estimer les performances du modèle choisi. Des solutions basées sur des critères faisant intervenir un terme supplémentaire de pénal-

isation dépendant de la complexité du modèle ont été proposées. On peut citer les critères d'Akaike AIC [Akaike 1973], d'information bayésienne BIC [Schwarz 1978] ou encore la méthode de la longueur de description minimale MDL [Rissanen 1978]. Sur le même principe, on retrouve d'autres solutions qui introduisent des termes de régularisation tel l'algorithme LARS qui utilise le principe des chemins de régularisation [Efron *et al.* 2004].

2.4.2 Apprentissages supervisé et non supervisé

Dans la littérature deux contextes se distinguent principalement en apprentissage automatique :

- **L'apprentissage supervisé** se focalise sur la recherche d'une fonction permettant de prédire la valeur prise par une variable d'intérêt à partir de l'observation d'un autre ensemble de variables appelées variables explicatives, la prédiction devant bien sûr être la meilleure possible [Hastie *et al.* 2006]. Le problème de diagnostic par reconnaissance des formes peut alors être posé formellement de la façon suivante. Nous supposons disposer de variables explicatives regroupées dans un vecteur forme $X \in \mathcal{X}$ à valeur dans R^P qui constituent les entrées du système de diagnostic, et d'une variable d'intérêt $Y \in \mathcal{Y}$ pouvant être continue dans le cas d'un problème de régression ou catégorielle, $\mathcal{Y} = \{c_1, \ldots, c_K\}$ dans le cas d'un problème de classification. Considérons un ensemble de réalisations (dit d'apprentissage) indépendantes et identiquement distribuées (*i.i.d.*) du couple $(X, Y) : \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, il s'agit d'estimer une fonction de prédiction $f(\mathbf{x}) : \mathcal{X} \to \mathcal{Y}$ reliant les entrées et la sortie et minimisant une fonction de coût $R(f(\mathbf{x}), y) : \mathcal{Y} \times \mathcal{Y} \to \mathrm{R}$ qui traduit les coûts associés à chaque type d'erreur de prédiction. L'évaluation de la fonction ainsi déterminée s'effectue sur de nouvelles entrées, dites de test, n'ayant pas servi à l'apprentissage.
- **L'apprentissage non supervisé** vise à trouver une structure cohérente au sein d'un ensemble de données afin d'en faciliter l'interprétation, l'analyse et la représentation [Ghahramani 2004, Hastie *et al.* 2006]. Dans ce cas l'ensemble de données est supposé ne contenir que les variables explicatives $X : \{(\mathbf{x}_1, \ldots, \mathbf{x}_N)\}$. La variable relative à la sortie est qualifiée de *donnée manquante* en opposition à l'appellation *donnée complète* utilisée en apprentissage supervisé. Dans ce contexte, on trouve les méthodes statistiques exploratoires qui cherchent à décrire les données dans un espace de faible dimension (souvent en deux dimensions) pouvant amener à une représentation graphique simplifiée des données. D'autres approches à vocation décisionnelle ont comme objectif d'organiser les données en classes homogènes. Pour cela, elles s'appuient soit sur des modèles probabilistes, soit sur des approches géométriques.

D'autres problèmes d'apprentissage à mi-chemin entre ces deux extrêmes ont motivé la communauté scientifique. On peut citer : l'apprentissage semi-supervisé où l'on dispose à la fois d'exemples parfaitement labellisés et d'exem-

ples dont l'étiquette est inconnue [Chapelle *et al.* 2006], l'apprentissage partiellement supervisé où chaque individu appartient à un ensemble de classes possibles [Ambroise & Govaert 2000, Hüllermeier & Beringer 2005], ou encore l'apprentissage sur labels imprécis qui manipule des grandeurs dites *fonctions de masse* pour quantifier les appartenances aux classes [Côme *et al.* 2009, Denoeux 2010].

2.4.3 Méthodes discriminatives versus génératives

Une autre manière d'organiser le domaine de l'apprentissage statistique est de distinguer les méthodes discriminatives et les méthodes génératives. Dans une approche discriminative, on cherche à modéliser la loi de probabilité conditionnelle d'appartenance aux classes sachant les observations $p(y|\mathbf{x})$. Une approche générative quant à elle cherche à modéliser la distribution jointe du couple entrée-sortie $p(y, \mathbf{x})$ et en déduit ensuite les distributions conditionnelles des sorties y par application de la loi de Bayes $p(y|\mathbf{x}) = p(\mathbf{x}, y)/p(\mathbf{x})$ où $p(\mathbf{x})$ est la densité marginale sur les observations. Alors que les approches discriminatives cherchent à positionner directement les frontières entre les classes et possèdent un lien étroit avec l'apprentissage statistique supervisé, les modèles génératifs, quant à eux, tentent de décrire la distribution statistique des données à l'intérieur des classes et sont souvent utilisés dans un contexte d'apprentissage non supervisé. Ceci leur confère une capacité naturelle à prendre en compte une information incomplète grâce en particulier aux modèles à variables latentes.

La critique faite parfois aux méthodes génératives est qu'en estimant d'abord les probabilités jointes, elles tentent de résoudre un problème plus complexe que celui qui est posé. Modéliser la génération des données est en effet une étape intermédiaire à la tâche finale de classification ou de régression. Cette complexité est à opposer à l'efficacité des méthodes discriminatives qui, s'attachent à résoudre directement le problème en déterminant les probabilités a posterori. A l'inverse, les approches génératives, de par leur principe, posent le problème de génération de données et paraissent alors plus flexibles et adaptées pour incorporer des connaissances a priori sur les données ou sur le problème posé. En estimant les densités marginales des données, des méthodes dédiées à la détection de données aberrantes ou de nouveautés ont par exemple été proposées dans [Bishop & Lasserre 2007, Tarassenko 1995]. Si les méthodes discriminatives peuvent être vues comme des modèles pouvant certes être très performants mais où les relations entre variables ne sont pas explicites, de récents travaux tentent d'inverser cette tendance et proposent d'inclure des informations a priori dans les modèles discriminatifs [Jin & Liu 2005].

Le dernier point de comparaison est lié à la taille de la base d'apprentissage. Dans le cas d'un modèle simple (classifieur de Bayes naïf), il a été montré dans [Bouchard 2005] qu'à taille d'échantillons donnée, l'estimateur génératif donnait de meilleurs résultats que l'estimateur discriminatif. Par contre, si les données sont de dimension importante, les méthodes génératives nécessitent de disposer d'une large base de données afin d'estimer les distributions conditionnelles

avec une précision raisonnable et les méthodes discriminatives peuvent s'avérer supérieures. Ces deux *écoles* d'apprentissage ont donné lieu à des travaux bien distincts, mais on peut noter certains travaux de recherche qui cherchent à les combiner [Jaakkola & Haussler 1998, Jebara 2004, Lasserre *et al.* 2006].

Parmi les méthodes de classification ou de régression discriminatives les plus populaires, on trouve la régression logistique [Hastie *et al.* 2006, Hosmer 1989], les réseaux de neurones [Rosenbalt 1958], les machines à vecteurs support [Vapnik 1999], les arbres de décision [Breiman *et al.* 1984]. Dans la catégorie des modèles probablistes génératifs, on peut citer les modèles de Markov cachés [Rabiner & Juang 1993], les modèles de mélange [McLachlan & Peel 2000], les réseaux bayésiens [Jensen 2001].

2.4.4 Approche discriminative pour le diagnostic des CdV

La méthode développée pour le diagnostic des CdV dans le cadre d'une approche discriminative avait pour principal but de s'affranchir du modèle physique du système [Debiolles 2007].

Comme cela a été vu précédemment, le CdV peut être considéré comme un système global composé de plusieurs sous-systèmes répartis spatialement. L'inspection d'un tel système peut fournir un ensemble de signatures élémentaires spatialement liées, *i.e.* une signature élémentaire dépend non seulement de l'état d'un sous-système particulier mais également de l'état des sous systèmes situés en amont. Ainsi, la méthode de diagnostic devait notamment permettre de lever les verrous suivants :

- Détecter le sous-système défectueux en tenant compte de la dépendance spatiale entre sous-systèmes. La décision quant à la présence de défaut ou non dans un sous-système donné, doit se faire dans un contexte global et non local qui risque de conduire à des décisions erronées.
- Le nombre de sous systèmes est variable, rendant indispensable une approche générique de diagnostic.

La méthodologie développée [Oukhellou *et al.* 2010] reposait sur la construction de plusieurs classifieurs locaux (un par sous-système) et sur la combinaison de leurs décisions respectives via la théorie des fonctions de croyances [Shafer 1976]. Le résultat de cette combinaison produisait la décision finale qui renseignait sur la présence et la localisation du défaut dans le système global. L'étape de fusion avait pour but de limiter l'impact d'éventuels conflits entre sous-classifieurs et de conduire ainsi à des décisions plus robustes. Cette approche permettait à la fois de détecter le sous système défectueux et de le localiser (figure 2.11). En revanche, elle ne pouvait détecter qu'un seul système défectueux et ne permettait pas d'estimer de la gravité du défaut.

Afin de combler ces lacunes, une méthode fusionnant les résultats des deux méthodes de diagnostic précédentes (diagnostic à base de modèle physique et diagnostic

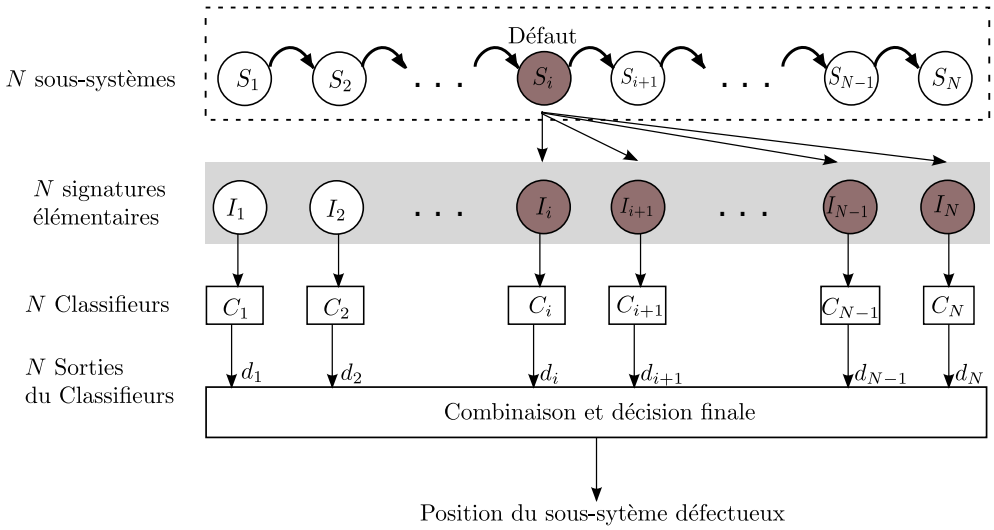


FIGURE 2.11 – *Architecture générale du système de diagnostic.*

par fusion de classifieurs) a été proposée [Debiolles 2007]. Cette combinaison a permis d'améliorer la qualité du diagnostic, principalement dans le cas multi-défauts. Toutefois, la méthode avait des difficultés à quantifier la valeur de l'ensemble des défauts et souffrait de la sensibilité de certains paramètres physiques difficiles à quantifier.

2.4.5 Approche générative pour le diagnostic des CdV

Dans le cadre des CdV, la formulation des modèles génératifs constitue un avantage dans la mesure où ils permettent d'incorporer plus facilement des connaissances a priori sur la structuration des données. Ces derniers peuvent ainsi être utilisés pour tirer partie des dépendances et des indépendances pouvant être spécifiées entre les variables d'intérêts pour le diagnostic et les observations décrivant les différentes portions du signal *Icc*. En effet, les connaissances physiques disponibles sur le système permettent par exemple d'affirmer que la défaillance d'un condensateur entraîne une variation du signal en aval de celui-ci mais pas en amont (figure 2.8).

En particulier les travaux menés par E. Côme [Côme 2009] se sont particulièrement intéressés à l'introduction, durant la phase d'apprentissage d'un modèle génératif, d'informations a priori sur le mécanisme de génération de données, ainsi que sur les labels des observations.

Concernant les labels, la problématique de diagnostic a été abordée dans le cadre semi-supervisé et partiellement supervisé où l'on suppose que les labels des différentes observations sont connus, mais entachés d'imprécision et d'incertitude. Les points suivants justifient un tel intérêt :

- Concernant les a priori sur la structure des données, il est possible de les

incorporer dans les modèles paramétriques génératifs lors de la définition de la famille de densités de probabilités. Le système CdV est lui même structurés selon son déroulé spatial. La prise en compte d'une telle information peut conduire à des modèles très pertinents comme peuvent l'être les modèles graphiques [Rabiner & Juang 1993].

- Concernant les a priori sur la labellisation, il faut garder à l'esprit que même si l'on dispose de grandes quantités de données, celles-ci ne sont généralement pas simples à labelliser car cela nécessite un travail coûteux et fastidieux. En revanche, dans les domaines industriels il existe souvent une riche expertise que l'on peut mettre à profit durant la labellisation. Dans ce cas, il est intéressant d'intégrer dans l'apprentissage des aspects liés aux réalités des données récoltées dans des situations réelles (difficulté et coût de labellisation par des experts, risque d'erreur, ... etc) et de proposer des algorithmes d'apprentissage prenant en compte une information même partielle sur les étiquettes des données.

Une solution a ainsi été proposée pour la formalisation et la résolution de ce problème à partir de données étiquetées de manière imparfaite [Côme 2009]. Dans un tel contexte, la problématique de diagnostic porte sur l'estimation de variables latentes liées aux défauts à partir de variables observées extraites des signaux d'inspection. L'objectif étant de proposer un cadre d'apprentissage de modèles génératifs à variables latentes lorsque l'information sur les classes d'origine des individus servant à l'apprentissage est partielle. 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 données. Plus précisément, il a été montré au travers de jeux de données jouets comment des labels prenant la forme de fonctions de masse de croyance [Shafer 1976], pouvaient être utilisés pour estimer les paramètres d'un modèle de mélange grâce à un critère utilisable au-delà du cadre probabiliste.

2.5 Positionnement des travaux

Dans le cadre de la reconnaissance des formes, les approches discriminatives et génératives diffèrent dans leur façon directe et indirecte de résoudre le problème de diagnostic. 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. Conjointement, lorsque la complexité des systèmes à diagnostiquer est importante, la prise en compte d'informations supplémentaires sur la structure des données peut s'avérer indispensable et une modélisation générative permet une représentation naturelle du système. Par ailleurs, les approches présentées précédemment ont encouragé une démarche en ce sens.

Cette thèse adopte également une approche générative pour le diagnostic des CdV mais traite la problématique du diagnostic d'un point de vue différent. Une

première partie considère le problème dans un contexte non supervisé et cherche à le modéliser de façon à prendre en compte les particularités physiques de l'application d'une part, et d'autre part les éventuels changement au niveau de la politique de maintenance. Depuis quelques années la surveillance des systèmes tels que les CdV entraîne des inspections régulières où des données sont collectées sur les mêmes objets. Face à ces données « historisées », il semble pertinent de développer des outils dédiés au suivi temporel. Le diagnostic est alors envisagé de manière dynamique plutôt que statique. Là où la décision était posée à chaque instant en n'utilisant uniquement que les informations prélevées au même instant, l'idée est ici de mettre à profit l'aspect temporel des données stockées en prenant en compte en supplément plusieurs inspections passées pour inférer la classe de fonctionnement à l'instant courant.

Une seconde partie des travaux s'est focalisée sur l'utilisation de données réelles pour mettre au point un outil de diagnostic automatique des CdV. Dans le but d'identifier au mieux les différents types de défauts et devant l'impossibilité de disposer de bases de signaux réels parfaitement labellisées, nous avons jugé nécessaire de mettre à profit des connaissances d'experts pour étiqueter les données d'apprentissage. Bien souvent l'étiquetage des données est une tâche fastidieuse et coûteuse, et les étiquettes d'appartenance aux classes sont en conséquence entachées d'imprécision et d'incertitude. Toutefois, lorsque des avis précis sont difficiles à obtenir, l'utilisation d'étiquettes imprécises et incertaines peut être envisagée et offre ainsi un cadre partiellement supervisé pour l'apprentissage du modèle [Côme 2009]. Dans le cas présent, les labels ayant été fournis par plusieurs experts, nous avons exploré l'intérêt de fusionner les différents avis via la théorie des fonctions de croyance pour améliorer la qualité du diagnostic sur données réelles. L'idée est d'obtenir la labellisation la plus fiable possible grâce aux capacités qu'offre cette théorie en terme de représentation d'informations et de règles pour la combinaison [Shafer 1976][Dempster 1967].

2.6 Conclusion

A travers ce chapitre, nous avons présenté 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, le circuit de voie (CdV). Après avoir décrit l'application, ses enjeux et ses particularités, nous avons fait une présentation de chacune des stratégies précédemment adoptées pour le diagnostic des CdV. Dans toutes ces approches l'opération de diagnostic est bâtie autour de l'analyse d'un signal enregistré grâce à un véhicule d'inspection spécifique.

La première méthode présentée est basée sur une approche à base de modèle où un modèle physique du CdV est supposée connue [Debiolles 2007]. Cette première approche a montré des résultats encourageants par rapport à la détection de défauts mais s'est montrée moins performante pour la localisation de ces derniers et l'estimation de leur gravité. Pour ces raisons et dans le but de s'affranchir de la contrainte

liée au modèle physique, une autre famille de méthodes de diagnostic a été considérée : les méthodes à base de reconnaissance des formes (RdF). La seconde méthode présentée traite la problématique de diagnostic à base de RdF par une approche discriminative dans un cadre supervisé [Debiolles 2007]. Basée sur l'utilisation d'un classifieur pour chaque condensateur et sur l'idée de fusionner leurs décisions afin de détecter les défauts, cette stratégie ne permettait par de traiter le cas multi-défauts ce qui constituait une faiblesse. Par la suite, une approche générative a été envisagée pour le diagnostic [Côme 2009]. Le recours à des modèles génératifs a ainsi permis de proposer des solutions plus performantes en considérant la problématique dans un cadre partiellement supervisé où des connaissances imparfaites sont prises en compte dans le modèle pour en améliorer les performances.

Cette présentation a permis de mettre en lumière les avantages et les inconvénients des précédents travaux et de positionner notre démarche. Cette thèse traite d'une part la problématique du diagnostic dans un cadre non supervisé en considérant l'aspect dynamique des données à travers une approche générative en se basant sur des données « historisées ». D'autre part, ces travaux ont participé à l'élaboration d'un outil de diagnostic opérationnel pour des signaux d'inspection réels. Dans ce cadre, une approche partiellement supervisée a été adoptée. Les chapitres suivants sont consacrés à la description de l'ensemble de ces propositions.

Chapitre 3

Approche générative pour le diagnostic

Sommaire

3.1 Introduction

Au cours de cette thèse nous nous sommes intéressés aux modèles génératifs pour la mise en œuvre d'outils de diagnostic sur un système complexe. L'avantage d'une approche générative est de pouvoir formaliser la problématique du diagnostic au travers d'états cachés liés aux classes de fonctionnement recherchées, et représentant les grandeurs d'intérêt. On parle alors de modèles à variables latentes. Une telle formulation du modèle permet d'incorporer plus facilement des connaissances a priori sur la structuration des données et de tirer partie des dépendances et indépendances qui peuvent exister entre les variables d'intérêts et les observations faites sur le système.

Ce chapitre est consacré aux modèles à variables latentes et aux algorithmes qui permettent d'en estimer les paramètres aussi bien dans un cas discret que dans un cas continu. Après une présentation des modèles graphiques, utiles pour la suite du

chapitre, nous détaillerons différents modèles à variables latentes. Les modèles de mélanges et les modèles de Markov cachés qui intègrent une variable latente discrète seront abordés ainsi que l'analyse en composantes indépendantes pour les variables latentes continues. Les approches décrites dans ce chapitre sont bien connues dans la littérature et sont présentées dans le but d'introduire les outils qui seront utilisés dans la suite du document.

3.2 Modèles à variables latentes

Les modèles à variables latentes sont des modèles statistiques qui contiennent des variables aléatoires dont les valeurs des réalisations ne peuvent être observées. Les propriétés des ces variables latentes doivent être déduites en utilisant un modèle statistique qui les relie à des variables observées.

Les premiers modèles statistiques à variables latentes ont été introduits dans les sciences humaines dès le début du 20ème siècle avec l'analyse factorielle [Spearman 1904, Thurstone 1947]. Actuellement ils sont utilisés pour une multitude d'applications dans différents domaines : traitement de la parole et chaîne de Markov cachée [Rabiner & Juang 1993], traitement des images et champ de Markov caché [Besag 1974], psychométrie et analyse factorielle [Spearman 1904], ... etc. En apprentissage statistique et principalement en contexte non supervisé, ces derniers constituent une solution élégante au problème véhiculé par ce cadre là, à savoir trouver des structures pertinentes au sein des données [Ghahramani 2004].

Cette section sera consacrée aux modèles faisant intervenir une unique variable latente discrète, tels que les modèles de mélanges et les modèles de Markov cachés, ainsi qu'aux modèles intégrant des variables latentes continues, qui nous ont été utiles pour le diagnostic du système. Dans ce qui suit nous allons présenter différents modèles à variables latentes et 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.

3.2.1 Modèles graphiques

Depuis les premiers travaux datant des années 1980 avec l'introduction des réseaux bayésiens [Pearl 1988], les modèles graphiques ont bénéficié d'un intérêt croissant dans différents domaines et notamment en reconnaissance des formes. Issus de travaux communs entre la théorie des probabilités et la théorie des graphes, ils constituent un outil naturel pour modéliser l'incertitude et la complexité d'un problème [Jordan 2004].

Beaucoup de systèmes probabilistes classiques issus de domaines tels que les statistiques, la théorie de l'information, la reconnaissance des formes ou encore la mécanique statistique, sont en fait des cas particuliers du formalisme plus général que constituent les modèles graphiques. Dans ce domaine, on peut citer les modèles de mélanges, les modèles de Markov cachés, les filtres de Kalman ou encore l'analyse

factorielle [Jordan 1999]. Ainsi, les modèles graphiques sont un moyen efficace de voir tous ces systèmes comme des instances d'un formalisme commun sous-jacent. L'avantage principal étant que les techniques développées pour certains domaines peuvent alors être aisément transférées à un autre domaine et être exploitées plus facilement [Jordan 2006].

Plus formellement, 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. 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 sa factorisation.

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 $pa(x), X \in S$ 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 :

$$p(x_1, x_2, \ldots, x_L) = \prod_{j=1}^{L} p(x_j | pa(x_j)). \tag{3.1}$$

Les conventions utilisées dans cette thèse pour représenter les divers modèles graphiques que nous rencontrerons sont détaillées en figure 3.1. Nous avons utilisé des couleurs différentes pour représenter les variables observées (fond gris) et les variables non observées (fond blanc). La forme carrée sera associée à une variable discrète et une forme ronde à une variable continue. Enfin, nous ferons parfois apparaître explicitement les paramètres du modèle.

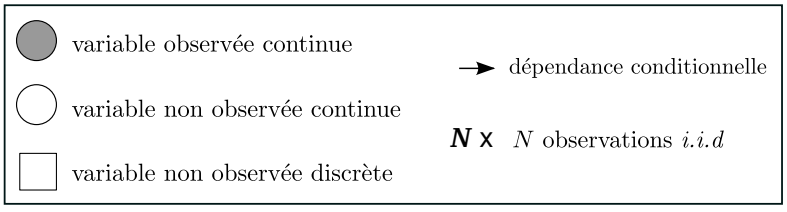


FIGURE 3.1 – *Conventions pour la représentation des modèles graphiques*

3.2.2 Modèles de mélanges

Les modèles de mélanges [McLachlan & Peel 2000] sont des modèles à variables latentes particulièrement adaptés à la modélisation de distributions de probabilités pour des populations hétérogènes. Ces derniers font l'hypothèse de l'existence de différentes sous-populations ou classes au sein des données, où les individus appartenant à une même classe constituent alors un échantillon de la même loi de probabilité.

Définition du modèle

Soit Y une variable latente prenant ses valeurs dans un ensemble discret à K composantes $\mathcal{Y} = \{1, \ldots, K\}$. La densité de la variable aléatoire X, où $\mathbf{x}$ est observée, est définie dans un cadre général comme étant :

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}|y = k), \tag{3.2}$$

où chaque élément $\pi_k = p(y = k)$ représente la probabilité qu'un point choisi aléatoirement appartienne à la classe k. Ces quantités sont non-négatives $\pi_k > 0, \forall k$ et vérifient $\sum_{k=1}^{K} \pi_k = 1$. Elles définissent ainsi les proportions de chacune des classes dans la population globale. La densité par rapport à chaque composante du mélange peut être écrite sous la forme d'un modèle paramétrique $f(\mathbf{x}|y = k) = f(\mathbf{x}; \boldsymbol{\theta}_k)$ tels que les $\boldsymbol{\theta}_k$ sont les paramètres des densités conditionnelles. Les différents paramètres intervenant dans la définition du modèle global se résument ainsi aux proportions des classes et aux paramètres des densités conditionnelles utilisées pour tirer les variables observées $\psi = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. La figure 3.2 présente le modèle graphique de génération des données pour un modèle de mélange avec les différents paramètres intervenant dans le modèle.

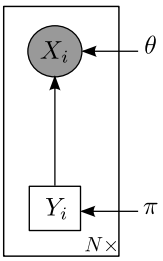


FIGURE 3.2 – *Modèle graphique de génération des données d'un modèle de mélange*

Les méthodes couramment utilisées pour l'estimation des paramètres dans les cas des modèles de mélanges sont la maximisation de la vraisemblance [McLachlan & Peel 2000] et les méthodes bayésiennes (Maximum A Posteriori (MAP)) où une distribution a priori est supposée pour les paramètres du modèle [Stephens 1997]. Dans le cadre de nos travaux, nous considérons uniquement

l'approche par maximum de vraisemblance. L'algorithme d'optimisation utilisé pour effectuer l'estimation du maximum de vraisemblance des paramètres est l'algorithme d'Espérance-Maximisation (EM) [Dempster *et al.* 1977] qui sera abordé dans la section suivante. Son objectif est de maximiser la vraisemblance, ou de façon équivalente, la log-vraisemblance des données observées en fonction des paramètres du modèle :

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

Suivant les lois conditionnelles utilisées pour tirer les observations, les modèles de mélanges peuvent être adaptés à différents problèmes. Le choix le plus classique pour la forme des lois conditionnelles quand les observations sont à valeur dans R^P, est sans aucun doute celui correspondant à la loi normale multivariée :

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\theta}_k) &= \boldsymbol{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \frac{1}{(2.\pi)^{\frac{P}{2}} |\det(\boldsymbol{\Sigma}_k)|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right), \end{aligned} \tag{3.4}$$

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. On parle alors de modèles de mélanges Gaussiens.

3.2.3 Algorithme EM

L'algorithme EM, [Baum *et al.* 1970, Dempster *et al.* 1977], est la solution classique aux problèmes de maximisation de la vraisemblance dans le cadre des modèles à variables latentes. L'idée de base consiste à maximiser la vraisemblance complète (données observées et données manquantes) plutôt que la vraisemblance avec données manquantes. Il met ainsi en œuvre les méthodes d'estimation généralement employées lorsque toutes les données sont observées, alors même qu'elles ne le sont pas. Pour palier ce problème de données manquantes, l'algorithme EM procède itérativement selon 2 étapes, les étapes E et M pour *Expectation* et *Maximisation* :

- **Etape E** : Cette étape construit une distribution de probabilité sur les valeurs pouvant être prises par les variables latentes $p(y|\mathbf{x}_i; \boldsymbol{\psi}^{(q)})$ pour chacun des individus en utilisant les données observées et les estimés courants des paramètres.
- **Etape M** : Cette étape estime les paramètres du modèle en utilisant les distributions de probabilité définies à l'étape précédente par maximisation de la vraisemblance.

Le point de départ de l'algorithme EM est la relation entre la vraisemblance complète et la vraisemblance avec données manquantes. La décomposition de la loi conditionnelle des variables latentes y connaissant les variables observées $\mathbf{x}$ et la

valeur courante des paramètres $\boldsymbol{\psi}$ sous la forme d'un rapport entre une loi jointe et une probabilité a priori conduit à la relation :

$$p(y|\mathbf{x};\boldsymbol{\psi}) = \frac{p(\mathbf{x},y;\boldsymbol{\psi})}{p(\mathbf{x};\boldsymbol{\psi})}, \tag{3.5}$$

que l'on peut également réécrire sous la forme :

$$p(\mathbf{x};\boldsymbol{\psi}) = \frac{p(\mathbf{x},y;\boldsymbol{\psi})}{p(y|\mathbf{x};\boldsymbol{\psi})}. \tag{3.6}$$

En passant au logarithme, nous obtenons l'expression suivante pour la log-vraisemblance marginale :

$$\log\left(p(\mathbf{x};\boldsymbol{\psi})\right) = \log\left(p(\mathbf{x},y;\boldsymbol{\psi})\right) - \log\left(p(y|\mathbf{x};\boldsymbol{\psi})\right). \tag{3.7}$$

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 :

$$\begin{aligned}\mathrm{E}[\log\left(p(\mathbf{x};\boldsymbol{\psi})\right)|X=\mathbf{x},\boldsymbol{\psi}=\boldsymbol{\psi}^{(q)}] = \\ \mathrm{E}[\log\left(p(\mathbf{x},Y;\boldsymbol{\psi})\right)|X=\mathbf{x},\boldsymbol{\psi}=\boldsymbol{\psi}^{(q)}] \\ - \mathrm{E}[\log\left(p(Y|\mathbf{x};\boldsymbol{\psi})\right)|X=\mathbf{x},\boldsymbol{\psi}=\boldsymbol{\psi}^{(q)}],\end{aligned} \tag{3.8}$$

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

$$\mathrm{E}[\log\left(p(\mathbf{x};\boldsymbol{\psi})\right)|X=\mathbf{x},\boldsymbol{\psi}=\boldsymbol{\psi}^{(q)}] = \log\left(p(\mathbf{x};\boldsymbol{\psi})\right) = \mathcal{L}(\boldsymbol{\psi};\mathbf{x}). \tag{3.9}$$

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

$$\mathcal{L}(\boldsymbol{\psi};\mathbf{x}) = Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(q)}) - H(\boldsymbol{\psi},\boldsymbol{\psi}^{(q)}), \tag{3.10}$$

tels que :

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

Le premier terme Q désigne la fonction auxiliaire et correspond à l'espérance conditionnelle de la log-vraisemblance des données complètes ou complétées. Cette vraisemblance, notée $\mathcal{L}_c$, est appelée vraisemblance complète. L'algorithme EM est basé sur la preuve que la maximisation de cette quantité suffit à faire croître la vraisemblance grâce à l'inégalité de Jensen [Cover & Thomas 1991, pages 25-30] et que la maximisation de celle-ci fait tendre la vraisemblance vers un maximum local [Dempster *et al.* 1977, Wu 1983, Xu & Jordan 1996].

Considérations pratiques

L'algorithme EM convergeant vers un maximum local de la vraisemblance, la valeur de l'estimateur à la convergence est dépendante de l'initialisation. En pratique, il est d'usage de lancer plusieurs fois l'algorithme avec des initialisations aléatoires différentes et de choisir ensuite celle qui conduit à la plus grande vraisemblance. D'autres stratégies d'initialisation plus élaborées ont été également proposées dans [Biernacki *et al.* 2003]. Aussi, un test de convergence est nécessaire pour l'arrêt de l'algorithme EM. Différentes solutions peuvent être mises en œuvre pour cela [McLachlan & Peel 2000, page 52] ; elles peuvent être basées sur l'évolution de la vraisemblance, l'évolution des probabilités a posteriori, ou bien encore celle des paramètres.

Plusieurs extensions de cet algorithme ont été proposées dans la littérature. Celles-ci peuvent chercher à accélérer la convergence ou répondre au problème des maxima locaux. On retrouve des versions classifiantes de l'algorithme CEM [Celeux & Govaert 1992], stochastiques SEM [Celeux & Diebolt 1988] ou encore des versions autorisant l'apprentissage « en ligne » où les données sont traitées au fur et à mesure de leur disponibilité [Titterington 1984, Wang & Zhao 2002, Cappé & Moulines 2007, Samé *et al.* 2007]. Pour un état de l'art complet des différentes extensions, on pourra consulter l'ouvrage de McLachalan et Krishnan [McLachlan & Krishnan 1996].

Algorithme EM pour les modèles de mélanges

On considère ici l'ensemble $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ comme étant l'ensemble de N données observées multidimensionnelles et $\mathbf{y} = (y_1, \ldots, y_N)$ comme l'ensemble des variables latentes associées prenant leurs valeurs dans un ensemble discret à K composantes $\mathcal{Y} = \{1, \ldots, K\}$. La log-vraisemblance de l'ensemble des paramètres du modèle $\boldsymbol{\psi}$ pour les données complétées est donnée par :

$$\mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{y}) = \log \prod_{i=1}^{N} p(\mathbf{x}_i, y_i; \boldsymbol{\psi}), \tag{3.11}$$

Sachant que $y_i \in \{1, \ldots, K\}$, cette expression peut être réécrite de la façon suivante :

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{y}) &= \sum_{i=1}^{N} \log \prod_{k=1}^{K} [p(y_i = k) p(\mathbf{x}_i | y_i = k; \boldsymbol{\theta}_k)]^{\mathbb{1}_{\{y_i = k\}}} \qquad (3.12) \\ \mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{y}) &= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}_{\{y_i = k\}} \log \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k), \end{aligned}$$

où $\mathbb{1}_{\{y_i = k\}}$ est la fonction indicatrice qui vaut 1 si $y_i = k$ ($\mathbf{x}_i$ est généré par la k ème composantes du mélange) et 0 sinon. Les deux étapes de l'algorithme EM sont alors réalisées de la façon suivante :

Etape E : cette étape consiste à calculer l'espérance conditionnelle de la log-vraisemblance complétée $\mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{y})$. L'espérance conditionnelle désigne la fonction Q et est donnée par :

$$\begin{aligned} Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \mathrm{E}\left[\mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{y}) | \mathbf{X}; \boldsymbol{\psi}^{(q)}\right] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathrm{E}\left[\mathbb{1}_{\{y_i=k\}} | \mathbf{x}_i; \boldsymbol{\psi}^{(q)}\right] \log \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} p(y_i = k | \mathbf{x}_i; \boldsymbol{\psi}^{(q)}) \log \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log f(\mathbf{x}_i; \boldsymbol{\theta}_k) \end{aligned} \tag{3.13}$$

D'après l'expression de Q et quelque soit la forme paramétrique des densités conditionnelles postulées, l'étape E exige donc simplement le calcul des probabilités a posteriori $p\left(y_i = k | \mathbf{x}_i; \boldsymbol{\psi}^{(q)}\right)$ notées t_{ik}. Celles-ci représentent la probabilité que l'observation i provienne de la classe k connaissant les variables observées et la valeur courante des paramètres $\boldsymbol{\psi}^{(q)}$, q étant l'itération courante. Cette dernière peut s'exprimer de la façon suivante :

$$t_{ik}^{(q)} = p(y_i = k | \mathbf{x}_i; \boldsymbol{\psi}^{(q)}) = \frac{\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{l=1}^{K} \pi_l^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_l^{(q)})}, \tag{3.14}$$

On notera que les variables $\mathbb{1}_{\{y_i=k\}}$ étant binaires, le calcul des espérances conditionnelles revient à calculer les probabilités conditionnelles pour ses mêmes variables $\mathrm{E}[\mathbb{1}_{\{y_i=k\}} | \mathbf{x}_i; \boldsymbol{\psi}^{(q)}] = p(y_i = k | \mathbf{x}_i; \boldsymbol{\psi}^{(q)})$.

Etape M : dans cette étape, il s'agit de maximiser la fonction auxiliaire Q (3.13) par rapport au vecteur de paramètres $\boldsymbol{\psi}$. Il est possible de décomposer cette fonction selon les paramètres à optimiser :

$$Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) = Q_{\boldsymbol{\pi}}(\boldsymbol{\pi}, \boldsymbol{\psi}^{(q)}) + \sum_{k=1}^{K} Q_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k, \boldsymbol{\psi}^{(q)}) \tag{3.15}$$

telle que :

$$Q_{\boldsymbol{\pi}}(\boldsymbol{\pi}, \boldsymbol{\psi}^{(q)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \log \pi_k \tag{3.16}$$

$$Q_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k, \boldsymbol{\psi}^{(q)}) = \sum_{i=1}^{N} t_{ik}^{(q)} \log p(\mathbf{x}_t | y_t = k; \boldsymbol{\theta}_k) \tag{3.17}$$

La maximisation de $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)})$ par rapport à $\boldsymbol{\psi}$ se fait en maximisant séparément $Q_{\boldsymbol{\pi}}(\boldsymbol{\pi}, \boldsymbol{\psi}^{(q)})$ et $Q_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k, \boldsymbol{\psi}^{(q)})$ pour $k = 1, \ldots, K$. La maximisation de $Q_{\boldsymbol{\pi}}(\boldsymbol{\pi}, \boldsymbol{\psi}^{(q)})$ par rapport à $\boldsymbol{\pi}$ telle que $\sum_k \pi_k = 1$ est un problème d'optimisation sous contrainte résolu en utilisant les multiplicateurs de Lagrange. Les proportions du mélange π_k sont calculées indépendamment des autres paramètres des densités conditionnelles postulées $\boldsymbol{\theta}_k$. Leur mise à jour s'effectue selon la relation (Annexe A.1) :

$$\pi_k^{(q+1)} = \frac{\sum_{i=1}^{N} t_{ik}^{(q)}}{N}. \tag{3.18}$$

Cette formule reste valable quelque soit la forme des densités conditionnelles. Les estimateurs des paramètres $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K$ sont obtenus en calculant la dérivée de Q par rapport à ces derniers et en annulant celle-ci :

$$\sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik}^{(q)} \frac{\partial \log(\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k))}{\partial \boldsymbol{\theta}} = 0 \tag{3.19}$$

Les formules de mise à jour diffèrent selon les lois conditionnelles et doivent être dérivées au cas par cas.

Algorithme EM pour les modèles de mélanges gaussiens

Dans le cas d'un mélange de gaussiennes (3.4), la log-vraisemblance des données complétées (3.20) prendra la forme suivante :

$$\mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}_{\{y_i = k\}} \log \pi_k \, \boldsymbol{\mathcal{N}}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.20}$$

Ainsi, étant donné le vecteur initial des paramètres $\psi = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ où $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, l'algorithme EM alterne entre les étapes E et M jusqu'à convergence vers un maximum local de la fonction de log-vraisemblance. Comme précédemment, l'étape E calcule les probabilités a posteriori (3.14), qui dans le cas d'un mélange de gaussiennes s'écriront :

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} \boldsymbol{\mathcal{N}}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l^{(q)} \boldsymbol{\mathcal{N}}(\mathbf{x}_i; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}, \tag{3.21}$$

L'étape M quant à elle permet de mettre à jour les proportions (3.18) et les paramètres des densités conditionnelles $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ qui optimisent la fonction auxiliaire Q (3.13). Dans le cas de mélanges gaussiens, la maximisation de la fonction Q par rapport aux paramètres $\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}_k$ mène aux formules de mise à jour suivantes [McLachlan & Peel 2000] :

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{(q)}} \sum_{i=1}^{N} t_{ik}^{(q)} \mathbf{x}_i \tag{3.22}$$

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{(q)}} \sum_{i=1}^{N} t_{ik}^{(q)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})'(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)}) \tag{3.23}$$

L'algorithme EM prend alors une forme simple lorsque des gaussiennes sont utilisées pour modéliser les différentes composantes (cf. algorithme 1)

Algorithme 1: pseudo-code de l'algorithme EM pour les modèles de mélanges gaussiens

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 $k \in \{1, \ldots, K_j\}$ **faire**

$t_{ik}^{(q)} = \frac{\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{l=1}^{K} \pi_l^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_l^{(q)})}$,

 # Etape M

 # maximisation de la fonction auxiliaire

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

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

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

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

$q = q + 1$

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

Dans la cas des modèles de mélanges, les observations sont considérées comme indépendantes et identiquement distribuées (i.i.d.). Lorsque les données représentent une séquence d'observations et sont donc ordonnées en fonction du temps il est possible de faire appel à des modèles qui permettent de relaxer cette hypothèse tels que les modèles de Markov cachés. La section suivante leur est consacrée.

3.2.4 Modèles de Markov cachés (HMM)

Les Modèles de Markov Cachés (Hidden Markov Model ou HMM) sont des modèles probabilistes décrivant un processus markovien et caché. Largement utilisés dans de nombreux domaines d'application, y compris la reconnaissance vocale, l'analyse d'image, la prédiction des séries temporelles, ... etc [Rabiner 1989]. Les HMM sont des modèles à variables latentes particulièrement adaptés à la modélisation de données séquentielles où les échantillons successifs ne peuvent être considérés comme indépendants. Un tel modèle peut donc être vu comme une généralisation du modèle de mélange en relâchant l'hypothèse d'indépendance et où la séquence des états cachés est modélisée par une chaîne de Markov. La figure 3.3 donne une représentation graphique d'un HMM.

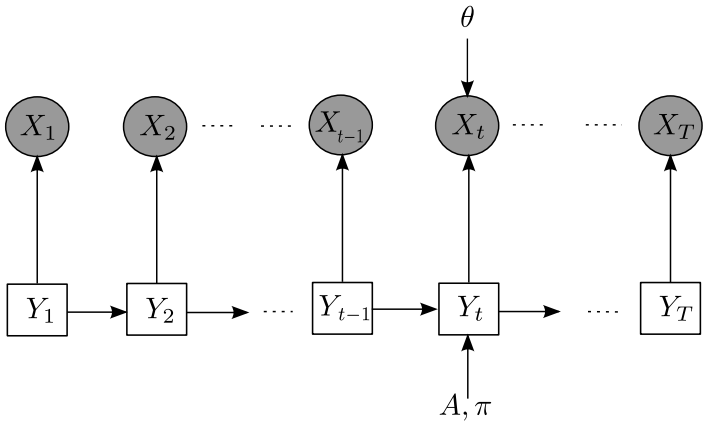


FIGURE 3.3 – *Modèle graphique de génération des données d'un HMM*

Notons $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ la séquence d'observation où les données multidimensionnelles $\mathbf{x}_t$ sont les données observées à l'instant t, et notons $\mathbf{y} = (y_1, \ldots, y_T)$ la séquence d'états cachés où y_t est une réalisation de la variable aléatoire discrète Y_t qui prend ses valeurs dans l'ensemble fini $\mathcal{Y} = \{1, \ldots, K\}$. Un HMM est entièrement défini par :

- la distribution initiale $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ où $\pi_k = p(y_1 = k)$ avec $k = 1, \ldots, K$;
- la matrice de transition probabiliste A où $A_{lk} = p(y_t = k | y_{t-1} = l)$ avec $t = 2, \ldots, T$ et $\sum_{k=1}^{K} A_{lk} = 1$ pour $l, k = 1, \ldots, K$ qui définie la probabilité de transition d'un état l à l'instant $t-1$ vers un état k à l'instant t ;
- l'ensemble des paramètres $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ des densités de probabilité conditionnelles des données observées $p(\mathbf{x}_t | y_t = k)$ pour $t = 1, \ldots, T$ et $k = 1, \ldots, K$ (ces probabilités sont également appelées les *probabilités d'émission).*

Compte tenu de la dépendance qui existe entre les états, la distribution d'une séquence d'états $\mathbf{y} = (y_1, \ldots, y_T)$ est donnée par :

$$p(\mathbf{y};\boldsymbol{\pi},A) = p(y_1;\boldsymbol{\pi})\prod_{t=2}^{T} p(y_t|y_{t-1};A). \tag{3.24}$$

Avec l'indépendance des observations conditionnellement à la séquence des états cachés, la distribution conditionnelle de la séquence observée est donnée par :

$$p(\mathbf{X}|\mathbf{y};\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{x}_t|y_t;\boldsymbol{\theta}). \tag{3.25}$$

On obtient alors la distribution conjointe suivante (également désignée par la vraisemblance sur les données complètes) :

$$\begin{aligned} p(\mathbf{X},\mathbf{y};\boldsymbol{\theta}) &= p(\mathbf{y};\boldsymbol{\pi},A)p(\mathbf{X}|\mathbf{y};\boldsymbol{\theta}) \\ &= p(y_1;\boldsymbol{\pi})p(\mathbf{x}_1|y_1;\boldsymbol{\theta})\prod_{t=2}^{T} p(y_t|y_{t-1};A)p(\mathbf{x}_t|y_t;\boldsymbol{\theta}). \end{aligned} \tag{3.26}$$

Les Modèles de Markov Cachés peuvent être classés en fonction des propriétés de leur chaîne de Markov cachée et en fonction du type de la distribution d'émission des états. Les HMM homogènes concernent les modèles pour lesquels la chaîne de Markov cachée est une matrice de transition fixe. Les HMM non-homogènes [Diebold *et al.* 1994, Hughes *et al.* 1999] concernent le cas où une dépendance temporelle sur les probabilités de transition est considérée. Parfois, dans certaines applications, on peut vouloir modéliser un phénomène dans lequel les états se succèdent de gauche à droite d'une manière successive selon leurs indices, par exemple dans les signaux de parole [Rabiner & Juang 1993]. Dans ce genre de cas il est possible d'imposer certaines restrictions sur le modèle en posant des contraintes au niveau de la matrice de transition [Rabiner & Juang 1993, Rabiner 1989]. Les HMM d'ordre supérieur désignent quant à eux les HMM où l'état actuel ne repose pas uniquement sur l'état précédent mais sur un historique fini des états précédents du HMM [Churchill 1989, Muri 1997].

HMM Gaussien

Dans un tel modèle, on considère des probabilités d'émission gaussienne. Ainsi, la distribution de chaque variable observée $\mathbf{x}_t$ conditionnellement aux états y_t est donnée par :

$$p(\mathbf{x}_t|y_t = k;\boldsymbol{\theta}_k) = \boldsymbol{\mathcal{N}}(\mathbf{x}_t;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \tag{3.27}$$

où $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$. Les paramètres d'un modèle de Markov caché peuvent être estimé avec une approche par maximum de vraisemblance à l'aide de l'algorithme EM.

Algorithme EM pour les modèles de Markov cachés

Soit $\boldsymbol{\psi} = (\boldsymbol{\pi}, A, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ le vecteur des paramètres à estimer dans le cas d'un HMM. L'estimation des paramètres se fait par la maximisation de la log-vraisemblance des données observées en fonction des paramètres :

$$\begin{aligned}\mathcal{L}(\boldsymbol{\psi};\mathbf{X}) &= \log p(\mathbf{X};\boldsymbol{\psi}) = \log \sum_{\mathbf{y}} p(\mathbf{X},\mathbf{y};\boldsymbol{\psi}) \\ &= \log \sum_{y_1,\ldots,y_T} p(y_1;\boldsymbol{\pi}) p(\mathbf{x}_1|y_1;\boldsymbol{\theta}) \prod_{t=2}^{T} p(y_t|y_{t-1};A) p(\mathbf{x}_t|y_t;\boldsymbol{\theta}). \quad (3.28)\end{aligned}$$

La log-vraisemblance étant difficile à maximiser directement, l'algorithme EM [Dempster *et al.* 1977], connu sous le nom de l'algorithme Baum-Welch dans le cas des HMM [Baum *et al.* 1970], est généralement utilisé pour la maximisation.

La vraisemblance des données complètes pour une configuration particulière de la séquence d'états $\mathbf{y}$ et des observations $\mathbf{X}$ (3.26) peut être réécrite :

$$\begin{aligned}p(\mathbf{X},\mathbf{y};\boldsymbol{\psi}) &= \prod_{k=1}^{K} p(y_1 = k;\pi)^{\mathbb{1}\{y_1=k\}} \\ &\quad \times \prod_{t=2}^{T}\prod_{k=1}^{K}\prod_{l=1}^{K} p(y_t = k|y_{t-1} = l;A)^{\mathbb{1}\{y_{t-1}=l,y_t=k\}} \\ &\quad \times \prod_{t=2}^{T}\prod_{k=1}^{K} p(\mathbf{x}_t|y_t = k;\boldsymbol{\theta}_k)^{\mathbb{1}\{y_t=k\}} \\ &= \prod_{k=1}^{K} \pi_k^{\mathbb{1}\{y_1=k\}} \prod_{t=2}^{T}\prod_{k=1}^{K}\prod_{l=1}^{K} A_{lk}^{\mathbb{1}\{y_{t-1}=l,y_t=k\}} \prod_{t=2}^{T}\prod_{k=1}^{K} p(\mathbf{x}_t|y_t = k;\boldsymbol{\theta}_k)^{\mathbb{1}\{y_t=k\}}, \quad (3.29)\end{aligned}$$

où $\mathbb{1}$ est la fonction indicatrice.

Le logarithme de l'équation précédente (3.29) permet d'obtenir la log-vraisemblance des données complètes ci-dessous :

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\psi};\mathbf{X},\mathbf{y}) &= \sum_{k=1}^{K} \mathbb{1}_{\{y_1=k\}} \log \pi_k \\ &\quad + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \mathbb{1}_{\{y_{t-1}=l,y_t=k\}} \log A_{lk} \\ &\quad + \sum_{t=2}^{T}\sum_{k=1}^{K} \mathbb{1}_{\{y_t=k\}} \log p(\mathbf{x}_t|y_t = k;\boldsymbol{\theta}_k). \quad (3.30)\end{aligned}$$

Etape E : L'étape consiste à calculer l'espérance conditionnelle de la log-vraisemblance complétée :

$$
\begin{aligned}
Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) &= \mathrm{E}\left[\mathcal{L}_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{y}) | \mathbf{X}; \boldsymbol{\psi}^{(q)}\right] \\
&= \sum_{k=1}^{K} \mathrm{E}\left[\mathbb{1}_{\{y_1=k\}} | \mathbf{X}; \boldsymbol{\psi}^{(q)}\right] \log \pi_k \\
&\quad + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{l=1}^{K} \mathrm{E}\left[\mathbb{1}_{\{y_{t-1}=l, y_t=k\}} | \mathbf{X}; \boldsymbol{\psi}^{(q)}\right] \log A_{lk} \\
&\quad + \sum_{t=2}^{T} \sum_{k=1}^{K} \mathrm{E}\left[\mathbb{1}_{\{y_t=k\}} | \mathbf{X}; \boldsymbol{\psi}^{(q)}\right] \log p(\mathbf{x}_t | y_t = k; \boldsymbol{\theta}_k) \\
&= \sum_{k=1}^{K} p(y_1 = k | \mathbf{X}; \boldsymbol{\psi}^{(q)}) \log \pi_k \\
&\quad + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{l=1}^{K} p(y_t = k | y_{t-1} = l | \mathbf{X}; \boldsymbol{\psi}^{(q)}) \log A_{lk} \\
&\quad + \sum_{t=2}^{T} \sum_{k=1}^{K} p(y_t = k | \mathbf{X}; \boldsymbol{\psi}^{(q)}) \log p(\mathbf{x}_t | y_t = k; \boldsymbol{\theta}_k) \\
&= \sum_{k=1}^{K} \gamma_{1k} \log \pi_k + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{l=1}^{K} \xi_{tlk} \log A_{lk} + \sum_{t=2}^{T} \sum_{k=1}^{K} \gamma_{tk} \log p(\mathbf{x}_t | y_t = k; \boldsymbol{\theta}_k),
\end{aligned}
\tag{3.31}
$$

où :

- $\gamma_{tk} = p(y_t = k | \mathbf{X}; \boldsymbol{\psi}^{(q)})$, pour $t = 1, \ldots, T$ et $k = 1, \ldots, K$, représente la probabilité a posteriori de l'état k à l'instant t sachant la séquence totale des observations et l'estimation courante des paramètres $\boldsymbol{\psi}^{(q)}$.
- $\xi_{tlk} = p(y_t = k | y_{t-1} = l | \mathbf{X}; \boldsymbol{\psi}^{(q)})$, pour $t = 2, \ldots, T$ et $l, k = 1, \ldots, K$, est la probabilité a posteriori jointe de l'état l à l'instant $t-1$ et de l'état k à l'instant t sachant la séquence totale des observations et l'estimation courante des paramètres $\boldsymbol{\psi}^{(q)}$.

Cette étape nécessite le calcul des probabilités a posteriori γ_{tk} et ξ_{tlk}. Ces probabilités sont calculées par l'intermédiaire de la procédure *forward-backward*. La procédure forward consiste à calculer de façon récursive les probabilités :

$$
\alpha_{tk} = p(\mathbf{x}_1, \ldots, \mathbf{x}_t, y_t = k; \boldsymbol{\psi}), \tag{3.32}
$$

où α_{tk} est la probabilité d'observer la séquence partielle $(\mathbf{x}_1, \ldots, \mathbf{x}_t)$ se terminant avec l'état k à l'instant t. La log-vraisemblance (3.28) peut alors être calculée d'après les probabilités forward : $\log p(\mathbf{X}; \boldsymbol{\psi}) = \log \sum_{k=1}^{K} \alpha_{Tk}$ [Rabiner & Juang 1993]. La procédure backward calcule les probabilités :

$$
\beta_{tk} = p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T, y_t = k; \boldsymbol{\psi}), \tag{3.33}
$$

où β_{tk} est la probabilité d'observer le reste de la séquence $(\mathbf{x}_{t+1}, ..., \mathbf{x}_T)$ sachant que l'on commence par l'état k à l'instant t. Les probabilités α_{tk} et β_{tk} sont calculées récursivement par l'algorithme comme suit (pour plus de détails voir annexe A.3) :

- $\alpha_{1k} = \pi_k p(\mathbf{x}_1|y_1 = k; \boldsymbol{\theta}_k)$ pour $t = 1$ et $k = 1, \ldots, K$,
- $\alpha_{tk} = \sum_{l=1}^{K} \alpha_{(t-1)l} A_{lk} p(\mathbf{x}_t|y_t = k; \boldsymbol{\theta}_k)$ pour $t = 2, \ldots, T$ et $k = 1, \ldots, K$,

et :

- $\beta_{Tk} = 1$ pour $t = T$ et $k = 1, \ldots, K$,
- $\beta_{tl} = \sum_{k=1}^{K} \beta_{(t+1)k} A_{lk} p(\mathbf{x}_{t+1}|y_{t+1} = k; \boldsymbol{\theta}_k)$ pour $t = T-1, \ldots, 1$ et $l = 1, \ldots, K$.

Dans la pratique, le calcul récursif des α_{tk} et des β_{tk} implique des multiplications répétées par des nombres faibles ce qui peut provoquer des problèmes de précision numérique. Pour éviter ce genre de situation, une normalisation est nécessaire lors des calculs, voir [Rabiner 1989, Rabiner & Juang 1993] pour plus de discussions. Les probabilités a posteriori sont ensuite exprimées en fonction des probabilités *forward* et *backward* comme suit :

$$\gamma_{tk}^{(q)} = \frac{\alpha_{tk}\beta_{tk}}{\sum_{l=1}^{K} \alpha_{tl}\beta_{tl}}, \tag{3.34}$$

et :

$$\xi_{tlk}^{(q)}(lk) = \frac{\alpha_{(t-1)l}^{(q)} A_{lk}^{(q)} p(\mathbf{x}_t|y_t = k; \boldsymbol{\psi}^{(q)}) \beta_{tk}^{(q)}}{\sum_{l=1}^{K}\sum_{k=1}^{K} \alpha_{(t-1)l}^{(q)} A_{lk}^{(q)} p(\mathbf{x}_t|y_t = k; \boldsymbol{\psi}^{(q)}) \beta_{tk}^{(q)}} . \tag{3.35}$$

Etape M : Lors de cette étape, la valeur des paramètres $\boldsymbol{\psi}$ est mise à jour par le calcul des paramètres $\boldsymbol{\psi}^{(q+1)}$ maximisant l'espérance de Q par rapport à $\boldsymbol{\psi}$. La fonction Q (3.30) est décomposée alors de la façon suivante :

$$Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) = Q_{\boldsymbol{\pi}}(\boldsymbol{\pi}, \boldsymbol{\psi}^{(q)}) + Q_A(A, \boldsymbol{\psi}^{(q)}) + \sum_{k=1}^{K} Q_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k, \boldsymbol{\psi}^{(q)}), \tag{3.36}$$

telle que :

$$Q_{\boldsymbol{\pi}}(\boldsymbol{\pi}, \boldsymbol{\psi}^{(q)}) = \sum_{k=1}^{K} \gamma_{1k}^{(q)} \log \pi_k \tag{3.37}$$

$$Q_A(A, \boldsymbol{\psi}^{(q)}) = \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \xi_{tlk}^{(q)} \log A_{lk} \tag{3.38}$$

$$Q_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k, \boldsymbol{\psi}^{(q)}) = \sum_{t=2}^{T} \gamma_{tk}^{(q)} \log p(\mathbf{x}_t|y_t = k; \boldsymbol{\theta}_k) \tag{3.39}$$

La maximisation de $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)})$ par rapport à $\boldsymbol{\psi}$ se fait en maximisant séparément $Q_{\boldsymbol{\pi}}(\boldsymbol{\pi}, \boldsymbol{\psi}^{(q)})$ par rapport à $\boldsymbol{\pi}$, $Q_A(A, \boldsymbol{\psi}^{(q)})$ par rapport à A et $Q_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k, \boldsymbol{\psi}^{(q)})$ par rapport à $\boldsymbol{\theta}_k$ pour $k = 1, \ldots, K$.

La maximisation de $Q_{\boldsymbol{\pi}}(\boldsymbol{\pi}, \boldsymbol{\psi}^{(q)})$ par rapport à $\boldsymbol{\pi}$ telle que $\sum_k \pi_k = 1$ est un problème d'optimisation sous contrainte résolu en utilisant les multiplicateurs de Lagrange. Les valeurs maximisant Q_A correspondent au nombre moyen de transitions de l'état l à l'état k relativement au nombre de passages dans l'état l. Les formules de mise à jour correspondantes sont données par :

$$\pi_k^{(q+1)} = \gamma_{1k}^{(q)} \tag{3.40}$$

$$A_{lk}^{(q+1)} = \frac{\sum_{t=2}^{T} \xi_{tlk}^{(q)}}{\sum_{t=2}^{T} \gamma_{tk}^{(q)}} \tag{3.41}$$

Les estimateurs des paramètres $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ sont obtenus en calculant la dérivée de Q par rapport à ces derniers et en annulant celle-ci, comme dans le cas des modèles de mélanges les formules de mise à jour diffèrent selon les lois conditionnelles et doivent être dérivées au cas par cas. Dans le cas des HMM gaussiens on obtient :

$$Q_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k, \boldsymbol{\psi}^{(q)}) = \sum_{t=2}^{T} \gamma_{tk}^{(q)} \log \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.42}$$

et ainsi :

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{(q)} \mathbf{x}_t}{\sum_{t=1}^{T} \gamma_{tk}^{(q)}} \tag{3.43}$$

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{(q)} (\mathbf{x}_t - \boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_t - \boldsymbol{\mu}_k^{(q+1)})'}{\sum_{t=1}^{T} \gamma_{tk}^{(q)}} \tag{3.44}$$

Dans la cas d'un HMM qui considère des probabilités d'émission gaussiennes, l'algorithme EM prend alors une forme simple (cf. algorithme 2).

Algorithme 2: pseudo-code de l'algorithme EM pour les HMM gaussiens

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

 # Initialisation

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

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

 # Etape E

 # Calcul des probabilités a posteriori

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

$\alpha_{1k} = \pi_k p(\mathbf{x}_1 | y_1 = k; \boldsymbol{\theta}_k)$

pour tous les $t \in \{2, \ldots, T\}$ **faire**

$\alpha_{tk} = \sum_{l=1}^{K} \alpha_{(t-1)l} A_{lk} p(\mathbf{x}_t | y_t = k; \boldsymbol{\theta}_k)$

$\beta_{Tk} = 1$

pour tous les $t \in \{T-1, \ldots, 1\}$ **faire**

$\beta_{tl} = \sum_{k=1}^{K} \beta_{(t+1)k} A_{lk} p(\mathbf{x}_{t+1} | y_{t+1} = k; \boldsymbol{\theta}_k)$

pour tous les $t \in \{1, \ldots, T\}$ **faire**

$$\gamma_{tk}^{(q)} = \frac{\alpha_{tk}\beta_{tk}}{\sum_{l=1}^{K} \alpha_{tl}\beta_{tl}}$$

$$\xi_{tlk}^{(q)} = \frac{\alpha_{(t-1)l}^{(q)} A_{lk}^{(q)} p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi}^{(q)}) \beta_{tk}^{(q)}}{\sum_{l=1}^{K} \sum_{k=1}^{K} \alpha_{(t-1)l}^{(q)} A_{lk}^{(q)} p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi}^{(q)}) \beta_{tk}^{(q)}}$$

 # Etape M

 # Maximisation de la fonction auxiliaire

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

$$\pi_k^{(q+1)} = \gamma_{1k}^{(q)}, \qquad A_{lk}^{(q+1)} = \frac{\sum_{t=2}^{T} \xi_{tlk}^{(q)}}{\sum_{t=2}^{T} \gamma_{tk}^{(q)}}$$

$$\mu_k^{(q+1)} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{(q)} \mathbf{x}_t}{\sum_{t=1}^{T} \gamma_{tk}^{(q)}}, \qquad \Sigma_k^{(q+1)} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{(q)} (\mathbf{x}_t - \mu_k^{(q+1)})(\mathbf{x}_t - \mu_k^{(q+1)})'}{\sum_{t=1}^{T} \gamma_{tk}^{(q)}}$$

$q = q + 1$

Résultat : Paramètres estimés : $\widehat{\boldsymbol{\psi}}^{ml}$, probabilités a posteriori : $\widehat{\gamma}_{tk}^{ml}$ et $\widehat{\xi}_{tlk}^{ml}$

3.2.5 Modèles à variables latentes continues

Jusqu'à présent les modèles présentés dans ce chapitre supposaient l'existence de différentes sous-populations aux propriétés différentes pour décrire le jeu de données et modélisaient 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 modélisation par des groupes ou des classes n'est pas toujours pertinente 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 soient centrées, ces modèles dans leur version linéaire sont définis par l'équation suivante :

$$\mathbf{x} = H\mathbf{z} + \xi, \tag{3.45}$$

où ξ est une réalisation de Ξ un bruit indépendant de $\mathbf{z}$; la matrice H est une matrice de taille $P \times L$ qui lie les variables latentes représentées dans cette équation par le vecteur $\mathbf{z}$ de dimension L et les variables observées, représentées par le vecteur $\mathbf{x}$ de taille P. L'objectif des modèles à variables latentes continues est de fournir une estimation de la matrice de mixage H et des variables latentes $\mathbf{z}$, également inconnues dans ce problème, à partir des seules observations de X sous certaines hypothèses. Ces hypothèses peuvent différer selon les modèles.

Les modèles à variables latentes continues peuvent également 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.

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 (3.45), il apparait clairement qu'un changement d'échelle d'une variable latente peut être compensé en divisant la colonne de H 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 l'unité. 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 H sont permutées elles aussi.

Les principaux modèles à variables latentes continues diffèrent par les hypothèses qu'ils postulent quant à la distribution des variables latentes et à celle du bruit. Lorsque les variables latentes sont considérées gaussiennes, on retrouve des modèles tels que l'Analyse Factorielle (FA pour Factor Analysis) pour un bruit gaussien de matrice de variance-covariance diagonale [Spearman 1904, Bartholomew & Martin 1999] ou l'Analyse en Composantes Principales Probabiliste (PPCA pour Probabilist Principal Component Analysis) [Tipping & Bishop 1997,

Roweis 1998] dans le cas d'un bruit isotrope (matrice de variance-covariance proportionnelle à l'identité). Le cas de variables latentes gaussiennes sans bruit correspond à l'Analyse en Composantes Principales, méthode très populaire en analyse de données [Pearson 1901, Hotelling 1933]. L'analyse en composantes indépendantes (ICA pour Independent Component Analysis) quant à elle, ne postule pas une forme gaussienne pour les variables latentes mais pose une hypothèse plus forte : l'indépendance de celles-ci. La section suivante revient plus en détail sur ce type de modèle.

3.3 Analyse en composantes indépendantes (ICA)

L'analyse en composantes indépendantes (ACI ou ICA pour Independent Component Analysis) 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]. Ces travaux ont ouvert la voie à la formulation du problème de *séparation aveugle de sources*. Une diffusion importante de la méthode au sein de la communauté internationale dans les années 1990 avec des travaux tels que [Comon 1994, Bell & Sejnowski 1995b, Cardoso 1997] a permis une meilleure compréhension des bases théoriques de cette méthode avec la mise au point d'algorithmes efficaces [Amari *et al.* 1996, Cardoso & Laheld 1996, Cardoso 1999, Hyvärinen 1999]. Différents ouvrages ont depuis été consacrés à cette méthode [Hyvärinen *et al.* 2001, Roberts & Everson 2001, Jutten & Comon 2007a, Jutten & Comon 2007b].

L'exemple de la soirée de cocktail est la description classique du problème de séparation aveugle de sources : On suppose que P micros sont placés à différents endroits d'une pièce où L 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. Le problème consiste alors à retrouver les L signaux de paroles des différents locuteurs à partir des P enregistrements.

Du fait de la diversité des domaines d'applications de l'analyse en composantes indépendantes, les termes sources, facteurs ou variables latentes seront employés indistinctement dans cette section.

3.3.1 Problématique

L'Analyse en composantes indépendantes définit un modèle génératif pour des données observées issues de variables aléatoires multivariées. Dans le modèle, les variables observées sont supposées être des mélanges linéaires de certaines variables latentes inconnues, tel que le processus de mixage est également inconnu. L'ICA cherche à estimer des variables latentes mutuellement indépendantes et le processus de mixage sous-jacent à partir des variables observées uniquement.

L'hypothèse d'indépendance est une hypothèse forte qui 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 [Hyvärinen *et al.* 2001].

Le modèle de l'ICA dans sa version de base suppose une relation déterministe entre variables latentes et variables observées avec de plus un nombre identique de variables observées et de variables latentes. Le modèle est alors de la forme :

$$\mathbf{x} = H\,\mathbf{z}, \tag{3.46}$$

où la matrice de mixage H est de taille $L \times L$.

Certaines méthodes [Moulines *et al.* 1997, Attias 1999, Ikeda 2000] s'appuient sur le modèle défini par l'équation (3.45), mais la formulation (3.46) du modèle est la plus utilisée. 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 possible d'utiliser une analyse en composantes principales pour séparer l'information utile du bruit, en ne conservant que les L 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 (3.46) est tout d'abord que si la matrice H est non-singulière, l'équation peut être inversée :

$$\mathbf{z} = H^{-1}\,\mathbf{x} = G\,\mathbf{x}. \tag{3.47}$$

Ceci conduit à rechercher non plus la matrice de mixage H, mais la matrice de démixage G 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 la propriété suivante.

Proposition 3.1 ***(Densité d'une transformation linéaire)***
Soient deux variables aléatoires telles que $\mathbf{x} = H.\mathbf{z}$*, avec* H *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 (cf. annexe A.4) :*

$$f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(H)|} g^{\mathcal{Z}}(H^{-1}\,\mathbf{x}). \tag{3.48}$$

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.

3.3.2 ICA et information mutuelle

Dans le modèle d'ICA, l'indépendance statistique des sources est un concept de base bien plus puissant que celui de la décorrélation. L'information mutuelle fournissant une mesure de l'indépendance entre un ensemble de variables, sa minimisation constitue une solution naturelle au problème lié à l'ICA. Les différentes méthodes proposées dans la littérature, en particulier depuis les années 1990 cherchent toutes implicitement ou explicitement à maximiser une mesure d'indépendance en lien avec l'information mutuelle. On citera notamment : la minimisation de l'information mutuelle [Amari *et al.* 1996], la maximisation de la vraisemblance [Moulines *et al.* 1997, Attias 1999], la maximisation de l'écart à la normalité [Hyvärinen *et al.* 2001, chap. 8], la décorrélation non linéaire, [Hérault *et al.* 1985, Bach & Jordan 2003], la diagonalisation de tenseurs [Cardoso 1999]. Cette section présente principalement les concepts liés à la minimisation de l'information en insistant sur les liens qui peuvent être établis avec les autres méthodes. Il est possible de consulter les ouvrages [Hyvärinen *et al.* 2001] et [Sodoyer 2004] pour plus de détails.

Dans le cadre de la théorie de l'information introduite par [Shannon 1948], l'information mutuelle peut être exprimée à travers différents concepts :

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

$$HT(Z) = -\mathrm{E}[\log(f^{\mathcal{Z}}(Z))] = -\int_{\mathcal{Z}} f^{\mathcal{Z}}(\mathbf{z}) \log(f^{\mathcal{Z}}(\mathbf{z}))\mathrm{d}\mathbf{z}\,. \tag{3.49}$$

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

$$HT(Z) = HT(Z_1) + HT(Z_2)\,. \tag{3.50}$$

- **La divergence de Kullback-Leibler :** Cet outil de la théorie de l'information permet de quantifier l'écart entre deux densités de probabilités. 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 :

$$KL(f^{\mathcal{Z}}||g^{\mathcal{Z}}) = \int_{\mathcal{Z}} f^{\mathcal{Z}}(\mathbf{z}) \log\left(\frac{f^{\mathcal{Z}}(\mathbf{z})}{g^{\mathcal{Z}}(\mathbf{z})}\right) \mathrm{d}\mathbf{z}\,. \tag{3.51}$$

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

Soit une variable aléatoire Z définie sur un espace produit $\mathcal{Z} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_L$. En notant la densité de probabilité jointe sur $\mathcal{Z} : f^{\mathcal{Z}}$ et $f^{\mathcal{Z}_j}$, $j \in \{1, \ldots, L\}$ les densités marginales, l'information mutuelle de cette variable aléatoire est donnée par :

$$IM(Z) = KL(f^{\mathcal{Z}} || \prod_{j=1}^{L} f^{\mathcal{Z}_j}) \; . \tag{3.52}$$

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

$$IM(Z) = \sum_{j=1}^{L} HT(Z_j) - HT(Z) \; . \tag{3.53}$$

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 est toujours positive et vaut zéro si et seulement si les variables sont statistiquement indépendantes. Elle permet ainsi de mesurer l'indépendance entre différentes variables aléatoires et fournit un principe d'estimation. Dans le contexte de l'analyse en composantes indépendantes, ceci revient à rechercher la matrice de démixage G telle que :

$$\widehat{G} = \arg\min_{G} \widehat{IM}(G.\mathbf{X}'), \tag{3.54}$$

où $\widehat{IM}(G.\mathbf{X}')$ 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 (3.47), il est possible, en utilisant l'équation (3.48) reliant la densité de Z à la densité de X et les définitions données précédemment, d'exprimer l'information mutuelle sous la forme suivante :

$$IM(Z) = \sum_{j=1}^{L} HT(Z_j) - HT(X) - \log(|\det(G)|), \tag{3.55}$$

En prenant en considération le fait que les variables latentes sont décorrélées ($\mathrm{E}[ZZ'] = \mathbf{I}$), puisqu'elles sont supposées indépendantes, on peut écrire :

$$IM(Z) = \sum_{j=1}^{L} HT(Z_j) - HT(X) + cste, \tag{3.56}$$

car :

$$\begin{aligned} \det(G \; \Sigma_x \; G') &= \det(G)\det(\Sigma_x)\det(G') = \det(\mathbf{I}) = 1 && (3.57) \\ \Rightarrow \det(G) &= cste \; . && (3.58) \end{aligned}$$

La relation (3.56) montre que les variations de l'information mutuelle en fonction de la matrice de démixage G choisie, dépendent uniquement des variations des entropies marginales, car l'entropie $HT(X)$ 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 (3.56) et (3.49) [Jutten & 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 de construire, dans le contexte de l'analyse en composante indépendantes, des estimateurs extrêmement simples de l'entropie, [Hyvärinen *et al.* 2001, p. 113-115]. En remplaçant dans cette approximation les cumulants par leurs estimés empiriques, il est possible de construire un critère mesurant l'indépendance d'un ensemble de variables.

L'information mutuelle peut également être mise en relation avec le principe du maximum de vraisemblance. La section suivante est consacrée à cette dernière approche.

3.3.3 Estimation par maximum de vraisemblance

La vraisemblance du modèle de l'analyse en composantes indépendantes s'écrit en utilisant (3.48) et l'hypothèse d'indépendance des variables latentes :

$$L(G;\mathbf{X}) = \prod_{i=1}^{N} |\det(G)| \left(\prod_{j=1}^{L} f^{\mathcal{Z}_j}\left((G\mathbf{x}_i)_j\right) \right). \tag{3.59}$$

En passant au logarithme, nous obtenons :

$$\mathcal{L}(G;\mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{L} \log\left(f^{\mathcal{Z}_j}\left((G\mathbf{x}_i)_j\right)\right) + N\log(|\det(G)|), \tag{3.60}$$

ce qui en divisant par N donnerait une forme identique au signe près au critère d'approximation de l'information mutuelle, lorsque les densités des sources sont fixées :

$$\frac{1}{N}\mathcal{L}(G;\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} \log\left(f^{\mathcal{Z}_j}\left((G\mathbf{x}_i)_j\right)\right) + \log(|\det(G)|). \tag{3.61}$$

$$\widehat{IM}(Z) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} \log\left(f^{\mathcal{Z}_j}\left((G\mathbf{x}_i)_j\right)\right) + \log(|\det(G)|) + cste. \tag{3.62}$$

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. De cette façon, la vraisemblance est exprimée en fonction des paramètres du modèle et l'estimation de ces paramètres peut être réalisé à l'aide d'un algorithme du type gradient [Bell & Sejnowski 1995b].

Optimisation par l'algorithme du gradient naturel

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 la dérivée de la log-vraisemblance par rapport à la matrice de démixage G (cf. annexe A.5). Celle-ci est donnée par :

$$\frac{\partial \mathcal{L}(G;\mathbf{X})}{\partial G} \propto (G^{-1})' - \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(G\mathbf{x}_i)\mathbf{x}_i', \tag{3.63}$$

avec :

$$\mathbf{g}(\mathbf{z}) = [g_1(z_1), \ldots, g_L(z_L)]' \quad \text{et} \quad g_j(z_j) = \frac{-\partial \log(f^{\mathcal{Z}_j}(z_j))}{\partial z_j}. \tag{3.64}$$

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

$$G^{(q+1)} = G^{(q)} + \tau \left(\left((G^{(q)})^{-1} \right)' - \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(G^{(q)}\,\mathbf{x}_i)\mathbf{x}_i' \right), \tag{3.65}$$

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 [Nocedal & Wright 1999].

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 $G'G$, ce qui conduit à la règle de mise à jour suivante pour la matrice de démixage :

$$G^{(q+1)} = G^{(q)} + \tau \left(\mathbf{I} - \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(\mathbf{z}_i^{(q)})\mathbf{z}_i'^{(q)} \right) G^{(q)}, \tag{3.66}$$

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

Densités des variables latentes

Comme celà a été précisé précédemment, étant donné l'expression de la vraisemblance (3.61), il est nécessaire que les fonctions de densité $f^{\mathcal{Z}_j}$ soient exactement

définies. Deux solutions sont communément possibles, dans certains cas ces dernières peuvent être connues et sont alors directement utilisées pour la maximisation de la vraisemblance. Si ce n'est pas le cas, il est possible d'approximer les densités des composantes indépendantes par une famille de densités avec un nombre réduit de paramètres. Il s'agit alors de déterminer si la densité est de nature *supergaussienne* ou *subgaussienne* à partir du critère suivant :

$$\text{sign}(\text{E}\{z_j g_j(z_j) - g'_j(z_j)\}) = \begin{cases} + & \text{densité supergaussienne} \\ - & \text{densité subgaussienne} \end{cases} \tag{3.67}$$

Ce critère mesure la forme de la fonction de densité au même titre que le moment d'ordre quatre (kurtosis ou coefficient d'aplatissement), et le choix quant à la nature de la densité peut être assimilé au fait qu'une distribution pointue possède un excès d'aplatissement positif et qu'une distribution aplatie possède un excès d'aplatissement négatif, voir [Hyvärinen *et al.* 2001, p.201-207].

Dans l'algorithme de Bell et Sejnowski [Bell & Sejnowski 1995b, Bell & Sejnowski 1995a], les fonctions de décorrélation non linéaires suivantes sont utilisées pour représenter les densités : $g_j(z_j) = -2\tanh(z_j)$ pour les composantes indépendantes supergaussiennes et $g_j(z_j) = \tanh(z_j) - z_j$ pour les composantes indépendantes subgaussiennes, avec pour critère $\phi_j = \text{E}\{z_j g_j(z_j) - g'_j(z_j)\} = \text{E}\{-\tanh(z_j)z_j + (1 - \tanh(z_j)^2)\}$ (cf. algorithme 3).

Algorithme 3: pseudo-code de l'ICA sans bruit par maximisation de la vraisemblance via la méthode du gradient naturel.

Données : Matrice de données centrée : $\mathbf{X}$

 # Initialisation de la matrice de démixage

$G^{(0)} = rand(L, L), q = 0$

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

  # Mise à jour des sources

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

  pour tous les $j \in \{1, \ldots, L\}$ **faire**

    # Mise à jour du moment non plynomial

    $\phi_j = \mathrm{E}\{-\tanh(z_j)z_j + (1 - \tanh(z_j)^2)\}$

    # Mise à jour des fonctions non linéaires

    si $\phi_j > 0$ **alors**

      $g_j(z_j) = -2\tanh(z_j)$

    sinon

      $g_j(z_j) = \tanh(z_j) - z_j$

  # Calcul du gradient naturel (3.66)

  $\Delta G^{(q)} = (\mathbf{I} - \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(\mathbf{z}_i^{(q)})\mathbf{z}_i'^{(q)})G^{(q)}$

  # Mise à jour de la matrice de démixage

  $G^{(q+1)} = G^{(q)} + \tau.\Delta G^{(q)}$

  $q = q + 1$

Résultat : Matrice de démixage et variables latentes estimées : $\widehat{G}^{ml}$, $\widehat{\mathbf{Z}}^{ml}$

3.4 Analyse en facteurs indépendants (IFA)

3.4.1 Description du modèle IFA

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

$$f^{\mathcal{Z}_j}(z_j) = \sum_{k=1}^{K_j} \pi_k^j \mathcal{N}(z_j; \mu_k^j, \nu_k^j), \tag{3.68}$$

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

$$\boldsymbol{\psi} = (G, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^L, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^L, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^L), \tag{3.69}$$

et la log-vraisemblance du modèle devient :

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}) = N \log(|\det(G)|) + \sum_{i=1}^{N} \sum_{j=1}^{L} \log \left(\sum_{k=1}^{K_j} \pi_k^j \mathcal{N}\left((G\mathbf{x}_i)_j, \mu_k^j, \nu_k^j \right) \right). \tag{3.70}$$

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 G 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 du modèle. 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.

3.4.2 Apprentissage du modèle IFA

Dans le cadre des modèles de mélanges, les composantes de la fonction de décorrélation non linéaire $\mathbf{g}$ nécessaire pour maximiser la vraisemblance par rapport à G sont données par :

$$\begin{aligned} \frac{-\partial \log(f^{\mathcal{Z}_j}(z_j))}{\partial z_j} &= \frac{-\partial \log\left(\sum_{k=1}^{K_j} \pi_k^j \mathcal{N}(z_j; \mu_k^j, \nu_k^j) \right)}{\partial z_j} \\ &= \sum_{k=1}^{K_j} t_k^j(z_j) \frac{(z_j - \mu_k^j)}{\nu_k^j}, \end{aligned} \tag{3.71}$$

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

$$t_k^j(z_j) = \frac{\pi_k^j \mathcal{N}(z_j; \mu_k^j, \nu_k^j)}{\sum_{l=1}^{K_j} \pi_l^j \mathcal{N}(z_j; \mu_l^j, \nu_l^j)}. \tag{3.72}$$

Nous avons vu précédemment comment faire croître la vraisemblance en modifiant G, 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 G est fixée. L'algorithme suivant (cf. algorithme 4) récapitule sur ces différents éléments et montre comment ceux-ci peuvent être combinés pour estimer tous les paramètres intervenant dans le modèle.

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

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

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

 # Mise à jour des sources

$\mathbf{z}_i = G.\mathbf{x}_i$

 # Etape E

 # Mise à jour des paramètres des sources / EM

pour tous les $j \in \{1, \ldots, L\}$ *et* $k \in \{1, \ldots, K_j\}$ **faire**

$$t_{ik}^{j(q)} = \frac{\pi_k^{j(q)} \mathcal{N}(z_{ij}; \mu_k^{j(q)}, \nu_k^{j(q)})}{\sum_{l=1}^{K_j} \pi_l^{j(q)} \mathcal{N}(z_{ij}; \mu_l^{j(q)}, \nu_l^{j(q)})}, \quad \forall i \in \{1, \ldots, N\}$$

 # Etape M

 # Mise à jour des paramètres des sources

pour tous les $j \in \{1, \ldots, L\}$ *et* $k \in \{1, \ldots, K_j\}$ **faire**

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

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

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

 # Calcul du gradient naturel (3.66)

$\Delta G = (\mathbf{I} - \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(\mathbf{z}_i^{(q)}) \mathbf{z}_i'^{(q)}) G^{(q)}$

 # Mise à jour de la matrice de démixage

$G^{(q+1)} = G^{(q)} + \tau.\Delta G$

$q = q + 1$

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

Algorithme axé sur la matrice de mixage

La principale difficulté pour l'introduction de contraintes sur l'indépendance statistique de certaines variables latentes vis à vis de certaines variables observées réside dans la reformulation du problème d'estimation de l'ICA en fonction de la matrice de mixage et non plus en fonction de la matrice de démixage comme c'est le cas classiquement [Hyvärinen *et al.* 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 $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$ grâce à la relation suivante (cf. annexe A.4) :

$$p^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(H)|} \prod_{j=1}^{L} f^{\mathcal{Z}_j}\left((H^{-1}\,\mathbf{x})_j\right). \tag{3.73}$$

La vraisemblance de la matrice de mixage H est alors donnée par :

$$L(H;\mathbf{X}) = \prod_{i=1}^{N} \frac{1}{|\det(H)|} \left(\prod_{j=1}^{L} f^{\mathcal{Z}_j}((H^{-1}\mathbf{x}_i)_j) \right), \tag{3.74}$$

pour un jeu de données i.i.d. $\mathbf{X}$. Il est alors possible d'écrire la log-vraisemblance d'une matrice de mixage H quelconque aussi bien que d'une matrice de démixage G quelconque :

$$\mathcal{L}(H;\mathbf{X}) = -N\log(|\det(H)|) + \sum_{i=1}^{N}\sum_{j=1}^{L} \log\left(f^{\mathcal{Z}_j}((H^{-1}\mathbf{x}_i)_j)\right), \tag{3.75}$$

$$\mathcal{L}(G;\mathbf{X}) = N\log(|\det(G)|) + \sum_{i=1}^{N}\sum_{j=1}^{L} \log\left(f^{\mathcal{Z}_j}((G\mathbf{x}_i)_j)\right). \tag{3.76}$$

En supposant que les densités des sources $f^{\mathcal{Z}_1}, \ldots, f^{\mathcal{Z}_L}$ soient connues, il est possible de maximiser, l'une ou l'autre de ces fonctions par rapport à H ou G, en utilisant des méthodes de type gradient simple ou gradient naturel comme précédemment (section 3.3.3). Comme les contraintes considérées portent sur la matrice de mixage, il est plus aisé de travailler sur la log-vraisemblance de H (3.75) 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 A.6) :

$$\Delta H^{(q)} = (H^{(q)^{-1}})' \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i^{(q)}\right) \mathbf{z}_i'^{(q)} - \mathbf{I} \right), \tag{3.77}$$

$$H^{(q+1)} = H^{(q)} + \tau.\Delta H^{(q)}, \tag{3.78}$$

avec $\mathbf{z}_i^{(q)} = H^{(q)^{-1}}\mathbf{x}_i$ et $\mathbf{g}$ une fonction de $\mathrm{R}^L \to \mathrm{R}^L$ définie par :

$$\mathbf{g}(\mathbf{z}) = \left[\frac{-\partial \log(f^{\mathcal{Z}_1}(z_1))}{\partial z_1}, \ldots, \frac{-\partial \log(f^{\mathcal{Z}_L}(z_L))}{\partial z_L} \right]'. \tag{3.79}$$

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 (cf. annexe A.6) :

$$\Delta_{nat} H^{(q)} = H \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i' - \mathbf{I} \right), \tag{3.80}$$

$$H^{(q+1)} = H^{(q)} + \tau \Delta_{nat} H^{(q)}. \tag{3.81}$$

En considérant des formes paramétriques pour les densités des sources, il est possible d'effectuer une analyse en facteurs indépendants basée sur l'algorithme GEM (algorithme 4), en remplaçant 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. Le pseudo code de cet algorithme est donné ci-après (cf. algorithme 5).

Algorithme 5: pseudo-code de l'analyse en facteurs indépendants sans bruit, avec gradient naturel sur la matrice de mixage.

Données : Matrice de données centrée $\mathbf{X}$

 # Initialisation du vecteur de paramètres

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

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

 # Mise à jour des sources

$\mathbf{z}_i = {H^{(q)}}^{-1}.\mathbf{x}_i$

 # Etape E

 # Mise à jour des probabilités a posteriori

pour tous les $j \in \{1, \ldots, L\}$ *et* $k \in \{1, \ldots, K_j\}$ **faire**

$$t_{ik}^{j(q)} = \frac{\pi_k^{j(q)} \mathcal{N}(z_{ij}; \mu_k^{j(q)}, \nu_k^{j(q)})}{\sum_{k'=1}^{K_j} \pi_{k'}^{j(q)} \mathcal{N}(z_{ij}; \mu_{k'}^{j(q)}, \nu_{k'}^{j(q)})}, \quad \forall i \in \{1, \ldots, N\}$$

 # Etape M

 # Mise à jour des paramètres des sources

pour tous les $j \in \{1, \ldots, L\}$ *et* $k \in \{1, \ldots, K_j\}$ **faire**

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

$$\mu_k^{j(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{j(q)}} \sum_{i=1}^{N} t_{ik}^{j(q)} \mathbf{z}_{ij}$$

$$\nu_k^{j(q+1)} = \frac{1}{\sum_{i=1}^{N} t_{ik}^{j(q)}} \sum_{i=1}^{N} t_{ik}^{j(q)} (\mathbf{z}_{ij} - \mu_k^{j(q+1)})^2$$

 # Calcul du gradient naturel (3.80)

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

 # Mise à jour de la matrice de mixage (cf. annexe A.6)

$H^{(q+1)} = H^{(q)} + \tau.\Delta H$

$q \leftarrow q + 1$

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

3.5 Conclusion

Ce chapitre nous a permis de présenter différents outils utilisés dans nos travaux. Nous avons ainsi introduit les modèles statistiques à variables latentes qui intègrent des variables inobservées et pour lesquelles aucune information n'est disponible. Dans ce contexte, nous avons présenté les modèles de mélanges qui permettent de modéliser l'existence de sous-populations aux propriétés différentes dans un jeu de données où les observations sont supposées indépendantes. Pour les jeux de données séquentielles, nous avons présenté les modèles de Markov cachés qui intègrent l'aspect temporel au modèle. Nous avons également mis en avant les modèles à variables latentes continues qui permettent de trouver des espaces de représentation intéressants. Dans ce contexte, nous avons introduit les modèles d'analyse en composantes indépendantes et d'analyse en facteurs indépendants.

Nous verrons dans les prochains chapitres l'utilité des ces méthodes pour le diagnostic du système CdV. Le chapitre 4 est consacré à la mise en pratique de ces méthodes dans un contexte non supervisé. En prenant en compte certaines particularités de l'application telles que les informations a priori sur la structure du modèle ou encore l'aspect temporel de données prélevées successivement dans le temps, des extensions de ces méthodes peuvent être utiles dans un tel contexte. Le chapitre 5 décrit quant à lui l'approche adoptée pour le diagnostic des CdV à partir de signaux réels labellisés de façon imparfaite.

Chapitre 4

Extensions de l'Analyse en composantes indépendantes pour le diagnostic

Sommaire

4.1 Introduction

Ce chapitre est consacré à la mise en œuvre de méthodes non supervisées pour le diagnostic des CdV. Les méthodes en question sont basées sur un modèle génératif à variables latentes, l'Analyse en Composantes Indépendantes (ICA). En prenant en compte des informations supplémentaires sur les données et sur le problème à résoudre, deux approches sont proposées.

L'objet de la première approche repose sur la prise en compte et l'intégration d'informations a priori sur la structure du processus de mixage lors de l'apprentissage du modèle d'ICA. Deux extensions permettant de prendre en compte ce type d'informations sont envisagées dans nos travaux et seront présentées dans la première partie de ce chapitre. Dans les deux cas, l'objectif est de mettre en évidence les dépendances et indépendances pouvant exister entre les variables observées et les variables latentes afin d'améliorer la qualité du diagnostic. Les résultats obtenus montreront l'intérêt de tels modèles.

La seconde approche concerne l'apprentissage d'un modèle d'Analyse en Facteurs Indépendants (IFA) qui intègre une modélisation temporelle des données. Les données sont alors supposées être prélevées sur un même système successivement à travers le temps et à intervalles réguliers. Dans ce contexte, la problématique du diagnostic consiste à suivre l'évolution du système par rapport aux états de fonctionnement antérieurs. La méthode proposée modélise l'ensemble des données par un modèle d'Analyse en Facteurs Indépendants où chaque facteur est décrit comme étant une Chaîne de Markov Cachée. Les résultats obtenus sur une base de signaux simulés permettent d'illustrer l'avantage d'une telle approche dans le cadre d'un diagnostic de type « suivi de point fonctionnement ».

4.2 Modélisation de la problématique du diagnostic

Dans le cadre de l'application CdV, nous avons adopté une approche en accord avec les spécificités de l'application (cf. chapitre 2). Ainsi, les variables observées extraites des signaux de mesure sont supposées être des mélanges linéaires de variables latentes liées aux défauts. Cette relation peut être représentée par le modèle graphique ci-dessous (figure 4.1).

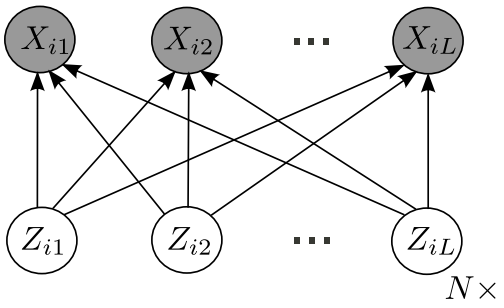


FIGURE 4.1 – *Modèle graphique génératif pour le diagnostic des CdV intégrant des variables latentes continues.*

Dans cette modélisation les variables latentes Z_{ij} sont des variables continues et représentent la valeur des capacités des condensateurs. Cette valeur reflète l'état de fonctionnement du $j$ème condensateur. Notons que les variables observées X_{ij} correspondent aux coefficients (b_{ij}, c_{ij}) du polynôme local approximant l'arche j du signal (figure 4.2). Deux coefficients seulement parmi trois sont utilisés en raison de la continuité nécessaire entre chaque polynôme (cf. chapitre 2).

Partant de cette représentation et en considérant le système précédent comme un système linéaire instantané, on peut représenter les liens entre variables observées et variables d'intérêt (latentes) par la relation :

$$\mathbf{x} = H\,\mathbf{z}, \tag{4.1}$$

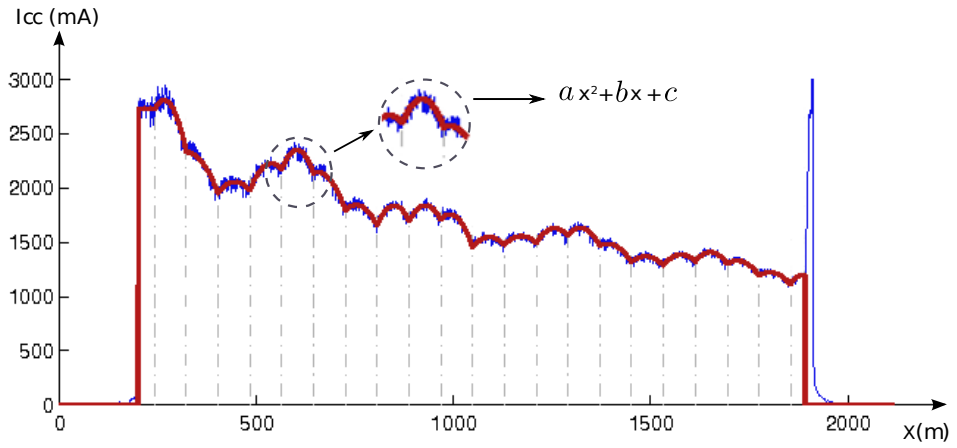


FIGURE 4.2 – *Coefficients extraits du signal d'inspection lors de l'étape de paramétrisation.*

où les variables observées représentent un mélange des variables latentes indiquant l'état de fonctionnement des condensateurs (les capacités). Dans un tel contexte, la tâche de diagnostic peut être considérée comme un problème de séparation de sources et une Analyse en Composantes Indépendantes peut être utilisée pour estimer le processus de mixage H ainsi que les variables latentes $\mathbf{z}$, uniquement à partir des variables observées $\mathbf{x}$.

Cette formulation du problème de diagnostic servira de base pour les propositions envisagées dans la suite du document.

4.3 Extensions du modèle ICA

L'ICA a suscité un grand nombre de travaux et plusieurs voies ont été explorées pour prendre en considération des informations supplémentaires sur le problème traité. L'idée sous-jacente est d'incorporer dans le modèle des connaissances a priori sur la structure du mélange ou sur les variables latentes recherchées en vue d'en améliorer l'estimation. On peut citer les travaux de [Moussaoui 2005, Bakir *et al.* 2006, Li *et al.* 2007] [Jutten & Comon 2007b, chap. 12] où des contraintes de positivité sont introduites dans l'estimation du modèle de l'ICA. Contraintes qui peuvent concerner les variables latentes, la matrice de mixage ou bien encore les deux et qui peuvent être introduites directement dans le problème d'optimisation ou au travers de la forme des densités postulées pour les sources.

D'autres méthodes de séparation de sources fondées sur la parcimonie ont également été proposées et sont particulièrement utiles dans le cas de mélanges sous-déterminés. A ce titre, on peut citer les travaux de [Jutten & Comon 2007b, chap.

10] où il s'agit d'appliquer une transformation de type ondelettes ou temps fréquence, pour augmenter la parcimonie tout en conservant la structure linéaire du mélange, puis d'estimer les sources parcimonieuses pour revenir ensuite aux sources initiales par la transformation inverse.

Aussi, certaines méthodes mettent en œuvre des contraintes sur les éléments de la matrice de mixage [Hyvärinen & Karthikesh 2002, Zhang & Chan 2006, Côme 2009] et permettent de préserver uniquement les liens qui interviennent dans le mélange. Ces méthodes ont été envisagées dans le cadre de notre application car elles correspondent parfaitement à la problématique étudiée.

En effet, dans le cas du système CdV les dépendances entre les variables observées et les variables latentes ont une signification spatiale qui devrait être considérée pour une meilleure représentation du problème de diagnostic. Ainsi, afin de tenir compte du fait qu'un condensateur d'accord n'influence pas le signal situé en amont de sa position, certaines connexions entre variables latentes et observées peuvent être éliminées (figure 4.3).

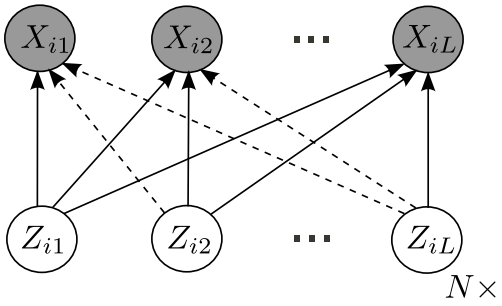


FIGURE 4.3 – *Modèle graphique génératif pour le diagnostic des CdV éliminant certaines connexions entre variables latentes et variables observées.*

Les extensions du modèle ICA détaillées dans la suite permettent d'introduire de telles informations directement sur la matrice de mixage.

4.3.1 Contraintes sur le processus de mixage

La présente extension a pour objectif de prendre en compte des informations supplémentaires sur le processus de mélange. Il s'agit plus particulièrement d'intégrer des hypothèses d'indépendance entre certaines variables latentes et observées (en plus de l'indépendance entre variables observées), hypothèses généralement issues de connaissances physiques sur le processus de mélange. Cette démarche, largement déployée dans le cadre de l'analyse factorielle [Bartholomew & Martin 1999] et plus particulièrement dans le domaine de la modélisation par équation structurelle [Bollen 1989], a été utilisée dans le cadre de l'IFA dans [Côme 2009] dans l'objectif d'estimer un modèle parcimonieux.

L'idée à travers cette approche est de considérer certaines hypothèses d'indépendance dans le mélange. Ces hypothèses sont de la forme :

$$X_h \perp\!\!\!\perp Z_p \tag{4.2}$$

ce qui signifie que la variable observée X_h est statistiquement indépendante de la variable latente Z_p. Postuler une telle hypothèse est équivalent à contraindre la forme de la matrice de mélange comme le montre la proposition suivante :

Proposition 4.1 *Dans le modèle de l'ICA sans bruit,*

$$X_h \perp\!\!\!\perp Z_p \Leftrightarrow H_{hp} = 0. \tag{4.3}$$

Preuve L'indépendance peut être définie comme :

$$X_h \perp\!\!\!\perp Z_p \Leftrightarrow f^{\mathcal{X}_h \times \mathcal{Z}_p}(x_h, z_p) = f^{\mathcal{X}_h}(x_h)\ f^{\mathcal{Z}_p}(z_p). \tag{4.4}$$

Dans le cas du modèle de l'IFA sans bruit, la densité de probabilité jointe $\mathcal{X}_h \times \mathcal{Z}_p$ est définie par :

$$\begin{aligned} f^{\mathcal{X}_h \times \mathcal{Z}_p}(x_h, z_p) &= \int_{\mathrm{R}^{L-1}} f^{\mathcal{X}_h \times \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_L}(x_h, z_1, \ldots, z_L) \prod_{l=1, l\neq p}^{L} \mathrm{d}z_l & (4.5)\\ &= \int_{\mathrm{R}^{L-1}} \prod_{j=1}^{L} f^{\mathcal{Z}_j}(z_j) \times \delta(x_h - H_{h.}\mathbf{z}) \prod_{l=1, l\neq p}^{L} \mathrm{d}z_l \\ &= f^{\mathcal{Z}_p}(z_p) \times \left(\int_{\mathrm{R}^{L-1}} \prod_{l=1, l\neq p}^{L} f^{\mathcal{Z}_l}(z_l) \times \delta(x_h - H_{h.}\mathbf{z}) \mathrm{d}z_l \right). & (4.6) \end{aligned}$$

On note que $H_{h.}$ est la $h$ième ligne de la matrice de mélange H. En utilisant cette hypothèse d'indépendance, on identifie :

$$X_h \perp\!\!\!\perp Z_p \Leftrightarrow f^{\mathcal{X}_h}(x_h) = \int_{\mathrm{R}^{L-1}} \prod_{l=1, l\neq p}^{L} f^{\mathcal{Z}_l}(z_l) \times \delta(x_h - H_{h.}\mathbf{z}) \mathrm{d}z_l, \tag{4.7}$$

où δ est la fonction de Dirac. L'intégrale ne doit pas dépendre de z_p (la $p$ième ligne de $\mathbf{z}$), ce qui n'est possible que si et seulement si H satisfait $H_{hp} = 0$. □

Sur le modèle graphique de génération de données de l'ICA présenté précédemment (figure 4.3), ceci revient à omettre les connexions en pointillé entre variables latentes et variables observées. De telles hypothèses auront pour conséquence de diminuer l'erreur d'estimation. On note cependant que ces hypothèses doivent concorder avec une réalité physique, faute de quoi les résultats seront bien évidemment dégradés.

La log-vraisemblance est alors maximisée sous la contrainte de nullité de certains coefficients de la matrice de mixage. Il suffit pour cela que la montée de gradient

soit effectuée sur ces seuls coefficients. L'initialisation et les règles de mise à jour sont donc données par :

$$\begin{aligned} H^{(0)} &= C \bullet H^{(0)} \\ H^{(q+1)} &= H^{(q)} + \tau\, C \bullet \Delta H^{(q)}, \end{aligned} \tag{4.8}$$

où $\bullet$ représente le produit d'Hadamard entre deux matrices (produit élément par élément), et C une matrice binaire définie par : $C_{hp} = 0$ si $X_h \perp\!\!\!\perp Z_p$, et $C_{hp} = 1$, sinon.

Postuler une hypothèse d'indépendance entre variables latentes et observées est équivalent à contraindre certains coefficients de la matrice de mixage à être nuls. Cette démarche semble naturelle dans un processus de génération de données et préférable à l'hypothèse alternative qui serait celle de contraindre la matrice de démixage. On notera au passage que les deux hypothèses de contrainte des matrices de mixage ou démixage ne sont pas équivalentes ; l'inverse d'une matrice contenant des valeurs nulles ne contient pas forcément des zéros aux mêmes emplacements.

La prise en compte des contraintes de structure sur la matrice de mixage impose une reformulation du problème d'estimation de l'ICA, celui-ci doit en effet s'effectuer en fonction de la matrice de mixage et non pas de la matrice de démixage comme c'est le cas habituellement. La log-vraisemblance par rapport à la matrice de mixage H est donnée par :

$$\mathcal{L}(H; \mathbf{X}) = -N |\det(H)| + \left(\sum_{i=1}^{N} \sum_{j=1}^{L} \log f^{\mathcal{Z}_j}((H^{-1}\mathbf{x}_i)_j) \right). \tag{4.9}$$

Les formules de mise à jour de cette matrice lors de l'utilisation d'un algorithme de gradient sont détaillées dans l'algorithme 6.

Algorithme 6: pseudo-code de l'ICA sans bruit par maximisation de la vraisemblance via la méthode du gradient naturel sur la matrice mixage.

Données : matrice de données centrée : $\mathbf{X}$, matrice de contraintes : C

 # Initialisation de la matrice de mixage

$H^{(0)} = rand(L, L), q = 0$

$H^{(0)} = C \bullet H^{(0)}$

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

  # Mise à jour des sources

  $\mathbf{z}_i = {H^{(q)}}^{-1}.\mathbf{x}_i$

  pour tous les $j \in \{1, \ldots, L\}$ **faire**

    # Mise à jour du moment non polynômial

    $\phi_j = \mathrm{E}\{-\tanh(z_j)z_j + (1 - \tanh(z_j)^2)\}$

    # Mise à jour des fonctions non linéaires

    si $\phi_j > 0$ **alors**

      $g_j(z_j) = -2\tanh(z_j)$

    sinon

      $g_j(z_j) = \tanh(z_j) - z_j$

  # Calcul du gradient naturel

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

  # Mise à jour de la matrice de mixage

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

  $q = q + 1$

Résultat : Matrice de mixage et variables latentes estimées : $\widehat{H}^{ml}$, $\widehat{\mathbf{Z}}^{ml}$

Remarque 4.1 ***(Pré-traitements)***
Hormis le centrage et la réduction, l'ICA avec contraintes sur la matrice de mixage n'autorise pas d'effectuer un pré-traitement des données de type blanchiment. Un tel pré-traitement aurait pour conséquence d'effacer les hypothèses d'indépendance entre variables latentes et observées.

4.3.2 Fonctions de pénalités

Une autre alternative pour tenir compte des dépendances spatiales consiste à appliquer des fonctions de pénalité sur les éléments de la matrice de mixage de la même façon qu'en régression [Fan & Li 2001]. Dans le cadre de l'ICA, ceci a pour but de réduire le nombre de coefficients intervenant dans le mélange et de mettre en avant la structure réelle du modèle en éliminant les éléments de la matrice qui expriment des dépendances trop faibles (figure 4.3). La matrice de mixage parcimonieuse est alors estimée à partir de la fonction objectif constituée de la log-vraisemblance associée à une fonction de pénalité P_λ où λ détermine le degré de pénalisation [Hyvärinen & Karthikesh 2002, Zhang & Chan 2006].

Ainsi, en incorporant une pénalité de type L_γ sur les coefficients de la matrice de mixage H_{lp}, la maximisation de la log-vraisemblance pénalisée est obtenue

en maximisant la log-vraisemblance du modèle d'ICA plus le terme de pénalité $\sum_{l,p=1}^{L} \lambda |H_{lp}|^{\gamma}$, où $\lambda \geq 0$ est un paramètre de complexité qui contrôle la réduction des coefficients : plus la valeur de λ est grande, plus les valeurs des coefficients sont proches de zéro.

Dans la littérature, l'influence des fonctions de pénalité a été largement étudiée. On s'intéressera ici aux pénalisations les plus fréquemment utilisées. Dans le cas où $\gamma = 2$ une pénalisation L_2 dite *ridge* [Hoerl & Kennard 1970] est appliquée et tend à réduire les coefficients les plus petits, mais ne permet pas de les définir comme nuls. La fonction de pénalité est alors donnée par :

$$\sum_{l,p=1}^{L} P_{\lambda}(H_{lp}) = \sum_{l,p=1}^{L} \lambda |H_{lp}|. \tag{4.10}$$

Le cas où $\gamma = 1$ correspond à une pénalisation du type *Lasso* pour *Least absolute shrinkage and selection operator*. Notée L_1, elle revient à introduire un a priori Laplacien sur les coefficients [Tibshirani 1996]. Cette dernière permet de régler automatiquement les coefficients non significatifs à 0 avec un λ approprié, mais a l'inconvénient d'influer également sur l'estimation des coefficients significatifs. La fonction de pénalité est alors définie par :

$$\sum_{l,p=1}^{L} P_{\lambda}(H_{lp}) = \sum_{l,p=1}^{L} \lambda |H_{lp}|^2. \tag{4.11}$$

Comme dans le cas précédent l'idée d'appliquer une pénalité sur les éléments de la matrice de mixage impose une formulation du problème d'ICA en fonction de la matrice de mixage. L'estimation de la matrice H se fait alors en maximisant la log-vraisemblance pénalisée $\mathcal{L}_{pen}$, constituée de la log-vraisemblance par rapport à la matrice de mixage exprimée précédemment (4.9) et de la fonction de pénalité associée P_{λ} :

$$\mathcal{L}_{pen}(H; \mathbf{X}) = -\frac{1}{N} \mathcal{L}(H; \mathbf{X}) - \sum_{l,p=1}^{L} P_{\lambda}(H_{lp}), \tag{4.12}$$

L'optimisation de la fonction objectif (4.12) se fait par la règle de mise à jour du gradient naturel suivante :

$$\Delta_{pen} H^{(q)} \quad = \quad H \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i' - \mathbf{I} - H' \mathbf{p}_{\lambda}(H_{lp}) \right), \tag{4.13}$$

où $\mathbf{p}_{\lambda}(H_{lp})$ définit la matrice dont l'élément (l, p) est calculé par $\dfrac{\partial P_{\lambda}(H_{lp})}{\partial H_{lp}}$.

Les algorithmes obtenus dans le cas d'une pénalisation L_1 et L_2 sont détaillés ci-dessous (cf. algorithme 7 et algorithme 8).

Algorithme 7: ICA sans bruit par maximisation de la vraisemblance avec pénalité L_1 via la méthode du gradient naturel sur la matrice mixage.

Données : matrice de données centrée : $\mathbf{X}$, degré de pénalisation : λ

 # Initialisation de la matrice de mixage

$H^{(0)} = rand(L, L), q = 0$

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

  # Mise à jour des sources

  $\mathbf{z}_i = {H^{(q)}}^{-1}.\mathbf{x}_i$

  pour tous les $j \in \{1, \ldots, L\}$ **faire**

    # Mise à jour du moment non polynômial

    $\phi_j = \mathrm{E}\{-\tanh(z_j)z_j + (1 - \tanh(z_j)^2)\}$

    si $\phi_j > 0$ **alors**

      $g_j(z_j) = -2\tanh(z_j)$

    sinon

      $g_j(z_j) = \tanh(z_j) - z_j$

  # Calcul du gradient naturel (4.13)

  $\Delta_{pen}H^{(q)} = H^{(q)}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}\left(\mathbf{z}_i^{(q)}\right)\mathbf{z}_i'^{(q)} - \mathbf{I} - \lambda H'\mathrm{sign}(H)\right)$

  # Mise à jour de la matrice de mixage

  $H^{(q+1)} = H^{(q)} + \tau.\Delta H^{(q)}$

  $q = q + 1$

Résultat : Matrice de mixage et variables latentes estimées : $\widehat{H}^{ml}$, $\widehat{\mathbf{Z}}^{ml}$

Algorithme 8: ICA sans bruit par maximisation de la vraisemblance avec pénalité L_2 via la méthode du gradient naturel sur la matrice mixage.

Données : matrice de données centrée : $\mathbf{X}$, degré de pénalisation : λ
 # Initialisation de la matrice de mixage
$H^{(0)} = rand(L, L), q = 0$
tant que *test de convergence* **faire**
  # Mise à jour des sources
  $\mathbf{z}_i = {H^{(q)}}^{-1}.\mathbf{x}_i$
  pour tous les $j \in \{1, \ldots, L\}$ **faire**
    # Mise à jour du moment non polynômial
    $\phi_j = \mathrm{E}\{-\tanh(z_j)z_j + (1 - \tanh(z_j)^2)\}$
    si $\phi_j > 0$ **alors**
      $g_j(z_j) = -2\tanh(z_j)$
    sinon
      $g_j(z_j) = \tanh(z_j) - z_j$
  # Calcul du gradient naturel (4.13)
  $\Delta_{pen}H^{(q)} = H^{(q)}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}\left(\mathbf{z}_i^{(q)}\right)\mathbf{z}_i'^{(q)} - \mathbf{I} - 2\lambda H'H\right)$
  # Mise à jour de la matrice de mixage
  $H^{(q+1)} = H^{(q)} + \tau.\Delta H^{(q)}$
  $q = q + 1$
Résultat : Matrice de mixage et variables latentes estimées : $\widehat{H}^{ml}$, $\widehat{\mathbf{Z}}^{ml}$

Remarque 4.2 ***(Pré-traitements)***
Comme dans le cas de l'ICA avec contraintes sur la matrice de mixage, l'ICA pénalisée n'autorise pas de pré-traitement des données hormis le centrage et la réduction afin de conserver les hypothèses d'indépendance entre variables latentes et variables observées.

4.3.3 Expérimentations et résultats

Afin d'étudier l'intérêt des extensions présentées précédemment, différentes expérimentations on été effectuées. Nous avons tout d'abord cherché à observer l'apport de la pénalisation sur le problème d'estimation de la matrice de mixage dans le cadre du diagnostic des CdV. Le but de ces premières expérimentations a été de déterminer le type et le degré de pénalisation offrant les meilleures performances en termes d'estimation. La suite des expérimentations a pour objectif de démontrer l'intérêt de prendre en compte la structure du modèle en comparant les résultats obtenus via les deux approches présentées précédemment et l'ICA standard pour le diagnostic des CdV.

Pour valider la pertinence des approches présentées, une base de données a dû être constituée afin d'envisager une approche non supervisée. Cette base de données a été construite 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 données nous disposons d'information sur les variables d'intérêts du problème et nous avons ainsi pu comparer les résultats à travers les différentes approches à de véritables valeurs. Nous présentons dans la section suivante la démarche utilisée pour constituer cette base de données.

Description des données

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 2300 Hz. Les caractéristiques suivantes ont par contre été tirées de manière aléatoire :

- les capacités des condensateurs ;
- le coefficient de désadaptation λ de la voie (paramètre de nuisance) qui ajuste l'amplitude de l'onde sinusoïdale de grande longueur d'onde (≈ 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 :

- la première sous population correspond aux condensateurs en parfait état et répond aux spécifications données par le constructeurs. Elle représente 95% des condensateurs simulés. Elle suit une loi normale de moyenne égale à la valeur nominale des condensateurs posés en voie $22\mu F$ et de variance 1.26 telle que 95% de cette sous population réponde aux tolérances fournies par le constructeur ($\pm 10\%$) ;
- une seconde sous population correspond aux condensateurs défectueux. Cette classe représente dans la base 3% des condensateurs. Cette classe suit elle aussi une loi normale, la moyenne de celle-ci a été fixée à $10\mu F$ et de variance 6 ;
- la dernière sous population correspond aux condensateurs arrachés. Cette classe représente 2% des condensateurs. Tous les individus générés dans cette classe se sont vus attribué une capacité égale à 0.

En utilisant cette démarche, deux bases de données ont été constituées. La première base contenant 1000 signaux est destinée à l'apprentissage des paramètres de la méthode (base d'apprentissage). La seconde contient 1500 signaux CdV est prévue pour quantifier les performances des différentes expérimentations (base de test). On notera que ces bases de données nous placent dans un contexte non-supervisé, cadre qui avait été peu abordé lors des travaux précédents cette thèse pour le diagnostic des CdV.

Description du modèle

Dans le cadre de ces expérimentations, l'ensemble des approches envisagées pour le diagnostic des CdV reposent sur une ICA sans bruit. La formulation du modèle dans le cas de l'application CdV nécessite quelques précisions. 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 est égal à $2 \times L$, où L est le nombre de condensateurs du CdV et donc le nombre de variables latentes d'intérêt pour le diagnostic. Comme dans le cadre de l'ICA sans bruit le nombre de variables latentes doit correspondre au nombre de variables observées nous avons extrait $2 \times L$ variables latentes, la moitié d'entre elles correspondant aux états de dégradation des condensateurs et l'autre moitié à des variables de bruit.

Expérimentations

En ce qui concerne le protocole expérimental, nous avons essentiellement voulu mettre en évidence l'intérêt d'intégrer des informations a priori sur la structure du mélange pour l'estimation des variables latentes. Dans l'objectif de quantifier l'apport des différents modèles d'ICA envisagés pour le diagnostic, les performances ont été évaluées sur la base de test à l'aide de deux indicateurs : les coefficients de corrélation entre les capacités réelles et leur estimation d'une part et la parcimonie de la matrice de mixage d'autre part. Les résultats présentés ont été obtenus en choisissant la permutation qui conduit à la meilleure corrélation moyenne.

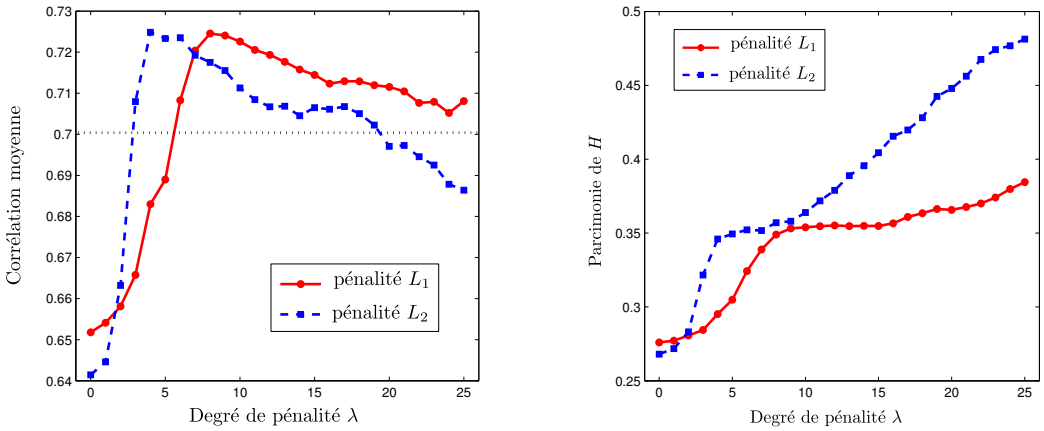
Dans le cas de l'ICA pénalisée, deux fonctions de pénalité ont été testées avec différents degrés de pénalisation afin de sélectionner un modèle performant en termes d'estimation. La figure 4.4 présente les résultats de l'ICA pénalisée dans le cas d'une pénalité L_1 puis avec une pénalité L_2, lorsque le degré de pénalisation λ varie entre 0 et 25. Ces résultats sont obtenus en moyennant 30 exécutions de chaque algorithme et avec différentes initialisations de la matrice de mixage H. Nous présentons d'une part, la moyenne du coefficient de corrélation (en valeur absolue) entre les sources estimées et les capacités réelles sur l'ensemble des condensateurs et d'autre part, l'évolution de la parcimonie de la matrice de mixage, le tout en fonction du degré de pénalisation.

La mesure de parcimonie utilisée permet de quantifier l'énergie contenue au niveau des éléments d'un vecteur [Hoyer 2004]. Celle-ci prend la valeur 1 si et seulement si H contient un unique élément non nul et prend la valeur 0 si tous les éléments sont égaux. Elle est calculée par la formule :

$$Parcimonie(H) = \frac{\sqrt{L \times 2L} - \left(\sum |H_{lp}|\right) / \sqrt{\sum H_{lp}^2}}{\sqrt{L \times 2L} - 1}. \tag{4.14}$$

D'après les résultats de la figure 4.4, la parcimonie de la matrice de mixage augmente de manière continue en fonction du degré de pénalisation, alors que la

corrélation augmente jusqu'à un certain niveau pour diminuer ensuite. Par conséquent, pour une pénalité de type L_2 un degré de pénalisation égal à 4 devrait être suffisant, quant à la pénalité de type L_1, le paramètre λ devrait être fixé à 8. Par ailleurs, le modèle intégrant avec pénalité de type L_2 avec $\lambda = 4$ semble parvenir à de meilleures performances que la pénalisation L_1 en termes de degré de corrélation et de niveau de parcimonie. Ceci peut être justifié par le fait qu'une pénalité de type L_1 a tendance à influer également sur l'estimation des coefficients significatifs du mélange.



(a) *(b)*

FIGURE 4.4 – *Valeur absolue de la corrélation entre les sources estimées et les capacités réelles moyennée sur l'ensemble des condensateurs en fonction du degré de pénalisation λ (a) et mesure parcimonie de la matrice de mixage en fonction du degré de pénalisation en fonction du degré de pénalisation λ (b). Les résultats sont obtenus sur une moyenne de 30 initialisations différentes.*

Nous avons ensuite cherché à comparer les différents approches étudiées ici afin d'en mesurer la pertinence : le modèle d'ICA standard, le modèle avec contraintes sur la matrice de mixage et le modèle d'ICA pénalisée (avec une pénalité L_2 et un degré de pénalisation de 4). Les performances ont été évaluées en fonction de la corrélation obtenue entre chaque source estimée et réelle. Le tableau 4.1 et la figure 4.5 donnent les résultats en terme de moyennes et d'écarts types sur les valeurs absolues des coefficients de corrélation entre les sources estimées et les capacités des 18 condensateurs pour chacun des trois modèles. Les différentes solutions ont été obtenues sur 30 initialisations différentes de chacun des algorithmes.

Globalement, on peut noter que les performances des modèles d'ICA avec parcimonie sont meilleures que celles du modèle standard. Les corrélations entre les capacités réelles des condensateurs et leur estimation sont plus significatives (> 0.7) et largement améliorées pour les sources situées en fin de circuit de voie (17 et 18).

TABLE 4.1 – *Résultats de l'ICA standard, avec contraintes de structure puis avec pénalité. Moyennes des valeurs absolues des coefficients de corrélation entre les sources estimées et les capacités des 18 condensateurs sur 30 initialisations différentes de la matrice de mixage.*

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

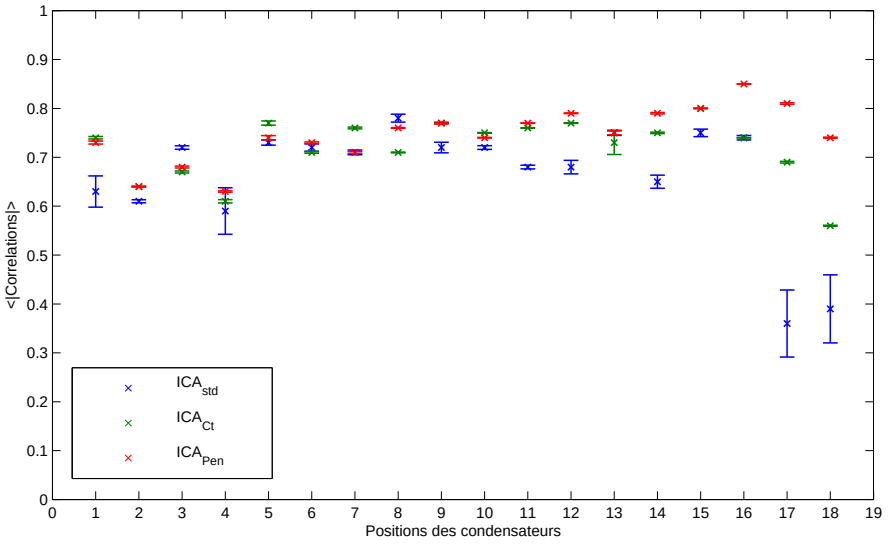
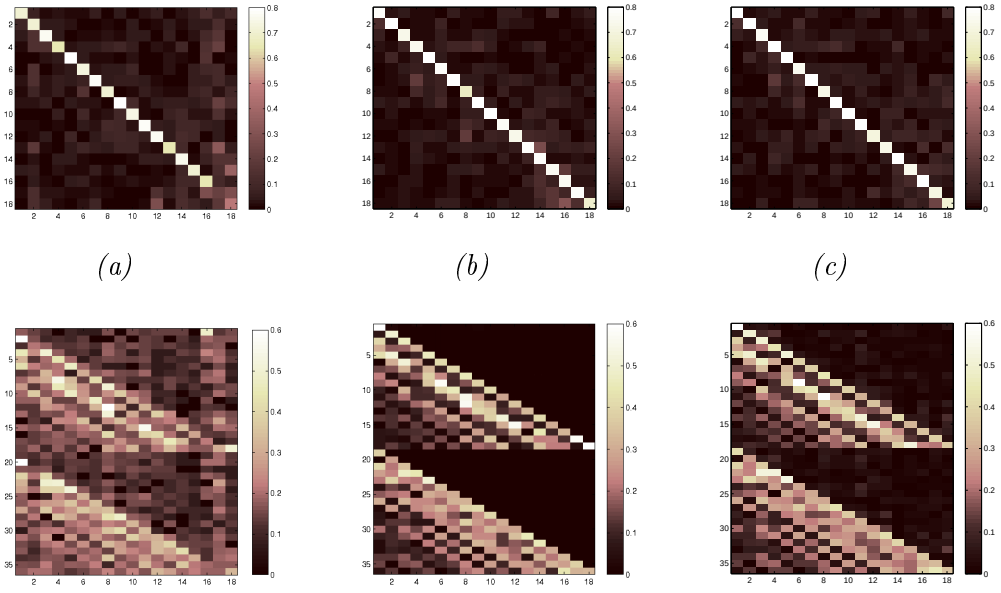


FIGURE 4.5 – *Moyennes et écarts types obtenus à partir des valeurs absolues des coefficients de corrélation entre les sources estimées et les capacités des 18 condensateurs sur 30 initialisations différentes de la matrice de mixage.*

Enfin, nous nous sommes intéressés à la structure de la matrice de mixage et à sa concordance par rapport aux liens connus entre les variables du système. La figure 4.6 donne une représentation de la matrice de mixage obtenue en sélectionnant la meilleur estimation parmi les 30 initialisations précédentes, dans le cas d'un modèle d'ICA standard, dans le cas de l'ICA avec contraintes sur la matrice de mixage et dans le cas d'une ICA pénalisée.

Dans la figure 4.6, la structure de la matrice de mixage est clairement identifiée dans le cas de l'ICA avec contraintes *(e)* dans le cas de l'ICA pénalisée *(f)*. Sachant que l'on cherche à estimer L sources pour $2 \times L$ variables observées, la matrice H est ainsi triangulaire inférieure par bloc et traduit bien les relations amont-aval visibles sur les données d'inspection. L'état d'un condensateur n'influence donc que les variables situées en aval (entre le condensateur et le récepteur). Le modèle d'ICA standard a quant à lui plus de difficultés à reconstituer la structure des liens entre les variables *(d)* et obtient ainsi de moins bonnes performances en termes d'estimation des sources, en particulier pour les sources en fin de CdV qui sont plus difficiles à estimer *(a)*.



(a) *(b)* *(c)*

(d) *(e)* *(f)*

FIGURE 4.6 – *Valeur absolue de la corrélation entre les valeurs réelles des capacités et leur estimation obtenue sur les données de test dans le cas du modèle d'ICA traditionnel (a), du modèle avec contraintes (b) et du modèle pénalisé (c), et valeur absolue de la matrice de mixage correspondante estimée dans le cas du modèle d'ICA traditionnel (d), du modèle avec contraintes (e) et du modèle pénalisé (f).*

En conclusion, les modèles d'ICA avec contraintes sur la matrice de mixage et le modèle pénalisé semblent nettement plus efficace dans l'identification de la structure de la matrice de mixage et par là même des connexions entre variables latentes et variables observées. Par ailleurs, nous avons pu constater que prendre en compte des informations liées aux dépendances entre variables permet d'obtenir des résultats plus pertinents en matière d'estimation des sources, que ce soit en intégrant directement des contraintes sur la matrice de mixage ou en cherchant à les mettre en exergue par le biais de pénalités. Ceci représente un avantage certain dans le

cadre de l'application CdV, car la structure triangulaire inférieure de la matrice de mélange peut permettre de réordonner la solution estimée et de résoudre ainsi les problèmes de permutation des sources.

4.4 Analyse en facteurs indépendants temporelle

L'aspect dynamique du système est également un *a priori* pouvant être pris en compte à travers un modèle tel que l'ICA. En effet face à une surveillance régulière des systèmes CdV, il parait intéressant de vouloir prendre en compte l'aspect temporel induit par des inspections réalisées successivement sur un même objet. Dans un tel contexte, l'idée est de pouvoir observer le comportement dynamique d'un système afin de détecter les événements anormaux.

Pour ce faire nous nous sommes orientés vers une méthodologie basée sur un modèle d'IFA intégrant une modélisation temporelle des sources. L'idée d'intégrer l'aspect dynamique des variables latentes en séparation de sources, a été abordée par différentes approches dans la littérature. Dans [Pearlmutter & Parra 1996, Pearlmutter & Parra 1997], l'aspect dynamique des sources est pris en compte dans le cadre de l'ICA en incorporant une information temporelle au niveau de la densité des sources ou par une modélisation auto-régressive de celles-ci. Une autre approche basée sur une représentation des sources par des processus auto-régressifs généralisés a été proposée dans [Penny *et al.* 2000], dans cette approche des modèles de Markov cachés (HMM) ont de plus été intégrés au modèle pour modéliser les transitions de l'instant $t-1$ à l'instant t. Un autre modèle intégrant l'aspect temporel à l'aide de HMM a également été présenté dans [Attias 2000] comme une extension au modèle d'IFA [Attias 1999]. C'est ce dernier qui sera considéré dans nos travaux.

4.4.1 Principe

Comme pour l'IFA classique présentée au chapitre précédent (section 3.4), l'IFA avec sources temporelles [Attias 2000], considère les variables observées comme des mélanges linéaires de variables latentes mutuellement indépendantes, appelées également facteurs. Toutefois, dans cette approche on considère également l'aspect temporel des sources et chaque source est décrite par un modèle de Markov caché (HMM). Notons $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ la séquence des T échantillons qui constituent les données observées et $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$ les données latentes. Dans une formulation sans bruit, le modèle est alors défini par :

$$\mathbf{x}_t = H\mathbf{z}_t, \tag{4.15}$$

où $\mathbf{x}_t = (x_{t1}, \ldots, x_{tL})$ est le vecteur de données observées à l'instant t, $\mathbf{z}_t = (z_{t1}, \ldots, z_{tL})$ est le vecteur de données latentes à l'instant t avec $t = 1, \ldots, T$ et H est la matrice de mixage de taille $L \times L$.

Nous avons pu voir au chapitre précédent le modèle classique d'IFA où les distributions des sources sont modélisées par des mélanges de Gaussiennes (MoG) afin de pouvoir approximer des densités arbitraires. Dans la cas de l'IFA temporelle [Attias 2000], chaque source est décrite par un HMM afin d'intégrer l'aspect temporel des sources. La figure suivante (figure 4.7) donne une représentation graphique du modèle global.

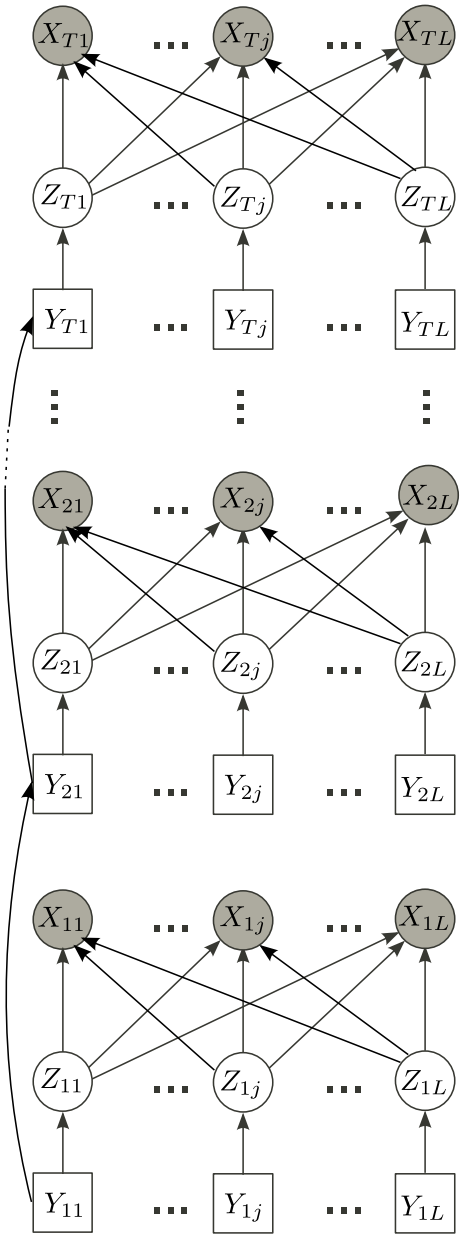


FIGURE 4.7 – *Modèle graphique génératif pour l'Analyse en Facteurs Indépendants avec données temporelles.*

Le modèle proposé dans [Attias 2000] considère des probabilités d'émission gaussiennes et chaque source $\mathbf{z}_j = (z_{1j} \ldots, z_{Tj})$ est supposée être régie par une chaîne de Markov de premier ordre à K_j états $\mathbf{y}_j = (y_{1j}, \ldots, y_{Tj})$. La distribution de chaque source z_{tj} conditionnellement à l'état y_{tj} est alors donnée par :

$$p(z_{tj}|y_{tj} = k) = \mathcal{N}(z_{tj}; \mu_k^j, \nu_k^j). \tag{4.16}$$

Le vecteur des paramètres du modèle précédent se résume donc à :

$$\boldsymbol{\psi} = (H, A^j, \boldsymbol{\pi}^j, \boldsymbol{\theta}^j), \tag{4.17}$$

où :

- H est la matrice de mixage ;
- A^j représente la matrice de transition pour la source j telle que chaque élément est donné par $A_{lk}^j = p(y_{tj} = k|y_{(t-1)j} = l)$ avec $l, k = 1, \ldots, K_j$;
- $\boldsymbol{\pi}^j = (\pi_1^j, \ldots, \pi_{K_j}^j)$ est la distribution initiale de la source j telle que $\pi_k^j = p(y_{1j} = k)$ pour $k = 1, \ldots, K_j$;
- $\boldsymbol{\theta}^j = (\theta_1^j, \ldots, \theta_{K_j}^j)$ définit l'ensemble des paramètres de la densité de probabilité conditionnelle de la source j. Dans le cas présent il s'agit de probabilités d'émissions gaussiennes $\theta_k^j = (\mu_k^j, \theta_k^j)$ pour $k = 1, \ldots, K_j$.

4.4.2 Estimation du modèle

Comme dans le modèle IFA classique (section 3.4), le vecteur de paramètres $\boldsymbol{\psi}$ peut être estimé à l'aide du maximum de vraisemblance :

$$L(\boldsymbol{\psi}; \mathbf{X}) = \frac{1}{|\det(H)|^T} \prod_{j=1}^{L} p(\mathbf{z}_j, \boldsymbol{\psi}), \tag{4.18}$$

La log-vraisemblance des données observées s'écrit alors :

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}) &= \log L(\boldsymbol{\psi}; \mathbf{X}) \\
&= -T \log(|\det(H)|) + \log \prod_{j=1}^{L} p(\mathbf{x}_j; \boldsymbol{\psi}) \\
&= -T \log(|\det(H)|) + \sum_{j=1}^{L} \log \sum_{\mathbf{y}_j} p(\mathbf{z}_j, \mathbf{y}_j; \boldsymbol{\theta}^j) \\
&= -T \log(|\det(H)|) + \sum_{j=1}^{L} \log \sum_{\mathbf{y}_j} p(\mathbf{y}_j; \boldsymbol{\pi}^j, A^j) p(\mathbf{z}_j|\mathbf{y}_j; \boldsymbol{\theta}^j) \\
&= -T \log(|\det(H)|) \\
&\quad + \sum_{j=1}^{L} \log \sum_{\mathbf{y}_j} p(y_{1j}; \boldsymbol{\pi}^j) \prod_{t=2}^{T} p(y_{tj}|y_{(t-1)j}; A^j) \prod_{t=1}^{T} \mathcal{N}(z_{tj}; \boldsymbol{\mu}^j, \boldsymbol{\nu}^j).
\end{aligned} \tag{4.19}$$

Dans cette approche, l'idée de modéliser l'aspect temporel des sources par le biais des HMM constitue un avantage pour l'apprentissage des paramètres du modèle, dans la mesure où l'optimisation de la log-vraisemblance des données peut être réalisée à travers le même type d'algorithme d'optimisation utilisé dans le cadre de l'IFA classique. On se tourne ainsi vers une stratégie d'optimisation alternée. L'algorithme du gradient naturel permet d'optimiser la log vraisemblance par rapport à H lorsque les paramètres des sources sont fixés. Inversement, pour une matrice H donnée, l'algorithme EM permet de maximiser la log-vraisemblance par rapport aux paramètres de chaque source.

L'algorithme du gradient naturel utilise quant à lui la dérivée de la log-vraisemblance par rapport à la matrice de mixage. La règle de mise à jour est alors donnée par :

$$\Delta_{nat} H = H'H\Delta_{nat} H = H(\frac{1}{T}\sum_{t=1}^{T} \mathbf{g}(\mathbf{z}_t)\mathbf{z}'_t - \mathbf{I}), \tag{4.20}$$

Par ailleurs, la mise à jour des paramètres de chacune des sources $j \in \{1, \ldots, L\}$ est réalisée à l'aide des formules de mise à jour classiques suivantes [Attias 2000] :

$$\begin{aligned} \pi_k^j &= \gamma_{1k}^j, & A_{lk}^j &= \frac{\sum_{t=2}^{T} \xi_{tlk}^j}{\sum_{t=2}^{T} \gamma_{tk}^j}, \\ \mu_k^j &= \frac{\sum_{t=1}^{T} \gamma_{tk}^j z_{tj}}{\sum_{t=1}^{T} \gamma_{tk}^j}, & \nu_k^j &= \frac{\sum_{t=1}^{T} \gamma_{tk}^j (z_{tj} - \mu_k^j)^2}{\sum_{t=1}^{T} \gamma_{tk}^j}. \end{aligned} \tag{4.21}$$

où :

- la quantité $\gamma_{tk}^j = p(y_{tj} = k|\mathbf{z}_j; \boldsymbol{\psi}^{(q)})$ représente la probabilité a posteriori d'être dans l'état k à l'instant t sachant la séquence totale de la source j et l'estimation courante des paramètres $\boldsymbol{\psi}^{(q)}$;
- la quantité $\xi_{tlk}^j = p(y_{tj} = k|y_{(t-1)j} = l|\mathbf{z}_j; \boldsymbol{\psi}^{(q)})$, pour $t = 2, \ldots, T$ et $l, k = 1, \ldots, K$, est la probabilité a posteriori jointe d'être à l'état l à l'instant $t-1$ et à l'état k à l'instant t sachant la séquence totale de la source j et l'estimation courante des paramètres $\boldsymbol{\psi}^{(q)}$.

Afin d'obtenir les probabilités γ_{tk}^j et ξ_{tlk}^j, une procédure *forward-backward* est nécessaire pour chaque source. Cette dernière consiste à calculer de façon récursive les probabilités *forward* $\alpha_{tk}^{j(q)} = p(z_{1j}, \ldots, z_{tj}, y_{tj} = k; \boldsymbol{\psi}^{(q)})$ et *backward* $\beta_{tk}^{j(q)} = p(z_{1j}, \ldots, z_{(t+1)j}, y_{Tj} = k; \boldsymbol{\psi}^{(q)})$. Les probabilités a posteriori sont obtenues pour chaque source à l'aide des formules :

$$\gamma_{tk}^j = \frac{\alpha_{tk}^j \beta_{tk}^j}{\sum_{l=1}^{K_j} \alpha_{tl}^j \beta_{tl}^j}, \tag{4.22}$$

$$\xi_{tlk}^j = \frac{\alpha_{(t-1)l}^j A_{lk}^j p(z_{tj}|y_{tj} = k_j) \beta_{tk}^j}{\sum_{l=1}^{K_j} \sum_{k=1}^{K_j} \alpha_{(t-1)l}^j A_{lk}^j p(z_{tj}|y_{tj} = k) \beta_{tk}^j}. \tag{4.23}$$

L'ensemble des principales étapes de cet algorithme sont données ci-dessous pour l'estimation de la matrice de mixage du modèle d'IFA avec sources temporelles par maximisation de la log-vraisemblance via la méthode du gradient naturel (cf. algorithme 9).

Algorithme 9: Pseudo-code de l'IFA avec sources temporelles par maximisation de la log-vraisemblance via la méthode du gradient naturel sur la matrice de mixage

Données : Matrice de données centrée $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$

 # Initialisation du vecteur de paramètres

$\psi^{(0)} = (H^{(0)}, A^{(0)}, \boldsymbol{\pi}^{1(0)}, \ldots, \boldsymbol{\pi}^{L(0)}, \boldsymbol{\mu}^{1(0)}, \ldots, \boldsymbol{\mu}^{L(0)}, \boldsymbol{\nu}^{1(0)}, \ldots, \boldsymbol{\nu}^{L(0)})$, $q = 0$

tant que *Condition de convergence* **faire**

 # Mise à jour des sources

$\mathbf{z}_t = H^{-1(q)}\mathbf{x}_t$

 # Mise à jour des paramètres des sources (EM)

 # Etape E : Calcul des probabilités a posteriori γ_{tk}^j *et* ξ_{tlk}^j

pour tous les $j \in \{1, \ldots, L\}$ *et* $l, k \in \{1, \ldots, K_j\}$ **faire**

pour tous les $t \in \{1, \ldots, T\}$ **faire**

$p(z_{tj}|y_{tj} = k) = \mathcal{N}(z_{tj}; \mu_k^j, \nu_k^j)$

 # Procedure Forward

$\alpha_{1k}^{j(q)} = \pi_k^{j(q)} p(z_{1j}|y_{1j} = k; \theta_k^j)$

pour tous les $t \in \{2, \ldots, T\}$ **faire**

$\alpha_{tk}^{j(q)} = \sum_{l=1}^{K_j} \alpha_{(t-1)l}^{j(q)} A_{lk}^{j(q)} p(z_{tj}|y_{tj} = k; \theta_k^j)$

 # Procedure Backward

$\beta_{Tk}^{j(q)} = 1$

pour tous les $t \in \{T-1, \ldots, 1\}$ **faire**

$\beta_{tl}^{j(q)} = \sum_{k=1}^{K_j} \beta_{(t+1)k}^{j(q)} A_{lk}^{j(q)} p(z_{(t+1)j}|y_{(t+1)j} = k; \theta_k^j)$

pour tous les $t \in \{1, \ldots, T\}$ **faire**

$\gamma_{tk}^{j(q)} = \frac{\alpha_{tk}^{j(q)} \beta_{tk}^{j(q)}}{\sum_{l=1}^{K_j} \alpha_{tl}^{j(q)} \beta_{tl}^{j(q)}}$, $\xi_{tlk}^{j(q)} = \frac{\alpha_{(t-1)l}^{j(q)} A_{lk}^{j(q)} p(z_{tj}|y_{tj}=k_j) \beta_{tk}^{j(q)}}{\sum_{l=1}^{K_j} \sum_{k=1}^{K_j} \alpha_{(t-1)l}^{j(q)} A_{lk}^{j(q)} p(z_{tj}|y_{tj}=k) \beta_{tk}^{j(q)}}$

 # Etape M : Mise à jour des paramètres des sources

pour tous les $i \in \{1, \ldots, L\}$ *et* $l, k \in \{1, \ldots, K_j\}$ **faire**

$\pi_k^{j(q+1)} = \gamma_{1k}^{j(q)}$, $A_{lk}^{j(q+1)} = \frac{\sum_{t=2}^{T} \xi_{tlk}^{j(q)}}{\sum_{t=2}^{T} \gamma_{tk}^{j(q)}}$

$\mu_k^{j(q+1)} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{j(q)} z_{tj}}{\sum_{t=1}^{T} \gamma_{tk}^{j(q)}}$, $\nu_k^{j(q+1)} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{j(q)} (z_{tj} - \mu_k^{j(q+1)})^2}{\sum_{t=1}^{T} \gamma_{tk}^{j(q)}}$

 # Calcul du gradient naturel et Mise à jour de H

$\Delta H = (H^{(q)})(\frac{1}{T} \sum_{t=1}^{T} \mathbf{g}(\mathbf{z}_t^{(q)}) \mathbf{z}_t'^{(q)} - \mathbf{I})$

$H^{(q+1)} = H^{(q)} + \tau . \Delta H$

$q = q + 1$

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

4.4.3 Expérimentations et résultats

Les expérimentations menées dans cette partie ont pour objectif de montrer l'intérêt de prendre en compte l'aspect temporel des données pour l'estimation des sources dans le cadre du diagnostic des CdV. Afin d'étudier la question, une base de données temporelles a été simulée à partir du modèle électrique du CdV [Aknin *et al.* 2003, Oukhellou *et al.* 2006]. Grâce à cette base de données nous disposons d'information sur les variables d'intérêts du problème et nous avons ainsi pu comparer les résultats à travers les différentes approches à de véritables valeurs. Nous présentons dans la section suivante la démarche utilisée pour constituer cette base de données.

Description des données

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 2300 Hz. Trois états de dégradations ont été pris en compte pour fixer les paramètres des lois utilisées pour tirer les capacités, ces dernières ont été tirées en séquences en fonction des paramètres suivants :

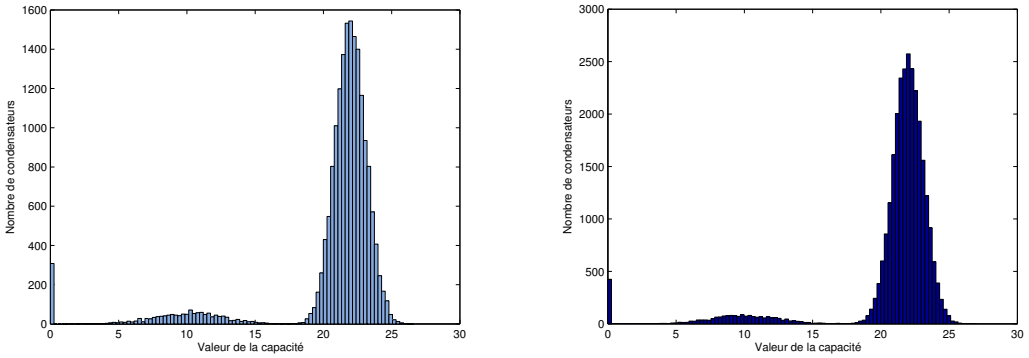
- La loi initiale utilisée supposait que 95% de la population des condensateurs étaient en bon fonctionnement à l'état initial $[0.95, 0.3, 0.2]$ et répond aux spécifications données par le constructeur. Elle suit une loi normale de moyenne égale à la valeur nominale des condensateurs posés en voie $22\mu F$ et de variance 1.26 telle que 95% de cette sous population réponde aux tolérances fournies par le constructeur ($\pm 10\%$) ; la seconde sous population correspond aux condensateurs défectueux. Cette classe représente dans la base 3% des condensateurs. Cette classe suit elle aussi une loi normale dont la moyenne a été fixée à $10\mu F$ et de variance 6 ; la dernière sous population correspond aux condensateurs arrachés. Elle représente 2% des condensateurs. Tous les individus générés dans cette classe se sont vus attribué une capacité égale à 0.
- La matrice représentative des probabilités de transitions entre les états a été fixée de façon empirique. Aucune information relative au vieillissement des condensateurs sur les circuits de voie n'étant disponible, les probabilités de transition ont été définies comme suit :

$$A = \begin{pmatrix} 0.95 & 0.04 & 0.01 \\ 0.5 & 0.45 & 0.05 \\ 0.8 & 0 & 0.2 \end{pmatrix},$$

  où A_{lk} est la probabilité de transition de l'état l vers l'état k telle que $l, k \in \{1, 2, 3\}$ avec 1 pour l'état de bon fonctionnement, 2 pour l'état de dégradation intermédiaire et 3 pour l'état de dégradation grave (condensateur absent).

En utilisant cette démarche, deux bases de données ont été constituées (figure 4.8). La première base contenant 1000 signaux CdV est destinée à servir lors de

l'apprentissage des paramètres de la méthode (base d'apprentissage). La seconde contenant 1500 signaux CdV est prévue pour quantifier les performances des différentes expérimentations (base de test).



(a) *(b)*

FIGURE 4.8 – *Répartition des condensateurs par rapport à la valeur de leur capacité dans la base de données d'apprentissage (a) et la base de données de test (b).*

Description du modèle

L'approche envisagée pour le diagnostic sur données temporelles reposent sur une formulation sans bruit du modèle. Nous avons donc adopté dans le cadre de son application au CdV la même démarche que lors des expérimentations précédentes. Le nombre de variables latentes devant correspondre au nombre de variables observées $2\times L$, nous avons extrait $2\times L$ 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 sont supposées régies par une chaîne de Markov à trois états, qui correspondent aux trois modes de fonctionnement suivants : sans défaut, avec défaut intermédiaire et avec défaut grave. Les variables de bruit ont quant à elles été modélisées à l'aide de variables aléatoires gaussiennes.

Expérimentations

Les expérimentions présentées dans cette partie ont pour but de mettre en évidence l'apport du modèle temporel d'abord dans un cadre d'une approche non supervisée, puis en observant l'évolution de ses performances en présence de labels. Dans l'objectif de quantifier l'apport de cette approche pour le diagnostic, les performances ont été évaluées sur la base de test en observant, les coefficients de corrélation entre les capacités réelles et leur estimation par condensateur d'une part et d'autre part, l'évolution de la corrélation moyenne sur l'ensemble des sources en fonction de la quantité de données labellisées.

Dans le cadre de l'approche totalement non supervisée, le modèle d'IFA avec HMM a été comparé au modèle d'ICA et d'IFA classiques en terme d'estimation des sources. Le tableau 4.2 et la figure 4.9 présentent les degrés de corrélation (en valeur absolue) obtenus entre les sources estimées et les capacités réelles par condensateur et en fonction de la permutation qui conduit à la meilleure corrélation moyenne. Ces résultats sont obtenus en moyennant 30 exécutions de chaque algorithme et avec différentes initialisations de la matrice de mixage H. Dans la figure 4.9 les écarts types sont également représentés.

TABLE 4.2 – *Résultats obtenus à partir du modèle de l'ICA, de l'IFA et de l'IFA avec HMM sur la base de test des données temporelles. Moyennes des valeurs absolues des coefficients de corrélation entre les sources estimées et les capacités des 18 condensateurs sur 30 initialisations différentes de la matrice de mixage.*

Sources	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$< \lvert r_{\hat{c}_i,c_i} \rvert >_{ICA}$	0.83	0.76	0.61	0.71	0.78	0.72	0.71	0.77	0.54
$< \lvert r_{\hat{c}_i,c_i} \rvert >_{IFA}$	0.96	0.95	0.86	0.88	0.88	0.87	0.76	0.80	0.81
$< \lvert r_{\hat{c}_i,c_i} \rvert >_{IFA-HMM}$	0.97	0.96	0.94	0.90	0.91	0.91	0.86	0.85	0.85
Sources	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
$< \lvert r_{\hat{c}_i,c_i} \rvert >_{ICA}$	0.62	0.56	0.65	0.57	0.61	0.53	0.46	0.40	0.41
$< \lvert r_{\hat{c}_i,c_i} \rvert >_{IFA}$	0.76	0.78	0.73	0.72	0.76	0.73	0.70	0.46	0.30
$< \lvert r_{\hat{c}_i,c_i} \rvert >_{IFA-HMM}$	0.84	0.82	0.81	0.82	0.81	0.80	0.76	0.41	0.30

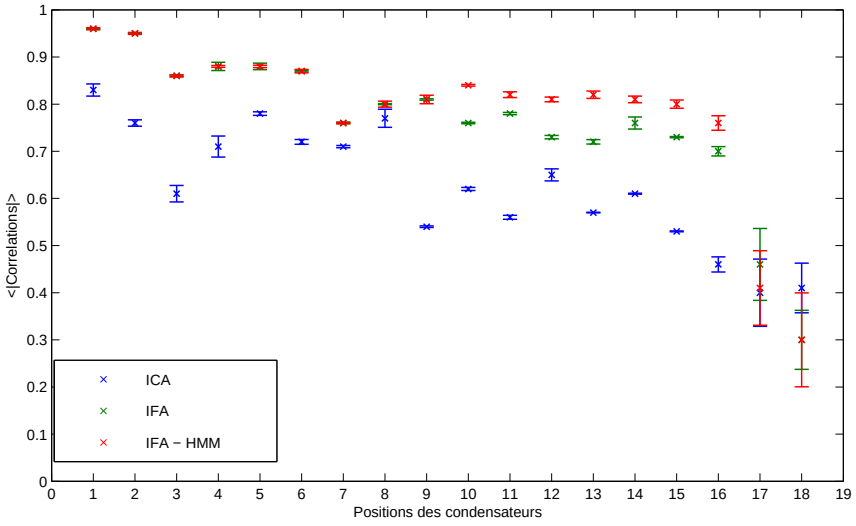


FIGURE 4.9 – *Moyennes et écarts types obtenus à partir des valeurs absolues des coefficients de corrélation entre les sources estimées et les capacités des 18 condensateurs sur 30 initialisations différentes de la matrice de mixage.*

A travers ces résultats (tableau 4.2 et figure 4.9), on peut noter que les performances du modèle d'IFA avec HMM sont meilleures que celles du modèle d'ICA standard et du modèle IFA classique. Les corrélations entre les capacités réelles des condensateurs et leur estimation sont entre à 0.76 et 0.97 pour les plus significatives. Toutefois, les sources situées en fin de circuit de voie (17 et 18) restent difficiles à estimer correctement. Ceci peut être expliqué par le fait qu'à l'extrémité du CdV le signal Icc est influencé par les états de fonctionnement de tous les condensateurs en amont.

Dans la suite des expérimentations nous nous sommes intéressés à la possibilité d'inclure des labels dans le modèle d'IFA avec HMM. En effet, la couche discrète de ce modèle permet de modéliser l'état de fonctionnement des condensateurs et des labels pourraient améliorer ses performances. Afin d'étudier les performances de la méthode dans différents contextes, nous avons fait varier le pourcentage de CdV labellisés de 0% à 100% en utilisant 500 signaux de la base d'apprentissage. 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 moyenne étant prise sur l'ensemble des condensateurs et sans permutations. La figure 4.10 présente l'évolution de ce critère pour l'IFA avec HMM ainsi que pour l'IFA classique.

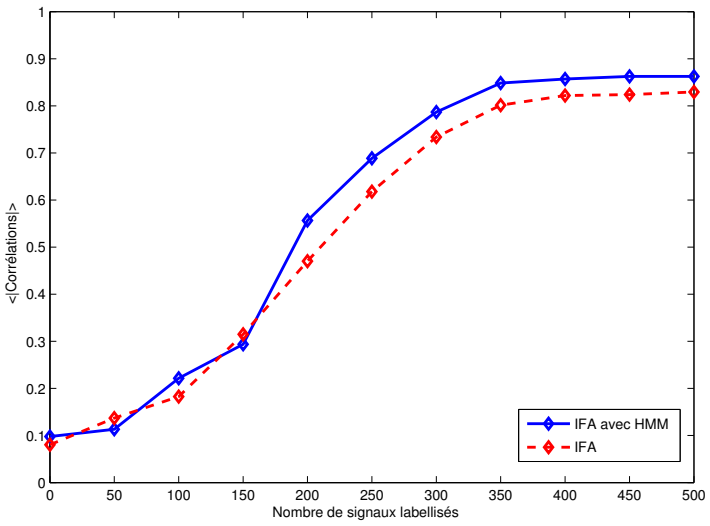


FIGURE 4.10 – *Évolution de la moyenne des coefficients de corrélations entre les sources estimées et les capacités des condensateurs, en fonction du nombre de signaux labellisés. Les différentes solutions ont été obtenues à l'aide du modèle de l'IFA avec HMM et de l'IFA classique à partir de la même initialisation de la matrice H.*

La figure 4.10 met en évidence les avantages d'une approche partiellement supervisée dans le cas des deux modèles. Les deux méthodes semblent suivre les mêmes

tendances avec un léger avantage dans le cas de l'IFA avec HMM. Avec un faible pourcentage de signaux labellisés, la méthode ne permet pas de retrouver les sources dans un ordre correct, d'où les valeurs médiocre du critère sur les premiers points de la courbe. Puis la courbe montre une croissance forte avec une stagnation à partir de 350 signaux labellisés. L'étiquetage permet donc 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 (> 350). Ainsi, il ne semble pas nécessaire d'étiqueter l'ensemble des CdV réels pour obtenir des performances intéressantes. Dans le cas de l'IFA avec HMM 350 signaux sont suffisants pour obtenir des résultats intéressants.

Nous présentons en figure 4.11 des exemples des résultats obtenus avec l'IFA avec HMM sur le jeu de données de 500 signaux dans les trois situations suivantes : non supervisée, avec 200 signaux labellisés et avec 400 signaux labellisés. 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 sur la base de données de test de 1500 signaux.

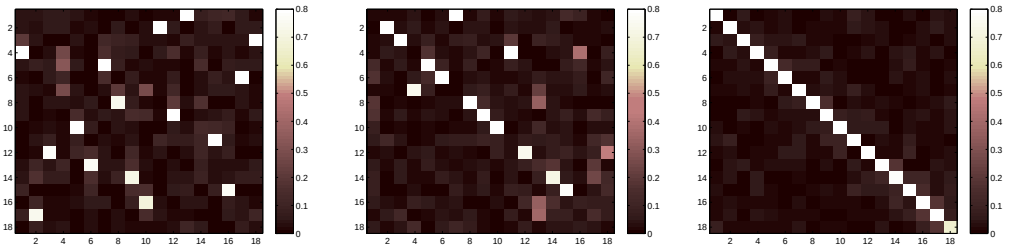


FIGURE 4.11 – *Indétermination du modèle d'IFA avec HMM par rapport aux permutations des sources : matrice des valeurs absolues des corrélations entre les sources estimées et les capacités réelles des condensateurs dans le cas non supervisé (a), avec 200 signaux labellisés (b) et avec 400 signaux labellisés (c).*

Nous pouvons observer sur la figure 4.11, l'intérêt de la labellisation pour résoudre le problème de l'indétermination par rapport aux permutations des sources. 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).

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 lorsqu'une partie des données utilisées est labellisée (meilleur contraste au niveau des matrices). On notera notamment, que même les sources situées en fin de circuit de voie arrivent à être détectées à partir d'une certaine quantité de signaux labellisés (> 350).

En conclusion, l'IFA avec HMM semble plus performante qu'une approche par ICA classique pour une estimation plus fidèle des sources et permet dans le cas de l'introduction de labels, d'une part d'améliorer cet estimation et d'autre part de lever le problème de l'indétermination du modèle par rapport aux permutations des sources. Toutefois, l'idée d'utiliser une telle approche pour le diagnostic des CdV reste peu envisageable pour l'instant, la quantité de données nécessaires pour un apprentissage correct du modèle n'étant pas actuellement disponible.

4.5 Conclusion

Ce chapitre a permis de présenter certaines méthodes pour le diagnostic des CdV dans un contexte non supervisé. La première concernait l'introduction d'information a priori au niveau des liens entre variables latentes et variables observées. Ceci en imposant des contraintes de nullité sur la matrice de mixage ou en intégrant des fonctions de pénalité sur les éléments de cette dernière. Des expériences sur données de circuit de voie simulées nous ont permis de montrer l'influence positive de ce type de contraintes sur les résultats de l'ICA, lorsque ces hypothèses sont étayées évidement.

La seconde contribution proposée visait à intégrer une représentation séquentielle des données dans le modèle pour le cas de données temporelles. La méthode utilisée proposait de modéliser l'ensemble des données par un modèle d'IFA où chaque facteur est décrit comme étant un HMM. Les résultats obtenus sur une base de données temporelles simulée ont permis de montrer l'intérêt d'une telle approche. Nous avons, par ailleurs exploré dans le cadre de cette approche, l'idée d'intégrer quelques labels lors de l'apprentissage et avons pu constater l'avantage qui peut être apporté pour le diagnostic.

Chapitre 5

Diagnostic avec étiquetage incertain et données réelles

Sommaire

5.1 Introduction

Les méthodes présentées dans le chapitre précédent sont essentiellement destinées à traiter des données entièrement non étiquetées. Toutefois, dans le cas de l'application circuit de voie, la quantité de données nécessaire à un bon apprentissage du modèle est trop importante pour que cela puisse être envisagé dès à présent. Par ailleurs, l'idée de travailler sur des données entièrement étiquetées n'est pas toujours réalisable, la récolte et l'obtention de telles étiquettes représente souvent un travail coûteux et fastidieux. Néanmoins, la prise en compte de connaissances expertes pour étiqueter les données peut constituer une réelle alternative.

Afin de répondre aux besoins pratiques rencontrés communément, des approches qualifiées de semi-supervisées [Grandvalet & Bengio 2005, Chapelle *et al.* 2006] et partiellement supervisées [Ambroise & Govaert 2000] ont été largement étudiées.

La possibilité d'utiliser des bases d'apprentissage constituées de données en partie étiquetées [McLachlan 1977] ou l'intégration de données partiellement étiquetées [Ambroise *et al.* 2001] sont des alternatives permettant de combler le coût de l'annotation par des experts. Les performances pour le diagnostic peuvent largement être améliorées en conséquence.

Il est également important de prendre en considération une autre difficulté rencontrée lors du travail d'étiquetage liée à la détermination des classes. En effet, il n'est pas toujours aisé pour un expert d'identifier de manière précise et certaine à quelle classe un individu peut appartenir. Dans le cas des signaux de circuit de voie, cette tâche peut s'avérer très délicate compte tenu de la variation de l'apparence des signaux et de la frontière parfois trop mince entre l'apparition d'un défaut grave ou la dégradation continue d'un défaut intermédiaire.

Ce chapitre décrit l'approche adoptée pour le diagnostic des circuits de voie à partir de signaux réels et prenant en compte les contraintes citées précédemment. Les signaux en question ont été fournis par la SNCF dans le cadre du projet ANR-PREDIT du nom de DIAGHIST, dans l'objectif de mettre en place une politique de maintenance prédictive. Les signaux en question correspondent aux inspections réalisées sur la ligne à grande vitesse LN3 (Paris-Lille) et ont servi à la constitution d'un base de données destinée à l'apprentissage des paramètres du modèle prévu pour le diagnostic.

Afin d'exploiter au mieux la quantité de données disponibles, une approche partiellement supervisée a été mise en œuvre. Pour ce faire, une campagne d'étiquetage a été organisée avec pour objectif de présenter une certaine quantité de signaux présélectionnés à plusieurs experts. Lors de cette opération, les experts étaient autorisés à étiqueter les données en émettant des avis imprécis et incertains lorsqu'il était difficile de prendre une décision franche. Par la suite, le modèle d'IFA partiellement supervisée a été utilisé pour le diagnostic, avec un apprentissage réalisé à partir des données étiquetées par la combinaison des différents avis via la théorie des fonctions de croyance [Dempster 1967, Shafer 1976].

La suite du document détaille certains éléments de cette théorie et démontre l'apport de son utilisation dans le cadre du diagnostic des circuits de voie.

5.2 Théorie des fonctions de croyance

La modélisation d'informations incertaines a longtemps été fortement liée à la théorie des probabilités qui consiste à représenter tout état de connaissance par une mesure de probabilité. Cependant, dans le cas d'informations incertaines, il peut être difficile de donner une mesure proprement dite pour représenter l'ignorance. Ainsi, ces dernières décennies ont vu apparaître d'autres théories telles que la théorie des possibilités initialement proposée par Zadeh [Zadeh 1978] puis considérablement développée par Dubois et Prade [Dubois & Prade 1988], ou la théorie des fonctions de croyance introduite par Dempster dans le cadre de ses travaux sur

les probabilités inférieures et supérieures [Dempster 1967]. Par la suite, Shafer a montré l'intérêt des fonctions de croyance pour la modélisation de connaissances incertaines [Shafer 1976]. Plus tard, ce modèle a été étendu par Smets sous le nom de Modèle des Croyances Transférables (MCT ou TBM pour Tranferable Belief Model) [Smets 1990a, Smets & Kennes 1994].

La théorie des fonctions de croyance est particulièrement avantageuse pour le traitement et la représentation d'informations imprécises et incertaines. Elle constitue, dans le cas discret, une alternative attrayante à la théorie des probabilités qui modélise avant tout les incertitudes ou à la théorie des possibilités qui modélise surtout les imprécisions.

5.2.1 Formalisme du Modèle des Croyances Transférables

Le Modèle des Croyances Transférables (TBM) est un cadre formel défini pour la représentation et la combinaison de connaissances partielles (imprécises et incertaines). Le TBM repose sur l'utilisation de fonctions de croyance afin de modéliser l'état de connaissance d'une source d'information donnée relativement à une certaine question. Lorsque plusieurs sources d'information sont disponibles, celle-ci pouvant être redondantes ou contradictoires, le TBM offre la possibilité de les fusionner pour améliorer l'analyse de la question. Un ensemble d'opérateurs définis dans le cadre du modèle permet la combinaison de ces fonctions.

Le TBM a pour particularité de traiter l'information selon deux niveaux. Il fait l'hypothèse que l'étape de raisonnement dans l'incertain et l'étape de décision sont des tâches de natures différentes. Deux niveaux distincts sont considérés :

- Le niveau crédal qui permet la représentation et la manipulation des informations ;
- Le niveau pignistique qui concerne la prise de décision en fonction éventuellement du risque ou du gain lié à cette décision.

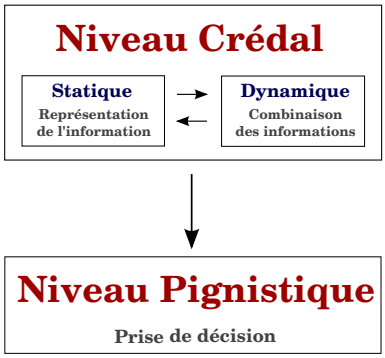


FIGURE 5.1 – *Représentation du fonctionnement global du Modèle des Croyances Transférables*

5.2.2 Modélisation de l'information

La théorie des fonctions de croyance est fondée sur la définition d'un cadre de discernement Ω composé de n hypothèses exclusives et exhaustives $\Omega = \{\omega_1, ..., \omega_n\}$ dans le cas discret. Toute information imparfaite est représentée par une fonction de masse m^Ω définie sur 2^Ω (l'ensemble des parties de Ω) et à valeurs dans $[0, 1]$ telle que :

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1. \tag{5.1}$$

Les sous-ensembles A de masse non nulle sont appelés éléments focaux. La quantité $m^\Omega(A)$ représente la part de croyance allouée à la proposition A par une source donnée. L'expression bba pour *basic belief assignment* est souvent utilisée pour les fonctions de masse.

Une fonction de masse diffère d'une distribution de probabilité par le fait qu'elle donne la possibilité d'allouer de la croyance à des sous-ensembles de Ω et non uniquement à ses singletons. Affecter une masse à une proposition A, différente d'une hypothèse singleton, exprime le fait qu'aucune hypothèse plus spécifique n'est accréditée par les éléments d'évidence disponibles. C'est cette possibilité qui offre une importante souplesse pour la représentation des connaissances.

Une fonction de masse est dite :

- ***normalisée*** si la masse sur l'ensemble vide est nulle $m^\Omega(\emptyset) = 0$;
- ***dogmatique*** si Ω n'est pas un élément focal ;
- ***vide*** si Ω est l'unique élément focal ($m^\Omega(\Omega) = 1$) ;
- ***simple*** si elle contient au plus un ensemble focal en plus de Ω ;
- ***catégorique*** si elle est à la fois simple et dogmatique ;
- ***bayésienne*** si les éléments focaux sont des singletons. La fonction m^Ω est alors équivalente à une distribution de probabilités.

Une fonction de masse simple telle que $m^\Omega(A) = 1 - w$, $\forall A \neq \Omega$ et $m(\Omega) = w$ peut être notée A^w. Ainsi, la fonction de masse vide est notée A^1 pour tout $A \subset \Omega$, et la fonction de masse catégorique est notée A^0 pour tout $A \neq \Omega$.

Degré d'ignorance et ignorance totale

Dans la théorie des fonctions de croyance la notion de degré d'ignorance a une représentation différente de celle adoptée dans le cadre probabiliste. Il est possible d'exprimer l'ignorance quant à la réalisation d'un événement A par une masse sur l'ensemble Ω. La fonction de masse vide définie par $m^\Omega(\Omega) = 1$ traduit l'ignorance totale. En revanche, la représentation d'une connaissance précise et certaine consiste à attribuer toute la masse à un singleton de Ω.

Exemple 5.1 *(Degré d'ignorance)*
Prenons l'exemple d'une course de trois chevaux : c_1 , c_2 , c_3. La question : « Qui va gagner la course ? » est posée à 2 experts. L'expert 1 connaît bien les chevaux participants et considère que les trois sont de même niveau. En revanche l'expert 2 découvre ces chevaux et n'a pas d'idée précise sur la question.
***Modélisation du problème dans le cadre probabiliste** :*
Expert 1 : $p_1(c_1) = p_1(c_2) = p_1(c_3) = \frac{1}{3}$
Expert 2 : $p_2(c_1) = p_2(c_2) = p_2(c_3) = \frac{1}{3}$
Ici deux états de connaissance totalement différents ont une seule et même représentation.
***Modélisation dans le cadre de la théorie des fonctions de croyance** :*
Expert 1 : $m_1(\{c_1\}) = m_1(\{c_2\}) = m_1(\{c_3\}) = \frac{1}{3}$
Expert 2 : $m_2(\{c_1, c_2, c_3\}) = 1$
Il est possible de modéliser explicitement l'ignorance, les deux états de connaissance ne sont pas confondus.

Hypothèse du monde ouvert

Une des particularités qu'offre la théorie des fonctions de croyance est la notion de *monde ouvert*. Le cadre de discernement représente le *monde*. Lorsque ce dernier est exhaustif, il contient nécessairement l'hypothèse valable pour répondre au problème. Dans ce cas, la fonction de masse m est normalisée ; on parle aussi de *monde fermé*. Cette condition imposée à l'origine par Shafer [Shafer 1976] peut être relâchée grâce à l'hypothèse du *monde ouvert* qui stipule que le cadre de discernement Ω peut être incomplet ou non exhaustif [Smets 1990a, Smets & Kennes 1994]. L'ensemble vide $\emptyset$ est alors interprété comme une hypothèse non clairement définie dans Ω, et la quantité $m^\Omega(\emptyset)$ est la part de croyance allouée au fait que la réponse à la question se trouve hors du cadre de discernement.

Plausibilité et crédibilité

À partir d'une fonction de masse notée m, d'autres fonctions équivalentes peuvent être définies pour représenter la même information mais sous une forme différente. Les fonctions de plausibilité et de crédibilité sont les principales fonctions associées.

Plausibilité : Cette fonction est définie de 2^Ω vers $[0, 1]$ indique dans quelle mesure les informations données par une source soutiennent la proposition A. Elle est constituée de la somme des masses des éléments ne contredisant pas A et donne ainsi une tendance maximale ou optimiste :

$$Pl^\Omega(A) = \sum_{B \cap A \neq \emptyset} m^\Omega(B), \quad \forall A \subseteq \Omega. \tag{5.2}$$

La fonction $pl : \Omega \to [0, 1]$ définie par $pl(\omega) = Pl(\{\omega\})$ pour tout $\omega \in \Omega$ est appelée *fonction de contour*.

Crédibilité : Cette fonction est également définie de 2^Ω vers $[0,1]$, elle quantifie la part de croyance totale pouvant soutenir la proposition A. Obtenue à partir de la somme des masses attribuées aux éléments impliquant A, elle donne ainsi une tendance minimale ou pessimiste :

$$Bel^\Omega(A) = \sum_{\emptyset \neq B \subseteq A} m^\Omega(B), \quad \forall A \subseteq \Omega. \tag{5.3}$$

Ces deux fonctions, sont liées par la relation suivante :

$$Pl^\Omega(A) = 1 - Bel^\Omega(\overline{A}). \tag{5.4}$$

Exemple 5.2 *(Plausibilité et crédibilité)*
Considérons le cas d'un patient qui aurait développé des symptômes pouvant suggérer trois maladies différentes $mal1$, $mal2$ *et* $mal3$. *Le médecin qui examine le patient donne le diagnostic suivant : « Le malade a peu de chance d'avoir contracté la maladie 3, mais il est difficile de déterminer s'il s'agit de la maladie 1 ou 2 ». Le cadre de discernement* $\Omega = \{mal1, mal2, mal3\}$ *est constitué des trois hypothèses :*
$mal1$ *: le patient est atteint de la maladie 1 ;*
$mal2$ *: le patient est atteint de la maladie 2 ;*
$mal3$ *: le patient est atteint de la maladie 3.*
Les informations fournies pas le médecin peuvent être représentées dans le cadre de la théorie des fonctions de croyance par :

TABLE 5.1 – *Représentation de l'information dans le cadre de la théorie des fonctions de croyance. Distribution de la fonction de masse et des fonctions de plausibilité et crédibilité associées.*

A	$m^\Omega(A)$	$Pl^\Omega(A)$	$Bel^\Omega(A)$
$\emptyset$	0	0	0
$\{mal1\}$	0	0.9	0
$\{mal2\}$	0	0.9	0
$\{mal3\}$	0.1	0.1	0.1
$\{mal1, mal2\}$	0.9	0.9	0.9
$\{mal1, mal3\}$	0	1	0.1
$\{mal2, mal3\}$	0	1	0.1
Ω	0	1	1

5.2.3 Fusion d'informations

De manière globale la fusion de données décrit l'ensemble des techniques qui permettent d'agréger différentes informations complémentaires, redondantes ou incomplètes pour obtenir une meilleure connaissance de l'environnement étudié. Dans

le cadre du TBM, l'étape de fusion représente la partie dynamique du niveau crédal et a pour objectif d'exploiter l'ensemble des connaissances disponibles avant le niveau décisionnel.

La théorie des fonctions de croyance offre plusieurs opérateurs qui permettent de combiner les informations issues de sources multiples. Ces règles de combinaison ont pour but d'agréger les connaissances des différentes sources pour construire une information globale pertinente. Afin de conserver un maximum d'information, les règles en question sont définies au niveau crédal et manipulent les fonctions de croyance durant l'étape de combinaison.

Combinaison conjonctive

Soient m_1^Ω et m_2^Ω deux fonctions de masses issues de deux sources fiables distinctes et définies sur un même cadre de discernement Ω. La combinaison conjonctive de m_1^Ω et de m_2^Ω, notée $m_{1\textcircled{\cap}2}^\Omega$, est définie par :

$$m_{1\textcircled{\cap}2}^\Omega(C) = \sum_{A\cap B=C} m_1^\Omega(A) m_2^\Omega(B), \quad \forall C \subseteq \Omega. \tag{5.5}$$

L'opérateur de combinaison conjonctive joue un rôle central en théorie des fonctions de croyance. Il permet de transférer la masse sur des sous-ensembles de cardinalité plus faible, ce qui peut être interprété comme une spécialisation [Smets 2002]. Cet opérateur est associatif, commutatif et non idempotent $m_1^\Omega \textcircled{\cap} m_1^\Omega \neq m_1^\Omega$. Ainsi, il ne peut être utilisé pour combiner des sources non distinctes. Il admet comme élément neutre la fonction de masse vide et comme élément absorbant la fonction de masse telle que $m^\Omega(\emptyset) = 1$.

La quantité $m_{1\textcircled{\cap}2}^\Omega(\emptyset)$ est représentative du *degré de conflit* entre m_1^Ω et m_2^Ω et résulte des intersections vides entre hypothèses incompatibles. Dans les applications de fusion de données, il est souvent intéressant de pouvoir mettre en évidence le désaccord entre les éléments d'information et d'en interpréter les raisons en fonction de l'application. Toutefois, afin de préserver l'hypothèse du monde fermé, il est possible dans certains cas de répartir le conflit en normalisant le résultat de la combinaison [Dempster 1967, Dempster 1968].

Combinaison conjonctive normalisée

Soient m_1^Ω et m_2^Ω deux fonctions de masses issues de deux sources distinctes, supposées fiables et définies sur un même cadre de discernement Ω. La combinaison conjonctive normalisée de m_1^Ω et de m_2^Ω, notée $m_{1\oplus 2}^\Omega$, est définie par :

$$m_{1\oplus 2}^\Omega(C) = \begin{cases} \frac{m_{1\textcircled{\cap}2}^\Omega(C)}{1-m_{1\textcircled{\cap}2}^\Omega(\emptyset)} & \text{si } \forall C \subseteq \Omega, C \neq \emptyset \\ 0 & \text{si } C = \emptyset. \end{cases} \tag{5.6}$$

La règle de combinaison conjonctive normalisée est également appelée *règle de combinaison de Dempster*.

Conditionnement

Le conditionnement d'une fonction de masse m^Ω par un sous ensemble $B \subseteq \Omega$, consiste à redistribuer la masse des propositions possibles sur celles issues de l'intersection avec B.

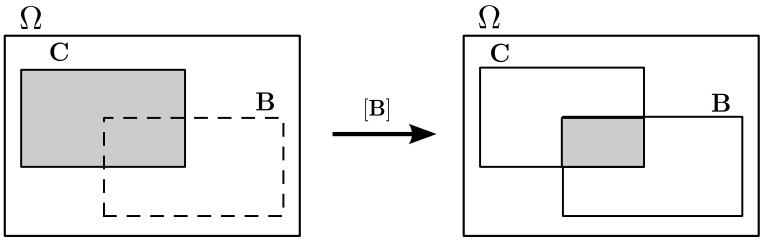


FIGURE 5.2 – *Conditionnement*

Ainsi, toute masse affectée à $C \subseteq \Omega$ est transférée sur $C \cap B$. La fonction de masse conditionnelle à B non normalisée, notée $m^\Omega[B](.)$, est définie par :

$$m^\Omega[B](A) = \begin{cases} \sum_{C \cap B = A} m^\Omega(C) & \text{si } A \subseteq B, \\ 0 & \text{sinon.} \end{cases} \tag{5.7}$$

Le conditionnement est un cas particulier de la combinaison conjonctive. Il revient à effectuer une combinaison conjonctive avec une fonction de masse *catégorique* $m^\Omega(B) = 1$. Il est donc particulièrement adapté aux cas où une information certaine est disponible pour la fusion.

Combinaison disjonctive

Soient m_1^Ω et m_2^Ω deux fonctions de masses issues de deux sources distinctes considérées comme non fiables et définies sur le cadre de discernement Ω. La combinaison disjonctive de m_1^Ω et de m_2^Ω notée $m_{1\textcircled{\cup}2}^\Omega$ est définie par :

$$m_{1\textcircled{\cup}2}^\Omega(C) = \sum_{A \cup B = C} m_1^\Omega(A) m_2^\Omega(B), \quad \forall C \subseteq \Omega. \tag{5.8}$$

La combinaison disjonctive est utile dans les cas où il est difficile de quantifier la fiabilité des sources [Dubois & Prade 1986, Smets 1993]. Cet opérateur permet de transférer la masse sur des ensembles focaux de cardinalité plus élevée, ce qui peut être considéré comme une procédure de généralisation des fonctions de masses [Smets 2002]. La combinaison disjonctive est une règle associative, commutative mais non idempotente, qui admet comme élément neutre $m^\Omega(\emptyset) = 1$ et comme élément absorbant la fonction de masse vide.

Combinaison conjonctive prudente

Les opérateurs introduits précédemment sont utiles dans les cas où les sources sont considérées comme indépendantes. Dans le cas où cette hypothèse n'est pas véri-

fiée, la règle de combinaison prudente $\textcircled{\wedge}$ introduite dans [Denoeux 2008] est adaptée à la combinaison de fonctions de masse issues de sources dépendantes. Cette règle peut être appliquée à toute fonction de masse non dogmatique, mais sera rappelée ici uniquement pour les fonctions de masse *séparables*, i.e., décomposables en combinaisons conjonctives de fonctions de masse simples [Shafer 1976][Smets 1995]. Soient m_1^Ω et m_2^Ω deux fonctions de masse données par :

$$\begin{aligned} m_1^\Omega &= \textcircled{\cap}_{A\subset\Omega} A^{w_1(A)}, \\ m_2^\Omega &= \textcircled{\cap}_{A\subset\Omega} A^{w_2(A)}, \end{aligned}$$

où $A^{w_1(A)}$ et $A^{w_2(A)}$ sont des fonctions de masse simples, telles que $w_1(A) \in (0,1]$ et $w_2(A) \in (0,1]$, pour tout $A \subset \Omega$. Leur combinaison par la règle prudente est définie par :

$$(m_1^\Omega \textcircled{\wedge} m_2^\Omega)(A) = \textcircled{\cap}_{A\subset\Omega} A^{w_1(A)\wedge w_2(A)}, \tag{5.9}$$

où $\wedge$ représente l'opérateur min.

Cette règle de combinaison a pour but d'éviter de comptabiliser deux fois l'information issue de sources non distinctes grâce à la propriété d'*idempotence* : $m \textcircled{\wedge} m = m$ pour tout m. Elle est par ailleurs commutative et associative.

5.2.4 Opérations sur le cadre de discernement

Les mécanismes pour la combinaison des fonction de croyance présentés jusque là, reposent sur l'hypothèse que les fonctions de masse à combiner sont définies sur le même cadre de discernement. Toutefois, cette condition n'étant pas toujours valable, des opérations permettant d'adapter les cadres de discernement sont possibles.

Raffinement et grossissement

L'opération de *raffinement* consiste à exprimer une fonction de masse définie sur un cadre de discernement Ω dans un autre cadre de discernement plus fin Θ. Elle revient ainsi à considérer que chaque singleton de Ω est représentatif d'un ensemble d'hypothèses plus détaillé de Θ (figure 5.3).

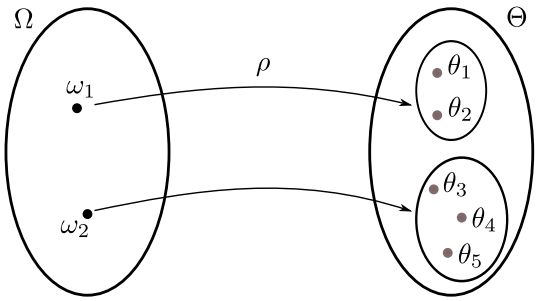


FIGURE 5.3 – *Raffinement du cadre de discernement* Ω.

Soient Ω et Θ deux ensembles finis. L'application ρ définie de 2^Ω vers 2^Θ est un raffinement si elle vérifie les propriétés suivantes :

1. L'ensemble $\{\rho(\{\omega\}), \omega \in \Omega\} \subseteq 2^\Theta$ est une partition de Θ.
2. Quelque soit $A \subseteq \Omega$, $\rho(A) = \bigcup_{\omega \in A} \rho(\{\omega\})$.

La redistribution des masses de croyance initialement définies dans Ω sur Θ est donnée par :

$$m^\Omega(\rho(A)) = m^\Theta(A), \quad \forall A \subseteq \Omega. \tag{5.10}$$

Si Θ est un *raffinement* de l'ensemble Ω, alors Ω est un *grossissement* de Θ. Le grossissement est donc l'opération qui permet de réduire le cadre de discernement en le rendant plus grossier.

Extension vide et marginalisation

L'*extension vide* permet de combiner des fonctions de masse définies sur des cadres de discernement différents Ω et Θ. L'opération consiste à étendre les fonctions de masse sur un espace commun $\Omega \times \Theta$ appelé espace produit dans le but de pouvoir les combiner. L'extension d'une fonction de masse m^Ω, notée $m^{\Omega\uparrow\Omega\times\Theta}$, est définie par :

$$m^{\Omega\uparrow\Omega\times\Theta}(B) = \begin{cases} m^\Omega(A) & \text{si } B = A \times \Theta \text{ pour un } A \subseteq \Omega, \\ 0 & \text{sinon.} \end{cases} \tag{5.11}$$

Ainsi, la combinaison conjonctive de deux fonctions de masses m^Ω et m^Θ peut être obtenue en combinant leurs extensions vides sur $\Omega \times \Theta$ (figure 5.4) :

$$m_1^\Omega \textcircled{\cap} m_2^\Theta = m_1^{\Omega\uparrow\Omega\times\Theta} \textcircled{\cap} m_2^{\Theta\uparrow\Omega\times\Theta}. \tag{5.12}$$

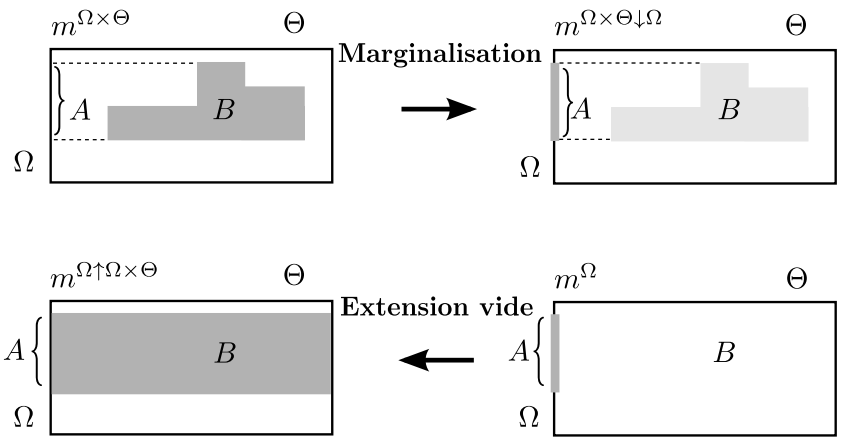


FIGURE 5.4 – *Marginalisation et extension vide*

La *marginalisation* permet d'effectuer l'opération inverse (figure 5.4). Lorsque le cadre de discernement d'une fonction de masse est défini par un espace produit $\Omega \times \Theta$, il est possible de passer à une fonction masse définie sur un des sous-espaces. Par exemple, la marginalisation sur Ω d'une fonction de masse $m^{\Omega\times\Theta}$ initialement définie sur l'espace produit est donnée par :

$$m^{\Omega\times\Theta\downarrow\Omega}(A) = \sum_{\{B\subseteq\Omega\times\Theta|(B\downarrow\Omega)=A\}} m^{\Omega\times\Theta}(B), \tag{5.13}$$

où $(B \downarrow \Omega)$ représente la projection de B sur Ω, définie par :

$$(B \downarrow \Omega) = \{\omega \in \Omega \mid \exists\, \theta \in \Theta, (\omega, \theta) \in B\}, \tag{5.14}$$

5.2.5 Notion d'indépendance

Différentes notions d'indépendance peuvent être définies dans le cadre de la théorie des fonctions de croyance [Ben Yaghlane *et al.* 2000], la forme la plus simple d'indépendance étant l'*indépendance cognitive* [Shafer 1976] définie ci-dessous.

Indépendance cognitive

Soient m^Ω et m^Θ les fonctions de masse définies sur Ω et Θ et soit Pl^Ω et Pl^Θ les fonctions de plausibilité correspondantes. Les cadres Ω et Θ sont dits *cognitivement indépendants* [Shafer 1976, page 149] par rapport à $m^{\Omega\times\Theta}$ si les égalités suivantes sont vérifiées :

$$Pl^{\Omega\times\Theta}(A \times B) = Pl^\Omega(A) Pl^\Theta(B), \tag{5.15}$$

pour tout $A \subseteq \Omega$ et $B \subseteq \Theta$ et où m^Ω et m^Θ représentent la marginalisation de la fonction de masse $m^{\Omega\times\Theta}$ sur Ω, respectivement sur Θ.

Comme le montre Shafer [Shafer 1976], cette propriété signifie qu'un nouvel élément d'évidence sur l'une des variables n'affecte pas nos croyances dans l'autre variable.

5.2.6 Prise de décision

L'intérêt de la fusion d'informations est d'obtenir une fonction de croyance unique qui synthétise l'ensemble des connaissances disponibles en vue de l'étape de prise de décision. Le fait de combiner des fonctions de croyance permet de rester à un niveau crédal et de conserver un maximum d'informations dans la fonction de masse résultante.

La prise de décision consiste à choisir, parmi un ensemble fini d'hypothèses possibles, celle qui répond le mieux au problème posé. Les fonctions de plausibilité et de crédibilité peuvent servir de support en sélectionnant l'hypothèse ayant le degré de doute le plus faible ou l'hypothèse ayant le degré de crédibilité le plus élevé.

Toutefois, la décision pouvant émaner du maximum de plausibilité ou du maximum de crédibilité est souvent trop optimiste ou trop pessimiste. Ainsi, le modèle de croyance transférable s'appuie pour la phase de décision sur la fonction de probabilité pignistique qui constitue un compromis entre les deux approches [Smets 1990b, Smets 2005].

Transformation pignistique

La transformation pignistique permet de convertir une fonction de masse en distribution de probabilité. Cette dernière est obtenue en répartissant à parts égales la masse d'une proposition A sur les hypothèses élémentaires contenues dans A, pour tout $A \subseteq \Omega$. La transformation pignistique d'une fonction de masse m^Ω est représentée par la fonction $betP_m^\Omega$ définie de Ω dans $[0,1]$ telle que :

$$betP_m^\Omega(\omega) = \sum_{\{A \neq \emptyset \,,\, A \subseteq \Omega \,,\, \omega \in A\}} \frac{m(A)}{|A|(1 - m^\Omega(\emptyset))}, \quad \forall \omega \in \Omega. \tag{5.16}$$

Remarque 5.1 *Il est à noter que les fonctions de plausibilité, crédibilité et probabilité pignistique sont des fonctions croissantes pour l'inclusion. Par conséquent la décision doit être considérée uniquement sur les hypothèses singletons.*

5.2.7 Gestion du conflit

Le problème du conflit dans la théorie des fonctions de croyance est dû à l'incompatibilité entre certaines hypothèses qui entraîne des intersections vides lors de la combinaison conjonctive. Les sources d'information pouvant s'exprimer sur 2^Ω, l'apparition de conflit est très probable. De plus, l'ensemble vide étant absorbant par l'opérateur conjonctif, le degré de conflit a tendance à croître avec le nombre de sources d'information à fusionner. Plusieurs stratégies sont possibles pour la gestion du conflit :

Cadre de discernement incomplet

Dans cette approche il est considéré que le conflit ne peut provenir que d'un problème mal posé ou autrement dit d'un cadre de discernement non exhaustif, conformément à l'hypothèse du monde ouvert [Smets 1990a]. Toute la masse conflictuelle est affectée à l'ensemble vide, la combinaison correspondant ainsi à la combinaison conjonctive (équation 5.5).

Afin de rester dans un monde fermé, une approche similaire consiste à introduire un nouvel élément e dans le cadre de discernement, destiné à supporter toute la masse conflictuelle [Yager 1983] :

$$\begin{aligned} m^\Omega(A) &= m_{\textcircled{\cap}}^\Omega(A), \quad \forall A \neq \emptyset, \\ m^\Omega(e) &= m_{\textcircled{\cap}}^\Omega(\emptyset). \end{aligned} \tag{5.17}$$

Affaiblissement

Ce processus consiste à associer un indice de fiabilité aux sources d'information afin de redéfinir leur fonction de masse [Shafer 1976]. L'affaiblissement d'une fonction de masse m^Ω est défini comme la pondération de chaque masse $m^\Omega(A)$ de la distribution par un coefficient $\tau \in [0,1]$ appelé *coefficient d'affaiblissement*. La redistribution des masses est obtenue par :

$$\begin{aligned} {}^\tau m^\Omega(A) &= \tau \cdot m^\Omega(A), \qquad \forall A \subset \Omega, \\ {}^\tau m^\Omega(\Omega) &= \tau \cdot m^\Omega(\Omega) + (1-\tau). \end{aligned} \tag{5.18}$$

Après affaiblissement, une partie de la masse est donc reportée sur l'élément d'ignorance Ω. Une source considérée comme totalement fiable ne doit pas être affaiblie ; par conséquent, plus la source est fiable plus τ est proche de 1.

Répartition du conflit

Dans cette approche, le conflit généré est expliqué par un manque de fiabilité des sources d'information. Les opérateurs présentés par [Yager 1987] et par [Dubois & Prade 1988] partent de ce principe et reposent sur l'idée de répartir le conflit lors de la combinaison.

Dans le cas de l'opérateur de Yager, il est supposé que l'une des sources intervenant dans la combinaison est fiable sans pouvoir définir laquelle. Ainsi, la solution étant obligatoirement dans le référentiel Ω, Yager propose d'attribuer la masse conflictuelle à l'ensemble Ω. La fonction de masse résultante m_Y^Ω est alors obtenue comme suit :

$$\begin{aligned} m_Y^\Omega(A) &= m_{\textcircled{\cap}}^\Omega(A), \quad \forall A \subset \Omega, A \neq \emptyset, \\ m_Y^\Omega(\Omega) &= m_{\textcircled{\cap}}^\Omega(\Omega) + m_{\textcircled{\cap}}^\Omega(\emptyset), \\ m_Y^\Omega(\emptyset) &= 0. \end{aligned} \tag{5.19}$$

La combinaison proposée par Dubois et Prade consiste à gérer le conflit partiel en répartissant toute masse conflictuelle (affectée par deux sources à des hypothèses incompatibles) sur l'ignorance partielle représentée par l'union des ces hypothèses. Cette règle est donnée par :

$$\begin{aligned} m_{DP}^\Omega(A) &= m_{1\textcircled{\cap}2}^\Omega(A) + \sum_{\substack{B \cup C = A \\ B \cap C = \emptyset}} m_1^\Omega(B) m_2^\Omega(C), \quad \forall A \subseteq \Omega, A \neq \emptyset, \\ m_{DP}^\Omega(\emptyset) &= 0. \end{aligned} \tag{5.20}$$

Notons que d'autres règles de combinaison permettant de gérer le conflit ont été proposées par la suite [Florea *et al.* 2006, Martin & Osswald 2007].

A ce stade, l'essentiel des principes relatifs à la théorie de fonctions de croyance a été introduit. La suite du chapitre détaille leur utilisation dans le cadre de la mise au point d'un modèle pour le diagnostic des CdV de la ligne à grande vitesse TGV Nord (LN3).

5.3 Apprentissage partiellement supervisé sur données réelles

Nous avons pu constater au chapitre précédent l'influence de facteurs tels que la quantité de données disponibles ou le pourcentage de données étiquetées sur l'apprentissage du modèle et de ce fait sur les performances du diagnostic. Dans le but d'optimiser la procédure de diagnostic des circuits de voie, plusieurs points ont été considérés afin de définir l'approche adéquate pour l'apprentissage à partir des données fournies par la SNCF.

Le premier point consistait à définir l'approche la plus adaptée compte tenu des données disponibles. Nous avons ainsi envisagé une méthode reposant sur des résultats détaillés dans les travaux de thèse de E. Côme [Côme 2009]. Ces derniers portaient sur l'évolution des performances du modèle IFA en fonction de la quantité de données étiquetées. Il nous a donc semblé pertinent de postuler un tel modèle qui d'une part permet une estimation de la gravité des défauts sur les condensateurs, et qui d'autre part suppose les individus comme indépendants ce qui laisse la possibilité d'utiliser beaucoup plus de signaux que dans le cas temporel (section 4.4).

Le second point reposait sur l'avantage que peut représenter la présence de labels lors de la phase d'apprentissage du modèle. En effet, compte tenu de la quantité de données présentes, un apprentissage non supervisé était à éviter. Toutefois, l'idée d'un contexte totalement supervisé impliquait un travail considérable d'étiquetage beaucoup trop coûteux et fastidieux dans le cas des CdV. Il paraissait donc judicieux de mettre à profit l'expertise existante pour fournir des connaissances partielles sur la présence et l'apparition de défaut sur les condensateurs. L'étiquetage obtenu avec une telle approche est alors considéré comme imprécis et incertain mais peut constituer un apport non négligeable lors de l'apprentissage des paramètres du modèle. Nous nous sommes tournés en conséquence vers une formulation partiellement supervisée du modèle IFA pour l'estimation des paramètres.

Le troisième point envisagé reposait donc sur l'idée de faire étiqueter les données réelles destinées à l'apprentissage par différents experts, en permettant à ces derniers d'émettre des labels imprécis et incertains quant à l'état de fonctionnement des composants des CdV. En prenant en compte les difficultés liées à l'opération d'étiquetage et aux conséquences d'une labellisation trop imprécise, nous avons fait le choix de fusionner les différents avis afin d'exploiter au mieux des labels entachés d'incertitude et d'imprécision.

Cette section décrit le modèle d'IFA partiellement supervisée utilisé lors de l'apprentissage, détaille la démarche adoptée pour l'étiquetage des données et présente

les résultats obtenus dans le cadre du diagnostic des circuits de voie.

5.3.1 IFA partiellement supervisée

L'apprentissage du modèle IFA repose généralement sur la seule observation des mélanges. Le contexte d'apprentissage est alors non supervisé. Dernièrement, certains travaux [Côme 2009, Côme *et al.* 2009, Denoeux 2010] ont permis d'étendre ce type de modèles afin de prendre en compte des informations partielles sur les individus.

Dans sa représentation, le modèle d'IFA partiellement supervisée postulé ici correspond à celui de l'IFA classique (figure 5.5). Chaque source est modélisée par un mélange de plusieurs composantes gaussiennes qui représentent les différents états discrets de la source. Toutefois dans sa formulation, il a pour avantage de pouvoir intégrer lors de l'apprentissage, les connaissances disponibles sur la composante d'origine de certains individus (celles-ci pouvant être imprécises et incertaines).

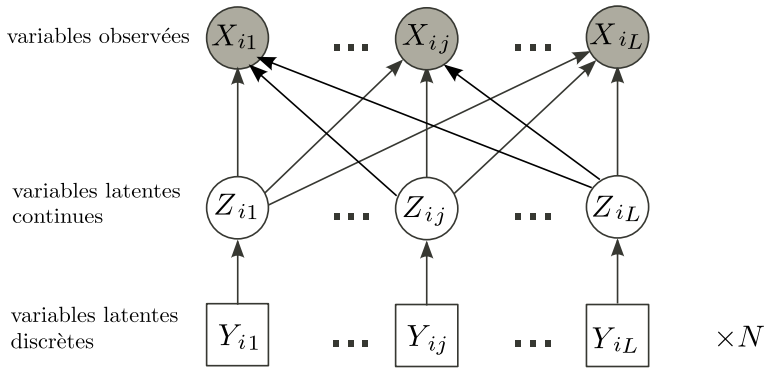


FIGURE 5.5 – *Modèle graphique génératif pour l'Analyse en Facteurs Indépendants (IFA).*

Définition du modèle

Rappelons que dans le cadre de l'IFA sans bruit, il s'agit d'estimer le vecteur de paramètres $\boldsymbol{\psi}$ contenant la matrice de mixage H et les paramètres des densités des sources. Soit :

$$\boldsymbol{\psi} = (H, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^L, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^L, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^L). \tag{5.21}$$

On considère ici le modèle d'IFA dans un cadre d'apprentissage partiellement supervisé, où une connaissance partielle sur la composante d'origine de certains échantillons est disponible sous la forme d'une fonction de croyance. Dans le cas général, nous allons considérer un ensemble d'apprentissage de la forme :

$$\mathcal{M} = \{(\mathbf{x}_1, m_1^1, \ldots, m_1^L), \ldots, (\mathbf{x}_N, m_N^1, \ldots, m_N^L)\}, \tag{5.22}$$

où chaque individu i est décrit par les observations contenues dans le vecteur $\mathbf{x}_i$ et un ensemble de fonctions de masse $m_i^1, \ldots, m_i^L$ spécifiant notre connaissance sur la composante d'origine de l'individu i pour chacune des L sources. Chaque fonction de masse m_i^j est définie sur le cadre de discernement $\mathcal{Y}^j = \{1, \ldots, K_j\}$ composé de toutes les classes possibles pour la source j.

Il faut souligner que, dans le modèle considéré ici, deux types d'incertitude sont présentes : l'incertitude *aléatoire* induite par le processus aléatoire de génération des données $\mathbf{x}_i$, et l'incertitude *épistémique* induite par la connaissance imparfaite de l'appartenance aux classes du mélange. Ce genre de problème d'estimation a été étudié dans le cas particulier des modèles de mélanges dans [Côme *et al.* 2009] et dans un cadre plus général dans [Denoeux 2010]. Dans les deux cas, une généralisation de la fonction de vraisemblance a été présentée ainsi qu'une extension de l'algorithme EM pour sa maximisation.

Notons $\mathbf{x}_i^c = (\mathbf{x}_i, y_{i1}, \ldots, y_{iL})$ les données complètes telles que $\mathbf{x}_i \in \mathrm{R}^L$ sont les variables observées et $y_{ij} \in \mathcal{Y}^j$, $\forall j \in \{1, \ldots, L\}$ sont les variables latentes discrètes codant l'appartenance aux classes du mélange. L'incertitude aléatoire induite par le processus de génération aléatoire permet d'émettre l'hypothèse d'indépendance stochastique entre les réalisations comme dans l'IFA classique :

$$f(\mathbf{X}^c; \boldsymbol{\psi}) = \prod_{i=1}^{N} f(\mathbf{x}_i^c; \boldsymbol{\psi}), \tag{5.23}$$

où $\mathbf{X}^c = (\mathbf{x}_1^c, \ldots, \mathbf{x}_N^c)$ est le vecteur des observations complètes et $f(\mathbf{x}_i^c)$ la fonction de densité d'une observation complète selon le modèle de l'IFA :

$$f(\mathbf{x}_i^c; \boldsymbol{\psi}) = \frac{1}{|\det(H)|} \prod_{j=1}^{L} \prod_{k=1}^{K_j} \left(\pi_k^j \mathcal{N}((H^{-1}\mathbf{x})_j; \mu_k^j, \nu_k^j) \right)^{\mathbb{1}\{y_{ij}=k\}}. \tag{5.24}$$

En outre, l'incertitude épistémique induite par la perception imparfaite de l'appartenance à l'une des classes du mélange permet de poser l'hypothèse d'indépendance cognitive suivante (5.15) :

$$pl(\mathbf{X}^c) = \prod_{i=1}^{N} pl_i(\mathbf{x}_i^c) = \prod_{i=1}^{N} \prod_{j=1}^{L} pl_i^j(y_{ij}), \tag{5.25}$$

où $pl(\mathbf{X}^c)$ est la plausibilité que le vecteur des observations complètes soit égale à $\mathbf{X}^c$, $pl_i(\mathbf{x}_i^c)$ est la plausibilité que les données complètes pour l'échantillon i soit $\mathbf{x}_i^c$ et $pl_i^j(y_{ij})$ est la plausibilité que la source j de l'échantillon i soit générée à partir de la composante y_{ij}.

Il est à noter que les hypothèses (5.23) et (5.25) ne sont pas liées : la première est une propriété du processus aléatoire de génération des données, tandis que la deuxième se rapporte au processus d'observation qui est incertain. Sous ces deux

hypothèses et selon [Denoeux 2010], la log-vraisemblance des données observées peut être écrite comme :

$$\mathcal{L}(\boldsymbol{\psi}\ ;\mathcal{M}) = \sum_{i=1}^{N} \log \mathrm{E}_{\boldsymbol{\psi}}[pl_i(\mathbf{x}_i^c)] = \sum_{i=1}^{N} \log \int_{\mathcal{X}} f(\mathbf{x}_i^c;\boldsymbol{\psi}) pl_i(\mathbf{x}_i^c) d\mathbf{x}^c \tag{5.26}$$

$$= -N \log(|\det(H)|) + \sum_{i=1}^{N}\sum_{j=1}^{L} \log\left(\sum_{k=1}^{K_j} pl_{ik}^j \pi_k^j \mathcal{N}((H^{-1}\mathbf{x}_i)_j; \mu_k^j, \nu_k^j)\right) \tag{5.27}$$

où $pl_{ik}^j = pl_i^j(k)$ est la plausibilité (calculée à partir de la fonction de masse m_i^j) que l'échantillon i de la variable latente j appartienne à la classe k.

Ce critère doit être maximisé par rapport à $\boldsymbol{\psi}$ pour l'estimation des paramètres. Une extension de l'algorithme EM appelé E^2M pour *Evidential* EM peut être utilisée pour effectuer cette tâche [Denoeux 2010]. Le paragraphe suivant présente cette extension pour le modèle de l'IFA.

Algorithme EM pour l'apprentissage partiellement supervisé de l'IFA

Comme l'algorithme EM classique, l'algorithme E^2M utilise la log-vraisemblance des données complètes, qui est égale à la quantité suivante dans le modèle de l'IFA :

$$\mathcal{L}(\boldsymbol{\psi};\mathbf{X}^c) = -N\log(|\det(H)|) + \sum_{i=1}^{N}\sum_{j=1}^{L}\sum_{k=1}^{K_j} \mathbb{1}_{\{y_{ij}=k\}} \log\left(\pi_k^j \mathcal{N}((H^{-1}\mathbf{x}_i)_j;\mu_k^j,\nu_k^j)\right). \tag{5.28}$$

Notons $f(\mathbf{x}_i^c|\mathcal{M};\boldsymbol{\psi}^{(q)})$ la fonction de densité conditionnelle obtenue en combinant $\mathcal{M}$ avec la fonction de densité des données complètes $f(\mathbf{x}_i^c;\boldsymbol{\psi})$ en utilisant la règle de Dempster [Denoeux 2010]. L'espérance conditionnelle de $\mathcal{L}(\boldsymbol{\psi};\mathbf{X}^c)$ par rapport à $f(\mathbf{x}_i^c|\mathcal{M};\boldsymbol{\psi}^{(q)})$ définit la fonction auxiliaire $Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(q)})$ qui sera maximisée lors de l'étape M de l'algorithme :

$$Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(q)}) = \mathrm{E}_{\boldsymbol{\psi}^{(q)}}\left[\mathcal{L}(\boldsymbol{\psi};\mathbf{X}^c)|\mathcal{M}\right] \tag{5.29}$$

$$= -N\log(|\det(H)|) + \sum_{i=1}^{N}\sum_{j=1}^{S}\sum_{k=1}^{K_j} t_{ik}^j \log\left(\pi_k^j \mathcal{N}((H^{-1}\mathbf{x}_i)_j;\mu_k^j,\nu_k^j)\right), \tag{5.30}$$

où t_{ik}^j est la probabilité a posteriori que l'échantillon i appartienne à la classe k pour la variable latente j compte tenu des observations $\mathbf{x}_i$, du label m_i^j et de l'estimation courante du vecteur des paramètres $\boldsymbol{\psi}^{(q)}$ où q est l'itération. Dans l'étape E de l'algorithme, ces probabilités a posteriori sont calculées comme suit :

$$t_{ik}^{j(q)} = \frac{pl_{ik}^j \pi_k^{j(q)} \mathcal{N}(z_{ij}^{(q)};\mu_k^{j(q)},\nu_k^{j(q)})}{\sum_{l=1}^{K_j} pl_{ik'}^j \pi_l^{j(q)} \mathcal{N}(z_{ij}^{(q)};\mu_l^{j(q)},\nu_l^{j(q)})}, \tag{5.31}$$

avec $z_{ij}^{(q)} = ((H^{(q)})^{-1}\mathbf{x}_i)_j$.

Les t_{ik}^j sont les seules quantités qui doivent être calculées lors de l'étape E de l'algorithme. Ces dernières diffèrent des probabilités a posteriori habituelles par la seule présence des termes pl_{ik}^j.

Au cours de l'étape M, la maximisation de $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)})$ conduit à des solutions analytiques semblables aux formules classiques de mises à jour pour les proportions, les moyennes et les variances des composantes du mélange :

$$\pi_k^{j(q+1)} = \frac{1}{N}\sum_{i=1}^{N} t_{ik}^{j(q)}, \tag{5.32}$$

$$\mu_k^{j\,(q+1)} = \frac{\sum_{i=1}^{N} t_{ik}^{j(q)} z_{ij}^{(q)}}{\sum_{i=1}^{N} t_{ik}^{j(q)}}, \tag{5.33}$$

$$\nu_k^{j(q+1)} = \frac{\sum_{i=1}^{N} t_{ik}^{j}(q)\left(z_{ij}^{(q)} - \mu_k^{j\,(q+1)}\right)^2}{\sum_{i=1}^{N} t_{ik}^{j(q)}}. \tag{5.34}$$

L'ensemble des étapes nécessaires à la mise en œuvre de cette méthode est résumé dans l'algorithme suivant (cf. algorithme 10).

Algorithme 10: pseudo-code de l'IFA sans bruit partiellement supervisée en utilisant l'algorithme $\mathrm{E}^2\mathrm{M}$ pour l'optimisation par rapport aux paramètres des densités des sources et une montée de gradient naturel pour l'optimisation par rapport à la matrice de mixage.

Données : Matrice de données centrée $\mathbf{X}$, Labels $\{pl_i^j\}_{i=1\ldots N,\, j=1\ldots L}$

 # Initialisation du vecteur de paramètres

$\boldsymbol{\psi}^{(0)} = (H^{(0)}, \boldsymbol{\pi}^{1(0)}, \ldots, \boldsymbol{\pi}^{L(0)}, \boldsymbol{\mu}^{1(0)}, \ldots, \boldsymbol{\mu}^{L(0)}, \boldsymbol{\nu}^{1(0)}, \ldots, \boldsymbol{\nu}^{L(0)}), \quad q = 0$

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

 # Mise à jour des sources

$\mathbf{z}_i = {H^{(q)}}^{-1}.\mathbf{x}_i$

 # Etape E

 # Mise à jour des probabilités a posteriori

pour tous les $j \in \{1, \ldots, L\}$ *et* $k \in \{1, \ldots, K_j\}$ **faire**

$$t_{ik}^{j(q)} = \frac{pl_{ik}^j \pi_k^{j(q)} \mathcal{N}(z_{ij}; \mu_k^{j(q)}, \nu_k^{j(q)})}{\sum_{l=1}^{K_j} pl_{il}^j \pi_l^{j(q)} \mathcal{N}(z_{ij}; \mu_l^{j(q)}, \nu_l^{j(q)})}, \quad \forall i \in \{1, \ldots, N\}$$

pour tous les $j \in \{1, \ldots, L\}$ *et* $k \in \{1, \ldots, K_j\}$ **faire**

 # Etape M

 # Mise à jour des paramètres des sources

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

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

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

 # Calcul du gradient naturel (3.80)

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

 # Mise à jour de la matrice de mixage (cf. annexe A.6)

$H^{(q+1)} = H^{(q)} + \tau.\Delta H$

$q \leftarrow q + 1$

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

L'approche présentée précédemment fournie un cadre d'apprentissage très général pour ce type de modèles lorsque l'information sur les classes d'origine des différents individus est partielle. En effet, selon le choix des fonctions de masse, cette formulation permet d'exprimer les cas supervisé, partiellement supervisé et non supervisé.

5.3.2 Acquisition des données et des labels

Afin de constituer l'ensemble des données destinées à l'apprentissage du modèle IFA partiellement supervisée, les relevés de signalisation réalisés par la rame de mesures IRIS320 et nécessaires à la reconstitution des signaux Icc nous ont été transmis par les services de la SNCF. Les données en question sont issues des tournées d'inspection réalisées sur la ligne TGV Nord entre janvier 2008 et janvier 2010 à raison d'une tournée tous les 15 jours.

A partir de chaque tournée, l'ensemble des signaux relatifs aux CdV de la ligne LN3 ont été extraits et stockés dans une base MySql contenant des informations sur l'infrastructure du réseau ferroviaire français. Bien qu'une grande quantité de données ait pu être recueillie, la plupart des signaux fournis étaient sans défaut, compte tenu de la politique de maintenance systématique actuellement en cours et des exigences liées à la sécurité ferroviaire.

Par conséquent, seule une partie des signaux disponibles a été sélectionnée pour les expériences. Les signaux présentant des défauts et considérés comme plus pertinents ont été choisis en priorité. Il est important de noter que cette sélection ne pénalise en rien la représentation des cas sans défaut dans l'ensemble de données, étant donné qu'il n'y a généralement pas plus d'un ou deux condensateurs défectueux par circuit de voie. Cette sélection a été réalisée sur les signaux *Icc* prélevés sur les CdV ayant 2300 Hz pour fréquence de fonctionnement et a permis de constituer une base de 422 signaux d'inspection.

L'ensemble des signaux constituant la base ont été présentés à quatre experts SNCF via une application dédiée à l'opération d'étiquetage. Dans le cadre de cette opération, trois classes ont été prises en compte pour le diagnostic. Les classes en question correspondent aux principaux états de fonctionnement des condensateurs : sans défaut, défaut intermédiaire et défaut majeur. A travers l'interface prévue pour cette opération (figure 5.6) il leur a été indiqué de :

1. Visualiser les signaux Icc un à un ;
2. Indiquer pour chaque condensateur, son appartenance potentielle à l'une des trois classes de fonctionnement sans exclusivité entre les classes ;
3. Préciser un degré de confiance par rapport à leur décision.

Ainsi, l'imprécision et l'incertitude des labels sont le reflet de l'hésitation quant au type de défaut et du doute quant à la présence effective de défaut. Le degré de confiance émis sert à caractériser ce doute.

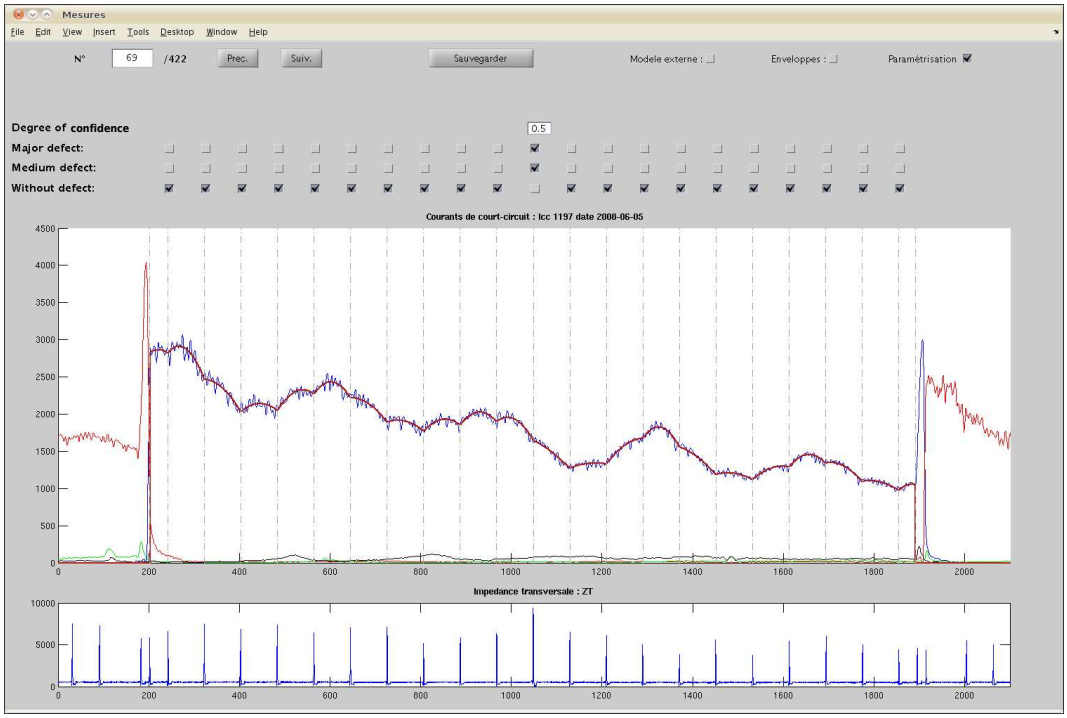


FIGURE 5.6 – *Interface de l'application présentée aux experts pour l'étiquetage des condensateurs à partir de chaque signal de la base d'apprentissage.*

Exemple 5.3 *(Étiquetage incertain et imprécis)*
Sur l'exemple d'étiquetage détaillé dans la figue 5.6, l'expert émet un avis à la fois imprécis et incertain quant à l'état de fonctionnement du 11*ème condensateur. En effet, celui-ci est imprécis quant au type de défaut en sélectionnant les classes qui correspondent à un défaut intermédiaire et à un défaut majeur. Il exprime également son incertitude en définissant un degré de confiance de* 0.5 *sur cette décision.*

Dans le cas présent, le principal intérêt d'un étiquetage partiel est d'avoir une représentation suffisamment fidèle de l'avis de chaque expert. Les différents avis peuvent par la suite être agrégés à l'aide des règles de combinaison afin de construire un étiquetage global plus robuste.

Par ailleurs, aucune réalité terrain n'étant disponible sur l'état des condensateurs, une autre expertise des signaux a été réalisée par nos soins dans un souci d'évaluation des résultats du diagnostic. Pour reprendre ce qui a été annoncé précédemment, à partir de l'ensemble des tournées réceptionnées sur une durée de deux ans, nous avons sélectionné un à un les signaux destinés à être étiquetés. Lors de cette sélection nous pouvions bénéficier d'un recul temporel qui permettait d'observer l'apparition et la progression des défauts sur les condensateurs. En conséquence, l'étiquetage en question correspondait à un étiquetage classique, pouvant être considéré

sans incertitude ou imprécision. En effet, l'incertitude et l'imprécision par rapport à la présence de défaut sur un signal observé de manière ponctuelle sont dissipées par l'historique et l'évolution des signaux relevés sur un même CdV. On notera toutefois, qu'il a fallu deux jours pour obtenir ces labels de référence, alors que les experts avaient seulement quelques heures pour l'opération d'étiquetage. Cet étiquetage de référence sera désigné par l'appellation REF dans la suite du document.

5.3.3 Prétraitements

Quelques prétraitements ont été nécessaires sur les labels avant de pouvoir les intégrer au processus d'apprentissage. En effet, une fois les avis d'experts obtenus, nous avons défini leur représentation dans le cadre de la théorie des fonctions de croyance pour pouvoir les fusionner par la suite.

Représentation des avis d'experts

Le cadre de discernement Ω défini dans le cas de l'application CdV est constitué des trois classes qui désignent les trois modes de fonctionnement pris en compte pour le diagnostic des condensateurs :

$$\Omega = \{\omega_0, \omega_1, \omega_2\}, \tag{5.35}$$

- ω_0 : le condensateur est sans défaut ;
- ω_1 : le condensateur comporte un défaut intermédiaire ;
- ω_2 : le condensateur a un défaut majeur.

L'état de connaissance d'un expert sur toute proposition $A \subseteq \Omega$ est donné par la fonction de masse notée m^Ω, telle que $m^\Omega(A)$ représente le degré de confiance alloué par l'expert à l'hypothèse exprimée. Ainsi, pour chaque condensateur, est construite une fonction de masse initialisée selon les degrés de confiance émis par les experts. La distribution des masses est liée aux classes sélectionnées par l'expert et au degré de confiance alloué. Notons que, pour un degré de confiance inférieur à 1, le complémentaire est considéré comme le degré d'ignorance et la masse est donc allouée à Ω.

Exemple 5.4 *(Fonction de masses représentative d'un avis d'experts)*
Dans le cas du 11*ème condensateur évoqué précédemment (figure 5.6), l'expert émet un avis imprécis sur la nature exacte du défaut en sélectionnant les deux classes : défaut intermédiaire et défaut majeur. Il associe de surcroît une confiance de* 0.5 *à cet avis. Selon le processus de représentation cité précédemment, cet étiquetage est traduit par l'allocation d'une masse de* 0.5 *sur l'ensemble* $\{\omega_1, \omega_2\}$ *et du complément* 0.5 *sur l'ensemble* Ω. *Le tableau 5.2 résume la distribution des masses par rapport aux éléments de* 2^Ω.

Ainsi, tout signal étiqueté par un expert i (avec $i \in \{1, \ldots, 4\}$), aura permis de définir une fonction de masse m_i^Ω pour chaque condensateur du CdV associé à ce

TABLE 5.2 – *Distribution des masses selon les classes sélectionnées par l'expert et le degré de confiance alloué*

A	$m^{\Omega}(A)$
$\emptyset$	0
$\{\omega_0\}$	0
$\{\omega_1\}$	0
$\{\omega_2\}$	0
$\{\omega_0, \omega_1\}$	0
$\{\omega_0, \omega_2\}$	0
$\{\omega_1, \omega_2\}$	0.5
Ω	0.5

signal. Les différentes fonctions de masse obtenues pour un même condensateur, ont par la suite été combinées dans le cadre de la théorie des fonctions de croyance pour obtenir un avis global.

Fusion des avis d'experts

Les quatre fonctions de masse obtenues pour chaque condensateur ont été combinées à l'aide des règles conjonctive, disjonctive et conjonctive prudente définies, respectivement, par les équations (5.5), (5.8) et (5.9). Le choix de ces trois règles est lié aux questions d'indépendance et de fiabilité des sources d'information. En particulier, il a été montré que la règle prudente permettait d'obtenir de bons résultats pour la combinaison d'information dépendantes [Ha-Duong 2008, Quost *et al.* 2011]. Elle s'avère ainsi appropriée pour combiner les opinions de plusieurs experts partageant des connaissances communes.

Dans le cas des règles conjonctives, les situations de conflit ont été traitées selon le principe de répartition proposé par Yager (5.19) après avoir combiné l'ensemble des fonctions de masse. Les fonctions de contour associées au résultat des combinaisons ont ensuite été utilisées pour estimer les paramètres du modèle IFA partiellement supervisée.

Le tableau 5.3 présente un exemple de résultats obtenus en combinant des fonctions de masse fournies par les experts selon la règle de combinaison utilisée. Les fonctions de contour associées sont données dans le tableau 5.4.

5.3.4 Évaluation des performances

Afin de mesurer l'impact et l'apport de chacune des règles de fusion, la base de données contenant 422 signaux a été associée à sept étiquetages différents, provenant

TABLE 5.3 – *Représentation des avis des quatre experts sous la forme de fonctions de masse et le résultat de leur combinaison avec les règles conjonctive, disjonctive et conjonctive prudente.*

A	$m_1^\Omega(A)$	$m_2^\Omega(A)$	$m_3^\Omega(A)$	$m_4^\Omega(A)$	$m_{\textcircled{\cap}}^\Omega(A)$	$m_{\textcircled{\cup}}^\Omega(A)$	$m_{\textcircled{\wedge}}^\Omega(A)$
$\emptyset$	0	0	0	0	0	0	0
$\{\omega_0\}$	0	0	0	0	0	0	0
$\{\omega_1\}$	0	0.8	0.9	0	0.98	0	0.9
$\{\omega_2\}$	0	0	0	0	0	0	0
$\{\omega_0, \omega_1\}$	0	0	0	0	0	0	0
$\{\omega_0, \omega_2\}$	0	0	0	0	0	0	0
$\{\omega_1, \omega_2\}$	0.5	0	0	0.9	0.019	0.3	0.09
Ω	0.5	0.2	0.1	0.1	0.001	0.7	0.01

TABLE 5.4 – *Fonctions de contour obtenues à partir de la combinaison des fonctions de masse détaillées dans le tableau 5.3 par les règles conjonctive, disjonctive et prudente.*

ω	$pl_{\textcircled{\cap}}(\omega)$	$pl_{\textcircled{\cup}}(\omega)$	$pl_{\textcircled{\wedge}}(\omega)$
$\{\omega_0\}$	0.001	0.7	0.01
$\{\omega_1\}$	1	1	1
$\{\omega_2\}$	0.02	1	0.1

de chacun des quatre experts et des trois règles de combinaison. Chacune de ces sept bases a été utilisée afin d'estimer les paramètres du modèle postulé pour le diagnostic, puis évaluée en terme de performances par rapport à l'étiquetage REF.

L'étiquetage REF a par ailleurs permis, dans un premier temps, d'observer la proportion de défauts présents dans la population des condensateurs et de confirmer ce qui avait été annoncé précédemment. Comme le montre le tableau 5.5, le nombre de défauts majeurs ou intermédiaires dans la base considérée est largement plus faible que le nombre de condensateurs sans défaut, entraînant une sous-représentation des deux classes de défauts.

Ceci est d'autant plus évident et problématique lorsque l'on considère le nombre de défauts par position de condensateur. La figure 5.7 montre la distribution des défauts (selon l'étiquetage REF) en fonction des positions des condensateurs pour les

TABLE 5.5 – *Répartition des condensateurs en fonction de leur classe d'appartenance et selon la l'étiquetage REF.*

	Nbr de sans défaut	Nbr de défauts intermédiaires	Nbr de défauts graves
Nbr de condensateurs	8102	126	106

circuits de voie dans la base des 422 signaux. On peut constater que pour certaines positions aucune observation de cas de défauts n'est disponible.

Cette représentation met clairement en évidence la sous-représentation des classes de défaut dans la base de données et la nécessité de rassembler davantage de cas de défaut pour pouvoir estimer les paramètres du modèle.

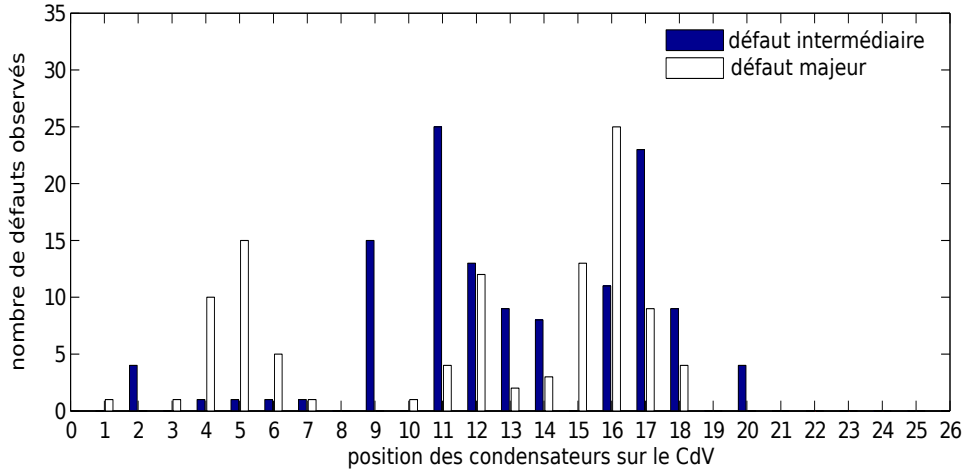


FIGURE 5.7 – *Nombre d'observations des cas de défaut intermédiaire (bleu) et de défaut grave (blanc) par condensateur sur des circuits de voie contenant* 25 *condensateurs et selon l'étiquetage REF.*

Pour surmonter ce problème, la base de données réelles a été complétée par des données simulées générées à l'aide d'un modèle électrique du CdV [Aknin *et al.* 2003, Oukhellou *et al.* 2006]. Cinq cents signaux bruités dont les défauts sont distinctement identifiés et labellisés ont été générés. Ces signaux correspondent à des circuits de voie de 25 condensateurs avec différentes valeurs de la capacité pour chaque condensateur. Afin d'obtenir une base de signaux simulés réalistes les capacités ont été générées selon un modèle de mélange reflétant les différentes sous-populations de condensateurs :

- La classe ω_0 représentant les condensateurs sans défaut avec un taux de 95%

des condensateurs par rapport à l'ensemble de la base ($\pi_1 = 0.95$). Elle suit une loi normale de moyenne égale à la valeur nominale des condensateurs posés en voie $\mu_1 = 22\mu$F et de variance telle que 95% de cette sous population réponde aux tolérances fournies par le constructeur (10%), c'est à dire $\nu_1 = (2.2/1.96)^2 \approx 1.26$;

- La classe ω_1 représentant les condensateurs avec défaut intermédiaire avec un taux de 3% des condensateurs par rapport à l'ensemble de la base ($\pi_2 = 0.03$). Elle suit une loi normale de la moyenne $\mu_2 = 10\mu$F et de variance $\nu_2 = 6$;
- La classe ω_2 représentant les condensateurs avec défaut grave avec un taux de 2% des condensateurs par rapport à l'ensemble de la base ($\pi_3 = 0.02$). Tous les individus générés dans cette classe se sont vus attribuer une capacité égale à 0.

Notons néanmoins que les données simulées ont été utilisées uniquement lors de la phase d'apprentissage du modèle et n'ont jamais servi à l'évaluation des performances du diagnostic. Une approche par validation croisée a été adoptée afin d'évaluer la performance de chaque étiquetage. La base de données a été décomposée aléatoirement en dix sous-ensembles. Neuf d'entre eux augmentés des 500 signaux simulés ont été utilisés pour l'apprentissage des paramètres du modèle IFA partiellement supervisée, et le sous-ensemble restant a été utilisé comme ensemble de test afin d'évaluer les performances obtenues avec les paramètres estimés. Ces deux étapes ont été répétées 10 fois, en désignant à chaque fois un sous-ensemble différent pour le test. Le résultat a été calculé sur la moyenne des 10 performances obtenues.

5.3.5 Expérimentations et résultats

Les résultats ont été analysés selon les prédictions faites à l'aide des modèles appris avec chacun des ensembles de labels concernant l'état de fonctionnement des condensateurs. Les matrices de confusion entre les classe définies par l'étiquetage REF et les classes estimées à l'aide de chacun des sept schémas d'étiquetage et pour l'ensemble des condensateurs contenus dans la base de données, sont données dans les tableaux 5.6 et 5.7. Les différentes classes réelles possibles et les différentes décisions possibles sont notées de la manière suivante :

- r_0 : le condensateur appartient à la classe ω_0,
- r_1 : le condensateur appartient à la classe ω_1,
- r_2 : le condensateur appartient à la classe ω_2,
- d_0 : le condensateur a été classé dans la classe ω_0,
- d_1 : le condensateur a été classé dans la classe ω_1,
- d_2 : le condensateur a été classé dans la classe ω_2.

Les décisions d_0, d_1 et d_2 ont été déterminées pour chaque condensateur par le maximum des probabilités a postériori (3.72) calculées sur tous les ensembles de tests en utilisant les paramètres estimés à l'aide de chacun des étiquetages.

Les résultats révèlent de bonnes performances de classification en dépit de quelques erreurs de classement dans le cas des classes voisines (ω_0 et ω_1 ainsi que ω_1 et ω_2). Les matrices de confusion obtenues dans le cas des experts fournissent certaines informations quant aux compétences de ces derniers (tableau 5.6). En effet, les experts 1 et 4 semblent mieux détecter les défauts majeurs, alors que les experts 2 et 3 semblent plus performants pour la détection des défauts intermédiaires. Avec la combinaison de ces expertises, il est possible d'améliorer la détection des deux types de défauts (tableau 5.7). Dans ce cas de figure, les meilleurs résultats ont été obtenus par la combinaison prudente, ce qui suggère que les opinions d'experts ne peuvent pas être considérées comme indépendantes.

TABLE 5.6 – *Matrices de confusion pour les décisions prises en fonction des bases d'apprentissage étiquetées par les experts.*

	r_0	r_1	r_2
d_0	98.8	33.1	2.1
d_1	0.9	51.1	6.9
d_2	0.2	15.8	90.9

(Expert 1)

	r_0	r_1	r_2
d_0	98.9	34.7	3.0
d_1	0.8	58.8	12.2
d_2	0.3	6.5	84.7

(Expert 2)

	r_0	r_1	r_2
d_0	98.7	22.1	2.1
d_1	1.1	63.6	13.8
d_2	0.2	14.3	84.1

(Expert 3)

	r_0	r_1	r_2
d_0	98.8	34.6	3.3
d_1	1.0	49.6	5.8
d_2	0.2	15.8	90.9

(Expert 4)

Les cas de confusion entre classes voisines (en particulier ω_1 et ω_2) peuvent être expliqués par deux facteurs. Tout d'abord, compte tenu du nombre total de condensateurs représentés dans la base de données, le nombre de défauts intermédiaires et majeurs reste trop faible par rapport aux cas sans défaut pour espérer un apprentissage fiable de ces deux classes. Par ailleurs, l'identification des défauts intermédiaires est un exercice particulièrement difficile en raison du caractère continu de l'état de dégradation réel. Dans les cas critiques, ces derniers peuvent être confondus avec les deux classes voisines (ω_0 d'une part et ω_2 d'autre part), ce qui réduit encore plus le taux de détection.

Afin de confirmer cette analyse, nous avons calculé le degré de conflit résultant de la combinaison conjonctive des fonctions de masse obtenues à partir des labels fournis par chacun des experts avec celles obtenues à partir de l'étiquetage REF. Comme le montre la figure 5.8, le degré de conflit est globalement très faible (< 0.03), ce qui

TABLE 5.7 – *Matrices de confusion pour les décisions prises en fonction des bases d'apprentissage étiquetées par les différents schémas de combinaison.*

	r_0	r_1	r_2
d_0	98.9	30.7	2.9
d_1	0.9	58.0	7.7
d_2	0.2	11.3	89.4

(Vote majoritaire)

	r_0	r_1	r_2
d_0	98.9	20.2	2.9
d_1	1.0	64.2	6.5
d_2	0.1	15.6	90.6

(Combinaison conjonctive)

	r_0	r_1	r_2
d_0	98.9	23.1	2.9
d_1	0.8	62.8	8.0
d_2	0.1	14.1	89.2

(Combinaison disjonctive)

	r_0	r_1	r_2
d_0	98.9	20.4	2.6
d_1	1.0	65.3	4.9
d_2	0.1	14.2	92.4

(Combinaison prudente)

est cohérent avec le nombre élevé de cas sans défaut dans la base de condensateurs (tableau 5.5).

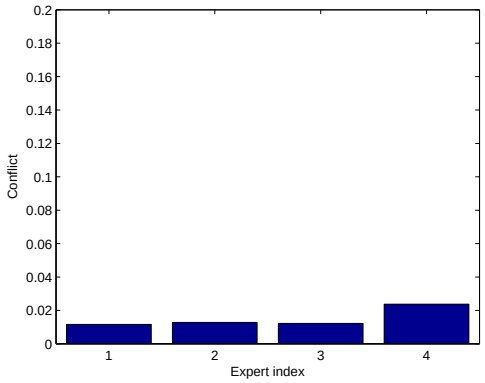


FIGURE 5.8 – *Degré de conflit moyen global entre chaque expert et l'étiquetage REF sur la totalité de la base de données.*

Toutefois, le conflit est plus élevé dans le cas des défauts majeurs (< 0.14) et l'est encore plus dans le cas de défauts intermédiaires (conflit entre 0.45 et 0.8), comme indiqué respectivement dans les figures 5.9 et 5.10.

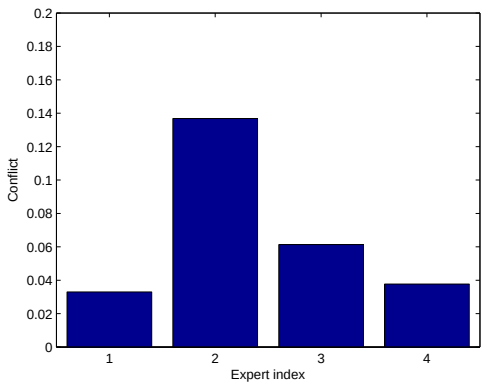


FIGURE 5.9 – *Degré de conflit moyen global entre chaque expert et l'étiquetage REF sur les cas de défaut majeur.*

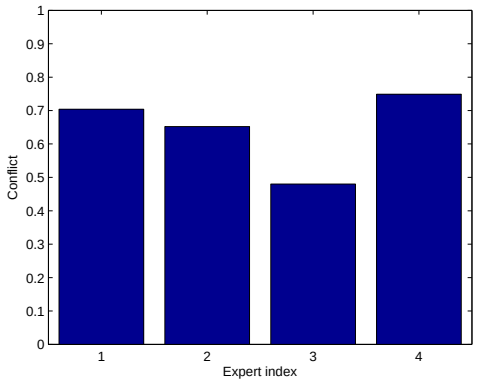


FIGURE 5.10 – *Degré de conflit moyen global entre chaque expert et l'étiquetage REF sur les cas de défaut intermédiaire.*

Les données simulées étant utilisées conjointement aux données réelles dans la cadre de ces travaux, nous avons vérifié si les données simulées seules n'étaient pas suffisantes pour obtenir de bonnes performances pour le diagnostic. Le tableau 5.8 montre les matrices de confusion pour les classifications obtenue sur les données réelles, en utilisant uniquement les données simulées lors de l'apprentissage du modèle (avec $N \in \{500, 900, 1500\}$ individus). Ces résultats montrent que les données simulées ne permettent pas à elles seules d'obtenir de bonnes performances de classification, même quand N est très grand. Ceci prouve que l'utilisation des signaux réels étiquetés par des experts est nécessaire afin d'atteindre de bonnes performances au niveau de la tâche de diagnostic.

Enfin, en regroupant les deux classes de défaut ($\omega_1 \cup \omega_2$), une matrice de détection peut être construite et les indicateurs de performance suivants peuvent alors être

TABLE 5.8 – *Matrices de confusion pour les décisions obtenues sur les données réelles à partir d'un apprentissage réalisé uniquement sur la base de données simulées de taille N.*

	ω_0	ω_1	ω_2
d_0	90.5	35.1	4.6
d_1	8.4	52.4	17.4
d_2	1.1	12.5	78.0

$N = 500$

	ω_0	ω_1	ω_2
d_0	91.0	28.4	5.0
d_1	8.4	58.1	14.2
d_2	0.6	13.5	80.8

$N = 900$

	ω_0	ω_1	ω_2
d_0	90.8	22.7	3.1
d_1	8.4	59.0	15.6
d_2	0.8	18.4	81.3

$N = 1500$

calculés :

- Le taux de bonne classification (BC), défini par la proportion de prédictions correctes :
$$BC = \frac{\#(d_0, r_0) + \#(d_1 \cup d_2, r_1 \cup r_2)}{N}; \tag{5.36}$$
- Le taux de bonne détection (BD), défini par la proportion de condensateurs défectueux correctement identifiés :
$$BD = \frac{\#(d_1 \cup d_2, r_1 \cup r_2)}{\#(d_0, r_1 \cup r_2) + \#(d_1 \cup d_2, r_1 \cup r_2)}; \tag{5.37}$$
- Le taux de faux négatif (FN), défini par la proportion de condensateurs défectueux détectés comme étant sans défaut :
$$FN = \frac{\#(d_0, r_1 \cup r_2)}{\#(d_0, r_1 \cup r_2) + \#(d_1 \cup d_2, r_1 \cup r_2)}. \tag{5.38}$$

Les résultats rapportés dans le tableau 5.9 montrent qu'une précision (taux de bonne classification) d'au moins 97% est atteinte quelque soit l'étiquetage utilisé pour l'apprentissage. Toutefois, les différents schémas de combinaison surpassent chacun des experts, notamment en termes de bonnes détections et de faux négatifs. Les règles conjonctive et prudente donnent de meilleurs résultats que la règle disjonctive.

Par ailleurs, l'apport d'un étiquetage partiel pour l'estimation des paramètres du modèle IFA peut être constaté en examinant les tableaux 5.6, 5.7 et 5.9. En effet, les résultats obtenus si l'on choisit d'ignorer les degrés de confiance alloués par les experts sur leurs avis, en se référant à l'opinion de la majorité, sont beaucoup plus faibles que ceux obtenus par les approches avec fusion. On notera également que ces résultats ne sont pas nécessairement meilleurs que les performances obtenues en utilisant les labels fournis par les experts individuellement (voir l'Expert 3 dans les tableaux 5.6 et 5.9).

TABLE 5.9 – *Taux de bonne classification (BC), taux de bonne détection (BD) et taux de faux négatifs (FN) correspondant au performance obtenues pour la diagnostic à partir des apprentissages réalisés avec chaque expert et chaque schéma de combinaison.*

	Expert1	Expert2	Expert3	Expert4	Maj.	⋂	⋃	⩓
BC	97.2%	97, 9%	98.5%	97.2%	98.2%	98.5%	98.1%	98.5%
BD	79.5%	79.5%	84.1%	78.2%	82.6%	87.5%	78.4%	87.9%
FN	20.3%	20.3%	12.8%	21.2%	17.3%	12.5%	13.6%	12.0%

5.4 Conclusion

Ce chapitre nous a permis de mettre en œuvre une méthode pour le diagnostic des CdV basée sur le modèle d'IFA partiellement supervisée. Dans le cadre de ce modèle l'estimation des paramètres est réalisée par maximisation du critère de vraisemblance généralisé, par le biais de l'algorithme E^2M introduit dans [Côme *et al.* 2009, Denoeux 2010].

L'apprentissage du modèle a été fait sur des données réelles fournies par la SNCF dans le cadre du projet ANR DIAGHIST destiné à fournir un outil de diagnostic pour l'aide à la maintenance et à la surveillance des circuits de voie. Afin de mener à bien ce projet, une campagne d'étiquetage des données a été organisée et a permis d'obtenir certaines informations destinées à être introduite dans la phase d'apprentissage du modèle IFA.

Les bons résultats obtenus sur des données réelles ont démontré l'intérêt de cette approche. Nous avons également pu observer l'utilité de la fusion de données, y compris en cas de conflit important, pour l'apprentissage et par conséquent pour le diagnostic.

Chapitre 6

Conclusion et perspectives

Ces travaux de thèse ont été consacrés au développement et à la mise en œuvre de méthodes de diagnostic pour une application dédiée à la signalisation ferroviaire. Le système en question est vu comme un système complexe où il s'agit de pister l'état des sous-composants dans l'objectif de mettre au point une politique de maintenance prévisionnelle. La tâche du diagnostic consiste alors à détecter, localiser et estimer la gravité des défauts éventuels.

Après avoir présenté l'application pratique à l'origine de la thèse, nous avons voulu positionner notre démarche par rapport aux approches existantes dans le domaine du diagnostic et aux travaux antérieurs (chapitre 2). Cette première partie justifie l'adoption d'une approche générative pour résoudre le problème du diagnostic dans le cas d'un système complexe. Le diagnostic porte alors sur l'estimation de variables latentes liées aux défauts à partir de variables observées extraites de signaux d'inspection. Le chapitre 3 présente les modèles à variables latentes pouvant intervenir dans ce cadre. La suite du mémoire décrit l'apport de cette thèse à proprement parler et met en avant les deux problématiques principales auxquelles nous nous sommes intéressés.

La première problématique porte sur le diagnostic du système dans un cadre non-supervisé (chapitre 4). Dans un tel contexte nous avons eu recours à des extensions des modèles d'analyse en composantes indépendantes (ICA) et d'analyses en facteurs indépendants (IFA) prenant en compte des connaissances sur la structure du modèle ou sur l'aspect temporel de données prélevées successivement dans le temps. Les résultats obtenus sur des données simulées nous ont permis de montrer l'intérêt de nos propositions.

Dans le cas de l'ICA, nous avons pu constater l'utilité d'intégrer des contraintes sur le modèle lorsque celles-ci traduisent des hypothèses vérifiées. Par ailleurs, dans le cas de données dynamiques nous avons montré qu'un modèle d'IFA, intégrant une hypothèse markovienne quant à la structure des données, pouvait apporter une amélioration à l'estimation des sources. Nous avons également évalué cette approche en intégrant quelques labels lors de l'apprentissage et avons constaté qu'elle permettait de réduire le nombre d'individus nécessaires à l'obtention d'un niveau de performance donné. Néanmoins, une certaine réserve est à émettre dans le cas de notre application compte tenu de la quantité de relevés d'inspection pouvant être obtenus sur un même système au cours temps. En effet, à l'heure actuelle il est

difficile de pouvoir envisager une telle approche sur des données réelles avec une fréquence d'inspection bimensuelle.

La seconde problématique abordée porte sur le diagnostic du système dans un cadre partiellement supervisé (chapitre 5). L'apprentissage du modèle a été réalisé sur des données réelles fournies par la SNCF et étiquetées de manière imprécise et incertaine par plusieurs experts. L'approche proposée reposait sur un modèle d'IFA partiellement supervisée. Nous avons eu recours à travers cette approche à la théorie des fonctions de croyance pour modéliser les différents avis afin de les intégrer lors de l'apprentissage. Nous avons par ailleurs constaté que la combinaison des avis permettait d'obtenir de meilleures performances pour le diagnostic.

Ces travaux peuvent être étendus à différents niveaux. Tout d'abord, dans le cadre des modèles d'ICA et d'IFA, une relation non linéaire entre variables latentes et variables observées peut être envisagée. Les travaux de [Taleb & Jutten 1999] proposent un modèle où les observations proviennent des transformations non linéaires de mélanges linéaires des sources avec une extension traitant de sources Markoviennes [Larue *et al.* 2004]. Par ailleurs, la modélisation du bruit proposée dans chacune des méthodes présentées pourrait être comparée à la solution alternative consistant à l'intégrer directement au modèle. Les approches variationnelles peuvent aussi être envisagées comme solution à ce problème [Lawrence & Bishop 2000, Choudrey & Roberts 2003].

D'autre part, la problématique abordée dans la dernière partie de la thèse ouvre d'autres perspectives. L'approche présentée repose sur l'élicitation des connaissances d'experts dans le cadre de la théorie des fonctions de croyance, qui constitue un problème important n'ayant pas reçu suffisamment d'attention jusqu'à présent. Dans le cadre de notre travail, seuls trois schémas de combinaison ont été évalués pour fusionner les avis d'experts, mais d'autres schémas pourraient également être envisagés en particulier pour la gestion du conflit [Martin & Osswald 2007]. Par ailleurs, une autre approche peut être adoptée pour limiter le conflit. Des taux d'affaiblissement pourraient être appris à partir des données et permettraient de tenir compte des compétences de chaque expert individuellement [Elouedi *et al.* 2004, Mercier *et al.* 2008]. Enfin, l'approche adoptée pour l'estimation des paramètres du modèle IFA à partir de données étiquetées de façon incertaine est évidemment très générale. Celle-ci peut être étendue à beaucoup d'autres problèmes impliquant un modèle génératif et des observations imparfaites. Par exemple, le modèle d'IFA avec HMM peut facilement être étendu à un tel cadre d'apprentissage [Ramasso 2009].

A

Annexes

A.1 Mise à jour des proportions lors de l'étape M de l'algorithme EM pour les modèles de mélange

La fonction auxiliaire Q est définie dans le cadre des modèles de mélange par :

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

avec :

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} \pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})} \tag{A.2}$$

Il est possible de décomposer cette fonction selon les paramètres à optimiser :

$$Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(q)}) = Q_\pi(\pi, \boldsymbol{\psi}^{(q)}) + \sum_{k=1}^{K} Q_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k, \boldsymbol{\psi}^{(q)}) \tag{A.3}$$

La mise à jour des proportions effectuée lors de l'étape M de l'algorithme EM, correspond à la maximisation de la fonction $Q_\pi(\pi, \boldsymbol{\psi}^{(q)})$ par rapport à $\boldsymbol{\pi}$. Afin de maximiser $Q_\pi(\pi, \boldsymbol{\psi}^{(q)})$ par rapport à $\boldsymbol{\pi}$ en prenant en considération la contrainte $\sum_{k=1}^{K} \pi_k = 1$, nous formons le lagrangien :

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik}^{(q)} \log(\pi_k) + \lambda.(1 - \sum_{k=1}^{K} \pi_k), \tag{A.4}$$

où λ est le multiplicateur de lagrange associé à la contrainte. En dérivant le lagrangien par rapport aux proportions nous obtenons :

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_k} = \frac{\sum_{i=1}^{N} t_{ik}^{(q)}}{\pi_k} - \lambda, \quad \forall k \in \{1, \ldots, K\}, \tag{A.5}$$

Pour maximiser Q_π 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 :

$$\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} = \lambda.(\pi_1 + \ldots + \pi_K) \tag{A.6}$$

et donc $\lambda = N$, ce qui permet d'obtenir la formule de mise à jour suivante en remplaçant λ par N dans (A.6) :

$$\pi_k^{(q+1)} = \sum_{i=1}^{N} t_{ik}^{(q)}/N. \tag{A.7}$$

A.2 Probabilités a posteriori pour un HMM

Dans le cadre des HMM, l'étape E de l'algorithme EM consiste à calculer l'espérance conditionnelle de la log-vraisemblance complétée. Cette étape nécessite le calcul des probabilités a posteriori γ_{tk} et ξ_{tlk} :

• La probabilité γ_{tk} représente la probabilité a posteriori de l'état k à l'instant t sachant la séquence totale des observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ et les paramètres du modèle $\boldsymbol{\psi}$ et est donnée par :

$$\begin{aligned}
\gamma_{tk} &= p(y_t = k|\mathbf{X};\boldsymbol{\psi}) \\
&= \frac{p(\mathbf{X}, y_t = k;\boldsymbol{\psi})}{p(\mathbf{X};\boldsymbol{\psi})} \\
&= \frac{p(\mathbf{X}|y_t = k;\boldsymbol{\psi})p(y_t = k;\boldsymbol{\psi})}{\sum_{l=1}^{K} p(\mathbf{X}|y_t = l;\boldsymbol{\psi})p(y_t = l;\boldsymbol{\psi})} \\
&= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_t|y_t = k;\boldsymbol{\psi})p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T|y_t = k;\boldsymbol{\psi})p(y_t = k;\boldsymbol{\psi})}{\sum_{l=1}^{K} p(\mathbf{x}_1, \ldots, \mathbf{x}_t|y_t = l;\boldsymbol{\psi})p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T|y_t = l;\boldsymbol{\psi})p(y_t = l;\boldsymbol{\psi})} \\
&= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_t, y_t = k;\boldsymbol{\psi})p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T|y_t = k;\boldsymbol{\psi})}{\sum_{l=1}^{K} p(\mathbf{x}_1, \ldots, \mathbf{x}_t, y_t = l;\boldsymbol{\psi})p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T|y_t = l;\boldsymbol{\psi})} \\
&= \frac{\alpha_{tk}\beta_{tk}}{\sum_{l=1}^{K} \alpha_{tl}\beta_{tl}}.
\end{aligned} \tag{A.8}$$

• La probabilité ξ_{tlk} est la probabilité jointe a posteriori de l'état l à l'instant $t-1$ et de l'état k à l'instant t sachant la séquence totale des observations $\mathbf{X}$ et des paramètres du modèle $\boldsymbol{\psi}$ et est donnée par :

$$\begin{aligned}
\xi_{tlk} &= p(y_t = k|y_{t-1} = l|\mathbf{X};\boldsymbol{\psi}) \\
&= \frac{p(y_t = k|y_{t-1} = l, \mathbf{X};\boldsymbol{\psi})}{p(\mathbf{X};\boldsymbol{\psi})} \\
&= \frac{p(y_t = k|y_{t-1} = l, \mathbf{X};\boldsymbol{\psi})}{\sum_{l=1}^{K}\sum_{k=1}^{K} p(y_t = k|y_{t-1} = l, \mathbf{X};\boldsymbol{\psi})} \\
&= \frac{p(\mathbf{X}|y_t = k, y_{t-1} = l;\boldsymbol{\psi})p(y_t = k, y_{t-1} = l;\boldsymbol{\psi})}{\sum_{l=1}^{K}\sum_{k=1}^{K} p(\mathbf{X}|y_t = k, y_{t-1} = l;\boldsymbol{\psi})p(y_t = k, y_{t-1} = l;\boldsymbol{\psi})} \\
&= \frac{\alpha_{(t-1)l}p(\mathbf{x}_t|y_t = k;\boldsymbol{\psi})\beta_{tk}A_{lk}}{\sum_{l=1}^{K}\sum_{k=1}^{K} \alpha_{(t-1)l}p(\mathbf{x}_t|y_t = k;\boldsymbol{\psi})\beta_{tk}A_{lk}}.
\end{aligned} \tag{A.9}$$

Le calcul de ses probabilités a posteriori nécessite le calcule des probabilités forward α_{tk} et backward β_{tk} pour $t \in \{1, \ldots, T\}$ et $k \in \{1, \ldots, K\}$ (cf. Annexe A.3).

A.3 Probabilités forward et backward

Les probabilités Forward-Backward qui peuvent être désignées par α_{tk} et β_{tk} respectivement, sont définies comme suit :

$$\alpha_{tk} = p(\mathbf{x}_1, \ldots, \mathbf{x}_t, y_t = k; \boldsymbol{\psi}), \tag{A.10}$$

qui représente la probabilité d'observer la séquence partielle $(\mathbf{x}_1, \ldots, \mathbf{x}_t)$ se terminant avec l'état k à l'instant t, et :

$$\beta_{tk} = p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T, y_t = k; \boldsymbol{\psi}), \tag{A.11}$$

qui est la probabilité d'observer le reste de la séquence $(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_1)$ sachant que l'on commence par l'état k à l'instant t.

Le calcul de ces probabilités peut se faire de la façon suivante :

$$\begin{aligned}
\alpha_{tk} &= p(\mathbf{x}_1, \ldots, \mathbf{x}_t, y_t = k; \boldsymbol{\psi}) \\
&= p(\mathbf{x}_1, \ldots, \mathbf{x}_t | y_t = k; \boldsymbol{\psi}) p(y_t = k; \boldsymbol{\psi}) \\
&= p(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1} | y_t = k; \boldsymbol{\psi}) p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi}) p(y_t = k; \boldsymbol{\psi}) \\
&= p(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, y_t = k; \boldsymbol{\psi}) p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi}) \\
&= \sum_{l=1}^{K} p(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, y_{t-1} = l, y_t = k; \boldsymbol{\psi}) p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi}) \\
&= \sum_{l=1}^{K} p(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, y_t = k | y_{t-1} = l; \boldsymbol{\psi}) p(y_{t-1} = l; \boldsymbol{\psi}) p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi}) \\
&= \sum_{l=1}^{K} p(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1} | y_{t-1} = l; \boldsymbol{\psi}) p(y_t = k | y_{t-1} = l; \boldsymbol{\psi})) p(y_{t-1} = l; \boldsymbol{\psi}) p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi}) \\
&= \sum_{l=1}^{K} p(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, y_{t-1} = l; \boldsymbol{\psi}) p(y_t = k | y_{t-1} = l; \boldsymbol{\psi})) p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi}) \\
&= \left[\sum_{l=1}^{K} \alpha(t-1) l A_{lk} \right] p(\mathbf{x}_t | y_t = k; \boldsymbol{\psi})
\end{aligned} \tag{A.12}$$

et :

$$
\begin{aligned}
\beta_{tl} &= p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T, y_t = l; \boldsymbol{\psi}) \\
&= \sum_{k=1}^{K} p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T, y_{t+1} = k | y_t = l; \boldsymbol{\psi}) \\
&= \sum_{k=1}^{K} p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T | y_{t+1} = k, y_t = l; \boldsymbol{\psi}) p(y_{t+1} = k | y_t = l; \boldsymbol{\psi}) \\
&= \sum_{k=1}^{K} p(\mathbf{x}_{t+2}, \ldots, \mathbf{x}_T | y_{t+1} = k, y_t = l; \boldsymbol{\psi}) p(y_{t+1} = k | y_t = l; \boldsymbol{\psi}) p(\mathbf{x}_{t+1} | y_{t+1} = k; \boldsymbol{\psi}) \\
&= \sum_{k=1}^{K} p(\mathbf{x}_{t+2}, \ldots, \mathbf{x}_T | y_{t+1} = k; \boldsymbol{\psi}) p(y_{t+1} = k | y_t = l; \boldsymbol{\psi}) p(\mathbf{x}_{t+1} | y_{t+1} = k; \boldsymbol{\psi}) \\
&= \sum_{k=1}^{K} \beta_{(t+1)k} A_{lk} p(\mathbf{x}_{t+1} | y_{t+1} = k; \boldsymbol{\psi}). \qquad \text{(A.13)}
\end{aligned}
$$

Les probabilités α_{tk} et β_{tk} sont alors calculées récursivement par les procédures suivantes :

Procédure Forward

- $\alpha_{1k} = p(\mathbf{x}_1, y_1 = k; \boldsymbol{\psi}) = p(y_1 = k) p(\mathbf{x}_1 | y_1 = k; \boldsymbol{\theta}_k) = \pi_k p(\mathbf{x}_1 | y_1 = k; \boldsymbol{\theta}_k)$ pour $t = 1$ et $k = 1, \ldots, K$.
- $\alpha_{tk} = \sum_{l=1}^{K} \alpha_{(t-1)l} A_{lk} p(\mathbf{x}_t | y_t = k; \boldsymbol{\theta}_k)$ pour $t = 2, \ldots, T$ et $k = 1, \ldots, K$.

Procédure Backward

- $\beta_{Tk} = 1$ pour $t = T$ et $k = 1, \ldots, K$.
- $\beta_{tl} = \sum_{k=1}^{K} \beta_{(t+1)k} A_{lk} p(\mathbf{x}_{t+1} | y_{t+1} = k; \boldsymbol{\theta}_k)$ pour $t = T - 1, \ldots, 1$ et $l = 1, \ldots, K$.

A.4 Densité d'une transformation linéaire

Soit $\mathbf{x}, \mathbf{z}$ deux vecteurs de dimension L reliés par la relation déterministe suivante :

$$\mathbf{x} = \mathbf{h}(\mathbf{z}). \tag{A.14}$$

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 :

$$f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(J_h(\mathbf{h}^{-1}(\mathbf{x})))|} f^{\mathcal{Z}}\left(\mathbf{h}^{-1}(\mathbf{x})\right), \tag{A.15}$$

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 $\mathbf{x} = H\mathbf{z}$ inversible, nous obtenons donc :

$$f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(H)|} f^{\mathcal{Z}}\left(H^{-1}\mathbf{x}\right). \tag{A.16}$$

Ces résultats peuvent être trouvés dans [Hyvärinen *et al.* 2001, p. 35-36].

A.5 Gradient de la log-vraisemblance de l'ICA par rapport à la matrice de démixage

La log vraisemblance d'une matrice de démixage G dans le cadre de l'ICA sans bruit est donnée par :

$$\mathcal{L}(G;\mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{L} \log\left(f^{\mathcal{Z}_j}\left((G\mathbf{x}_i)_j\right)\right) + N\log(|\det(G)|), \tag{A.17}$$

en supposant que les densités des sources $f^{\mathcal{Z}_1},\ldots,f^{\mathcal{Z}_L}$ sont connues.

Pour calculer le gradient de $\mathcal{L}(G;\mathbf{X})$ par rapport à G 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 & Pedersen 2008, p. 8]) :

$$\frac{\partial\log(|\det(X)|)}{\partial X_{lk}} = (X^{-1})_{kl}, \tag{A.18}$$

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 :

$$\begin{aligned}\frac{\partial\mathcal{L}(G;\mathbf{X})}{\partial G_{lk}} &= N(G^{-1})_{kl} + \sum_{i=1}^{N}\frac{\partial\log\left(f^{\mathcal{Z}_l}\left((G\mathbf{x}_i)_l\right)\right)}{\partial G_{lk}} \\ &= N(G^{-1})_{kl} - \sum_{i=1}^{N}\mathbf{x}_{ik}g_l\left((G\mathbf{x}_i)_l\right),\end{aligned} \tag{A.19}$$

avec $g_l(z)$ l'opposé de la dérivée du logarithme de la densité de la source l :

$$g_l(z) = \frac{-\partial\log(f^{\mathcal{Z}_l}(z))}{\partial z} \tag{A.20}$$

En prenant des notations matricielles, nous pouvons définir la fonction $\mathbf{g}$:

$$\begin{aligned}\mathbf{g} &: \mathrm{R}^L \to \mathrm{R}^L \\ \mathbf{g}(\mathbf{z}) &= \left[\frac{-\partial\log(f^{\mathcal{Z}_1}(z_1))}{\partial z_1},\ldots,\frac{-\partial\log(f^{\mathcal{Z}_L}(z_L))}{\partial z_L}\right]'.\end{aligned} \tag{A.21}$$

Ce qui nous permet d'obtenir la matrice contenant la dérivée de la log-vraisemblance par rapport à chaque coefficient de G :

$$\begin{aligned}\frac{\partial\mathcal{L}(G;\mathbf{X})}{\partial G} &= N(G^{-1})' - \sum_{i=1}^{N}\mathbf{g}\left(G\mathbf{x}_i\right)\mathbf{x}_i' \\ &\propto (G^{-1})' - \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}\left(G\mathbf{x}_i\right)\mathbf{x}_i'.\end{aligned} \tag{A.22}$$

A.6 Gradient de la log-vraisemblance de l'ICA par rapport à la matrice de mixage

La log-vraisemblance d'une matrice de mixage H dans le cadre de l'ICA sans bruit est donnée par :

$$\mathcal{L}(H;\mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{L} \log\left(f^{\mathcal{Z}_j}\left((H^{-1}\mathbf{x}_i)_j\right)\right) - N\log(|\det(H)|), \tag{A.23}$$

en supposant que les densités des sources $f^{\mathcal{Z}_1}, \ldots, f^{\mathcal{Z}_L}$ sont connues.

Pour calculer le gradient de $\mathcal{L}(H;\mathbf{X})$ par rapport à H, 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 (voir [MacKay 1996], [Petersen & Pedersen 2008, p. 8]), cette dérivée est donnée par :

$$\frac{\partial (X^{-1})_{jp}}{\partial X_{lk}} = -(X^{-1})_{jl}(X^{-1})_{kp}, \tag{A.24}$$

En utilisant (A.18) et (A.24) nous obtenons donc la dérivée de la log-vraisemblance par rapport à un élément de l,k de la matrice de mixage :

$$\begin{aligned}
\frac{\partial \mathcal{L}(H;\mathbf{X})}{\partial H_{lk}} &= -N(H^{-1})_{kl} + \sum_{i=1}^{N}\sum_{j=1}^{L} \frac{\partial \log\left(f^{\mathcal{Z}_j}\left((H^{-1}\mathbf{x}_i)_j\right)\right)}{\partial H_{lk}} \\
&= -N(H^{-1})_{kl} + \sum_{i=1}^{N}\sum_{j=1}^{L} \frac{\partial \log\left(f^{\mathcal{Z}_j}\left(\sum_{p=1}^{L}(H^{-1})_{jp}(\mathbf{x}_i)_p\right)\right)}{\partial H_{lk}} \\
&= -N(H^{-1})_{kl} + \sum_{i=1}^{N}\sum_{j=1}^{L} \left(\sum_{p=1}^{L}(H^{-1})_{jl}(H^{-1})_{kp}(\mathbf{x}_i)_p\right) g_j\left((H^{-1}\mathbf{x}_i)_j\right) \\
&= -N(H^{-1})_{kl} + \sum_{i=1}^{N}(H^{-1}\mathbf{x}_i)_k \sum_{j=1}^{L}(H^{-1})_{jl} g_s\left((H^{-1}\mathbf{x}_i)_j\right) \\
&= -N(H^{-1})_{kl} + \sum_{i=1}^{N}(H^{-1}\mathbf{x}_i)_k \sum_{j=1}^{L}(H^{-1})_{jl}(\mathbf{g}(H^{-1}\mathbf{x}_i))_s \\
&= -N(H^{-1})_{kl} + \sum_{i=1}^{N}(H^{-1}\mathbf{x}_i)_k \left((H^{-1})^t \mathbf{g}(H^{-1}\mathbf{x}_i)\right)_l,
\end{aligned} \tag{A.25}$$

avec g_j l'opposé de la fonction score de la source j (A.20) et $\mathbf{g}$ le vecteur contenant les opposées des fonctions scores de toutes les sources (A.21). La dérivée matricielle de la log-vraisemblance de l'ICA par rapport à la matrice de mixage est donc donnée

par :

$$\begin{aligned}
\Delta H &\propto \frac{\partial \mathcal{L}(H;\mathbf{X})}{\partial H} \propto -N.(H^{-1})' + \sum_{i=1}^{N} (H^{-1})' \mathbf{g}\left(H^{-1}.\mathbf{x}_i\right) (H^{-1}\mathbf{x}_i)' \\
&\propto -(H^{-1})' + \frac{1}{N}\sum_{i=1}^{N} (H^{-1})' \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i' \\
&\propto (H^{-1})' \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i' - \mathbf{I} \right), \qquad \text{(A.26)}
\end{aligned}$$

avec $\mathbf{z}_i = H^{-1}\mathbf{x}_i$. Le gradient naturel $\Delta_{nat} H$ correspondant est quant à lui donné par [Lewicki *et al.* 1997, Amari *et al.* 1996] :

$$\Delta_{nat} H = HH'\Delta H = H \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{g}\left(\mathbf{z}_i\right) \mathbf{z}_i' - \mathbf{I} \right). \qquad \text{(A.27)}$$

Bibliographie

[Adrot 2000] O. Adrot. *Diagnostic à base de modèles incertains utilisant l'analyse par intervalles : l'approche bornante.* PhD thesis, Institut National Polytechnique de Lorraine, 2000. 17

[Akaike 1973] H. Akaike. *Information theory and an extension of the maximum liekelihood principle.* In Proceedings of the Second International Symposium on Information Theory, 1973. 21

[Aknin *et al.* 2003] P. Aknin, L. Oukhellou et F. Vilette. *Track circuit diagnosis by automatic analysis of inspection car measurements.* WCRR, 2003. 14, 70, 81, 111

[Alcorta Garcia & Frank 1996] E. Alcorta Garcia et Frank. *Analysis of a class of dedicated observer schemes to sensor fault isolation.* UK ACC International Conference, Exter, UK, 1996. 17

[Amari *et al.* 1996] S. Amari, A. Cichocki et H. H. Yang. *A New Learning Algorithm for Blind Signal Separation.* In Proceedings of the 8th Conference on Advances in Neural Information Processing Systems (NIPS), volume 8, pages 757–763. MIT Press, 1996. 47, 49, 52, 129

[Ambroise & Govaert 2000] C. Ambroise et G. Govaert. *EM algorithm for partially known labels.* In Proceedings of the 7th Conference of the International Federation of Classication Societies (IFCS), pages 161–166. Springer, 2000. 22, 87

[Ambroise *et al.* 2001] C. Ambroise, Denoeux T., G. Govaert et Ph. Smets. *Learning from an imprecise teacher : probabilistic and evidential approaches.* In Proceedings of the 10th International symposium on applied stochastic models and data analysis (ASMDA), volume 1, pages 100–105, 2001. 88

[Attias 1999] H. Attias. *Independent Factor Analysis.* Neural Computation, vol. 11, no. 4, pages 803–851, 1999. 48, 49, 55, 76

[Attias 2000] H. Attias. *Independent factor analysis with temporally structured factors.* In Proceedings of the 12th Conference on Advances in Neural Information Processing Systems (NIPS), pages 386–392. MIT Press, 2000. 76, 77, 78, 79

[Bach & Jordan 2003] F. R. Bach et M. Jordan. *Kernel independent component analysis.* Journal of Machine Learning Research, vol. 3, pages 1–48, 2003. 49

[Bakir *et al.* 2006] 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.* In Proceedings of the IEEE International Conference on Image Processing, pages 3237–3240, 2006. 63

[Bartholomew & Martin 1999] D. J. Bartholomew et K. Martin. Latent variable models and factor analysis. Arnold, London, 1999. Seconde édition. 46, 64

[Basseville & Benveniste 1986] M. Basseville et A. Benveniste. Detection of abrupt changes in signals and dynamical systems. Springer Lecture Notes in Control and Information Sciences, 1986. 17

[Basseville & Nikiforov 1993] M. Basseville et Igor V. Nikiforov. Detection of abrupt changes : theory and application. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. 17

[Basseville 1988] M. Basseville. *Detecting changes in signals and systems.* Automatica, vol. 24, no. 3, pages 309–326, 1988. 16

[Baum *et al.* 1970] 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, vol. 41, no. 1, pages 164–171, 1970. 33, 41

[Beard 1971] R.V. Beard. *Failure Accomodation in Linear System through Self-Reorganization.* Man Vehicle Lab., Report MVT-71-1, Massachussets Institute of Technology, 1971. 16

[Bell & Sejnowski 1995a] A. J. Bell et T. J. Sejnowski. *Fast Blind Separation Based on Information Theory.* In in Proc. Intern. Symp. on Nonlinear Theory and Applications (NOLTA), Las Vegas, pages 43–47, 1995. 53

[Bell & Sejnowski 1995b] A. J. Bell et T. J. Sejnowski. *An Information-Maximization Approach to Blind Separation and Blind Deconvolution.* Neural Computation, vol. 7, no. 6, pages 1129–1159, 1995. 47, 52, 53

[Bellman 1957] R. Bellman. Dynamic programming. Princeton university Press, 1957. 19

[Ben Yaghlane *et al.* 2000] B. Ben Yaghlane, Ph. Smets et K. Mellouli. *Independence concepts for belief functions.* In Proceedings 8th International Conference IPMU Information Processing and Management of Uncertainty in Knowledge-based Systems, volume 1, pages 357–364, Madrid, Spain, 2000. 97

[Besag 1974] J. Besag. *Spatial interaction and the statistical analysis of lattice systems.* Journal of the Royal Statistical Society B, vol. 35, pages 192–236, 1974. 30

[Biernacki *et al.* 2003] 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, vol. 41, pages 561–575, 2003. 35

[Bishop & Lasserre 2007] C. Bishop et J. Lasserre. *Generative or Discriminative ? Getting the best of both worlds.* Bayesian Statistics, vol. 8, pages 3–23, 2007. 22

[Bishop 2006] C. Bishop. Pattern recognition and machine learning. Springer, 2006. 18

[Blum & Langley 1997] Avrim L. Blum et Pat Langley. *Selection of relevant features and examples in machine learning.* Artif. Intell., vol. 97, pages 245–271, December 1997. 19

[Bollen 1989] K. A. Bollen. Structural equations with latent variables. Wiley, 1989. 64

[Bouchard 2005] G. Bouchard. *Generative models in supervised statistical learning with applications to digital image categorization and structural reliability.* PhD thesis, Université Joseph Fourier - Grenoble 1, 2005. 22

[Breiman *et al.* 1984] L. Breiman, J. H. Friedman, R. A. Olshen et C. J. Stone. Classification and Regression Trees. CRC Press, 1984. 23

[Canarail 2003] Canarail. *Evaluation préliminaire des tracés, des technologies et des coûts d'implantation inhérents à un train haute vitesse entre Montréal et la frontière américaine - Revue des technologies en utilisation commerciale.* Rapport technique 03-108, Ministère des transports du Québec, Juillet 2003. 7

[Cappé & Moulines 2007] O. Cappé et E. Moulines. *Online EM Algorithm for Latent Data Models*, 2007. 35

[Cardoso & Laheld 1996] J. F. Cardoso et B. Laheld. *Equivariant adaptive source separation.* IEEE Transactions on Signal Processing, vol. 44, no. 12, pages 3017–3030, 1996. 47

[Cardoso 1997] J. F. Cardoso. *Infomax and maximum likelihood for source separation.* IEEE Letters on Signal Processing, vol. 4, no. 4, pages 112–114, 1997. 47

[Cardoso 1999] J. F. Cardoso. *High-order contrasts for independent component analysis.* Neural Computation, vol. 11, no. 1, pages 157–192, 1999. 47, 49

[Celeux & Diebolt 1988] G. Celeux et J. Diebolt. *A random inputation principle : The Stochastic EM Algorithm.* Rapport technique 901, INRIA, 1988. 35

[Celeux & Govaert 1992] G. Celeux et G. Govaert. *A classification EM algorithm for clustering and two stochastic versions.* Computation Statistics and Data Analysis, vol. 14, pages 315–332, 1992. 35

[Chapelle *et al.* 2006] O. Chapelle, B. Schölkopf et A. Zien, editeurs. Semi-supervised learning. MIT Press, 2006. 22, 87

[Chen & Patton 1999] J. Chen et R.J. Patton. Robust model-based fault diagnosis for dynamic systems. Kluwer Academic Publishers, Norwell, MA, USA, 1999. 16

[Choudrey & Roberts 2003] R. A. Choudrey et S. J. Roberts. *Variational mixture of Bayesian independent component analyzers.* Neural Computation, vol. 15, pages 213–252, January 2003. 120

[Churchill 1989] G. Churchill. *Stochastic models for heterogeneous DNA sequences.* Bulletin of Mathematical Biology, vol. 51, pages 79–94, 1989. 40

[Côme *et al.* 2009] E. Côme, L. Oukhellou, T. Denoeux et P. Aknin. *Learning from partially supervised data using mixture models and belief functions.* Pattern Recognition, vol. 42, no. 3, pages 334–348, 2009. 22, 101, 102, 117

[Comon 1994] P. Comon. *Independent Component Analysis, a new concept ?* Signal Processing, vol. 36, no. 3, pages 287–314, 1994. Special issue on Higher-Order Statistics. 47

[Cover & Thomas 1991] T. M. Cover et J. A. Thomas. Elements of information theory. Wiley, 1991. 34

[Côme 2009] E. Côme. *Apprentissage de modèles génératifs pour le diagnostic de systèmes complexes avec labellisation douce et contraintes spatiales.* PhD thesis, Université de Technologie de Compiègne, 2009. 3, 24, 25, 26, 27, 64, 100, 101

[Debiolles 2007] 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. 3, 17, 23, 24, 26, 27

[Demartines & Hérault 1997] P. Demartines et J. Hérault. *Curvilinear component analysis : a self organizing neural network for non linear mapping of data sets.* IEEE Transactions on Neural Networks, vol. 8, no. 1, pages 148–154, 1997. 20

[Dempster *et al.* 1977] 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, Series B, vol. 39, no. 1, pages 1–38, 1977. 33, 34, 41

[Dempster 1967] A. P. Dempster. *Upper and Lower Probabilities Induced by a Multivalued Mapping.* Annals of Mathematical Statistics, vol. 38, no. 2, pages 325–339, 1967. 26, 88, 89, 93

[Dempster 1968] A. P. Dempster. *A generalization of Bayesian inference.* Journal of the Royal Statistical Society, vol. 30, pages 205–247, 1968. 93

[Denoeux 2008] T. Denoeux. *Conjunctive and Disjunctive Combination of Belief Functions Induced by Non Distinct Bodies of Evidence.* Artificial Intelligence, vol. 172, pages 234–264, 2008. 95

[Denoeux 2010] T. Denoeux. *Maximum Likelihood from Evidential Data : An Extension of the EM Algorithm.* In In C. Borgelt et al. (Eds), editeur, Combining Soft Computing and Statistical Methods in Data Analysis, volume 77 of *Advances in Soft Computing*, pages 181–188. Springer Berlin, Heidelberg, 2010. 22, 101, 102, 103, 117

[Diebold *et al.* 1994] F. Diebold, J.-H. Lee et G. Weinbach. *Regime Switching with Time-Varying Transition Probabilities.* Nonstationary Time Series Analysis and Cointegration. (Advanced Texts in Econometrics, C.W.J. Granger and G. Mizon, eds.), 1994. 40

[Dubois & Prade 1986] D. Dubois et H. Prade. *A set-theoretic view of belief functions : logical operations and approximations by fuzzy sets.* International Journal of General Systems, vol. 12, pages 193–226, 1986. 94

[Dubois & Prade 1988] D. Dubois et H. Prade. *Representation and combination of uncertainty with belief functions and possibility measures.* Computational Intelligence, vol. 4, no. 4, pages 244–264, 1988. 88, 99

[Dubuisson 1990] B. Dubuisson. Diagnostic et reconnaissance des formes. Hermès, 1990. 18

[Dubuisson 2001] B. Dubuisson. Diagnostic, intelligence artificielle et reconnaissance des formes. Hermès, 2001. 15

[Duda *et al.* 2000] R. O. Duda, P. E. Hart et D. G. Stork. Pattern classification (2nd edition). Wiley, 2000. 18, 19

[Efron & Tibshirani 1994] B. Efron et R. J. Tibshirani. An introduction to the bootstrap. Chapman & Hall, 1994. 20

[Efron *et al.* 2004] B. Efron, T. Hastie, I. Johnstone et R. J. Tibshirani. *Least angle regression.* Annals of Statistics, vol. 32, no. 2, pages 407–499, 2004. 21

[Elouedi *et al.* 2004] Z. Elouedi, K. Mellouli et Ph. Smets. *Assessing sensor reliability for multisensor data fusion within the transferable belief model.* Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on, vol. 34, no. 1, pages 782–787, 2004. 120

[Fan & Li 2001] J. Fan et R. Li. *Variable selection via nonconcave penalized likelihood and its oracle properties.* Journal of the American Statistical Association, vol. 96, no. 456, pages 1348–1360, 2001. 67

[Florea *et al.* 2006] M. C. Florea, J. Dezert, P. Valin, F. Smarandache et A-L. Jousselme. *Adaptative combination rule and proportional conflict redistribution rule for information fusion.* COGnitive systems with Interactive Sensors, vol. abs/cs/0604042, 2006. 99

[Frank 1990] P.M. Frank. *Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy - A survey and some new results.* Automatica, vol. 26, no. 3, pages 459–474, 1990. 16, 17

[Fukunaga 1990] Keinosuke Fukunaga. Introduction to statistical pattern recognition (2nd ed.). Academic Press Professional, Inc., San Diego, CA, USA, 1990. 19

[Ghahramani 2004] Zoubin Ghahramani. *Unsupervised learning.* In Advanced Lectures on Machine Learning, pages 72–112. Springer-Verlag, 2004. 21, 30

[Golub & Kahan 1965] G.H. Golub et W. Kahan. *Calculating the singular values and pseudo-inverse of a matrix.* Journal of the Society for Industrial and Applied Mathematics : Series B, Numerical Analysis, vol. 2, no. 2, pages 205–224, 1965. 19

[Grandvalet & Bengio 2005] Y. Grandvalet et Y. Bengio. *Semi-supervised learning by entropy minimization.* In Advances in Neural Information Processing Systems, volume 17, pages 529–536, 2005. 87

[Ha-Duong 2008] M. Ha-Duong. *Hierarchical fusion of expert opinions in the Transferable Belief Model, application to climate sensitivity.* International Journal of Approximate Reasoning, vol. 49, no. 3, pages 555–574, 2008. 109

[Hastie *et al.* 2006] T. Hastie, T. Tibshirani et J. Friedman. The elements of statistical learning, data mining, inference and prediction. Statistics. Springer, 2006. 20, 21, 23

[Hérault *et al.* 1985] J. Hérault, C. Jutten et B. Ans. *Détection de grandeur primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non-supervisé.* In Actes du 10ème colloque GRETSI, pages 1017–1022, 1985. 47, 49

[Hoerl & Kennard 1970] A.E. Hoerl et R.W Kennard. *Ridge regression : biased estimation for nonorthogonal problems.* Technometrics, vol. 12, pages 55–67, 1970. 68

[Hosmer 1989] D. W. Hosmer. Applied logistic regression. Wiley Sons, 1989. 23

[Hotelling 1933] H. Hotelling. *Analysis of a complex of statistical variables into principal components.* Journal of Educational Psychology, vol. 24, pages 417–441,498–520, 1933. 47

[Hoyer 2004] P.O. Hoyer. *Non-negative Matrix Factorization with Sparseness Constraints.* Journal of Machine Learning Research, vol. 5, pages 1457–1469, December 2004. 72

[Hughes *et al.* 1999] J.P. Hughes, P. Guttorp et S.P. Charles. *A non-homogeneous hidden Markov model for precipitation occurrence.* Applied Statistics, pages 15–30, 1999. 40

[Hüllermeier & Beringer 2005] E. Hüllermeier et J. Beringer. *Learning from Ambiguously Labeled Examples.* In Proceedings of the 6th International Symposium on Intelligent Data Analysis (IDA), pages 168–179, 2005. 22

[Hyvärinen & Karthikesh 2002] A. Hyvärinen et R. Karthikesh. *Imposing Sparsity on the Mixing Matrix in Independent Component Analysis.* Neurocomputing, vol. 49, no. 1, pages 151–162, 2002. 64, 67

[Hyvärinen *et al.* 2001] A. Hyvärinen, J. Karhunen et E. Oja. Independant component analysis. Wiley, 2001. 19, 47, 48, 49, 51, 53, 57, 126

[Hyvärinen 1999] A. Hyvärinen. *Fast and Robust fixed point algorithms for independant component analysis.* IEEE, Transaction on Neural Networks, vol. 10, no. 3, 1999. 47

[Ikeda 2000] S. Ikeda. ICA on noisy data : a factor analysis approach, pages 201–215. Springer, 2000. 48

[Isermann 1984] R. Isermann. *Process fault detection based on modeling and estimation methods.* Automatica, vol. 20, no. 3, pages 387–404, 1984. 16

[Isermann 1997] R. Isermann. *Supervision, Fault-detection and fault-diagnosis Methods- An introduction.* Control Engineering Practice, vol. 5, no. 5, pages 639–652, 1997. 15, 16

[Isermann 2006] R. Isermann. An introduction from fault detection to fault tolerance. Spinger, New York, NY, USA, 2006. 15

[Jaakkola & Haussler 1998] T. Jaakkola et D. Haussler. *Exploiting Generative Models in Discriminative Classifiers.* In In Advances in Neural Information Processing Systems 11, pages 487–493. MIT Press, 1998. 23

[Jackson 1991] J. E. Jackson. A user's guide to principal components. Wiley New York, 1991. 19

[Jebara 2004] T. Jebara. Machine learning : Discriminative and generative. Kluwer Academic (Springer), 2004. 23

[Jensen 2001] F.V. Jensen. Bayesian networks and decision graphs. Springer, 2001. 23

[Jin & Liu 2005] R. Jin et Y. Liu. *A Framework for Incorporating Class Priors into Discriminative Classification.* In Tu Ho, David Cheung et Huan Liu, editeurs, Advances in Knowledge Discovery and Data Mining, volume 3518 of *Lecture Notes in Computer Science*, pages 401–412. Springer Berlin / Heidelberg, 2005. 22

[Jolliffe 2002] I. T. Jolliffe. Principal component analysis. Springer, New York, NY, USA, 2002. 19

[Jordan 1999] M. I. Jordan. Learning in graphical models. MIT Press, 1999. 31

[Jordan 2004] M. I. Jordan. *Graphical models.* Statistical Science (Special Issue on Bayesian Statistics), no. 19, pages 140–155, 2004. 30

[Jordan 2006] M. I. Jordan. An introduction to graphical models. Berkeley, U. C., 2006. 31

[Jutten & Comon 2007a] C. Jutten et P. Comon, editeurs. Séparation de source 1, concepts de base et analyse en composantes indépendantes. Hermès, 2007. 47

[Jutten & Comon 2007b] C. Jutten et P. Comon, editeurs. Séparation de source 2, au-delà de l'aveugle et application. Hermès, 2007. 47, 51, 63, 64

[Larue *et al.* 2004] A. Larue, S. Hosseini et C. Jutten. *Markovian source separation in post-nonlinear mixtures.* In Proceedings of the International Conference on Independent Component Analysis, pages 702–709. Springer, 2004. 120

[Lasserre *et al.* 2006] J. Lasserre, C. Bishop et T. Minka. *Principled hybrids of generative and discriminative models.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006. 23

[Lawrence & Bishop 2000] N. Lawrence et C. Bishop. *Variational Bayesian Independent Component Analysis.* Rapport technique, University of Manchester, 2000. 120

[Lewicki *et al.* 1997] M. S. Lewicki, T. J. Sejnowski et H. Hughes. *Learning nonlinear overcomplete representations for efficient coding.* In Proceedings of the 10th Conference on Advances in Neural Information Processing Systems (NIPS), pages 815–821. MIT Press, 1997. 129

[Li *et al.* 2007] 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, vol. 48, no. 1-2, pages 83–97, 2007. 63

[MacKay 1996] D. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Non publié, 1996. 52, 127, 128

[Martin & Osswald 2007] A. Martin et C. Osswald. *Toward a combination rule to deal with partial conflict and specificity in belief functions theory.* In International Conference on Information Fusion, 2007. 99, 120

[McLachlan & Krishnan 1996] G. J. McLachlan et T. Krishnan. The em algorithm and extensions. Wiley, New York, 1996. 35

[McLachlan & Peel 2000] G. J. McLachlan et D. Peel. Finite mixture models. Wiley, 2000. 23, 32, 35, 37

[McLachlan 1977] G. J. McLachlan. *Estimating the linear discriminant function from initial samples containing a small number of unclassified observations.* Journal of the American Statistical Association, vol. 72, no. 358, pages 403–406, 1977. 88

[Mercier *et al.* 2008] D. Mercier, B. Quost et T. Denoeux. *Refined modeling of sensor reliability in the belief function framework using contextual discounting.* Information Fusion, vol. 9, no. 2, pages 246–258, 2008. 120

[Moulines *et al.* 1997] E. Moulines, J. Cardoso et E. Cassiat. *Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 5, pages 3617–3620, 1997. 48, 49, 55

[Moussaoui 2005] S. Moussaoui. *Séparation de sources non-négatives. Application au traitement des signaux de spectroscopie.* PhD thesis, Université Henri Poincaré - Nancy I, 2005. 63

[Muri 1997] F. Muri. *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN.* PhD thesis, Université Paris V, 1997. 40

[Nocedal & Wright 1999] J. Nocedal et S. J. Wright. Numerical optimization. Springer Series in Operations Research. Springer, 1999. 52

[Oukhellou *et al.* 2006] L. Oukhellou, P. Aknin et E. Delechelle. *Infrastructure system diagnosis using empirical mode decomposition and Hilbert transform.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 3, 2006. 13, 14, 70, 81, 111

[Oukhellou *et al.* 2010] L. Oukhellou, A. Debiolles, T. Denoeux et P. Aknin. *Fault diagnosis in railway track circuits using Dempster-Shafer classifier fusion.* Engineering Applications of Artificial Intelligence, no. 23, pages 117–128, 2010. 23

[Pearl 1988] J. Pearl. Probabilistic reasoning in intelligent systems : Networks of plausible inference. Morgan Kaufmann, 1988. 30

[Pearlmutter & Parra 1996] B.A. Pearlmutter et L.C. Parra. *A Context-Sensitive Generalization of ICA*. In International Conference on Neural Information Processing, 1996. 76

[Pearlmutter & Parra 1997] B.A. Pearlmutter et L.C. Parra. *Maximum likelihood blind source separation : A context-sensitive generalization of ICA*. In C. Mozer, M.I. Jordan et T. Petsche, editeurs, Advances in Neural Information Processing Systems, volume 9, pages 613–619. MIT press, 1997. 76

[Pearson 1901] K. Pearson. *On Lines and Planes of Closest Fit to Systems of Points in Space*. Philosophical Magazine, vol. 2, no. 6, pages 559–572, 1901. 47

[Penny *et al.* 2000] W. Penny, S. Roberts et R. Everson. *Hidden Markov Independent Components Analysis*. In M. Girolami, editeur, Advances in Independent Component Analysis. Kluwer Academic Publishers, 2000. 76

[Petersen & Pedersen 2008] K. B. Petersen et M. S. Pedersen. *The Matrix Cookbook*, 2008. 127, 128

[Quost *et al.* 2011] B. Quost, M-H. Masson et T. Denoeux. *Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules*. International Journal of Approximate Reasoning, vol. 52, pages 353–374, 2011. 109

[Rabiner & Juang 1993] L. Rabiner et B. H. Juang. Fundamentals of speech recognition. Signal Processing. Prentice Hall, 1993. 23, 25, 30, 40, 42, 43

[Rabiner 1989] L. Rabiner. *A tutorial on hidden markov models and selected applications in speech recognition*. In Proceedings of the IEEE, pages 257–286, 1989. 39, 40, 43

[Ramasso 2009] E. Ramasso. *Contribution of belief functions to hidden markov models with an application to fault diagnosis*. In Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6, 2009. 120

[Rissanen 1978] J. Rissanen. *Modeling by shortest data description*. Automatica, vol. 14, pages 465–471, 1978. 21

[Roberts & Everson 2001] S. Roberts et R. Everson, editeurs. Indepedent component analysis, principles and practices. Cambridge Univeristy Press, 2001. 47

[Rosenbalt 1958] F. Rosenbalt. *The perceptron : a probabilistic model for information storage and organization in the brain*. Psychological Review, vol. 65, pages 386–408, 1958. 23

[Roweis 1998] S. Roweis. *EM Algorithms for PCA and SPCA*. In M. I. Jordan, M. J. Kearns et S. A. Solla, editeurs, Proceedings of the 11th Conference on Advances in Neural Information Processing Systems (NIPS), volume 10. MIT Press, 1998. 47

[Samé *et al.* 2007] A. Samé, C. Ambroise et G. Govaert. *An online classification EM algorithm based on the mixture model.* Statistics and Computing, vol. 17, no. 3, pages 209–218, 2007. 35

[Schwarz 1978] G. Schwarz. *Estimating the number of components in a fnite mixture model.* Annals of Statistics, vol. 6, pages 461–464, 1978. 21

[Shafer 1976] G. Shafer. A mathematical theory of evidence. Princeton University Press, Princeton, 1976. 23, 25, 26, 88, 89, 91, 95, 97, 99

[Shannon 1948] C. E. Shannon. *A Mathematical theory of communication.* Bell Systems Techinal Journal, vol. 27, pages 379–423,623–656, 1948. 49

[Smets & Kennes 1994] Ph. Smets et R. Kennes. *The transferable belief model.* Artificial Intelligence, vol. 66, no. 2, pages 191–234, 1994. 89, 91

[Smets 1990a] Ph. Smets. *The Combination of Evidence in the Transferable Belief Model.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 5, pages 447–458, 1990. 89, 91, 98

[Smets 1990b] Ph. Smets. *Constructing the Pignistic Probability Function in a Context of Uncertainty.* In UAI '89 : Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence, pages 29–40, Amsterdam, The Netherlands, The Netherlands, 1990. North-Holland Publishing Co. 98

[Smets 1993] Ph. Smets. *Belief functions : The disjunctive rule of combination and the generalized Bayesian theorem.* International Journal of Approximate Reasoning, vol. 9, no. 1, pages 1 – 35, 1993. 94

[Smets 1995] Ph. Smets. *The canonical decomposition of a weighted belief.* In Proceedings of the 14th international joint conference on Artificial intelligence, volume 2, pages 1896–1901, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. 95

[Smets 2002] Ph. Smets. *The Application of the Matrix Calculus to Belief Functions.* International Journal of Approximate Reasoning, vol. 31, pages 1–30, 2002. 93, 94

[Smets 2005] Ph. Smets. *Decision making in the TBM : the necessity of the pignistic transformation.* International Journal of Approximate Reasoning, vol. 38, no. 2, pages 133–147, 2005. 98

[Sodoyer 2004] D. Sodoyer. *La séparation de sources audiovisuelles.* PhD thesis, Institut National Polytechnique de Grenoble, 2004. 49

[Spearman 1904] C. Spearman. *General intelligence, objectively determined and measured.* American Journal of psychology, vol. 15, pages 201–293, 1904. 30, 46

[Stephens 1997] M. Stephens. *Bayesian Methods for Mixtures of Normal Distributions.* PhD thesis, University of Oxford, 1997. 32

[Taleb & Jutten 1999] A. Taleb et C. Jutten. *Sources separation in post-nonlinear mixtures.* vol. 10, no. 47, pages 2807–2820, 1999. 120

[Tarassenko 1995] L. Tarassenko. *Novelty detection for the identification of masses in mammograms.* In Proceedingd of the 4th IEE International Conference on Artificial Neural Networks, volume 4, pages 442–447, 1995. 22

[Tenenhaus 1998] M. Tenenhaus. La regression pls : Theorie et pratique. Springer, Paris, 1998. 20

[Theodoridis & Koutroumbas 2006] S. Theodoridis et K. Koutroumbas. Pattern recognition. Academic Press, 2006. 19

[Thurstone 1947] L. L. Thurstone. Multiple factor analysis. University of Chicago Press, 1947. 30

[Tibshirani 1996] Robert Tibshirani. *Regression shrinkage and selection via the lasso.* J. Roy. Statist. Soc. Ser. B, vol. 58, no. 1, pages 267–288, 1996. 68

[Tipping & Bishop 1997] M. E. Tipping et C. Bishop. *Probabilistic principal component analysis.* Journal of the Royal Statistical Society, Series B, vol. 61, pages 611–622, 1997. 47

[Titterington 1984] D. M. Titterington. *Recursive parameter estimation using incomplete data.* Journal of the Royal Statistical Society, vol. 46, no. 2, pages 257–267, 1984. 35

[Uebel 1989] H. Uebel. *Signalling system for German high speed lines.* In Main Line Railway Electrification, 1989., International Conference on, pages 36–39, 1989. 7

[Vapnik 1999] V. N. Vapnik. The nature of statistical learning theory (information science and statistics). Springer, 1999. 23

[Wang & Zhao 2002] S. Wang et Y. Zhao. *Almost sure convergence of Titterington's recursive estimator for mixture models.* In Proceedings of the IEEE International Symposium on Information Theory, ISIT (2002), pages 306–313. Morgan Kaufmann, San Francisco, CA, 2002. 35

[Weber 1999] P. Weber. *Diagnostic de procédé par analyse des estimations paramétriques de modèles de représentation à temps discret.* PhD thesis, Institut National Polytechnique de Grenoble, France, 1999. 16

[Willsky 1976] A.S. Willsky. *A survey of design methods for failure detection in dynamic systems.* Automatica, vol. 12, pages 601–611, 1976. 16

[Wu 1983] C. F. Wu. *On the convergence Properties of the EM algorithm.* Annals of Statistics, vol. 11, pages 95–103, 1983. 34

[Xu & Jordan 1996] L. Xu et M. Jordan. *On Convergence Properties of the EM Algorithm for Gaussian Mixtures.* Neural Computation, vol. 8, no. 1, pages 129–151, 1996. 34

[Yager 1983] R. R. Yager. *Hedging in the combination of evidence.* Journal of Information and Optimization Sciences, vol. 4, no. 1, pages 73–81, 1983. 98

[Yager 1987] R. R. Yager. *On the Dempster-Shafer framework and new combination rules.* Informations Sciences, vol. 41, no. 2, pages 93–137, 1987. 99

[Zadeh 1978] L. A. Zadeh. *Fuzzy sets as a basis for a theory of possibility.* Fuzzy Sets Systems, vol. 1, 1978. 88

[Zhang & Chan 2006] K. Zhang et L. W. Chan. *ICA with Sparse Connections.* In Proceedings of Intelligent Data Engineering and Automated Learning Conference (IDEAL), pages 530–537. Springer, 2006. 64, 67

[Zhang *et al.* 1994] Q. Zhang, M. Basseville et A. Benveniste. *Early warning of slight changes in systems.* Automatica, Special issue on Statistical Methods in Signal Processing and Control, vol. 30, no. 1, pages 95–113, 1994. 17

[Zwingelstein 2002] G. Zwingelstein. Diagnostic des défaillances. Hermès Science Publications, 2002. 15, 17

Notations

Cette partie a pour but de résumer les principales notations utilisées dans ce document. De manière générale, les lettres capitales X, Y, Z, représentent des variables aléatoires, les lettres calligraphiées $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, représentent les domaines de définition de ces variables aléatoires. Les réalisations des variables aléatoires multidimensionnelles sont dénotées par des lettres minuscules grasses $\mathbf{x}, \mathbf{y}, \mathbf{z}$, les matrices contenant des données par des lettres grasses majuscules $\mathbf{X}$, une application linéaire par une lettre majuscule A, H, G. Enfin, les paramètres sont représentés par des lettres grecs minuscules θ, π et grasses lorsqu'il s'agit de vecteurs $\boldsymbol{\psi}, \boldsymbol{\theta}$.

Notations générales	
$L(\boldsymbol{\psi}; \mathbf{X})$	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 $\boldsymbol{\psi}$
$\mathrm{E}[X]$	espérance de X
$\mathrm{E}[Y \vert X = x]$	espérance de Y conditionnellement à $X = x$
$X \perp\!\!\!\perp Y$	X indépendante de Y
$f(\,.\,)$	notation générique d'une densité
$f(\,.\,; \boldsymbol{\psi})$	notation générique d'une densité paramétrée par $\boldsymbol{\psi}$
$\mathcal{N}(\,.\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	loi normale multidimensionnelle de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$
H, G	matrices de mixage et matrice de demixage
H_{ij}	intersection de la i^{e} ligne et de la j^{e} colonne de H
$H_{i.}$	i^{e} ligne de H
$H_{.j}$	j^{e} colonne de H
$\mathbf{x}$	vecteur colonne
$(\mathbf{x})_i$	i^{e} ligne de $\mathbf{x}$
$\mathbf{x}'$	transposé de $\mathbf{x}$
H^{-1}	inverse de H
$\mathrm{tr}(H)$	trace de H
$\det(H)$	déterminant de H
$\mathrm{sign}(H)$	matrice des signes des éléments de H

Indices

k	indice sur un ensemble de classes ou groupes
i, l	indices sur un ensemble de variables latentes continues
j, p	indices sur un ensemble de variables observées

Constantes du problème

N	nombre d'observations
T	nombre d'observations temporelles
L	nombre de variables observées ou de variables latentes
K	nombre de classes
K_j	nombre de classes pour la j^{e} variable latente

Variables

$\mathbf{x}$	observations à valeurs continues $\in \mathrm{R}^L$
$\mathbf{z}$	variables latentes continues $\in \mathrm{R}^L$
y	variable à valeurs discrètes $\in \{1, \ldots, K_j\}$

Théorie des fonctions de croyances

2^Ω	l'ensemble des parties de l'ensemble Ω
$\lvert\Omega\rvert$	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
$\textcircled{\cup}$	combinaison conjonctive
$\textcircled{\cap}$	combinaison disjonctve
$\textcircled{\wedge}$	combinaison conjonctive prudente
$\uparrow$	extension vide
$\downarrow$	marginalisation

Glossaire

CdV Circuit de Voie

EM Expectation Maximization (Espérence Maximisation)

E^{2}M Evidential EM

FA Factor Analysis (Analyse Factorielle)

HMM Hidden Markov Model (Modèle de Markov Caché)

ICA Independent Component Analysis (Analyse en Composantes Indépendantes)

IFA Independent Factor Analysis (Analyse en Facteurs Indépendants)

Icc Courant de court-circuit

JES Joints Électrique de Séparation

LGV Ligne à Grande Vitesse

TGV Train à Grande Vitesse

TVM Transmission Voie Machine

Liste des publications

Revues Internationales

- Z. Cherfi, L. Oukhellou, E. Côme, T. Denœux, P. Aknin. *Partially supervised Independent Factor Analysis using soft labels elicited from multiple experts : Application to railway track circuit diagnosis.* Soft Computing. Special Issue in Knowledge Extraction from Low Quality Data (dernière révision).

Conférences Internationales

- Z. Cherfi, L. Oukhellou, E. Côme, T. Denœux et P. Aknin. Using imprecise and uncertain information to enhance the diagnosis of a railway device. *International Conference on Nonlinear Mathematics for Uncertainty and Its Applications NLMUA*, Pékin, Chine, Septembre 2011.

- Z. Cherfi, L. Oukhellou, T. Denœux et P. Aknin. Using imprecise and uncertain information to enhance the diagnosis of a railway device. *The 3rd International Conference of the ERCIM Working Group on Computing & Statistics*, Londres, Décembre 2010.

- Z. Cherfi, E. Côme, L. Oukhellou et P. Aknin. Railway device diagnosis using sparse independent component analysis. Dans *Proceedings of the 17th European Signal Processing Conference EUSIPCO, Glasgow, Ecosse*, Août 2009.

- Z. Cherfi, L. Oukhellou et P. Aknin. Using independent componant analysis for the diagnosis of a large scale railway system. Dans *Proceedings of the International Conference on Control Monitoring CM09, Dublin*, Irlande, 2009.

- 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 8th International Conference on Machine Learning and Applications (ICMLA)*, San-Diego, Décembre 2008.

Conférences Francophones

- Z. Cherfi, L. Oukhellou, E. Côme, P. Aknin et T. Denœux. Supervision partielle par des experts d'une analyse en facteurs indépendants. Application au diagnostic d'un système ferroviaire *Colloque du groupe de recherche et d'étude en traitement du signal (GRETSI)*, Bordeaux, 2011.

- Z. Cherfi, L. Oukhellou, P. Aknin et T. Denœux. Analyse en composantes indépendantes parcimonieuse pour le diagnostic de systèmes répartis. Dans *Actes du 12ème Colloque du groupe de recherche et d'étude en traitement du signal (GRETSI)*, Dijon, 2009.

- L. Oukhellou, E. Côme, Z.L. Cherfi, P. Aknin, T. Denœux. Diagnostic de systèmes répartis à l'aide de modèles génératifs en contexte supervisé ou partiellement supervisé. *Workshope Surveillance, Sûreté et Sécurité des Grands Systèmes*, Troyes, 2007.

Résumé : Ce travail de thèse présente l'élaboration de méthodes de diagnostic pour un système complexe de l'infrastructure ferroviaire, le circuit de voie. La tâche de diagnostic porte sur l'estimation de variables latentes, liées aux défauts, à partir de variables observées, extraites de signaux d'inspection et les solutions proposées s'appuient sur une approche générative permettant de modéliser les liens et relations entre ces variables. Dans le première partie de ces travaux, des méthodes non supervisées ont été envisagées pour le diagnostic. Les approches développées dans ce contexte ont montré l'intérêt de prendre en compte certaines informations a priori sur la structure du modèle ou sur l'aspect temporel de données prélevées séquentiellement. La seconde partie de cette thèse porte sur le diagnostic du système dans un cadre partiellement supervisé et consistait à utiliser des données réelles étiquetées de manière imprécise et incertaine par plusieurs experts lors de l'apprentissage. L'approche proposée repose sur l'utilisation de la théorie des fonctions de croyance pour modéliser et combiner les différents avis avant de les intégrer au modèle statistique proposé. Les résultats obtenus ont permis de montrer l'intérêt d'une telle démarche pour le diagnostic.

Mots clés : Diagnostic, Apprentissage non supervisé, Apprentissage partiellement supervisé, Analyse en Composantes Indépendantes, Analyse en Facteurs Indépendants, Algorithme EM, Théorie des fonctions de croyance.

Title : Complex system diagnosis in unsupervised and partially supervised contexts. Application to railway track circuit.

Abstract : This thesis aims to develop diagnosis tools for a complex railway infrastructure system, namely the track circuit. The diagnosis task is based on the estimation of latent variables linked to the system component defects from observed variables extracted from inspection signals. The proposed solutions are based on a generative approach which allows to model the relationships between these variables. In the first part, unsupervised methods are considered for the diagnosis. The models developed in this context have shown the interest of taking into account some prior information regarding the model structure or the temporal aspect of the collected data. The second part of this thesis focuses on the diagnosis of the system in a partially supervised context. In this part, real data partially labelled by several experts are available. The adopted approach is then based on the theory of belief functions to represent and fuse the different opinions intended to be used during the learning phase. The obtained results demonstrate the interest of this proposition for the diagnosis task.

Keywords : Diagnosis, Unsupervised learning, Partially supervised learning, Independent Component Analysis , Independent Factor Analysis, EM algorithm, Theory of belief functions.