N° d'ordre : 05/2014-D /EL


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Electronique et d'Informatique

THESE

Présentée pour l'obtention du garde de **DOCTORAT EN SCIENCES**

En : **ELECTRONIQUE**

Spécialité : **Traitement du Signal et des Images**


Par : **ROUIGUEB Abdenebi**


## Thème

# Réseaux Bayésiens par l'analyse en composantes indépendantes. Application : Identification multimodale des individus.

Soutenue publiquement, le 15-06-2014, devant le jury composé de :

| | | |
|---|---|---|
| Amar AÏSSANI | Professeur à l'USTHB | Président |
| Salim CHITROUB | Professeur à l'USTHB | Directeur de thèse |
| Hicham BOUSBIA-SALAH | Maître de Conférences/A, à l'ENP | Examinateur |
| Amar REBBOUH | Maître de Conférences/A, à l'USTHB | Examinateur |
| Camel TANOUGAST | HDR à l'Université de Lorraine, France | Examinateur |
| Mustapha DJEDDOU | Maître de Conférences/A, à l'EMP | Examinateur |

# Remerciements

Tout d'abord, je remercie vivement **Salim CHITROUB**, Professeur à l'U.S.T.H.B, de m'avoir proposé le sujet de thèse et d'en être le directeur, pour m'avoir guidé tout au long de ma thèse. Je le remercie pour sa disponibilité et pour la partie importante des contributions présentées dans ce mémoire.

Je tiens à exprimer mes remerciements à **Ahmed BOURIDANE**, Professeur à l'université de Northumbria (Newcastle,UK), mon co-promoteur de thèse, pour sa disponibilité et pour ses conseils avisés. Je le remercie pour m'avoir accueilli au sein de son laboratoire ainsi que pour les conditions de travail qu'il m'a procurées.

Je tiens à remercier les membres du jury, **Hicham BOUSBIA-SALAH**, Maître de Conférences A à l'École Nationale Polytechnique, **Amar REBBOUH**, Maître de Conférences A à l'U.S.T.H.B, **Camel TANOUGAST**, HDR à l'Université de Lorraine (France) et **Mustapha DJEDDOU**, Maître de Conférences A à l'École Militaire Polytechnique de m'avoir fait l'honneur d'examiner ce travail.

Je remercie vivement **Amar AISSANI**, Professeur à l'U.S.T.H.B, de m'avoir fait l'honneur de présider mon jury de thèse.

Je remercie toute ma famille et particulièrement mes parents pour leur soutien, pour leurs encouragements et leur patience.

Enfin, je remercie toute personne ayant contribué de près ou de loin à la finalisation de cette thèse, je remercie tout le personnel du laboratoire Mathématiques Appliquées (E.M.P) et particulièrement son responsable, **Abdennour LABED**, pour ses précieux conseils et aides.

# Table des matières

**Introduction générale** **1**

1 Contexte et problématique . . . . . 1

2 Objectifs . . . . . 4

3 Contributions . . . . . 4

4 Publications . . . . . 5

5 Structure du mémoire . . . . . 6

**Chapitre 1 chapitre 1, Réseaux Bayésiens : Concepts, Algorithmes, Avantages et Limites** **7**

1 Introduction . . . . . 9

2 Préliminaires . . . . . 9

2.1 Indépendance statistique . . . . . 9

2.2 Calcul probabiliste . . . . . 10

3 Réseaux Bayésiens . . . . . 11

3.1 Définition des réseaux Bayésiens . . . . . 11

3.2 Condition de Markov . . . . . 12

3.3 La notion de d-séparation . . . . . 13

4 Apprentissage de paramètres . . . . . 13

4.1 Maximum de vraisemblance (MV) . . . . . 14

4.2 Maximum a posteriori (MAP) . . . . . 14
5 Apprentissage de structure . . . . . 15
5.1 Notion d'équivalence de Markov . . . . . 15
5.2 Algorithmes à base de tests d'indépendance conditionnelle . . . . . 16
5.3 Algorithmes à base de score . . . . . 16
6 Inférence . . . . . 17
6.1 État de l'art des algorithmes d'inférence . . . . . 17
6.2 Algorithme de l'arbre de jonction . . . . . 18
6.2.1 Potentiel . . . . . 19
6.2.2 Opérations sur les potentiels . . . . . 19
6.2.3 Construction de l'arbre de jonction . . . . . 20
6.2.4 Propagation de croyances sur l'arbre de jonction . . . . . 21
7 Limites des RBs et directions de recherche actuelle . . . . . 24
7.1 Complexité . . . . . 26
7.1.1 Complexité en espace mémoire . . . . . 26
7.1.2 Complexité temporelle d'algorithmes d'apprentissage . . . . 26
7.1.3 Complexité temporelle d'algorithmes d'inférence . . . . . 27
7.2 Difficultés liées à l'apprentissage . . . . . 27
7.2.1 Apprentissage de la structure . . . . . 27
7.2.2 Apprentissage des paramètres . . . . . 27
7.2.3 Problème des valeurs manquantes . . . . . 28
7.3 Problème des variables latentes . . . . . 28
7.4 Problème des variables continues . . . . . 30
8 Conclusion . . . . . 30

**Chapitre 2 Analyse en Composantes Indépendantes : Algorithmes et Applications 33**

1 Introduction . . . . . 35
2 Analyse en composantes indépendantes . . . . . 37

2.1 Définition de l'ACI . . . 37
2.2 Ambiguïtés de l'ACI . . . 38
2.3 Restrictions de l'ACI . . . 38
3 Independence statistique . . . 38
3.1 Mutuelle information . . . 38
3.2 Information mutuelle et entropie . . . 39
3.3 Dépendance, Corrélation et Gaussianité . . . 40
4 Fonctions objectif de l'ACI . . . 41
4.1 Maximisation de la non Gaussianité . . . 41
4.1.1 Théorème de la limite centrale . . . 41
4.1.2 Mesures de la non Gaussianité . . . 42
4.2 Minimisation de l'information mutuelle . . . 43
4.3 Maximisation de vraisemblance . . . 44
4.4 Méthodes tensorielles . . . 44
5 Prétraitement pour l'ACI . . . 45
5.1 Centrage . . . 45
5.2 Blanchiment (whitening) . . . 46
6 Algorithmes ACI populaires . . . 47
7 Applications de l'ACI . . . 47
7.1 Séparation aveugle de sources . . . 47
7.2 Réduction de bruit . . . 47
7.3 Extraction de paramètres . . . 49
7.4 Application en domaine biomédical . . . 50
8 Extensions et variantes de l'ACI . . . 50
8.1 ACI non linéaire . . . 50
8.1.1 Mélanges post-non linéaires (PNLs) . . . 51
8.1.2 Mélanges convolutifs . . . 51

8.2 Analyse en sous-espaces indépendants (ASI) . . . . . 52
8.3 ACI parcimonieuse . . . . . 52
8.4 ACI topographique . . . . . 53
8.5 Limites de l'ACI . . . . . 53
9 Conclusion . . . . . 54

**Chapitre 3 Analyse en Composantes Indépendantes, Réseaux Bayésiens et Modèles Causaux** **57**

1 Introduction . . . . . 58
2 RBs et modèles causaux . . . . . 59
2.1 Le modèle causal est un RB . . . . . 59
2.2 Causalité et dépendance statistique . . . . . 60
2.3 Inférence Bayésienne et inférence causale . . . . . 61
2.4 Causalité probabiliste . . . . . 61
2.5 Connaissances causales et construction des RBs . . . . . 63
3 ACI et causalité . . . . . 64
4 Combinaison des techniques des RBs et de l'ACI . . . . . 65
4.1 Approche Bayésienne pour l'ACI . . . . . 65
4.2 L'ACI pour la recherche des facteurs cachés . . . . . 66
4.3 Extraction des paramètres par l'ACI : application aux RBs . . . . . 66
5 Correspondance entre RB et ACI . . . . . 67
5.1 Relation entre le modèle RB et et l'ACI . . . . . 67
5.2 Similitudes entre RB et ACI . . . . . 69
5.3 Divergences entre RB et ACI . . . . . 69
5.4 Correspondance entre les structure de RB et de l'ACI linéaire . . . . . 69
5.4.1 De la structure ACI vers la structure RB . . . . . 69
5.4.2 Exemples . . . . . 71
5.4.3 Le passage de la structure RB vers la structure ACI . . . . . 71

6 ACI, RB et modèles fonctionnels . . . . . 73
6.1 Relation entre les MESs et les RBs . . . . . 73
6.2 ACI, MESs et RBs . . . . . 74
7 Conclusion . . . . . 76

**Chapitre 4 Analyse en composantes indépendantes et réseaux Bayésiens pour la représentation de données 77**

1 Introduction . . . . . 79
2 ACI pour la représentation de données . . . . . 80
2.1 Estimation de densité par l'ACI linéaire . . . . . 80
2.2 Estimation de densité par l'ACI non linéaire . . . . . 81
2.2.1 ACI non linéaire pour l'estimation de densité . . . . . 81
2.2.2 ACI locale . . . . . 82
2.2.3 Estimation de densité par l'ACI locale . . . . . 82
2.2.4 Validité de clustering . . . . . 84
2.3 Amélioration de la classification des vecteurs MFCCs . . . . . 84
2.3.1 Le clustering pour la reconnaissance du locuteur indépendante de texte . . . . . 84
2.3.2 Le modèle de monde en RAL indépendante de texte . . . . . 85
2.3.3 ACI pour l'analyse de la voix . . . . . 86
2.3.4 Notre proposition . . . . . 87
2.4 Reconnaissance du locuteur par la fusion de la voix et du mouvement des lèvres . . . . . 94
2.4.1 Algorithmes de clustering flou . . . . . 94
2.4.2 Système de classification proposé . . . . . 96
2.4.3 Préparation de données . . . . . 98
2.4.4 Résultats et discussions . . . . . 101
3 ACI locale et RB pour la représentation de données : application en estimation de densité de données de hautes dimensions . . . . . 104

3.1 ACI locale, hautes dimensions et sparsity . . . . . 104
3.1.1 ACI et hautes dimensions . . . . . 104
3.1.2 Le clustering et les hautes dimensions . . . . . 105
3.1.3 L'ACI et la sparsity . . . . . 106
3.2 Réseaux Bayésiens et ACI pour l'estimation de densité . . . . . 108
3.2.1 Apprentissage de structure . . . . . 108
3.2.2 Arbre de jonction pour la factorisation . . . . . 109
3.2.3 Estimation de densité en utilisant ACI locale et RBs . . . . . 110
3.3 Résultats expérimentaux . . . . . 110
3.3.1 Application en estimation de densité . . . . . 110
3.3.2 Application en classification supervisée . . . . . 112
4 Conclusion . . . . . 116

**Chapitre 5 Inférence sur le modèle ACI linéaire (application dans la fusion de scores multi-biométriques 119**

1 Introduction . . . . . 120
2 Énonce du problème . . . . . 121
3 Solution par les réseaux Bayésiens discrets . . . . . 122
4 Algorithme d'inférence sur le modèle ACI . . . . . 122
4.1 Modèle proposé (ACI-RB) . . . . . 122
4.2 Apprentissage des paramètres . . . . . 124
4.3 Algorithme d'inférence . . . . . 125
4.4 Expérimentation . . . . . 129
4.4.1 Méthodologie . . . . . 129
4.4.2 Datasets . . . . . 130
4.4.3 Résultats . . . . . 131
4.4.4 Analyse de complexité . . . . . 132
5 Application pour la fusion des scores multi-biométriques . . . . . 133

5.1 Fusion des scores multi-biométriques . . . 133
5.2 Scores absents ou de mauvaise qualité . . . 134
5.3 Approche proposée . . . 134

**Chapitre 6 Inférence Bayésienne sur les variables continues en utilisant la méthode ACI locale** **137**

1 Introduction . . . 139
2 Limites du modèle ACI-RB . . . 140
3 Propositions . . . 140
4 Ajustement d'histogrammes par des séquences polynomiales . . . 141
4.1 Approximation de la densité par des fonctions polynomiales . . . 141
4.2 Approximation d'une fonction par une spline . . . 143
4.2.1 Définitions . . . 143
4.2.2 Régression par des splines . . . 143
4.2.3 Choix des nœuds de jonction . . . 144
4.2.4 Contraintes sur les extrémités . . . 144
4.2.5 Régression polynomiale (un seul morceau) . . . 145
4.2.6 Algorithme de régression développé . . . 147
5 Algorithme d'inférence . . . 149
5.1 Apprentissage des paramètres . . . 149
5.2 Inférence . . . 151
6 Intégration de polynômes sur des polytôpes convexes . . . 153
6.1 Introduction . . . 153
6.2 Énonce du problème . . . 154
6.3 Intégration d'un polynôme sur un polytôpe par la triangulation . . . 155
6.4 Intégration d'un polynôme multivariable sur le simplexe standard (simplexe unitaire) . . . 155
6.5 Intégration d'un polynôme multivariable sur un simplexe (cas général) . 156

6.6 Amélioration proposée pour accélérer l'intégration . . . . . 158
6.6.1 Objectif . . . . . 158
6.6.2 Une constatation intéressante . . . . . 159
6.6.3 Proposition . . . . . 159
6.7 Conclusion . . . . . 162
7 Expérimentation . . . . . 163
7.1 Description de données de test . . . . . 163
7.2 Méthodologie de comparaison . . . . . 165
7.3 Résultats obtenus . . . . . 166
8 Conclusion . . . . . 168

**Conclusion générale et perspectives** **171**

**Bibliographie** **173**

**Table des figures** **185**

**Liste des tableaux** **189**

**Glossaire** **191**

# Introduction générale

## 1 Contexte et problématique

De nombreux modèles statistiques ont été conçus pour la représentation de connaissances et le raisonnement à partir de ces représentations. Les variables observées ne représentent en réalité que la partie visible de l'iceberg, l'autre partie émergée mais aussi si importante, correspondant à des facteurs cachés et inconnus, explique correctement les relations entre les variables observées. Malheureusement, l'inaccessibilité des facteurs cachés engendre une forme d'incertitude. Les modèles graphiques probabilistes, et plus précisément les réseaux Bayésiens, proposés par Judea Pearl dans les années 1980s, se sont montrés pratiques et assez formels pour la représentation de connaissances incertaines ainsi que pour le raisonnement probabiliste à partir des données incomplètes.

Brièvement, un réseau Bayésien (RB), $R = (G, P)$, est défini par un graphe $G$ acyclique et orienté où chaque sommet correspond à une variable aléatoire $X_i$ et par un ensemble de distributions de probabilités conditionnelles $P = \{P(X_i|Parents_{X_i}, i = 1, ..., n\}$. La composante graphique $G$ désigne la structure de relations de dépendances conditionnelles entre les variables, et offre un outil visuel et compréhensible à l'utilisateur. En outre, les distributions $P$ permettent de quantifier les dépendances incertaines en utilisant les probabilités.

La construction d'un RB débute par la détermination des paramètres qualitatifs ($G$) en se servant des informations fournies par l'expert et|ou éventuellement la possibilité de les estimer à partir de données par des algorithmes d'apprentissage automatique de structure des RBs. Ensuite, une fois la structure établie, l'apprentissage des paramètres quantitatifs locaux ($P$) peut être réalisé en utilisant des techniques d'estimation des distributions conditionnelles. Cette séparation permet d'agir en deux étapes et facilitant ainsi l'apprentissage du modèle RB.

Les RBs tels qu'ils ont été introduits par Pearl permettent une factorisation de la distribution conjointe en un produit de distributions conditionnelles de dimensions réduites; $P(X) = \prod_{i=1}^{n} P(X_i|Parents_{X_i})$. Cette propriété est fondamentale dans la mesure où elle est à la base du développement d'algorithmes efficaces d'inférence Bayésienne, qui calculent la probabilité a

posteriori de n'importe quelle variable du modèle à partir de l'observation même partielle des autres variables (raisonnement probabiliste). Il convient de noter que l'inférence sur les RBs relève d'un problème d'optimisation des calculs.

Les tables de probabilités conditionnelles (TPCs) sont habituellement utilisées pour la quantification d'un RB lorsqu'il implique seulement des variables discrètes. Plusieurs méthodes d'inférence exacte sur des RBs discrets de petite taille ont été proposées dans la littérature [Lauritzen and Spiegelhalter, 1988, Shenoy and Shafer, 1990]. Ces méthodes effectuent une sorte de calcul local (*local computation*), qui veut dire qu'elles calculent les distributions marginales sans calculer la distribution conjointe par l'exploitation de la propriété de factorisation des RBs. Les algorithmes d'inférence exacte ne sont pas pratiques lorsque le graphe est dense (contenant plusieurs arcs) et/ou large (contenant plusieurs sommets), ou lorsque le nombre de valeurs discrètes par variable est grand parce que la complexité en espace mémoire nécessaire pour la sauvegarde des TPCs et des conjointes locales et la complexité en temps de calcul d'algorithmes d'inférence deviennent trop élevées. En conséquence, les méthodes heuristiques d'inférence qui diminuent la complexité au détriment de la précision sont utiles dans de telles situations. En plus du problème de la complexité calculatoire, dans les grands RBs discrets, le problème de sur-apprentissage [Hawkins, 2004] des TPCs s'aggrave du fait que le nombre de paramètres des TPCs à estimer devient plus important par rapport aux données d'apprentissage. Par exemple, la taille d'une TPC d'un sommet ayant 03 parents et où chaque variable a 10 valeurs possibles est de $10^4 = 10000$, le nombre de vecteurs d'apprentissage doit être donc très supérieur à $10^4$ (e.g $10^5$) et assez représentatif de la population.

En contrepartie, l'impossibilité de la prise en compte des variables continues demeure une limitation majeure des RBs. La solution triviale et aussi générale consiste à discrétiser les domaines des variables continues [Kozlov and Koller, 1997, Friedman and Goldszmidt, 1996]. Malheureusement, certaines variables continues exigent une haute résolution de discrétisation et contraignent à faire un compromis entre la complexité et la précision. Comme alternative, des solutions sans recours à la discrétisation ont été proposées seulement pour quelques types restreints et spécifiques de RBs. Si nous n'optons pas pour la discrétisation en utilisant des TPCs (qui est une solution non paramétrique, générale, mais qui a une complexité élevée), nous sommes contraints de considérer des solutions (semi-)paramétriques ou restreintes à des types particuliers de variables continues.

Le modèle conditionnel Gaussien (CLG) [Lauritzen, 1992] est un type bien connu des RBs continus. Un CLG est un modèle RB hybride contenant des variables discrètes et des variables continues où la distribution conjointe des variables continues étant donnée la valeur des variables discrètes est supposée être une Gaussienne multivariable. En outre, il a la restriction que les sommets discrets n'admettent pas des parents continus dans le graphe. Des algorithmes d'inférence pour un tel type de modèles ont été proposés (voir [Lauritzen and Jensen, 2001] et [Cowell, 2005]). Une autre classe importante de RBs hybrides emploie des potentiels de mélanges d'exponentielles tronquées (MTE) [Moral et al., 2001]. Un MTE peut être perçu comme une solution médiane entre la discrétisation et la modélisation analytique, où le domaine des va-

riables est subdivisé en un ensemble d'hyper-rectangles (pièces) et un mélange d'exponentielles tronquées est employé pour approximer la conjointe à l'intérieur de chaque pièce. Cependant, l'apprentissage d'un MTE n'est pas une tâche facile. La recherche d'un MTE optimal de plusieurs variables est un problème délicat puisque cela implique une optimisation non linéaire dans un espace de dimension élevée pour chaque hyper-rectangle, qui risque de converger vers des solutions optimales locales [Romero et al., 2006]. En parallel, adopter une résolution élevée de discrétisation (plusieurs pièces de petites tailles) pour améliorer l'apprentissage entraîne une augmentation de la complexité d'inférence.

La modélisation de dépendances statistiques entre variables aléatoires fait l'objet d'intérêt des RBs et également d'un grand nombre d'outils d'analyse statistique de données tels que la méthode de l'analyse en composantes indépendantes (ACI). L'ACI est une méthode statistique émergente d'ordre supérieur pour l'analyse de données (des signaux) multidimensionnelles. Le concept a été initialement proposé par Comon [Comon, 1994]. Dans sa forme simple, son objectif consiste à décomposer les vecteurs de données observées $\mathbf{x}(t) = (x_1(t), x_2(t), ..., x_n(t))^T$ en une combinaison linéaire $\mathbf{x}(t) = A\mathbf{s}(t)$ des composantes $s_1(t), s_2(t), ..., s_m(t)$ qui sont statistiquement indépendantes, où $A$ est une matrice inversible de mélange. En plus de l'hypothèse de l'indépendance mutuelle, la séparation ACI n'est possible que si les variables observées sont non Gaussiennes et que $n \geq m$. L'ACI a été originalement utilisée pour résoudre le problème de séparation aveugle de sources (BSS), mais son domaine d'application est plus large que celui de la BSS.

D'un point de vue géométrique, l'analyse en composantes principales est un simple changement linéaire de base, $\mathbf{x} = A\mathbf{y}$, qui a pour effet une nouvelle représentation dans des directions correspondant à des variables $y_1, ..., y_m$ décorrélées et ordonnées selon la variance expliquée par chacune. De la même manière, l'ACI peut être vue comme un changement linéaire, $\mathbf{x} = A\mathbf{s}$, de base dont l'ambition est par ailleurs de révéler des facteurs $s_1, ..., s_m$ qui soient indépendants au sens plein, et non plus seulement au sens de la non corrélation. Le but d'appliquer l'ACI peut se résumer à la volonté d'aller vers un espace indépendant car l'indépendance des variables est une propriété très désirable en analyse de données, et parfois ça consiste à identifier des variables physiques (sources, causes) afin de fournir une description plus réaliste car les sources sont souvent faiblement dépendantes dans le monde réel.

Fondamentalement, les RBs sont utilisés pour le raisonnement probabiliste et l'ACI pour résoudre le problème BSS. Cependant, ces deux modèles ont quelques points en commun ; les deux modèles procèdent par la recherche, la quantification (la mesure) et la représentation des relations de dépendance statistique entre des variables aléatoires. Le contexte de cette thèse vise à exploiter les résultats récents de l'ACI dans la perspective d'améliorer certains aspects relatifs aux RBs, à savoir :

1. emploi des méthodes ACI pour l'apprentissage de la structure $G$ ou des distributions conditionnelles $P$ d'un RB ;
2. utilisation des paramètres ACI pour la quantification d'un RB ;

3. extension des RBs à des données non standard ;
4. proposition de nouveaux modèles combinant les RBs et l'ACI.

# 2 Objectifs

A l'instar des modèles CLG et MTE, l'objectif principal de cette thèse consiste à proposer de nouveaux algorithmes d'inférence Bayésienne sur des variables continues, non Gaussiennes et dont les relations de dépendance inter-variables peuvent être approximées par des modèles ACI. Sous ces hypothèses, les distributions de densité des composantes indépendantes et les matrices de mélanges ACI vont jouer le rôle du graphe $G$ et des distributions conditionnelles $P$ (TPCs en cas discret) dans un RB classique.

Pour atteindre cet objectif, il nous appartient d'abord de montrer que les modèles ACI peuvent être utilisés pour fournir des représentations acceptables d'un nombre assez large de données réelles. La deuxième étape serait de concevoir des algorithmes d'inférence Bayésienne sur les représentations ACI proposées. Concrètement, nous voulons proposer un modèle d'inférence Bayésienne sur des données continues non Gaussiennes quelconques et qui peuvent être dérivées à partir des applications diverses. Cependant, dans cette thèse, les tests de prédiction seront globalement focalisés sur des données biométriques telles que la reconnaissance automatique du locuteur par la voix ou par la voix et le mouvement des lèvres, et la fusion multimodale des scores biométriques.

# 3 Contributions

Nos principales contributions peuvent être scindées en deux volets. Dans un premier lieu, la proposition des modèles de représentation de données par l'ACI (points **a)** et **b**)). Dans un second lieu, le développement d'algorithmes d'inférence Bayésienne sur ces représentations (points **c)** et **d**)). Elles sont synthétisées dans la liste ci-dessous.

– **a)** Utilisation de la méthode ACI locale à base de clustering flou pour l'estimation de densité, l'estimateur proposé est combiné avec le classificateur discriminatif SVM afin d'améliorer la classification supervisée des vecteurs MFCCs et la classification des vecteurs MFCCs fusionnés avec les attributs labiaux (mouvement de lèvres) dans un contexte de reconnaissance automatique du locuteur. Cette contribution a fait l'objet de deux articles de conférences internationales [Rouigueb et al., 2013b, Rouigueb et al., 2012].

– **b)** Combinaison des méthodes d'apprentissage de structure des RBs et de l'ACI locale floue pour l'estimation de densité de données de hautes dimensions. Les algorithmes RBs sont responsables de l'apprentissage des relations d'indépendance conditionnelle inter-variables et ainsi la factorisation de la distribution conjointe globale selon un arbre de

jonction valide en une fonction de distributions conjointes de dimensions réduites, alors que l'ACI floue est responsable de l'estimation des conjointes de dimensions réduites. Les simulations numériques d'estimation de densité sont également complétées et consolidées par une classification supervisée sur quelques benchmarks standard de classification en hautes dimensions. Les résultats obtenus sont résumés dans un article publié en ligne dans le journal "*Intelligent Data Analysis* " de Iospress [Rouigueb et al., 2014b].

– **c)** Proposition d'un algorithme d'inférence Bayésienne sur les clusters des variables qui peuvent être approximées par un modèle ACI linéaire standard. Des histogrammes de probabilité ont été utilisés pour la modélisation de densité des composantes indépendantes. Cette deuxième contribution a fait l'objet d'un article qui a été publié en ligne dans le journal ("*Journal of Applied Statistics*") de Taylor & Francis [Rouigueb et al., 2014a].

– **d)** Proposition d'un algorithme d'inférence Bayésienne sur des variables continues et non Gaussiennes quelconques où le modèle ACI local à base de clustering a été utilisé pour l'ajustement de données. Cette contribution est une généralisation du travail `'c)'` à des dépendances inter-variables non linéaires. Une amélioration a été également proposée par l'emploi des fonctions splines pour l'approximation des $P(s_i)$, une séquence de peu de polynômes de degrés inférieurs, au lieu des histogrammes (utilisés dans `'c)'`). Une méthode d'intégration des fonctions polynomiales (produits de splines) sur des hyper-polytôpes pour être utilisée dans l'inférence a été proposée et elle a fait l'objet d'une communication nationale [Rouigueb et al., 2013a]. Plus D'enrichissements et d'autres expérimentations de cet algorithme d'inférence sont en train d'être complétés afin de résumer ces résultats dans un papier qui sera soumis pour publication dans un journal spécialisé.

# 4 Publications

- Communications dans des conférences internationales et nationales avec comité de lecture :
  – A. ROUIGUEB, S. CHITROUB et A. BOURIDANE, "Intégration des fonctions polynomiales sur des polytôpes de dimensions quelconques". 8ème Conférence sur le Génie Electrique CGE'08, EMP Algérie, 16-17 Avril 2013.
  – A. ROUIGUEB, S. CHITROUB et A. BOURIDANE, "Local ICA To Improve Text-Independent MFCCS Vectors Classification", 2013 IEEE International Workshop on Machine Learning for Signal Processing, sept. 22-25, 2013, Southampton, UK.
  – A. ROUIGUEB, S. CHITROUB et A. BOURIDANE, "Fuzzy Local ICA for Speaker Recognition Using Voice and Lip Motion", The World Congress on Engineering 2012 (WCE2012). The 2012 International Conference of Data Mining and Knowledge Engineering, London, UK.

  - A. ROUIGUEB, S. CHITROUB et A. BOURIDANE, "ACI Locale Floue pour la Reconnaissance du Locuteur en Combinant la Voix et les Mouvements des Lèvres, 11ème Colloque Africain sur la Recherche en Informatique et Mathématiques Appliquées 2012 (CARI'12), Alger.
- Publications dans des revues spécialisées :
  - A. ROUIGUEB, S. CHITROUB et A. BOURIDANE, "Density Estimation of High Dimensional Data Using ICA and Bayesian Networks", *Intelligent data analysis journal*, IOS Press Publishers, Volume 18 (02), 157-179.
  - A. ROUIGUEB, S. CHITROUB et A. BOURIDANE, "Bayesian Inference over ICA models: Application to Multibiometric score fusion with quality estimates", *Journal of applied statistics*, Taylor & Francis, Volume 41 (10), 2123-2140.

# 5 Structure du mémoire

Le reste de ce rapport de thèse est organisé comme suit :

Au premier chapitre, une introduction aux RBs est brièvement étalée. Les définitions ainsi que les principales approches et applications de la méthode ACI sont données dans le deuxième chapitre.

Le chapitre III sera consacré à la présentation d'une étude théorique sur la relation entre les RBs et l'ACI autour du concept de la causalité. Dans lequel, quelques points en communs, divergences, correspondance et possibilités de complémentarité entre ces deux modèles seront détaillés. Au quatrième chapitre, nous proposons des modèles de représentation de données multivariables en utilisant la méthode ACI. Des applications en classification supervisée et en estimation de densité sur des données hétérogènes (biométrie, son, image, sismiques,etc.) seront également montrées.

Le cinquième et le sixième chapitre seront consacrés à la présentation des algorithmes d'inférence proposés sur le modèle ACI linéaire et le modèle ACI local. Les résultats obtenus sont comparés avec ceux des RBs discrets à base des TPCs.

Chapitre 1

# chapitre 1, Réseaux Bayésiens : Concepts, Algorithmes, Avantages et Limites

## Sommaire


**1 Introduction . . . . . 9**

**2 Préliminaires . . . . . 9**

2.1 Indépendance statistique . . . . . 9

2.2 Calcul probabiliste . . . . . 10

**3 Réseaux Bayésiens . . . . . 11**

3.1 Définition des réseaux Bayésiens . . . . . 11

3.2 Condition de Markov . . . . . 12

3.3 La notion de d-séparation . . . . . 13

**4 Apprentissage de paramètres . . . . . 13**

4.1 Maximum de vraisemblance (MV) . . . . . 14

4.2 Maximum a posteriori (MAP) . . . . . 14

**5 Apprentissage de structure . . . . . 15**

5.1 Notion d'équivalence de Markov . . . . . 15

5.2 Algorithmes à base de tests d'indépendance conditionnelle . . . . . 16

5.3 Algorithmes à base de score . . . . . 16

**6 Inférence . . . . . 17**

6.1 État de l'art des algorithmes d'inférence . . . . . 17

6.2 Algorithme de l'arbre de jonction . . . . . 18

6.2.1 Potentiel . . . . . 19

6.2.2 Opérations sur les potentiels . . . . . 19

6.2.3 Construction de l'arbre de jonction . . . . . 20

6.2.4 Propagation de croyances sur l'arbre de jonction . . . . . 21

**7 Limites des RBs et directions de recherche actuelle . . . . . 24**

7.1 Complexité . . . . . 26
7.1.1 Complexité en espace mémoire . . . . . 26
7.1.2 Complexité temporelle d'algorithmes d'apprentissage . . . 26
7.1.3 Complexité temporelle d'algorithmes d'inférence . . . . . 27
7.2 Difficultés liées à l'apprentissage . . . . . 27
7.2.1 Apprentissage de la structure . . . . . 27
7.2.2 Apprentissage des paramètres . . . . . 27
7.2.3 Problème des valeurs manquantes . . . . . 28
7.3 Problème des variables latentes . . . . . 28
7.4 Problème des variables continues . . . . . 30
**8 Conclusion . . . . . 30**

# 1 Introduction

Ce premier chapitre permet d'introduire les différents concepts liés aux réseaux Bayésiens (RBs). Les RBs représentent un outil puissant et formel de décision car ils sont capables de combiner l'aspect quantitatif sous forme de probabilités conditionnelles et l'aspect qualitatif par la spécification graphique des relations entre variables observées, ils s'appuient simultanément sur la théorie de graphes et celle de calcul probabiliste, ce qui explique l'appellation "modèles graphiques probabilistes" par certains auteurs pour désigner les RBs.

La formalisation des RBs a commencé vers la fin des années 80 grâce notamment aux travaux de Judea Pearl [Pearl, 1988]. Depuis, l'intérêt des RBs dans la communauté de l'intelligence artificielle n'a cessé de grandir. L'idée clé consiste à explorer et à exploiter les relations d'indépendance statistique, plutôt que de s'intéresser à la structure causale qui ne peut pas être toujours dévoilée à partir de données.

Les travaux actuels liés aux RBs portent principalement sur les algorithmes d'inférence [Madsen and Jensen, 1999, Shafer and Shenoy, 1990, Lauritzen and Spiegelhalter, 1988, Jensen et al., 1990], sur l'apprentissage de la structure du réseau [Spirtes et al., 1993, Cheng et al., 1997a, Cheng et al., 1997b, Heckerman et al., 1995, Chickering et al., 1995, Cooper and Herskovits, 1992, de Campos et al., 2003], sur l'apprentissage des paramètres du réseau [Naim et al., 2004, Niculescu et al., 2006] ou sur des nouvelles extensions des modèles des RBs. A l'heure actuelle, les algorithmes développés d'inférence exacte ou approchée et ceux d'apprentissage ne sont pratiques que sur des RBs discrets et de taille raisonnable.

Dans la suite de ce chapitre, nous visons en particulier à donner une vue d'ensemble sur les RBs où nous essayons de mettre l'accent sur la définition de la notion des RBs, les méthodes de construction des RBs et leur application au raisonnement probabiliste. Nous procédons en quatre temps par la présentation de :

- Concepts et définitions liés aux RBs.
- Un état de l'art des méthodes communes d'apprentissage de structures et de paramètres des RBs.
- Algorithmes d'inférence Bayésienne appliqués aux RBs.
- Quelques limites ainsi que les nouveaux axes ouverts de recherche en rapport avec les RBs.

# 2 Préliminaires

## 2.1 Indépendance statistique

Considérons $X = \{X_i, ..., X_n\}$, un ensemble de $n$ variables aléatoires. Dans les modèles graphiques probabilistes, deux propriétés sont centrales : l'indépendance marginale et l'indépen-

dance conditionnelle.

**Définition 1 (*Indépendance statistique*)**

*Deux variables aléatoires $X_i$ et $X_j$ sont (statistiquement) indépendantes si et seulement si nous avons :*
$P(X_i|X_j) = P(X_i)$

L'indépendance entre deux variables $X_i$ et $X_j$, notée $X_i \perp X_j$, veut dire qu'une information sur les valeurs prises par la variable $X_j$ n'a pas d'impact sur notre connaissance sur la distribution de probabilité de la variable $X_i$. Il découle de cette définition que $X_i \perp X_j$ est valide si et seulement si $P(X_i, X_j) = P(X_i) \times P(X_j)$, nous avons donc

$$X_i \perp X_j \quad \equiv \quad P(X_i, X_j) = P(X_i) \times P(X_j) \quad \equiv \quad P(X_i|X_j) = P(X_i) \tag{1}$$

Une non égalité $P(X_i|X_j) \neq P(X_i)$ implique que $X_i$ et $X_j$ sont statistiquement dépendantes. L'indépendance est une notion symétrique, si $X_i$ est indépendante de $X_j$, alors $X_j$ est aussi indépendante de $X_i$.

**Définition 2 (*Indépendance statistique conditionnelle*)**

*Deux variables aléatoires $X_i$ et $X_j$ sont conditionnellement indépendantes sachant un sous ensemble de variables $S \subset (X - \{X_i, X_j\})$, notée $X_i \perp X_j|S$, si et seulement si nous avons :*
$P(X_i|X_j, S) = P(X_i|S)$

Le concept d'indépendance conditionnelle est symétrique; si $X_i$ est conditionnellement indépendante de $X_j$ sachant $S$ alors $X_j$ est également conditionnellement indépendante de $X_i$ sachant $S$. Par rapport à la distribution de la conjointe et les distributions marginales, l'indépendance conditionnelle $X_i \perp X_j|S$ est vérifiée si et seulement si $P(X_i, X_j|S) = P(X_i|S) \times P(X_j|S)$, nous avons

$$X_i \perp X_j|S \quad \equiv \quad P(X_i, X_j|S) = P(X_i|S) \times P(X_j|S) \tag{2}$$

Une non égalité implique que $X_i$ et $X_j$ sont conditionnellement dépendantes sachant $S$. Il est bien de mentionner que l'indépendance représente un cas spécial de l'indépendance conditionnelle, lorsque l'ensemble de conditionnement $S$ est vide.

## 2.2 Calcul probabiliste

**Théorème 1 (*Règle fondamentale*)**

*Pour tout couple x,y des variables aléatoires, nous avons la factorisation suivante:*

$$P(x, y) = P(x|y) \times P(y) \tag{3}$$

Cette règle nous montre comment nous pouvons calculer la distribution conjointe $P(x, y)$ lorsque nous disposons de la distribution marginale $P(y)$ et de la distribution conditionnelle $P(x|y)$.

**Théorème 2 (*Règle de Bayes*)**

*Pour tout couple x,y de variables aléatoires, nous avons :*

$$P(x|y) = \frac{P(y|x) \times P(x)}{P(y)} \tag{4}$$

La règle de Bayes nous fournit un moyen pour actualiser notre connaissance sur la distribution d'une variable $x$ sachant que nous avons reçu une information sur la distribution de $y$. Pour cette raison, $P(x)$ est appelée la probabilité *a priori* de $x$, alors que $P(x|y)$ est appelée la *probabilité postérieure* de $x$ sachant $y$ ; La probabilité conditionnelle $P(y|x)$ est appelée la *vraisemblance* de $y$ étant donné $x$.

# 3 Réseaux Bayésiens

## 3.1 Définition des réseaux Bayésiens

Actuellement, les modèles graphiques probabilistes, et plus particulièrement les RBs proposés par [Pearl, 1988] vers la fin des années 80, suscitent beaucoup d'intérêt au sein de la communauté de l'intelligence artificielle et de reconnaissance de formes. Ils sont notamment utilisés pour la représentation de connaissances incertaines ainsi que pour le raisonnement probabiliste à partir des données incomplètes.

**Définition 3 (*Réseaux Bayésiens*)**

*Un Réseau Bayésien est un modèle graphique probabiliste défini par le couple $(G, P)$, où :*

- *$G = (X, E)$ est un graphe orienté et sans circuit (DAG pour 'Directed Acyclic Graph') dont les sommets correspondent à un ensemble de variables aléatoires $X = \{X_1, ..., X_n\}$ et les arcs $E$ **représentent** les relations d'indépendance conditionnelle existantes entre les variables $X$.*
- *$P = \{P(X_i|Parents_{X_i}, i = 1, ..., n\}$ est un ensemble de distributions de probabilités conditionnelles, une distribution de $X_i$ pour chaque instance (état) de ses parents directs dans le graphe $Parents_{X_i}$.*

Les RBs permettent selon la structure du graphe $G$ une factorisation de la fonction de distribution de probabilité conjointe en un produit de distributions de probabilités conditionnelles de chaque nœud $X(i)$ conditionnellement à ses parents $Parents_{X_i}$ dans le graphe :

$$P(X) = \prod_{i=1}^{n} P(X_i|Parents_{X_i}) \tag{5}$$

Un RB discret comportant 05 variables est illustré sur la figure 1.1.

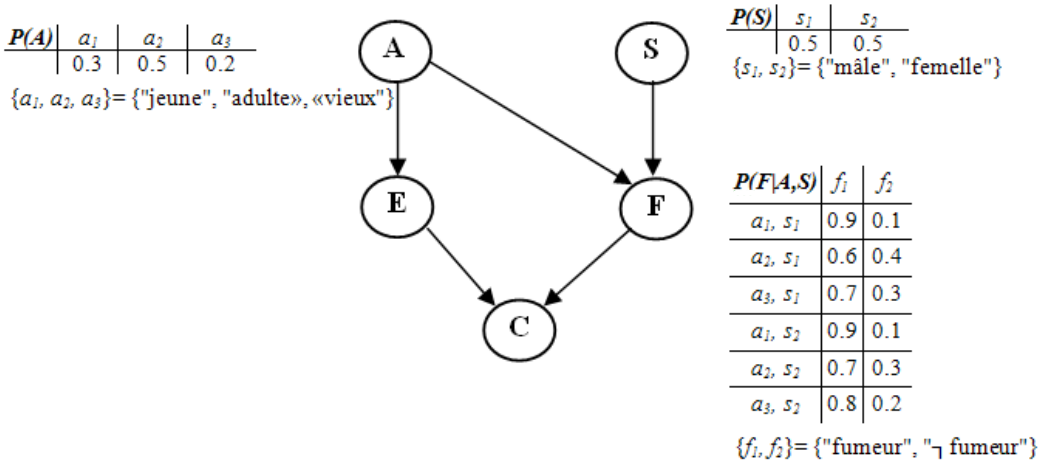



FIGURE 1.1 – Exemple d'un réseau Bayésien à variables discrètes.

Les tables de probabilités conditionnelles des variables A, S et F sont affichées. Les variables sont comme suit, A : Age, S : Sexe, F : Fumeur, E : Exposition aux produits toxiques et C : Cancer.

## 3.2 Condition de Markov

**Proposition 1 (*Pearl, 1988*)**

*Soit $X$ un ensemble de variables, $G$ un graphe orienté acyclique sur $X$ et $P$ une distribution de probabilités sur $X$. Le graphe $G$* **représente** *$P$ si et seulement si chaque variable de $X$ est indépendante pour $P$ de tous ses non descendants dans $G$ conditionnellement à ses parents dans $G$.*

La structure $G$ d'un RB, définie par l'ensemble de ses arcs $E$, exige que certaines contraintes d'indépendance doivent être vérifiées dans n'importe quelle distribution qui peut être représentée par $G$. La propriété de Markov, énoncée dans la Proposition 1, est une condition nécessaire et suffisante de représentation d'une distribution de probabilités $P$ par un graphe $G$. En d'autres termes, la décomposition de la conjointe selon l'équation (5) n'est possible que lorsque chaque nœud est conditionnellement indépendant des ses non descendants sachant ses parents directs dans le graphe. La condition de Markov suscite beaucoup de discussions contemporaines portant sur les RBs, elle a été utilisée pour la définition des RBs par plusieurs auteurs [Pearl, 1988, Williamson, 2005].

Dans l'exemple de la figure 1.1, parmi les relations f'indépendance conditionnelle qui découlent de la condition de Markov, nous avons les relations: $F \perp E|A$, $F \perp E|(A,S)$, $C \perp A|(E,F)$, $C \perp S|(E,F)$, $E \perp S|A$, $S \perp E$, $S \perp A$.

## 3.3 La notion de d-séparation

**Définition 4 (*d-séparation d'un chemin*)**

*Soit $G$ un graphe acyclique orienté défini sur un ensemble de variables $X$, un chemin $c$ est d-séparé par un sous-ensemble $S \subset X$ si au moins une des deux conditions suivantes est vérifiée :*

1. *$c$ contient une chaîne $X_i \rightarrow X_j \rightarrow X_k$ ou une connexion divergente $X_i \leftarrow X_j \rightarrow X_k$ telle que $X_j$ appartient $S$ ;*
2. *$c$ contient une connexion convergente $X_i \rightarrow X_j \leftarrow X_k$ telle que ni $X_j$ , ni aucun de ses descendants n'appartiennent pas à $S$ .*

**Définition 5 (*d-séparation d'un ensemble de variables*)**

*Soit $G$ un graphe orienté acyclique défini sur un ensemble de variables $X$ et $S$ , $U$ et $V$ trois sous-ensembles de $X$. L'ensemble $S$ d-sépare $U$ et $V$ dans $G$ si tout chemin partant d'une variable de $U$ vers une variable de $V$ est d-séparé par $S$ .*

**Théorème 3 (*Verma et Pearl, 1988*)**

*Soit $G$ un graphe acyclique orienté défini sur un ensemble de variables $X$ et soient $S$ , $U$ et $V$ trois sous-ensembles disjoints de $X$. $S$ d-sépare $U$ et $V$ dans $G$ si et seulement si $U$ est conditionnellement indépendant de $V$ sachant $S$ pour toute distribution de probabilités représentée par $G$.*

Pour toutes les distributions de probabilité qui peuvent être représentées par un graphe $G$, la d-séparation permet d'extraire toutes les indépendances probabilistes conditionnelles entre les différents sous-ensembles des variables d'un RB. Ce résultat fondamental se montre également très utile dans les algorithmes de construction de RBs à partir des résultats partiels de tests d'indépendance conditionnelle menés sur des sous-ensembles de variables du domaine.

A titre d'illustration, dans l'exemple de la figure 1.1, nous avons l'indépendance $S \perp C|(A, F)$. En effet, le sous-ensemble $A, F$ d-sépare bel et bien les deux variables $S$ et $C$. En examinant le graphe de près, il y a seulement deux chemins du nœud $S$ vers le nœud $C$ et qui sont : $S \rightarrow F \rightarrow C$ et $S \rightarrow F \leftarrow A \rightarrow E \rightarrow C$. Ces deux chemins sont d-séparés respectivement par les variables $F$ et $A$ (en utilisant la première clause de la Définition 4.

# 4 Apprentissage de paramètres

Un RB $B = (G, P)$ est quantifié à l'aide d'un ensemble de probabilités conditionnelles $P = \{P(X_i|Parents_{X_i}, i = 1, ..., n\}$. En domaine discret, les tables de probabilités conditionnelles (TPCs) sont souvent utilisées pour la modélisation de $P$. La TPC $\theta_i$ correspondant à la variable $X_i$ est une table multidimensionnelle dont la taille est égale à $Card(X_i) \times \prod Card(Parents_{X_i})$

(voir l'exemple de la figure 1.1). La case $\theta_{i,j,k} = P(X_i = x_k|Parents_{X_i} = px_j)$ est la probabilité que la variable $X_i$ soit à l'état $k$ étant donné que ses parents directs sont à l'état $j$. L'apprentissage des valeurs de $\theta_{i,:,:}$ $(i = 1, ..., n)$ est une des tâches principales qui relèvent des RBs. Généralement, selon la nature de données d'apprentissage, on distingue entre les cas suivants: données complètes et représentatives, données incomplètes (quelques valeurs des champs manquent), peu de données sont disponibles. Dans cette section, nous allons décrire brièvement les méthodes populaires utilisées pour apprendre les TPCs $\theta_i$.

## 4.1 Maximum de vraisemblance (MV)

Dans le cas où les données d'apprentissage sont abondantes et toutes les variables sont observées, la méthode de maximum de vraisemblance permet l'estimation des TPCs. Les paramètres optimaux du RB sont ceux qui maximisent la vraisemblance des données d'apprentissage, $\hat{\theta} = argmax_\theta(P(donnes|\theta))$ sont les paramètres estimés, et la solution optimale de MV est

$$\hat{\theta}^{MV}_{i,j,k} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}} \tag{6}$$

où $N_{i,j,k}$ est le nombre d'événements dans la base de données pour lesquels la variable $X_i$ est à l'état d'indice $k$ et ses parents sont dans la configuration d'indice $j$. (Voir la démonstration de ce résultat dans [Naim et al., 2004]).

## 4.2 Maximum a posteriori (MAP)

L'estimateur MAP suit une approche Bayésienne pour l'estimation de paramètres. Le MAP tente de trouver les paramètres $\theta$ les plus probables sachant que des données $D$ ont été observées, mais à la différence du MV, le MAP exploite aussi les distributions a priori sur les paramètres $\theta$. Selon la règle de Bayes, nous avons :

$$P(\theta|D) = \frac{P(D|\theta)}{P(D)} \times P(\theta) \tag{7}$$

où $P(D|\theta)$ est la vraisemblance des données d'apprentissage et $P(\theta)$ est la distribution a priori des paramètres. Si la distribution de données suit la loi multinomiale, la loi a priori conjuguée $P(\theta)$ est la distribution de Dirichlet et la solution optimale de l'estimateur MAP est:

$$\hat{\theta}^{MAP}_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k} - 1}{\sum_k (N_{i,j,k} + \alpha_{i,j,k} - 1)} \tag{8}$$

où les constantes numériques $\alpha_{i,j,k}$ sont les paramètres de la distribution de Dirichlet associée à la loi a priori $P(X_i = x_k|Parents_{X_i} = px_j)$. Plus de détails sur le MAP sont donnés dans [Pearl, 1988]. Il faut noter que (6) est un cas particulier de (8) pour des coefficients de Dirichlet égaux à 1 (sans informations a priori).

# 5 Apprentissage de structure

La construction de structure consiste à identifier l'ensemble $E$ d'arcs qui composent le graphe $G = (X, E)$ du RB. Le graphe causal où l'arc $x \rightarrow y$ signifie que $x$ cause $y$ est un RB et l'inverse n'est pas toujours vrai. Dans ce sens, la structure du RB peut être déduite facilement si les informations des relations causales inter-variables fournies par l'expert sont suffisantes. En cas d'absence ou d'insuffisance des informations de l'expert ce qui est le plus fréquent, l'apprentissage automatique de la structure à partir de données devient nécessaire. La construction du graphe $G$ du RB à partir d'un échantillon de données d'apprentissage est un des problèmes complexes relatifs aux RBs. L'apprentissage de la structure $E$ présente le compromis suivant: comment peut-on construire un réseau dont la taille est minimale (un graphe moins dense) et qui représente la distribution conjointe $P$?

La difficulté inhérente à l'apprentissage de la structure des RBs peut être expliquée par les deux points suivants. Premièrement, l'indépendance (ou dépendance) statistique conditionnelle est un concept facile à définir en théorie, mais la confirmation d'une telle indépendance entre un ensemble de variables aléatoires sur la base d'un échantillon empirique d'observations est une tâche très délicate et sujette aux erreurs d'estimation. Secondairement, le nombre de graphes possibles (espace de recherche) qui peuvent être formés à partir d'un ensemble de variables est super-exponentiel. Trouver la structure optimale c'est comme chercher une aiguille dans une balle de foin. Pour avoir une idée sur le gigantisme de l'espace de recherche, on peut former à partir de 4, 5, 10, 20, 30 et 40 nœuds, respectivement, 453, 29281, $4.7 \times 10^{17}$, $2.34 \times 10^{72}$, $2.71 \times 10^{158}$ et $1.12 \times 10^{276}$ graphes DAG différents [Asadi et al., 2008].

## 5.1 Notion d'équivalence de Markov

Parmi tout l'ensemble de DAGs possibles, un sous-ensemble des DAGs sont équivalents et décrivent les mêmes relations d'indépendance conditionnelle existantes entre les variables. L'objectif de l'apprentissage de la structure se résume à déterminer seulement un DAG de cette classe. Il est bien de noter que au maximum un seul DAG de la classe d'équivalence peut être causal. La définition 6 permet de donner une partie des DAGs qui sont équivalents à un DAG donné.

**Définition 6 (*classe d'équivalence de Markov*)**

*Deux RBs $B_1$ et $B_2$ appartiennent à la même classe d'équivalence de Markov s'ils représentent les mêmes relations d'indépendance conditionnelle.*

Verma et Pearl [Verma and Pearl, 1991] ont démontré que tous les DAGs équivalents possèdent le même squelette (graphe non orienté) et les mêmes V-structures (une V-structure est une connexion entre 03 variables en forme de la lettre V, $X_i \leftarrow X_j \rightarrow X_k$).

## 5.2 Algorithmes à base de tests d'indépendance conditionnelle

Les méthodes qui suivent cette approche commencent par la détermination à partir de données d'apprentissage les relations d'indépendance conditionnelle entre deux variables quelconques conditionnellement à un autre ensemble de variables. Ensuite, elles exploitent les résultats des tests statistiques d'indépendance pour en construire le graphe du RB d'une manière incrémentale. Les tests d'hypothèse d'indépendance de "Pearson's Chi-2" et de l'"Information Mutuelle" sont les plus utilisés. Parmi les méthodes populaires suivant cette approche, nous citons les algorithmes IC et IC* de Pearl et Verma [Pearl, 2000], les algorithmes SGS, PC, CI, FCI de [Spirtes et al., 1993], et l'algorithme BN-PC de [Cheng et al., 1997b, Cheng et al., 1997a]. Tous ces algorithmes opèrent généralement par les étapes suivantes:

1. Effectuer un ensemble suffisamment important de tests d'indépendances statistiques conditionnelles;
2. Construire un graphe non orienté à partir des résultats des tests obtenus en étape 1;
3. Détecter les V-structure (en ajoutant parfois si nécessaire d'autres tests d'indépendance conditionnelle);
4. Compléter les orientations des autres arêtes non encore dirigées (qui sont en dehors des V-structures détectées à l'étape 3), tout cela sans qu'on crée de nouvelles V-structures).

Puisque il est impossible de réaliser tous les tests possibles, on commence par la vérification d'un ensemble réduit de tests afin d'avoir une approximation grossière de la structure (étape 2), puis on ajoute des tests bien ciblés dans le but d'éliminer les confusions d'orientation d'arêtes restantes(étapes 3 et 4).

## 5.3 Algorithmes à base de score

A la différence des méthodes de la première approche basée sur les tests d'indépendance conditionnelle, les méthodes suivantes visent à chercher dans l'espace des graphes possibles une structure qui maximise une certaine fonction de score.

L'objectif des fonctions de scores est d'évaluer la qualité d'une structure $B$, la bonne structure est celle qui est plus simple et qui correspond le mieux possible aux données d'apprentissage $D$. En effet, ces fonctions de scores comportent souvent deux termes : la vraisemblance $L(D|B)$, et un deuxième terme qui mesure la complexité du modèle pour pénaliser les structures complexes (ayant beaucoup d'arcs). En plus de la définition d'un score, les méthodes de cette famille s'appuient sur un algorithme de recherche itératif qui converge vers une bonne structure.

Les fonctions de scores sont basées sur des principes différents. Parmi les fonctions courantes, nous citons la fonction dite K2, proposée par [Cooper and Herskovits, 1992], et la fonction BD (Bayesian Dirichlet), proposée par [Heckerman et al., 1995] comme une généralisation de K2. Les algorithmes de recherche les plus utilisés sont des méthodes de recherche locale

telles que [Chickering et al., 1995], [Cooper and Herskovits, 1992], [de Campos et al., 2003] et [Heckerman et al., 1995].

# 6 Inférence

## 6.1 État de l'art des algorithmes d'inférence

L'inférence dans les RBs consiste en un calcul numérique de la distribution marginale d'une variable d'intérêt $Q$ sachant que nous disposons des valeurs réelles prises par d'autres variables $E = e$, dites des évidences. Formellement, nous voulons évaluer $P(Q = q|E = e)$, l'impact des évidences sur notre connaissance sur $Q$, et qui peut être réécrit en utilisant la règle de Bayes comme suit :

$$\begin{aligned} P(Q = q|E = e) &= \frac{P(Q=q,E=e)}{P(E=e)} \\ &= \frac{P(Q=q,E=e)}{\sum_Q P(Q=q,E=e)} \end{aligned} \tag{9}$$

On distingue deux types d'évidences : (i) déterministe (*hard evidence*), où une variable instanciée $E_i \in E$ prend une valeur précise $E_i = e_i$. (ii) imprécise (*soft évidence*) où une variable instanciée $E_i \in E$ prend une distribution de probabilité $P(E_i)$. Dans ce chapitre, nous abordons le cas simple, celui de l'évidence déterministe.

L'inférence sur les RBs est un problème pur du calcul. Le grand défi durant l'évaluation de (9) est comment évaluer le numérateur d'une manière efficace, ce dernier ne peut pas être pratiquement évalué que par la marginalisation de la conjointe $P_X(X)$ sur les autres variables libres $L = \{X - \{Q \cup E\}\}$; $P(Q = q, E = e) = \sum_L P_X(Q = q, L, E = e)$. En effet, la solution triviale qui consiste au calcul de toutes les valeurs de la conjointe (par force brute) n'est pas une solution souhaitable à cause de la complexité exponentielle en mémoire et en temps, mais plutôt il faut investir dans la structure du graphe et opter pour une méthode du calcul local (*Local computation*) [Lauritzen and Spiegelhalter, 1988, Jensen et al., 1990] en définissant des fonctions locales qui impliquent un nœud et ses voisins (conjointes locales) et une stratégie de propagation.

Les méthodes d'inférence exacte peuvent être scindées en trois classes :

1. Elimination de variables, lors de la marginalisation, certaine variables peuvent être éliminées en évaluant des conjointes locales de petites dimensions, l'ordre de l'élimination des variables dépend de la requête d'inférence et de la structure du graphe. A titre d'exemple, si nous voulons calculer la marginale de la variable $D$ sur le RB défini par le graphe

$G = A \rightarrow B, A \rightarrow C, B \rightarrow D, C \rightarrow D$, nous avons :

$$\begin{aligned} P(D) &= \sum_A \sum_B \sum_C P(A,B,C,D) \\ &= \sum_A \sum_B \sum_C P(A).P(B|A).P(C|A).P(D|B,C) \quad (Factorisation\ selon\ G) \\ &= \sum_B \sum_C P(D/B,C) \sum_A P(A).P(B/A).P(C/A) \quad (Elimination\ de\ A\ avant\ B\ et\ C) \end{aligned}$$

2. Algorithmes de propagation de messages sur le RB, qui utilisent un mécanisme d'échange de messages le long des arcs du réseau. Le premier algorithme, proposé par [Pearl, 1982]en 1982 avant même le mûrement du concept des RBs, est restreint au cas des RBs dont le graphe est un arbre, une extension de cet algorithme aux graphes multi-arbres (*Polytrees*) est faite par [Kim and Pearl, 1983]. Par la suite, des techniques expliquant comment cet algorithme peut être utilisé pour des graphes plus généraux qui peuvent avoir des boucles (cycles non orientés) ont été proposées par [Pearl, 1988].
3. Algorithmes à base de regroupement des nœuds, ce sont des méthodes qui procèdent d'abord par des modifications importantes sur le graphe afin réorganiser les nœuds du RB en un arbre de cliques de variables (appelé *arbre de jonction*) [Lauritzen and Spiegelhalter, 1988], puis d'effectuer l'inférence par la propagation des messages de mise à jour entre cliques voisines. Des améliorations ont été proposées par [Jensen et al., 1990, Cowell et al., 1999], voir [Cowell et al., 1999] pour une riche vue d'ensemble sur cette approche. On note également que ces algorithmes représentent en réalité une variante des algorithmes d'élimination de variables dans lesquels l'ordre de l'élimination est guidé par l'arbre de jonction.

A cause de la complexité calculatoire excessive des méthodes d'inférence exacte sur des RBs complexes, des algorithmes de marginalisation approchée à base de méthodes stochastiques type MCMC (*Markov Chain Monte Carlo*)) [Henrion, 1988, Gilks et al., 1996] comprenant entre autres les échantillonneurs de Gibbs [Neal, 1993, York, 1992] ont reçu plus d'attention.

## 6.2 Algorithme de l'arbre de jonction

L'arbre de jonction (JA) est incontestablement la technique la plus populaire qui fournit une réponse exhaustive au problème de l'inférence sur les RBs [Jensen et al., 1990]. L'algorithme JA se déroule en quartes étapes : regroupement des nœuds en cliques, connexion des cliques afin d'avoir un arbre de jonction, propagation de l'information provenant des évidences via l'arbre de jonction et finalement réponse à une requête d'inférence donnée. Avant d'expliquer comment l'algorithme JA est utilisé par les RBs, il est judicieux d'introduire quelques éléments et définitions utiles concernant l'arbre de jonction.

**Définition 7 (*Arbre de jonction*)**

*Soit $I$ un ensemble fini de variables , et $\tau$ un ensemble de parties de $I$ telles que $\bigcup_{C\in\tau} C =$*

*I. On appelle arbre de jonction sur I tout arbre dont les nœuds sont des éléments dans C et vérifiant la propriété suivante :*
*Pour tout couple $(C_a, C_b)$ d'éléments de $\tau$ non adjacents, tout sommet intermédiaire $C_z$ sur une chaîne H joignant $C_a$ et $C_b$ vérifie la condition $C_a \bigcap C_b \subset C_z$.*

### 6.2.1 Potentiel

Durant la propagation des évidences via la structure du RB en utilisant des fonctions du calcul local, on manipule des tables multidimensionnelles qui sont définies sur des sous-ensembles de variables voisines dans le graphe du réseau et ces dernières ne correspondent pas toujours à des TPCs. Ces tables sont appelées des "potentiels".

**Définition 8 (*Potentiel*)**

*Un potentiel $\Phi$ est une fonction réelle sur un ensemble de variables X à valeur discrètes finies*

$$\Phi : espace(X) \mapsto R$$

Le domaine d'un potentiel, désigné par $dom(\Phi)$, est l'ensemble de toutes les configurations possibles de $X$. Par exemple, si les variables binaires $A$,$B$ peuvent prendre respectivement des valeurs dans $\{'y','n'\}$ et $\{'0','1'\}$, alors nous avons $dom(\Phi(A|B)) = dom(\Phi(A)) \bigotimes dom(\Phi(B))$= $\{('y','0'),('n','0'),('y','1'),('n','0')\}$.

### 6.2.2 Opérations sur les potentiels

**- Extension** Soit $U \subset W$ et soit $\Phi$ un potentiel défini sur $U$, le potentiel $\Phi$ est étendu au domaine $W$ de cette manière. Soit $x \in W$, alors on défini $\Phi(x) = \Phi(y)$ où $y$ est la projection de $x$ sur $U$.

**- Multiplication** Pour deux potentiels $\Phi_1$ et $\Phi_2$ définis respectivement sur les domaines $dom(\Phi_1)$ et $dom(\Phi_1)$, nous définissons leur produit $\Phi_1 \cdot \Phi_2$ sur le domaine $U = dom(\Phi_1) \bigcup dom(\Phi_2)$ par :

$$(\Phi_1.\Phi_2)(x) = \Phi_1(x) \times \Phi_2(x) \tag{10}$$

où $\Phi_1(x)$ et $\Phi_2(x)$ (le terme à droite de l'équation) sont d'abord étendus au domaine $U$.

**- Division** La division est définie de la même manière que la multiplication, à l'exception du cas de division par zéro qui est traité différemment.

$$(\Phi_1/\Phi_2)(x) = \Phi_1(x)/\Phi_2(x) \; si \; \Phi_2(x) \neq 0 \;\; et \;\; 0 \;\; ailleurs \tag{11}$$

**- Marginalisation** Soit $U \subset W$ et $\Phi$ un potentiel défini sur $W$. On défini $\Sigma_{W-U}\Phi$ la marginalisation du potentiel $\Phi$ sur $U$ par l'élimination des variables $W - U$, elle se calcule comme suit :

$$(\Sigma_{W-U}\Phi)(x) = \Sigma_{z \in W-U}\Phi(z.x) \tag{12}$$

| A | B | $\emptyset_1$ |
|---|---|---|
| y | y | 0.2 |
| y | n | 0.3 |
| n | y | 0.4 |
| n | n | 0.6 |

(a)

| A | $\emptyset_2$ |
|---|---|
| y | 0.5 |
| n | 0 |

(b)

| A | B | $\emptyset_2$ |
|---|---|---|
| y | Y | 0.5 |
| y | n | 0.5 |
| n | y | 0 |
| n | n | 0 |

(c)

| A | B | $\emptyset_1 . \emptyset_2$ |
|---|---|---|
| y | y | 0.10 |
| y | n | 0.15 |
| n | y | 0 |
| n | n | 0 |

(d)

| B | $\Sigma_A \emptyset_1$ |
|---|---|
| y | 0.6 |
| n | 0.9 |

(e)

FIGURE 1.2 – Exemple de manipulations de base sur les potentiels.

**(a)**: $\Phi_1(A, B)$. **(b)**: $\Phi_2(A)$. **(c)**: $\Phi_1(A, B)$, extension de $\Phi_2$ du $dom(A)$ au $dom(A, B)$. **(d)**: $\Phi_1 \cdot \Phi_2$, produit de $\Phi_1$ et $\Phi_2$ défini sur $dom(A, B)$. **(e)**: $\Sigma_A \Phi_1$, marginalisation de $\Phi_1$ sur $B$.

où $z \cdot x$ est un élément dans $W$ avec la projection de $x$ sur $U$ et la projection de $z$ sur $W - U$.

Un exemple qui illustre les opérations d'extension, de multiplication et de marginalisation est montré sur la figure 1.2.

### 6.2.3 Construction de l'arbre de jonction

Le but de cette partie est de transformer un RB quelconque en un arbre de cliques (une clique $C$ est un sous-graphe dont tous les nœuds sont interconnectés), tout cela en préservant toutes les relations d'indépendance conditionnelle exprimées par le RB initial. La construction est faite par la moralisation et la triangulation du graphe RB, puis par l'identification et la jonction des cliques.

**- Moralisation** Soit $G = (X, E)$ un DAG d'un RB. Le graphe moral $G^M = (X, E^M)$ est obtenu par le mariage des parents du même nœud, puis par la suppression des orientations d'arcs $E$. La moralisation définie ainsi implique qu'un nœud et ses parents seront contenus dans la mêmes cliques.

**- Triangulation** Un graphe non orienté est triangulé s'il ne contient pas de cycles de longueur supérieure à trois (03). Pour trianguler un graphe non orienté $G^M = (X, E^M)$, on ajoute une corde pour tout cycle de longueur au moins égale à quatre (04), une corde est une arête reliant deux sommets non consécutifs sur le cycle. La triangulation n'est pas unique, elle dépend de la séquence des cordes ajoutées.

**- Construction de l'arbre** Soit $G'$ un graphe non orienté (moralisé et triangulé). Pour la construction de l'arbre de jonction associé, on commence d'abord par l'identification de toutes les cliques maximales dans $G'$. Ensuite, la liaison des cliques est faite en vérifiant une propriété dite de l'intersection courante.

**Définition 9 (*propriété d'intersection courante (PIC)*)**

*Une séquence $(C_1, C_2, ..., C_k)$ de cliques est dite qu'elle vérifie la propriété d'intersection courante (PIC) si pour tout $1 < j \leq k$, il y a un $i < j$ tel que $C_j \bigcap (C_1, ..., C_{j-1}) \subseteq C_i$.*

Si les cliques d'un graphe décomposable sont ordonnées de telle sorte qu'elles vérifient cette propriété, alors l'Algorithme 1 peut être utilisé pour construire un arbre de jonction valide.

**Algorithme 1** [Cowell et al., 1999]

1. Ordonner l'ensemble de cliques $(C_1, C_2, ..., C_k)$ d'un graphe non orienté dans un ordre vérifiant la PIC.
2. Associer un nœud de l'arbre à chaque clique.
3. Pour $i = 2, ..., k$, ajouter un arc entre $C_i$ et $C_j$ où $j$ est n'importe quelle valeur dans $\{1, ..., i-1\}$ tel que $C_i \bigcap (C_1 \bigcup ... \bigcup C_{i-1}) \subseteq C_j$

Pour un graphe triangulé, un ordre vérifiant la PIC est toujours existant et donc la construction d'un arbre de jonction valide est toujours possible.

**Théorème 4 (*Existence de l'arbre de jonction*)**

*L'ensemble de cliques d'un graphe moralisé et triangulé peut toujours être muni de la structure de l'arbre de jonction [Smail, 2004].*

Vous trouver dans [Smail, 2004] une riche littérature sur les techniques de construction des arbres de jonction sur des RBs. On note que la propriété de l'intersection courante précise qu'il existe un tel ordre mais pas son unicité; un ou plusieurs arbres de jonction sont possibles. On préfère souvent l'arbre couvrant de poids maximum. D'abord, on construit un graphe des cliques où deux cliques sont connectées si elles contiennent des variables en commun, le poids attribué à l'arête de connexion est égal au nombre des variables en commun. La recherche d'un sous ensemble d'arêtes de poids maximum qui forme un arbre peut être résolu d'une manière exacte par l'algorithme glouton proposé par [Kruskal, 1956], en favorisant de connecter les cliques qui ont plusieurs variables partagées en premier lieu.

**Définition 10 (*Séparateur*)**

*Un séparateur est l'intersection entre deux cliques voisines.*

Un exemple illustrant les étapes de construction d'un arbre de jonction du RB 'Asia' [Lauritzen and Spiegelhalter, 1988] est montré sur la figure 1.3. L'arbre de jonction final où les séparateurs sont indiqués sous forme de rectangles est affiché sur la figure 1.3.(e).

### 6.2.4 Propagation de croyances sur l'arbre de jonction

Parmi les principales algorithmes d'inférence Bayésienne sur un RB("propagation de croyances") qui s'appuient sur l'arbre de jonction, nous citons la méthode [Lauritzen and Spiegelhalter,

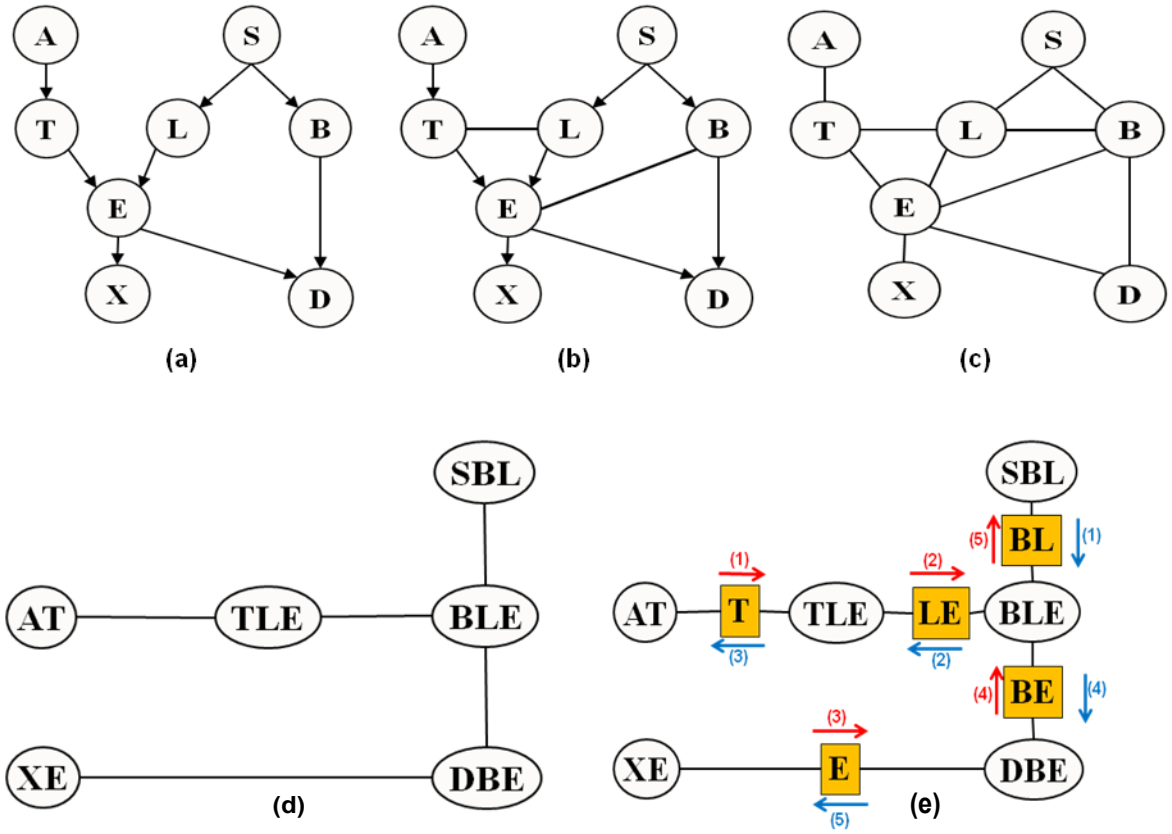



FIGURE 1.3 – Exemple de construction de l'arbre de jonction.

**(a)**: le RB 'Asia' proposé par [Lauritzen and Spiegelhalter, 1988]. **(b)**: Mariage des parents. **(c)**: Désorientation du graphe et triangulation. **(d)**: Construction de l'arbre de jonction. **(e)**: séparateurs sont affichés sous forme de rectangles, racine choisie de l'arbre: la clique 'SBL', le sens et l'ordre des messages de *la procédure de collecte* sont indiqués en rouge, le sens et l'ordre des message de *la procédure de distribution* sont indiqués en bleu.

1988], l'algorithme de propagation Shenoy&Shafer [Shafer and Shenoy, 1990], l'architecture "Hugin" [Jensen et al., 1990] et plus récemment la technique dite "Lazy propagation" [Madsen and Jensen, 1999]. Afin de bien illustrer le principe de cette approche à base de regroupement des nœuds sous forme d'un arbre de cliques, nous avons choisi d'expliquer avec un peu plus de détails la méthode "Hugin" qui est une méthode à la fois simple et générale, un exemple numérique d'inférence est également proposé dans ce contexte.

**Algorithme 2: Hugin [Madsen and Jensen, 1999]**

Un potentiel est associé à chaque clique et à chaque séparateur de l'arbre de jonction construit à partir du graphe du RB. L'inférence est réalisée en transmettant des messages de

mise à jour. Lorsqu'un élément de l'arbre de jonction (clique ou séparateur) reçoit un message, il modifie son potentiel et envoie un autre message dans l'autre sens de l'arbre.

**a) Initialisation** Avant d'utiliser l'arbre de jonction en inférence, les potentiels de tous les éléments de l'arbre doivent être initialisés correctement. Une des initialisations valides est la suivante. Les potentiels des séparateurs sont fixés au potentiel unité, tandis que les potentiels des cliques sont initialisés au produit des probabilités conditionnelles des variables dont les parents sont également contenus dans la même clique.

$$\begin{array}{rc} Séparateurs : & \forall x\ \Phi_S(x) = 1 \\ Cliques : & \Phi_C = \Pi_V P(V|Parents_V) \textbf{ tel que } \{V \cup Parents_V\} \subset C \end{array} \tag{13}$$

**b) Incorporation des évidences** L'introduction d'une évidence $E_i = e_i$ dans l'arbre de jonction peut être réalisée comme suit :

1. Trouver une clique $C$ qui contient la variable $E_i$ ;
2. Annuler toutes les lignes du potentiel $\Phi_C$ qui ne correspondent pas à $E_i = e_i$; $\Phi_C(z) = 0$ pour tout $E_i \neq e_i$, avec $z$ une extension de $E_i$ sur $C$.

**c) Projection** La procédure de projection d'une clique $C_i$ vers un séparateur $S$ adjacent sauvegarde le potentiel courant de $S$ et attribue un nouveau potentiel à $S$ :

$$\begin{array}{rcl} \Phi_S^{old} & \leftarrow & \Phi_S \\ \Phi_S & \leftarrow & \Sigma_{C_i - S} \Phi_{C_i}\ (marginalisation) \end{array} \tag{14}$$

**d) Absorption** La procédure d'absorption d'un séparateur $S$ vers une clique adjacente $C_j$ met à jour le potentiel de $C_j$ en utilisant les potentiels ancien et nouveau de $S$ :

$$\Phi_{C_j} \ \rightarrow \ \Phi_{C_j} \cdot \frac{\Phi_S}{\Phi_S^{old}} \tag{15}$$

**e) Propagation d'évidences** Après avoir achevé correctement l'initialisation et l'introduction des évidences, l'inférence revient à échanger des messages d'actualisation entre les cliques sous forme de fonctions projection-absorption le long de l'arbre de jonction. A la fin de la propagation, et pour que l'arbre soit consistent, il faut que chaque message soit envoyé sur les deux directions de chaque lien dans l'arbre. L'arbre est dite consistent si pour tout séparateur $S$ et leurs deux cliques adjacentes $C_i$ et $C_j$ nous avons $\Sigma_{C_i - S} \Phi_{C_i} = \Phi_S = \Sigma_{C_j - S} \Phi_{C_j}$.

Pour parvenir à une propagation complète, il est possible de choisir un nœud racine (le choix peut être arbitraire), puis d'exécuter la procédure de collecte des messages d'actualisation des feuilles vers la racine, ensuite d'exécuter la procédure de distribution dans le sens inverse, de la racine vers les feuilles. Pour assurer une propagation complète en deux parcours de l'arbre, durant la phase de collecte, une clique n'expédie pas son message à son père que lorsqu'elle reçoit tous les messages de ses fils, le père d'une clique est la clique adjacente menant à la racine et les fils sont les autres nœuds adjacents. Un exemple qui montre un ordre valide de propagation des massages est affiché sur la figure 1.3(e).

**f) Calcul de marginales** La réponse à une requête d'inférence (celle de l'Equation (9)) consiste en deux operations. D'abord, une marginalisation pour calculer la conjointe $P(Q, E)$. Pour ce faire, après avoir déroulé les deux procédures de propagation d'évidences (la collecte et la distribution), il suffit de trouver une clique $C$ contenant la variable $Q$, puis de marginaliser le potentiel de $C$, $\Phi_Q = \Sigma_{C-Q}\Phi_C$ (évaluation du numérateur de l'Equation (9)). Par la suite, une normalisation de $\Phi_Q$ pour avoir une probabilité marginale dont la somme est égale à 1 (une division par le dénominateur de l'équation 9). Nous avons donc

$$P(Q) = \frac{\Phi_Q}{\Sigma_Q \Phi_Q} \tag{16}$$

On note que s'il n y pas d'évidences, et après avoir achevé la propagation des messages d'actualisation, la marginalisation permet de calculer les marginales a priori. En outre, pour les mêmes évidences nous pouvons calculer les marginales de plusieurs variables sans refaire la propagation. Dans le cas de présence d'une nouvelle évidence, il suffit de l'introduire dans l'arbre précompilé et de refaire la propagation.

**Exemple d'execution de l'algorithme Hugin**

Pour bien éclaircir le principe de cet algorithme, nous avons proposé de montrer un exemple numérique d'inférence sur un RB simple $B = (G, P)$ où $G = (\{A, B, C, D\}, \{A \rightarrow B, A \rightarrow C, B \rightarrow D, C \rightarrow D\})$. Toutes les variables considérées sont binaires. Les TPCs associées aux variables, $P = \{P(A), P(C|A), P(B|A), P(D|B, C)\}$, sont montrées sur la figure 1.4(a). L'arbre de jonction correspondant à cet exemple contient deux cliques $C_1 = \{A, B, C\}$ et $C_2 = \{D, B, C\}$ liées par le séparateur $S = \{B, C\}$. Nous avons choisi la clique $C_2$ comme une racine de propagation des messages et nous voulons répondre à la requête $P(A|C = c_2, D = d_2)$ =? Voir la trace d'exécution de l'algorithme Hugin sur la figure 1.4.

# 7 Limites des RBs et directions de recherche actuelle

Il évident que les RBs ont apporté une révolution d'une grande échelle dans le domaine de l'intelligence artificielle et de l'analyse de données. Cette gratitude est leur attribuée du fait qu'ils permettent une représentation de connaissances souvent plus naturelle et assez lisible par opposition aux réseaux de neurones et aux arbres de décision par exemple. Particulièrement, le succès des RBs est dû à l'existence de plusieurs algorithmes bien fondés de construction des RBs et d'inférence probabiliste. Néanmoins, leur expansion demeure confrontée à certaines limites et plusieurs challenges liés aux RBs méritent d'être entamés. Dans cette section, nous allons discuter les principales limites et déficiences concernant les RBs, à savoir: la fiabilité d'apprentissage, la complexité spatio-temporelle d'algorithmes RBs (construction et inférence) et la restriction aux variables discrètes.

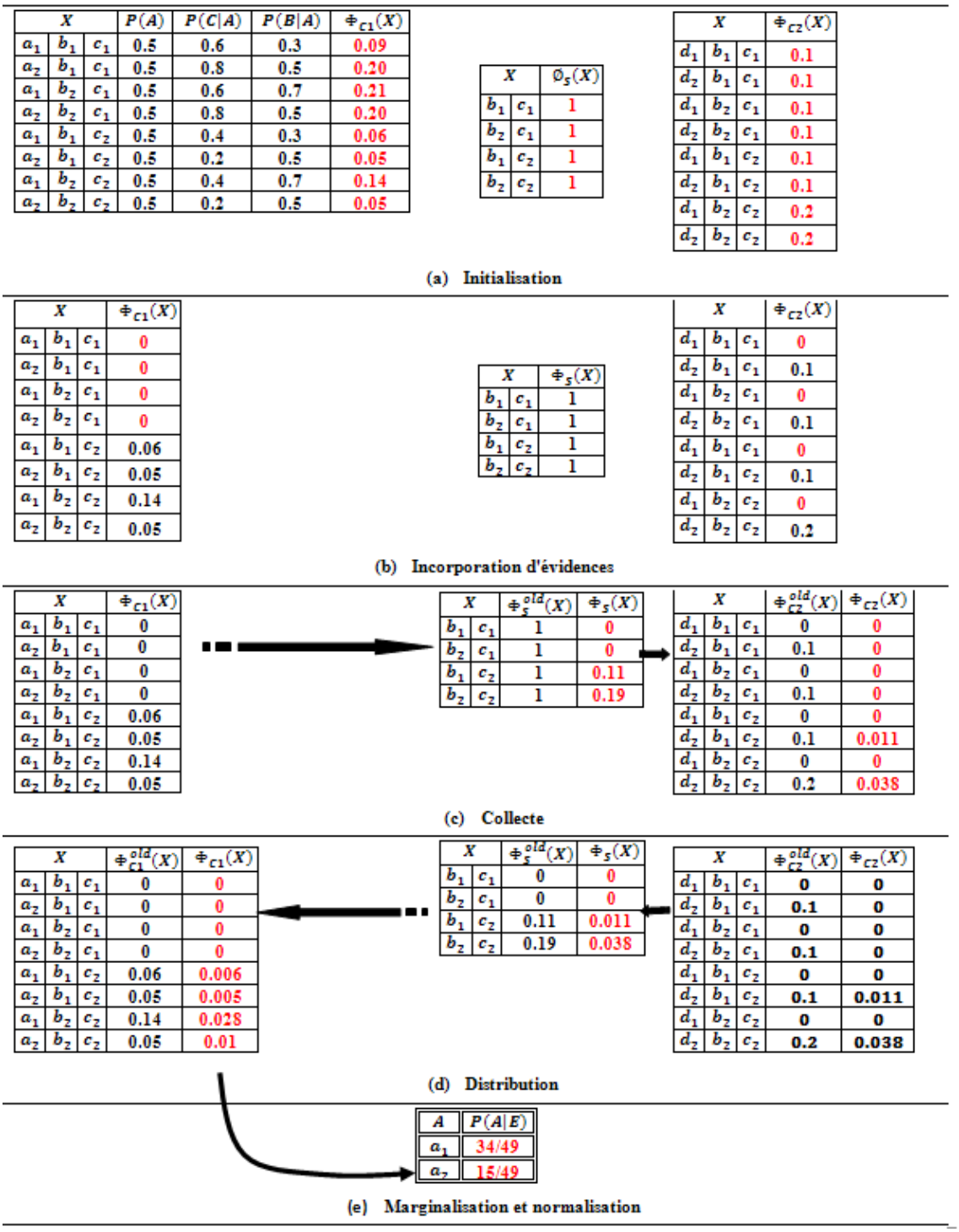



FIGURE 1.4 – Exemple d'inférence sur l'arbre de jonction en utilisant l'algorithme Hugin.

**(a)**: Initialisation des potentiels des cliques $C_1$, $C_2$ et de séparateur $S$. **(b)**: Introduction des évidences $E = \{C = c_2 \; et \; D = d_2\}$. **(c)**: Collecte des messages vers la racine $C_2$. **(d)**: Distribution des messages de la racine vers les feuilles. **(e)**: Marginalisation et normalisation $P(A|E) = \sum_{B,C} \Phi_{C_1}(X)/P(E)$.

## 7.1 Complexité

Un des inconvénients majeurs des RBs est la complexité exponentielle d'algorithmes d'apprentissage et d'inférence par rapport à la taille du RB. La complexité spatio-temporelle augmente significativement en fonction des trois paramètres suivants: (i) la taille du graphe, définie par le nombre des nœuds dans le graphe, (ii) la densité du graphe qui peut être mesurée par le nombre d'arcs par rapport à la taille du graphe, et précisément la complexité dépend plus de la taille de la clique maximale et (iii) la taille des domaines des variables, mesurée par le nombre de valeurs discrètes que peut prendre une variable (résolution de discrétisation).

### 7.1.1 Complexité en espace mémoire

Les valeurs des TPCs ou celles des potentiels ont besoin d'être sauvegardées en mémoire pour qu'on en puisse calculer les marginales. Pour une clique $C$ composée de $K$ variables, $C = \{V_1, ...., V_K\}$, l'espace mémoire nécessaire est $T = \prod_{i=1}^{K} card(V_i)$ cases mémoires tel que $card(V_i)$ est le nombre de valeurs possibles du nœud $V_i$. On constate bien que cet espace deviendra démesuré pour des cliques et des cardinalités relativement grandes.

### 7.1.2 Complexité temporelle d'algorithmes d'apprentissage

L'apprentissage de la structure des RBs est connu d'être un problème NP-complet [Chickering, 1996] et la quasi-majorité des méthodes développées pour cette fin sont basées sur des heuristiques.

En examinant de près les méthodes d'apprentissage, les méthodes à base des tests d'indépendance conditionnelle s'appuient sur un fondement théorétique solide. Ces méthodes exigent d'effectuer un grand nombre de tests statistiques d'indépendance de la forme $Test(A \perp B|Z)$, tester l'indépendance entre les variables $A$ et $B$ sachant un ensemble de variables conditionnés $Z$. Certes, on peut y former plusieurs triplets différents de tests conditionnels et à cause du temps élevé qu'un seul test nécessite, les méthodes de cette approche se limitent à réaliser seulement une partie réduite des tests possibles, et cela peut aboutir à des résultats non fiables. Cette approche a tendance de donner des bons résultats avec des RBs moins denses et avec des petits échantillons de données, cependant elle ne cadre pas très bien avec des RBs importants ou avec une grande quantité de données.

Pour les algorithmes basés sur le paradigme score-recherche, la tâche d'apprentissage est vue comme un problème d'optimisation combinatoire où la recherche est opérée dans l'espace des RBs. L'inconvénient y rencontré est que le nombre de structures possibles augmente super-exponentiellement en fonction du nombre de nœuds. [Robinson, 1973] quantifie le nombre à $O(n!2^{\binom{n}{2}})$ pour un problème de taille $n$. Les algorithmes de recherche de cette approche sont sensibles à l'initialisation, et puisque ils ne peuvent pas scruter tout l'espace de recherche,

souvent ils risquent de converger vers des optima locaux.

### 7.1.3 Complexité temporelle d'algorithmes d'inférence

L'inférence exacte dans un réseau quelconque est un problème NP-complet [Cooper, 1990]. Même les méthodes approchées d'inférence telles que Monte Carlo sont NP-complet [Dagum and Luby, 1993]. L'algorithme de l'arbre de jonction a une complexité exponentielle par rapport à taille de la clique maximale. En effet, l'inférence revient à faire des marginalisations, des multiplications et des divisions sur les potentiels de cliques. Par conséquent, la complexité d'inférence est proportionnelle aux tailles de cliques, en particulier celle de la clique maximale. Cet inconvénient de complexité excessive explique l'intérêt croissant des travaux de recherche récents consacrées aux méthodes approchées de marginalisation [Cooper and Herskovits, 1992, Neal, 1993, Gilks et al., 1996, Henrion, 1988]. La triangulation des graphes RBs est une étape indispensable afin d'accomplir l'inférence d'une manière efficace, soit explicitement lors de la construction de l'arbre de jonction, soit d'une manière implicite si une autre techniques d'inférence est utilisée. La tâche d'obtenir une triangulation optimale (en termes du nombre minimal de liens de triangulation ajoutés) a été prouvée d'être un problème NP-complet [Wen, 1991].

## 7.2 Difficultés liées à l'apprentissage

### 7.2.1 Apprentissage de la structure

Vu la complexité d'apprentissage de structures qui est hors d'atteinte, les méthodes à base de tests d'indépendance conditionnelle se limitent à un sous-ensemble réduit des tests possibles et cela peut compromettre la qualité d'apprentissage. Un autre problème important est que les tests d'indépendance sont sujets aux erreurs d'estimation, établir un test d'indépendance ($A \perp B|Z$) entre variables $A$ et $B$ sachant un ensemble $Z$ requiert que l'échantillon de données soit représentatif (identiquement et indépendamment distribué, et suffisamment grand) et que le type de test adopté soit approprié. Quant aux méthodes d'apprentissage à base de score, la validité de la fonction de score pose toujours un problème. Il est opportun de noter qu'il n y a pas une garantie que les algorithmes de recherche convergent vers une structure optimale, en effet, en fonction de l'initialisation la recherche risque de converger vers un optimum local. La complexité excessive de l'apprentissage explique bien l'intérêt des méthodes approchées qui ne cesse pas d'augmenter [Riggelsen, 2008].

### 7.2.2 Apprentissage des paramètres

L'apprentissage des paramètres est généralement réalisé par la méthode de maximum de vraisemblance ou par l'apprentissage Bayésien des paramètres (MAP) où seulement les distri-

butions a priori de Dirichlet ont été proposées pour les RBs discrets. Pour la première méthode, le problème de sur-apprentissage [Hawkins, 2004] est aperçu lorsque le nombre de paramètres (tailles des TPCs) est important par rapport à la quantité de données d'apprentissage. Pour atténuer le problème de sur-apprentissage si peu de données sont disponibles, la méthode d'inférence Bayésienne à base des lois a priori de Dirichlet est habituellement employée. Un des problèmes liés à cette alternative se rapporte à l'initialisation, il est souvent au delà de la capacité de l'expert de spécifier des distributions a priori complètes de Dirichlet. Un deuxième problème récurrent est qu'il est impossible d'incorporer des contraintes, fournies par l'expert, entre les paramètres du réseau Bayésien [Niculescu et al., 2006].

Comme n'importe quel problème d'apprentissage artificiel, il y a un compromis à gérer entre le pouvoir de généralisation et la simplicité du modèle. En apprentissage de structures ou de paramètres, les modèles les plus généralistes ayant plus de paramètres à estimer ne sont pas adéquats lorsque les données d'apprentissage sont insuffisantes. Par contre, les modèles simples ne tirent pas avantage de l'abondance de données.

### 7.2.3 Problème des valeurs manquantes

Dans les applications réelles, les bases de données peuvent contenir des observations dont certains champs (attributs) sont manquants. Écarter ces observations ne semble pas une solution appropriée car des informations pertinentes peuvent de même être supprimées. Généralement, les champs vides sont complétés par des valeurs ayant une vraisemblance maximale, ensuite les algorithmes classiques peuvent être utilisés. Deux techniques sont communément employées à savoir l'échantillonnage de Gibbs et l'algorithme EM (Expectation-Maximization). Les deux algorithmes supposent que les valeurs manquantes apparaissent aléatoirement et suivent une certaine distribution.

## 7.3 Problème des variables latentes

Les modèles de représentation de données entre autres les RBs font souvent la même hypothèse que les variables du problème étudié sont toutes observées (des variables à notre disposition et mesurables), malheureusement cette hypothèse est quelquefois assurée. L'incorporation de variables latentes dans certains modèles RBs donne des meilleurs résultats puisque cela permet de rendre le graphe du réseau lisible et moins dense, et par conséquent, de réduire la complexité considérablement. Cependant, la problématique est quand l'exploration des variables latentes est avantageuse? Et comment les identifier?

Un constat très intéressant est qu'un ensemble de variables mutuellement dépendantes (des variables qui sont contenues dans la même clique dans l'arbre de jonction correspondant) est généralement l'effet d'une variable (ou plusieurs) latente(s) qui joue le rôle d'une cause commune qui explique ces dépendances, l'ajout d'une telle variable latente diminue la taille de la clique

FIGURE 1.5 – Exemple d'introduction d'une variable latente.

**(a)**: Un réseau Bayésien dont la taille de la clique maximale est égale à 03. **(b)**: Nouveau réseau après introduction de la variable latente $H$. Taille de la clique maximale est égale à 02. Cette modification ainsi n'est permise que si $H$ est à l'origine des dépendances dans le graphe entre $A$, $B$ et $C$ dans **(a)**.

de cet ensemble, améliore la représentativité du réseau et réduit la complexité en mémoire et en temps d'inférence.

Certainement, il est préférable que l'expert examine s'il est possible d'ajouter des variables cachées ce qui permet de simplifier le RB. Dans certains cas, l'introduction des variables latentes artificielles (ne sont pas des variables physiques) peut s'avérer très utile, cela peut être effectué comme suit :

- Rechercher des possibilités des variables latentes potentielles et les connecter au réseau.
- Déterminer le nombre des valeurs des variables latentes ajoutées.
- Estimer les paramètres (TPCs des variables après connexion des variables latentes), en utilisant les algorithmes standard tels que l'algorithme EM.
- Faire un bilan de la modification apportée en termes de la complexité et de la représentativité des données par le nouveau réseau.

Quand nous envisageons l'exploration des variables latentes, nous étendons significativement la complexité d'algorithmes d'apprentissage et la taille du modèle à apprendre, mais en revanche, cela peut contribuer à accélérer les algorithmes d'inférence. Généralement, avec la présence des variables latentes, l'apprentissage est confronté au problème d'existence de plusieurs issues d'identification puisque il y a plusieurs modèles qui peuvent expliquer les données et beaucoup de paramètres à estimer dont ceux qui concernent les variables latentes. Pour remédier à ce problème d'indétermination, des restrictions sur les modèles explorés sont souvent faites. Par exemple, les variables latents sont généralement supposées être des parents des variables observées et être mutuellement indépendantes. D'autres restrictions peuvent être considérées, par exemple dans [Anandkumar et al., 2013], les relations entre les variables latentes et les variables observées sont supposées linéaires. [Binder et al., 1997] a proposé un algorithme d'apprentissage à base du gradient des paramètres des RBs contenant des variables latentes dont la structure est connue.

## 7.4 Problème des variables continues

Un arsenal large de théories et d'outils se rapportant aux RBs a été proposé seulement pour le cas où toutes les variables sont discrètes. Cependant, des situations dans lesquelles les variables continues existent toutes seules ou coexistent avec des variables discrètes dans le même phénomène sont assez courantes en pratique. Dans des tels cas, les variables continues sont généralement discrétisées, ensuite tous les algorithmes pour les variables discrètes peuvent être utilisés. Malheureusement, la discrétisation n'est pas gratuite et entraîne un surcoût. En effet, le compromis imposé est qu'avec une discrétisation faible on va perdre de l'information et avec une discrétisation élevée la complexité des algorithmes d'apprentissage et d'inférence augmente.

Dans ce contexte, le modèle CLG [Lauritzen and Jensen, 2001, Cowell, 2005] est une des premières solutions permettant l'inférence exacte à travers les techniques du calcul local dans les RBs hybrides composés de nœuds continus et de nœuds discrets. Dans lequel, les variables discrètes ne sont pas permises d'avoir des parents continus. Afin d'éviter cette restriction, [Koller et al., 1999] modélise la distribution des nœuds discrets qui ont des parents continus par des mélanges d'exponentielles tronquées (MTE pour *Mixture of Truncated Exponentials*). Dans le modèle MTE, le domaine de la conjointe est subdivisé en hypercubes (une faible discrétisation en quelque sorte), puis un mélange d'exponentielles propre à chaque hypercube est utilisé pour approximer la conjointe localement. Même les MTEs souffrent des problèmes classiques des RBs liés à la convergence de l'apprentissage et à la complexité en temps et en espace mémoire.

# 8 Conclusion

Ce chapitre nous a permis d'introduire plus précisément les RBs. Nous avons mis en évidence que ce type d'outils permet de modéliser les relations d'indépendance conditionnelle entre les variables d'un phénomène et qu'il permet également une manipulation formelle de l'incertitude de connaissances via le calcul probabiliste. Un des points forts des RBs est qu'ils disposent d'une lisibilité certaine comparés à d'autres paradigmes tels que les réseaux de neurones.

Les qualités des RBs font d'eux un choix prometteur pour la modélisation des problèmes complexes. On dénombre plusieurs applications des RBs dans des domaines variés où l'incertitude joue un rôle important. En effet, on en trouve des applications dans les domaines de : bioinformatique et prise de décision médicale, diagnostique de défauts et analyse de fiabilité de systèmes, traitement d'image et du signal, datamining et classification de données. Une collection d'exemples d'applications diverses des RBs est fournie dans [Mittal and Kassim, 2007].

Certainement, les RBs disposent de plusieurs avantages. Cependant, les difficultés liées à l'apprentissage ainsi que la complexité super-exponentielle des algorithmes d'apprentissage et d'inférence empêchent la généralisation de leur emploi. Il convient de noter que jusqu'à nos

jours, la prise en compte des variables continues représente une limite majeure aux RBs.

Chapitre 2

# Analyse en Composantes Indépendantes : Algorithmes et Applications

## Sommaire


**1 Introduction . . . . . 35**

**2 Analyse en composantes indépendantes . . . . . 37**

2.1 Définition de l'ACI . . . . . 37

2.2 Ambiguïtés de l'ACI . . . . . 38

2.3 Restrictions de l'ACI . . . . . 38

**3 Independence statistique . . . . . 38**

3.1 Mutuelle information . . . . . 38

3.2 Information mutuelle et entropie . . . . . 39

3.3 Dépendance, Corrélation et Gaussianité . . . . . 40

**4 Fonctions objectif de l'ACI . . . . . 41**

4.1 Maximisation de la non Gaussianité . . . . . 41

4.1.1 Théorème de la limite centrale . . . . . 41

4.1.2 Mesures de la non Gaussianité . . . . . 42

4.2 Minimisation de l'information mutuelle . . . . . 43

4.3 Maximisation de vraisemblance . . . . . 44

4.4 Méthodes tensorielles . . . . . 44

**5 Prétraitement pour l'ACI . . . . . 45**

5.1 Centrage . . . . . 45

5.2 Blanchiment (whitening) . . . . . 46

**6 Algorithmes ACI populaires . . . . . 47**

**7 Applications de l'ACI . . . . . 47**

7.1 Séparation aveugle de sources . . . . . 47

7.2 Réduction de bruit . . . . . 47

7.3 Extraction de paramètres . . . . . 49
7.4 Application en domaine biomédical . . . . . 50
**8 Extensions et variantes de l'ACI . . . . . 50**
8.1 ACI non linéaire . . . . . 50
8.1.1 Mélanges post-non linéaires (PNLs) . . . . . 51
8.1.2 Mélanges convolutifs . . . . . 51
8.2 Analyse en sous-espaces indépendants (ASI) . . . . . 52
8.3 ACI parcimonieuse . . . . . 52
8.4 ACI topographique . . . . . 53
8.5 Limites de l'ACI . . . . . 53
**9 Conclusion . . . . . 54**

# 1 Introduction

L'intérêt des modèles statistiques pour l'analyse et la représentation de données s'agrandit de plus en plus ces dernières années. Une des méthodes populaires qui ont été développées pour cette fin est l'analyse en composantes indépendantes (ACI). Le concept de l'ACI a été initialement proposé par [Comon, 1994] et les premiers algorithmes de l'ACI ont été employés pour résoudre le problème de séparation aveugle de sources [J.F.Cardoso, 1998].

L'exemple célèbre connu sous le nom de "cocktail-party" illustre un champ d'application de l'ACI en BSS. Imaginer que dans une pièce, deux personnes discutent simultanément et deux microphones sont installés à proximité dans des endroits différents. Ainsi, les deux microphones enregistrent effectivement deux signaux temporels différents, dénotés par $x_1(t)$, et $x_2(t)$ , avec $x_1$ et $x_2$ les amplitudes des signaux enregistrés à l'instant $t$. Les signaux enregistrés $x_1$ et $x_2$ sont relativement corrélés mais ils ne sont pas identiques car les microphones ne se trouvent pas à la même distance aux locuteurs et le signal acoustique s'affaiblit avec la distance parcourue de propagation. En effet, $x_1$ et $x_2$ sont une sommation pondérée des deux signaux vocaux émis par les locuteurs, que nous désignons par $s_1(t)$ et $s_2(t)$. Ce scénario peut être exprimé sous forme d'une équation linéaire

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \qquad \textit{Pour chaque} \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \qquad \textit{instant } t : 1, ..., T \end{aligned} \tag{1}$$

où $a_{11}$,$a_{12}$,$a_{21}$ et $a_{22}$ sont des constantes réelles qui dépendent des distances séparant les microphones aux locuteurs et $T$ est la taille des signaux. Sachant qu'on dispose seulement des signaux enregistrés $x_1(t)$ et $x_2(t)$, la reconstitution des signaux source $s_1(t)$ et $s_2(t)$ peut s'avérer très utile. Sans aucune connaissance a priori ou hypothèse sur les paramètres inconnus du modèle (1) qui sont : les coefficients de mélange $a_{11}$,$a_{12}$,$a_{21}$,$a_{22}$ et les sources $s_1(t)$ et $s_2(t)$, l'identification d'une solution de (1) est théoriquement impossible. En BSS, le mot "aveugle" signifie que nous avons peu d'hypothèses faibles sur les paramètres inconnus de mélange. Si les deux variables source $s_1$ et $s_2$ sont supposées statistiquement indépendantes et au plus une d'elles peut avoir une distribution Gaussienne, le modèle (1) devient identifiable avec certaines ambiguïtés légères [Comon and Jutten, 2010, Hyvärinen et al., 2001]. Quand la maximisation de l'indépendance statistique entre les sources $s_1(t)$ et $s_2(t)$ est employée comme un critère principal d'optimisation, le problème de la séparation aveugle de sources (1) devient équivalent à celui de l'ACI où chaque source représente une composante indépendante.

Étant donné un nuage de points de $T$ points en dimension $n$, du point de vue géométrique, l'ACI et l'analyse en composante principales (ACP) peuvent être perçues comme un simple changement de base (empirique), qui a pour effet une représentation dans des nouvelles directions. L'ACP trouve la direction qui exprime la plus grande variance totale ; puis la deuxième direction orthogonale à la précédente portant la plus grande variance restante et ainsi de suite pour les autres directions, et obtenant des composantes décorrélées qui sont susceptibles d'expliquer l'information contenue dans le nuage de départ en éliminant une grande part de re-

dondance. A la différence de l'ACP, l'objectif de l'ACI est de projeter le nuage des points de départ sur des directions indépendantes le maximum possible, l'indépendance est considérée dans le plein sens de l'indépendance statistique et pas résumée à la décorrélation, la corrélation constitue seulement une part de la dépendance statistique, (deux variables indépendantes sont décorrélées et l'inverse n'est pas toujours vrai), voir exemple des directions obtenues par l'ACP et l'ACI dans 2.1(b).

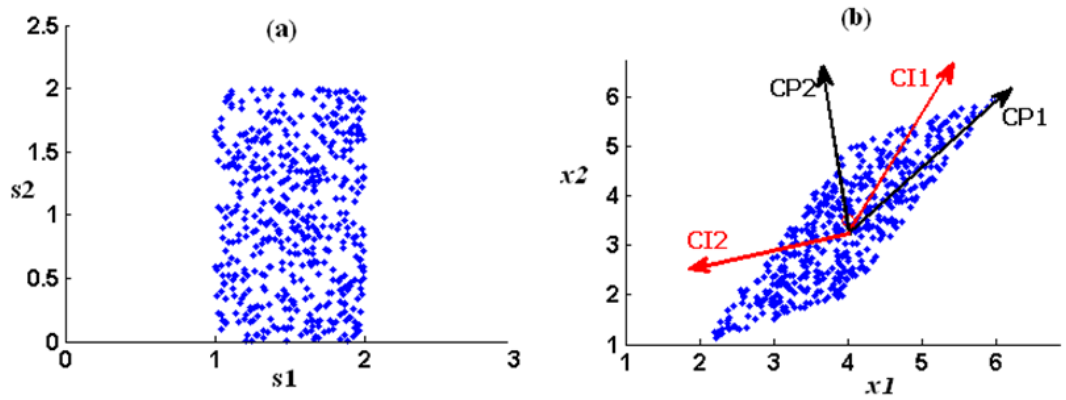



FIGURE 2.1 – Interprétation géométrique de l'ACI et l'ACP.

**(a)**: La distribution conjointe de deux composantes indépendantes $s_1$ et $s_2$ ayant des distributions uniformes. **(b)**: La distribution conjointe des mélanges observés $x_1$ et $x_2$, $\mathbf{x} = (x_1, x_2)^T = A\mathbf{s} = A(s_1, s_2)^T$ avec $A = [2\ 1; 1\ 2]$. Les directions des composantes principales sont $CP_1$ et $CP_2$ et les directions des composantes indépendantes sont $CI_1$ et $CI_2$.

L'ACI a été originalement développée pour traiter des problèmes qui sont similaires à celui du "cocktail-party". Cependant, avec l'élargissement sans cesse de son intérêt au sein de la communauté de traitement de signal et d'analyse de données, d'autres applications ont vu le jour. Elle a été employée avec succès pour dévoiler l'information pertinente de l'activité du cerveau à partir des enregistrements électriques type électroencéphalogramme (EEG) de l'activité cérébrale. L'ACI est aussi appliquée pour l'extraction des paramètres afin de fournir une représentation convenable de l'image, de l'audio ou d'autres types de données pour des applications diverses telles que le filtrage de bruit, la classification et la compression. En résumé, l'ACI standard s'agit d'une transformation linéaire et qui se révèle utile lorsque des traitements appliqués sur les composantes indépendantes (des variables latentes) sont plus favorables que sur les variables observées. Les principales transformations linéaires sont : Fourier, Haar, transformée en cosinus, ACP et chacune d'elles a ses propres favorables propriétés.

Dans la suite de ce chapitre, nous présentons une synthèse de l'ACI en exposant ses principales propriétés, un ensemble d'algorithmes et d'approches ACI populaires, ainsi que certaines applications et limites de l'ACI.

# 2 Analyse en composantes indépendantes

## 2.1 Définition de l'ACI

Un modèle à variables latentes peut être utilisé pour procurer une définition rigoureuse de l'ACI [Comon, 1994]. Soit $\mathbf{x}(t) = (x_1(t), x_2(t), ..., x_n(t))^T$ un signal observé de dimension $n$ à l'instant $t$ et $\mathbf{s}(t) = (s_1(t), s_2(t), ..., s_m(t))^T$ un signal sous-jacent des composantes indépendantes de dimension $m$ à l'instant $t$. Le modèle ACI conventionnel consiste en un mélange *linéaire* et *instantané* de la forme suivante

$$\mathbf{x}(t) = A\mathbf{s}(t) \tag{2}$$

avec

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ ... \\ x_n(t) \end{pmatrix} \quad \mathbf{s}(t) = \begin{pmatrix} s_1(t) \\ s_2(t) \\ ... \\ s_m(t) \end{pmatrix} \quad A = \begin{pmatrix} a_{11} & a_{21} & ... & a_{m1} \\ a_{12} & a_{22} & ... & a_{m2} \\ ... & ... & ... & ... \\ a_{1n} & a_{2n} & ... & a_{mn} \end{pmatrix}$$

Pour éviter le problème de sous-détermination, il est supposé qu'il y a plus d'observations que des composantes indépendantes; $n \geq m$. Parfois, la définition (2) est étendue pour inclure un terme de bruit Gaussien blanc et additif $\mu$; $\mathbf{x}(t) = A\mathbf{s}(t) + \mu$.

Le modèle ACI est un modèle paramétrique et génératif, c'est-à-dire qu'il décrit comment les données observées sont générées à partir d'un processus de mélange de composantes $s_i$. Les composantes indépendantes (CIs) sont des variables latentes qui ne sont pas observables directement. La matrice de mélange $A$ est également inconnue. Nous avons à disposition seulement le vecteur des données observées $\mathbf{x}$ et le but de l'ACI est d'estimer à la fois $A$ et $\mathbf{s}$.

L'estimation ACI est souvent faite en se basant sur des hypothèses générales (faibles). L'hypothèse simple qui est habituellement considérée est l'*indépendance statistique* des variables latentes $s_i$, on donnera une définition rigoureuse de l'indépendance statistique en Section 3.1. Le problème ACI standard est toujours paramétrique en $A$, et il est non paramétrique en $s$. Dans certains cas, il est paramétrique en $s$ conformément aux hypothèses appliquées aux densités de probabilité des $s_i$ (connaître la famille de distributions des $s_i$ simplifie la résolution du problème ACI). Dans un but de simplification, la matrice $A$ est souvent supposée carrée ($n = m$).

Le concept ACI est étroitement lié à la séparation aveugle de sources. L'ACI est un outil qui peut être vu comme une transformation (similaire à celle de l'ACP) vers un espace de CIs, ces dernières peuvent représenter des grandeurs physiques (sources) ou des variables artificielles. Cependant, la BSS est une application dans laquelle on veut reconstituer les sources (signaux originaux) en disposant de peu d'informations (peu d'hypothèses) sur la matrice de mélange et sur les sources. L'ACI correspond donc à la BSS lorsque l'indépendance statistique est l'hypothèse centrale de résolution en BSS. Dans le reste du manuscrit, nous utilisons les termes "composante indépendante", "source" et l'acronyme "CI" d'une manière interchangeable.

## 2.2 Ambiguïtés de l'ACI

Lorsque nous utilisons l'ACI pour la BSS avec les restrictions spécifiées dans la section 2.3, le modèle ACI standard, celui de l'Equation (2), présente les deux ambiguïtés suivantes :

**- Indétermination à l'échelle** On ne peut pas déterminer l'énergie (variance) des CIs, c'est-à-dire $s_{i\ estimée} = K \times s_{i\ originale}$ où $K$ est une constante scalaire. La justification pour cela est que les deux paramètres à estimer $A$ et $\mathbf{s}$ sont des inconnus. Si une source $s_i$ est multipliée par un scalaire $K$ et la ligne $i$ de la matrice $A$ est subdivisée par le même scalaire $K$, le problème (2), reste inchangé.

**- Indétermination à l'ordre** On ne peut pas déterminer l'ordre des CIs, la composante estimée d'indice $i$ peut correspondre à la source d'indice $j \neq i$. Formellement, une matrice de permutation $P(P_{i,j} \in \{0,1\} / \sum_{i=1}^{n} P_{i,j} = \sum_{j=1}^{n} P_{i,j} = 1)$ et son inverse peuvent être insérées dans le modèle (2) pour donner $\mathbf{x} = AP^{-1}P\mathbf{s}$, où $P\mathbf{s}$ est une simple permutation de $\mathbf{s}$, et qui correspond à la nouvelle matrice de mélange à estimer $AP^{-1}$. En effet, la solution de (2) peut être une permutation des sources originales.

## 2.3 Restrictions de l'ACI

Pour être plus précis, nous discutons ici les restrictions ou les conditions qui doivent être vérifiées pour que l'estimation du modèle ACI standard (2) soit en mesure de restituer les sources (application en BSS). D'abord, la restriction fondamentale pour que la séparation ACI soit possible est que les sources $s_i$ doivent être statistiquement indépendantes et que le mélange des sources suit le modèle (2), c'est-à-dire un mélange *instantané* et *linéaire* et pas un autre type de mélanges (convolutif, non linéaire, etc.). En outre, les sources doivent être non Gaussiennes ; et plus précisément, une seule composante au maximum peut être Gaussienne pour réussir l'ACI. Un exemple numérique ainsi que la justification pourquoi les variables Gaussiennes ne sont pas autorisées dans le modèle ACI standard sont donnés dans [Hyvärinen and Oja, 2000]. Enfin, pour que la solution ne soit pas sous-déterminée, le nombre de variables observées (non identiques) doit être supérieur ou égal au nombre de sources ($n \geq m$).

# 3 Independence statistique

## 3.1 Mutuelle information

Deux variables aléatoires $a$ et $b$ sont dites (statistiquement) dépendantes si leur information mutuelle est non nulle, $I(a,b) \neq 0$, voir la Section I.2.1 pour plus de détails.

Soit $Y = \{Y_1, ..., Y_n\}$ un vecteur aléatoire de dimension $n$, une mesure de l'indépendance statistique entre ses entrées $\{Y_1, ..., Y_n\}$ est la divergence de Kullback-Leibler (KLD) entre la conjointe $P(Y)$ et le produit des distributions marginales $\prod_{i=1}^{n} P(Y_i)$ [Cardoso, 2003]:

$$I(Y) = KLD[P(Y), \prod_{i=1}^{n} P(Y_i)] \tag{3}$$

où la divergence KLD entre une distribution de probabilité $Q(x)$ et une autre distribution $R(x)$ est définie comme suit :

$$KLD(Q, R) = \int_x Q(x) log(\frac{Q(x)}{R(x)}) dx \tag{4}$$

La divergence $KLD(Q, R)$ est toujours positive et s'annule seulement si $Q(x)$ et $R(x)$ sont identiques. $KLD(Q, R)$ est non symétrique et mesure la perte d'information si $Q$ est approximée par $R$

## 3.2 Information mutuelle et entropie

L'entropie est un concept basique de la théorie de l'information. L'entropie mesure combien une variable aléatoire est aléatoire et imprédictible, elle peut être interceptée comme la quantité d'information exprimée par une variable. Par conséquent, elle peut être utilisée pour la quantification de l'information mutuelle entre variables, voir [Cover and Thomas, 1991] pour une introduction à la théorie de l'information.

En remplaçant le logarithme dans la fraction (4) par une différence des logarithmes, on aura

$$I(Y) = \sum_{i=1}^{n} H(Y_i) - H(Y) \tag{5}$$

où $H(.)$ désigne l'entropie différentielle de Shannon pour les variables continues. L'entropie $H(x)$ se calcule comme suit :

$$H(x) = - \int_x P_x(x) log(P_x(x)) dx \tag{6}$$

En pratique, les méthodes ACI opèrent souvent en deux étapes : la première étape consiste à soustraire la moyenne $EY$ et le blanchiment, alors que durant la deuxième étape, on maximise une mesure d'indépendance. Le blanchiment s'agit d'une projection simple dont le but est d'avoir une matrice de covariance égale à l'identité $R_Y = I$. En d'autres termes, le blanchiment consiste en une transformation telle que les entrées du vecteur $Y$ soient décorrélées (covariance=0). Dans ce contexte, l'ACI est une méthode à minimisation d'entropie avec la contrainte de blanchiment [Cardoso, 2003].

Il y lieu de noter, que la propriété de blanchiment est préservée en appliquant n'importe quelle transformation orthogonale [Hyvärinen et al., 2001].

## 3.3 Dépendance, Corrélation et Gaussianité

Il existe un lien surprenant entre les notions de dépendance (statistique), de corrélation et de Gaussianité au sein du modèle ACI standard, $\mathbf{x} = A\mathbf{s}$ (nous définissons la Gaussianité comme une mesure du caractère Gaussien d'une variable). La compréhension de cette relation permet de concevoir les différentes approches classiques suivies par l'ACI.

♦ **Non Gaussianité** Il est montré dans [Cardoso, 2003] que la meilleure approximation Gaussienne de la distribution $P(Y)$ du vecteur aléatoire multidimensionnel $Y$ est la distribution Gaussienne $N(EY, R_Y)$, où $EY$ et $R_Y$ sont la moyenne et la matrice de covariance du vecteur $Y$, respectivement. La non Gaussianité $G(Y)$ d'un vecteur $Y$ est définie comme étant la divergence KLD entre la distribution de $Y$ et sa meilleure approximation Gaussienne [Cardoso, 2003] comme suit :

$$G(Y) = KLD[P(Y), N(EY, R_Y)] \tag{7}$$

Une petite valeur de $G(Y)$ indique que $Y$ a une distribution proche d'une Gaussienne.

♦ **Corrélation** Pour une distribution Gaussienne $N(EY, R_Y)$, la meilleure approximation par un produit de Gaussiennes est la distribution $N(EY, diag(R_Y))$, où $diag(R_Y)$ désigne la matrice diagonale avec les mêmes éléments de la diagonale que ceux de $R_Y$ [Cardoso, 2003]. Dans le sens de la divergence KLD, on définit la corrélation $C(Y)$ d'un vecteur $Y$ comme étant la divergence entre sa meilleure approximation Gaussienne, $N(EY, R_Y)$, et sa meilleure approximation en produit de Gaussiennes $N(EY, diag(R_Y))$ comme suit [Cardoso, 2003]

$$C(Y) = KLD[N(EY, R_Y), N(EY, diag(R_Y))] \tag{8}$$

♦ **Relation entre dépendance, corrélation et non Gaussianité** La relation qui lie les trois concepts de dépendance (information mutuelle), de corrélation et de non Gaussianité, tels qu'ils sont introduits, est la suivante [Cardoso, 2003]

$$I(Y) = C(Y) + (G(Y) - \sum_{i=1}^{n} G(Y_i)) \tag{9}$$

Il convient de noter que toutes les quantités figurant dans l'Equation (9) sont invariantes aux transformations de décalage et de changement d'échelle appliquées sur les entrées du vecteur $Y$ [Cardoso, 2003].

Si le vecteur $Y$ est Gaussien, alors nous avons $G(Y) = 0$ et ses marginales sont également des Gaussiennes, $G(Y_i) = 0$ pour $i = 1, ..., n$. La relation (9) devient alors $I(Y) = C(Y)$, ce constat est évident car la dépendance statistique se résume à la corrélation dans le cas Gaussien.

La relation (9) montre aussi que la différence entre la dépendance et la corrélation, $I(Y) - C(Y)$, est égale à la différence du caractère non Gaussien de la conjointe et la somme des

marginales, $G(Y) - \sum_{i=1}^{n} G(Y_i)$. En effet, les variables décorrélées ne sont que partiellement indépendantes et dans le contexte de l'ACI, la décorrélation n'est qu'une part de la séparation. Cela explique bien pourquoi un grand nombre de méthodes ACI commencent d'abord par le blanchiment (une transformation qui a comme effet d'avoir une matrice de variance-covariance $R_Y = \mathtt{I}$), ce qui représente en fait une décorrélation (éléments non diagonaux de $R_Y = 0$) plus une normalisation (éléments diagonaux de $R_Y = 1$).

# 4 Fonctions objectif de l'ACI

## 4.1 Maximisation de la non Gaussianité

### 4.1.1 Théorème de la limite centrale

L'indépendance statistique, le critère conventionnel à maximiser par l'ACI, est étroitement liée au concept de la (non) Gaussianité dans le contexte du modèle (2). En plus des développements rapportés en Section 3.3, l'emploi de la non Gaussianité comme un critère d'optimisation peut être aussi justifié par le théorème célèbre de la limite centrale. Ce théorème joue un rôle primordial en statistique, et particulièrement en BSS et en ACI.

**Théorème 5 (*Limite centrale*)**

*Soit $Z = \{z_1, ..., z_K\}$ un ensemble de $K$ variables aléatoires indépendantes de moyennes $\{u_1, ..., u_K\}$ et de variances $(\sigma_1^2, ..., \sigma_K^2)$, respectivement. Alors, pour un grand nombre de variables $K$, la variable $T$, définie ainsi :*

$$T = \Sigma_{j=1}^{K} z_j$$

*a une distribution approximativement Gaussienne de moyenne $\Sigma_j u_j$ et de variance $\Sigma_j \sigma_j^2$ .*

Selon ce théorème, la distribution d'une somme de variables aléatoires indépendantes tend toujours vers une distribution Gaussienne. Dans un mélange ACI typique, une composante $x_i$ du vecteur multivariable observé $\mathbf{x}$ a la forme suivante

$$T = \Sigma_{j=1}^{n} a_{ij} x_j$$

avec $a_{ij}$ $j = 1, ..., n$ des coefficients constants de mélange ACI. Le théorème de la limite centrale peut être généralisé au cas d'une sommation linéairement pondérée des variables aléatoires. Comme conséquence, même pour un nombre relativement petit de sources, la distribution de $P(x_i)$ est généralement plus proche d'une Gaussienne que les distributions $P(s_j)$ $j = 1, ..., n$.

En suivant ce raisonnement, la maximisation d'indépendance implique une maximisation de la non Gaussianité. Cela est un principe fondamental de l'ACI. Cet aboutissement s'accorde avec les résultats théoriques obtenus en Equation (9), après le blanchiment (la décorrélation

est une part de blanchiment), la indépendance devient égale à la différence du caractère non Gaussien des CIs $\Sigma_{i=1}^{n} G(s_i)$ et celui du vecteur observé $G(\mathbf{x})$. En effet, la Gaussianité telle qu'elle définie dans (7) est invariante sous les transformations linéaires [Cardoso, 2003], c'est-à-dire nous avons $G(\mathbf{x}) = G(\mathbf{s})$. Il découle de (9) que la dépendance (information mutuelle) $I(\mathbf{s})$, à minimiser dans le modèle ACI, entre les entrées du vecteur des CIs $\mathbf{s} = (s_1, ..., s_m)$ est

$$I(\mathbf{s}) = C(\mathbf{s}) - \Sigma_{i=1}^{m} G s_i + cst \tag{10}$$

avec $cst$ un terme constant sous n'importe quelle transformation linéaire (dans notre cas, $G(\mathbf{x}) = G(\mathbf{s}) = cst$). En suivant ce raisonnement, la minimisation de la fonction objectif $I(\mathbf{s})$ sous la contrainte de blanchiment revient à maximiser la non Gaussianité des CIs, c'est-à-dire à maxmiser $\Sigma_{i=1}^{m} G(s_i)$.

### 4.1.2 Mesures de la non Gaussianité

♦ **Kurtosis** Le kurtosis est une mesure classique de la non Gaussianité en utilisant des moments statistiques d'ordre supérieur (cumulant d'ordre 4 et la variance). Le kurtosis $Kurt(Y)$ d'une variable aléatoire $Y$ est défini par

$$Kurt(Y) = \frac{E(Y^4)}{[E(Y^2)]^2} - 3 \tag{11}$$

$Kurt(Y)$ est nul seulement si $Y$ est Gaussienne, il est positif pour les variables dites super Gaussiennes et négatif pour les variables dites sous Gaussiennes. Plus précisément, la non Gaussianité est mesurée par la valeur absolue du kurtosis. L'emploi du kurtosis est justifié essentiellement par sa simplicité et son estimation rapide. Cependant, il est sensible aux données aberrantes [Hyvärinen et al., 2001], une bonne estimation du kurtosis requiert souvent un échantillon de données de taille importante.

♦ **Néguentropie** La néguentropie d'une variable aléatoire $Y$ que nous noterons $J(Y)$ se définit par

$$J(Y) = H(Y_{gauss}) - H(Y) \tag{12}$$

où $Y_{gauss}$ est une variable aléatoire Gaussienne ayant la même variance que $Y$ et $H(.)$ est l'entropie de la variable paramètre telle qu'elle est définie dans (6).

Une variable Gaussienne dispose de la plus grande entropie parmi toutes les variables aléatoires ayant la même variance. Cela veut dire que l'entropie peut être utilisée comme une mesure de la non Gaussianité. La néguentropie, définie comme étant un différentiel de l'entropie (12), est toujours positive et elle s'annule seulement si $Y$ est Gaussienne.

L'avantage de l'usage de la néguentropie est qu'elle est en quelque sorte l'estimateur optimal de la non Gaussianité [Hyvärinen et al., 2001] en termes de performances. Néanmoins, un de ses inconvénients réside dans sa complexité exorbitante de calcul, notamment dans l'estimation de la distribution $P(Y)$ et dans le calcul de l'intégrale de l'entropie en Equation (6).

♦ **Approximation de la néguentropie en utilisant les cumulants** Le calcul de la néguentropie est une tâche complexe et non pratique dans la fonction objectif de l'ACI. Ainsi, les approximations sont souhaitables afin d'avoir des algorithmes plus efficaces. En faisant un développement de Gram-Charlier de la néguentropie autour d'une variable Gaussienne, il est démontré [Hyvärinen et al., 2001] qu'une approximation de la néguentropie d'une variable aléatoire de moyenne nulle et de variance 1, est obtenue en utilisant les cumulants d'ordre supérieurs par

$$J(Y) \approx E(Y^3)^2 + \frac{1}{48}[kurt(Y)]^2 \tag{13}$$

Cette approximation permet d'avoir un calcul simple et peu complexe de la néguentropie.

## 4.2 Minimisation de l'information mutuelle

Dans le mélange ACI standard, l'objectif principal est de trouver des sources de telle sorte qu'elles soient indépendantes le plus possible. Le critère de maximisation de l'indépendance statistique est équivalent à la minimisation l'information mutuelle entre les CIs. On rappelle que pour le mélange ACI standard de l'équation (2), $\mathbf{x}(t) = A\mathbf{s}(t)$, l'objectif consiste à estimer une matrice de séparation $W$ de telle sorte que le vecteur des CIs $\mathbf{y}(t) = W\mathbf{x}(t)$ soit le plus proche possible des sources $\mathbf{s}(t)$. En se basant sur les fondements de la théorie de l'information [Cover and Thomas, 1991], la relation entre les entropies conjointes de $\mathbf{y}$ et de $\mathbf{x}$ est donnée par :

$$H(\mathbf{y}) = H(y_1, ..., y_n) = H(x_1, ..., x_n) + \texttt{log}|\texttt{det}(W)| \tag{14}$$

Il est possible donc d'écrire l'information mutuelle entre les composantes de $\mathbf{y}$ comme un différentiel d'entropie comme suit :

$$\begin{array}{rcl} I(y_1, ..., y_n) & = & \sum_{i=1}^{n} H(y_i) - H(y_1, ..., y_n) \\ & = & \sum_{i=1}^{n} H(y_i) - H(x_1, ..., x_n) - \texttt{log}|\texttt{det}(W)| \end{array} \tag{15}$$

Ainsi, et puisque $H(x_1, ..., x_n) = cst$, minimiser l'information mutuelle $I(y_1, ..., y_n)$ est équivalent à minimiser la fonction objectif suivante

$$F(W) = \sum_{i=1}^{n} H(y_i) - \texttt{log}|\texttt{det}(W)| \tag{16}$$

Lorsque il y a la contrainte que les composantes $y_i$ sont décorrélées et leurs variances égales à l'unité, $var(y_i) = 1$, l'information mutuelle à minimiser devient [Hyvärinen and Oja, 2000, Hyvärinen et al., 2001]

$$I(y_1, ..., y_n) = Cst - \sum_{i=1}^{n} J(y_i) \tag{17}$$

où la constante $Cst$ ne dépend pas de $W$. L'équation (17) montre le lien entre la néguentropie et l'information mutuelle. (17) montre également que rechercher une transformation linéaire inversible ($W$) qui minimise l'information mutuelle est équivalent à trouver des directions qui maximisent la néguentropie. L'estimation ACI par minimisation de l'information mutuelle est équivalente à la maximisation de la somme des non Gaussianités des CIs estimées lorsque les composantes estimées sont supposées décorrélées [Hyvärinen et al., 2001].

## 4.3 Maximisation de vraisemblance

Le maximum de vraisemblance est une approche très répandue pour l'estimation des paramètres des modèles de représentation. Son emploi pour l'ACI est illustré dans l'algorithme "infomax" [Bell and Sejnowski, 1995].

La vraisemblance d'un échantillon de $N$ vecteurs de données $\{\mathbf{x}(1), ..., \mathbf{x}(N)\}$ s'écrit :

$$\texttt{log}(\ L(\ \{\mathbf{x}(1), ..., \mathbf{x}(N)\}, W\ )\ ) = \sum_{t=1}^{N} \sum_{i=1}^{n} \texttt{log} f_i(W_i^T \mathbf{x}(t)) + N\texttt{log}|\texttt{det}(W)| \tag{18}$$

où $W_i^T$ est la colonne $i$ de la matrice $W$. $f_i$ sont les fonctions de distribution de densité de composantes $s_i$. Dans cette approche, on cherche les densités $f_i$ dans une famille restreinte de fonctions de distributions. L'équation (18) est obtenue tout en supposant que les densités $f_i$ sont indépendantes.

## 4.4 Méthodes tensorielles

Les cumulants croisés généralisent le concept des matrices de covariance $cov(x^i, x^j)$. Le tenseur de cumulants d'ordre quatre est la matrice de dimension quatre dont les entrées sont données par les cumulant croisés d'ordre quatre $cum(x^i, x^j, x^k, x^l)$. Les cumulants croisés d'ordre 2, 3 et 4 d'une variable aléatoire vectorielle $x$ de moyenne nulle valent respectivement

$$\begin{aligned} cum(x^i, x^j) &= E(x^i x^j) \\ cum(x^i, x^j, x^k) &= E(x^i x^j x^k) \\ cum(x^i, x^j, x^k, x^l) &= E(x^i x^j x^k x^l) - E(x^i x^j)E(x^k x^l) \\ &= -E(x^i x^k)E(x^j x^l) - E(x^i x^l)E(x^j x^k) \end{aligned} \tag{19}$$

Le tenseur de cumulants s'agit d'un opérateur linéaire défini par les cumulants d'ordre quatre. Soit $F$ le tenseur d'ordre quatre, agissant sur des matrices $n \times n$ défini par

$$M, n \times n \rightarrow F_{i,j}(M) = \sum_{k,l} m_{k,l}\, cum(x^i, x^j, x^k, x^l) \tag{20}$$

où $m_{k,l}$ sont les éléments de la matrice $M$ à transformer. Une matrice propre de $F$ est une matrice $M$ telle que $F(M) = \lambda M$, $\lambda \in R$.

Il est intéressant d'expliquer le lien qui existe entre le tenseur de cumulants et l'estimation ACI. Supposons que les mélanges observés, $\mathbf{x}$, sont blanchis et que nous voulons estimer une matrice de mélange ACI, $A$, tel que $\mathbf{x} = A\mathbf{s} = W^{-1}\mathbf{s} = W^T\mathbf{s}$ (après le blanchiment, la recherche est limitée aux matrices orthogonales, et pour une matrice diagonale nous avons $W^{-1} = W^T$). Il est démontré dans [Hyvärinen et al., 2001] que toute matrice de la forme $M = w_m w_m^T$, avec $m = 1, ..., n$ et $w_m$ et le vecteur ligne d'indice $m$ de la matrice de séparation $W$, est une matrice propre du tenseur et les valeurs propres correspondantes sont données par le kurtosis des CIs. En plus, toutes les autres valeurs propres du tenseur sont nulles [Hyvärinen et al., 2001]. Par conséquent, on peut facilement obtenir les CIs à partir des matrices propres ; si les valeurs propres (kurtosis) non nulles sont distinctes, chaque matrice propre, $M = w_m w_m^T$, nous donne une seule colonne de la matrice de mélange blanchie $A = W^T$. Dans le cas où les valeurs propres ne sont pas distinctes, des considérations supplémentaires doivent être prises en compte pour construire $A = W^T$ [Hyvärinen et al., 2001].

Les cumulants d'ordre quatre contiennent toute l'information d'ordre quatre de la même manière que la matrice de covariance qui contient toute l'information d'ordre deux. Une des approche de l'ACI consiste donc à minimiser les cumulants croisées d'ordre quatre pour déterminer des composantes aussi indépendantes que possible.

La méthode bien connue JADE (Joint Approximate Diagonalization of Eigenmatrices), introduite dans [Cardoso and Souloumiac, 1993], adopte cette approche tensorielle de minimisation des cumulants croisés. Son principe consiste à chercher une matrice orthogonale $W$ qui diagonalise conjointement les matrices $Q_i = WF(M_i)W^T$ , où $M_i$ est une matrice propre de $F$, les matrices propres de $F$ sont calculées dans un stage antérieur. JADE est une des techniques principales proposées pour résoudre le problème des valeurs propres égales du tenseur de cumulants. Des définitions et des propriétés intéressantes concernant les cumulants ainsi qu'un panorama des méthodes tensorielles sont donnés dans [Hyvärinen et al., 2001].

# 5 Prétraitement pour l'ACI

## 5.1 Centrage

Un prétraitement simple et souvent favorable pour l'ACI consiste à centrer les données observées par la soustraction du vecteur moyenne $\mathbf{m} = E(\mathbf{x}), \overline{\mathbf{x}} = \mathbf{x} - E(\mathbf{m})$. Ainsi, l'entrée de l'ACI $\overline{\mathbf{x}}$ est un vecteur multidimensionnel de moyenne zéro. Ce prétraitement n'est pas indispensable mais il est très utile pour la simplification du calcul dans la majorité des algorithmes ACI.

Après l'estimation de la matrice de séparation $W$ en utilisant des données centrées, l'estimation des sources peut être complétée par l'addition du vecteur moyenne $W \times m$ aux sources estimées $\overline{\mathbf{s}}$ : $\mathbf{s} = \overline{\mathbf{s}} + W \times m$.

## 5.2 Blanchiment (whitening)

Un autre prétraitement utile dans la plupart des algorithmes ACI consiste à blanchir les données observées avant l'optimisation de la fonction objectif ACI mais après le centrage. Le blanchiment s'agit d'une transformation linéaire, $\tilde{\mathbf{x}} = P \times \overline{\mathbf{x}}$ où $P$ est une matrice de transformation de taille $n \times n$, afin d'avoir un vecteur $\tilde{\mathbf{x}}$ blanchi (white). Le vecteur de données $\tilde{\mathbf{x}}$ est dit blanchi si ses composantes sont mutuellement décorrélées et si sa matrice de covariance est égale à l'identité, formellement, si nous avons

$$E(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T) = I \tag{21}$$

Le blanchiment est toujours possible. Une des méthodes populaires pour le faire consiste en la détermination de la décomposition en valeurs propres EVD (eigen-value decomposition) de la matrice de covariance $E(\mathbf{x}\mathbf{x}^T) = VDV^T$ où $V$ est la matrice orthogonale des vecteurs propres de $E(\mathbf{x}\mathbf{x}^T)$ et $D$ est la matrice diagonale de ses valeurs propres. Il est démontré dans [Cardoso and Souloumiac, 1993] que la matrice de blanchiment est $P = VD^{-1/2}V^T$ et qu'après avoir effectué le blanchiment, la nouvelle matrice de mélange ACI à estimer $\tilde{A}$, qui correspond au nouveau modèle $\tilde{\mathbf{x}} = A\mathbf{s}$ , est une matrice orthogonale. Ce prétraitement est utile car il restreint la recherche au espace des matrices orthogonales (moins de paramètres à estimer), plutôt que l'espace de toutes les matrices possibles de taille $n \times n$. Une deuxième propriété importante est qu'après le blanchiment, la maximisation de l'indépendance statistique devient équivalente à la maximisation de la non Gaussianité (voir Section 3.3).

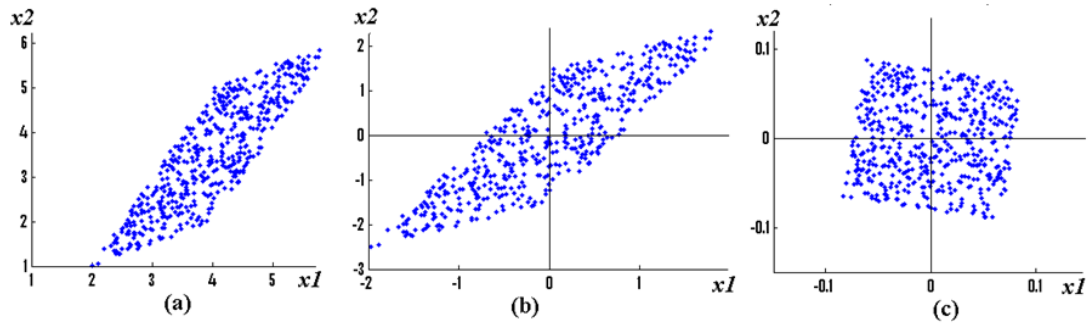



FIGURE 2.2 – Exemple de centrage et de blanchiment.

**(a)**: La distribution conjointe des mélanges observés $x_1$ et $x_2$, c'est le même échantillon de la figure 2.1(b). **(b)**: La conjointe de $x_1$ et $x_2$ après centrage. **(c)**: La conjointe de $x_1$ et $x_2$ après le centrage et le blanchiment.

# 6 Algorithmes ACI populaires

Un grand nombre d'algorithmes ACI ont été proposés. Ces algorithmes se distinguent essentiellement selon les hypothèses qui sont faites sur les CIs, la forme de la fonctions de mélange ACI, les fonctions objectif ACI à optimiser, les approximations adoptées, les méthodes d'optimisation. Le tableau 2.1 dresse un panorama des principales méthodes ACI.

# 7 Applications de l'ACI

L'ACI est originalement appliquée aux problèmes similaires à celui de le exemple célèbre du "cocktail party", après, son usage est rapidement étendu à la représentation et à l'analyse de données multidimensionnelles non-gaussiennes.Dans cette section, nous passons en revue certaines applications d'intérêt de l'ACI.

## 7.1 Séparation aveugle de sources

L'ACI est une méthode initialement conçue pour résoudre le problème de séparation aveugle de sources. Par le mot "aveugle" nous voulons dire que peu de faibles hypothèses sont faites sur les sources recherchées. Si seulement l'indépendance statistique entre les sources est considérée, le modèle ACI devient équivalent à la formulation de la BSS. Dans le cas du mélange BSS linéaire et instantané, le modèle ACI standard (2), défini dans [Comon, 1994], peut être résolu avec des indéterminations à l'échelle et à l'ordre sous quelques restrictions (voir les sections 2.2 et 2.3. Nous référons à [Cardoso, 1989, Bell and Sejnowski, 1995, Hyvärinen et al., 2001, J.F.Cardoso, 1998] pour un état de l'art riche sur le rapport entre l'ACI, qui peut être vue comme une méthode d'analyse de données, et la BSS, qui peut être vue comme une application.

Toutefois, quand la fonction du mélange BSS prend des formes générales (non linéaire par exemple), la solution de l'ACI basée seulement sur l'indépendance statistique n'est pas souvent unique. Par conséquent, des contraintes ou des hypothèses supplémentaires sont nécessaires pour que l'estimation ACI converge vers la solution BSS. Les contraintes peuvent être imposées sur les distributions de sources et|ou sur la famille de fonctions de mélange. Par exemple, pour la fonction du mélange ACI post-non linéaire qui prend la forme suivante : $x_i(t) = f_i[\sum_{j=1}^{n} a_{ij} s_j(t)], i = 1, ..., n$, il est montré dans [Taleb and Jutten, 1999] que ce modèle présente les mêmes indéterminations que le modèle ACI standard.

## 7.2 Réduction de bruit

Un des champs intéressants d'application de l'ACI est le filtrage de bruit. Plusieurs approches de réduction de bruit ont été proposées dans la littérature. Une d'elles consiste à appli-

Tableau 2.1 – Comparaison entre quelques méthodes ACI populaires.

| Algorithme ACI | Ordre statist. | Blanch. | Méthode d'optimisation | Fonction objectif | Quelques propriétés |
|---|---|---|---|---|---|
| FastICA [Hyvärinen et al., 2001] | 4 | Oui | – Algorithme itérative de point-fixe<br>– A gradient naturel | – Maximisation de non Gaussianité | – déflationniste (composante après composante) ou simultané (parallèle) |
| JADE [Cardoso and Souloumiac, 1993] | 4 | Oui | – Diagonalisation conjointe approximée | – Annulation des cumulants croisés d'ordre supérieur | – JADE généralisation de FOBI<br>– Restreinte aux petites dimensions |
| Infomax [Bell and Sejnowski, 1995] | 4 | Oui | – Approche neuronale | – Maximisation de l'information mutuelle | – Equivalente au maximum de vraisemblance |
| SOBI [Belouchrani et al., 1997] | 2 | Oui | – Diagonalisation conjointe approximée | – Minimisation des auto covariances | – Exploitant la structure temporelle |
| FOBI [Cardoso, 1989] | 4 | Oui | – Décomposition en valeurs propres de la matrice de corrélation blanchie | – Annulation des cumulants croisés d'ordre 4 | – les kurtosis des CIs doivent être différents<br>– a une faible complexité algorithmique. |
| Kernel ICA [Jordan and Bach, 2002] | 4 | Variable [1] | – Décomposition incomplète de Cholesky pour le calcul des matrices des vecteurs propres généralisés des matrices de Gram | – Minimisation d'une fonction de contraste | – Méthodes à noyau dans un espace de Hilbert à noyau reproduisant (RKHS) |
| Birth [Albera et al., 2003] | 6 | Non | – Décomposition en valeurs propres<br>– diagonalisation conjointe | – Annulation des cumulants croisés d'ordre 6 | – Mélanges sous-déterminés ($n \leq m$) |

(1) en fonction de la fonction noyau choisie (projection souvent non linéaire), un type d'indépendance est minimisé.

quer la méthode ACI bruitée dont le modèle s'écrit comme suit :

$$\mathbf{x} = A\mathbf{s} + \mathbf{b} \tag{22}$$

où le vecteur de bruit $\mathbf{b} = (b_1, ..., b_n)^T$ est souvent supposé additif, Gaussien et indépendant du vecteur des CIs, $\mathbf{s}$. Si le bruit est supposé engendré en amant (avec les sources), avant le mélange ACI, le modèle correspondant est

$$\mathbf{x} = A(\mathbf{s} + \mathbf{b}) = A\tilde{\mathbf{s}} = A\mathbf{s} + A\mathbf{b} \tag{23}$$

Les composantes bruitées, $\tilde{\mathbf{s}} = \mathbf{s} + \mathbf{b}$, sont non Gaussiennes, donc les méthodes ACI classiques peuvent être utilisées pour les estimer. Etant donné que le bruit est Gaussien et sa matrice de covariance est connue a priori, les composantes débruitées $\mathbf{s} = \tilde{\mathbf{s}} - \mathbf{b}$ peuvent être déduite, ainsi les variables observées après débruitage sont reconstituées, $\mathbf{x}_{débruits} = A\mathbf{s}$.

Une autre manière pour le débruitage est envisageable quand certaines composantes représentent le bruit à réduire. Ici, on distingue entre le cas de bruit non Gaussien et le cas de bruit Gaussien. Dans le premier cas, il faut un moyen comme l'intervention d'un expert pour identifier les CIs jugées comme un bruit afin de les supprimer. Par contre, dans le deuxième cas, on peut chercher directement les premières directions maximisant la non Gaussianité et les composantes minimisant la non Gaussianité (bruit) sont délaissées. Voir [Hyvärinen et al., 2001] pour plus de détails sur l'ACI débruitées.

## 7.3 Extraction de paramètres

La méthode ACI a été utilisée dans une multitude de travaux pour l'extraction des paramètres discriminants pour des tâches différentes telles que la régression et la classification, notamment dans le domaine du traitement d'images et de l'acoustique. A titre d'exemple, [Martiriggiano et al., 2005] a employé l'ACI à base de noyaux pour l'extraction des paramètres faciaux pour la reconnaissance de visages. D'abord, la matrice de données, $X$, est organisée de telle sorte que les lignes représentent les différents pixels et les colonnes représentent les différentes images. Ensuite, l'estimation ACI est faite en utilisant les données d'apprentissage $X_{app}$, $Y_{app} = W \times X_{app}$. Chaque colonne de $Y_{app}$ et de $Y_{test} = W \times X_{test}$ représente une instance de données d'apprentissage et de test de classification, respectivement. Un autre exemple d'extraction de paramètres de la parole en utilisant l'ACI pour la reconnaissance des phonèmes est illustré dans [Kwona and Lee, 2004], où les vecteurs des coefficients cepstraux fréquentiels de mel (MFCCs) représentent les entrées de l'ACI, et les vecteurs des CIs estimés représentent les vecteurs d'attributs à classifier.

Généralement, des tâches d'apprentissage telles que la classification et la régression sont plus performantes sur un espace de CIs que sur l'espace des variables observées, c'est cette hypothèse qui a été exploitée dans [Martiriggiano et al., 2005, Kwona and Lee, 2004]. Une autre

astuce légèrement différente pour l'extraction de paramètres pour la classification supervisée, consiste à inclure la variable classe comme une variable observée dans le modèle ACI, est décrite dans [Kwak et al., 2001], $\mathbf{x}(t) \oplus z(t) = (x_1(t), x_2(t), ..., x_n(t), z(t))^T$ représente le vecteur observé du modèle ACI.

## 7.4 Application en domaine biomédical

L'ACI et les techniques qui s'en dérivent sont intensivement appliquées pour le traitement des données biomédicales, en particulier les signaux EEG et les images FMRI.

L'électro-encéphalographie (EEG) est une technique d'exploration cérébrale qui mesure l'activité électrique du cerveau par des électrodes placées sur le cuir chevelu souvent représentée sous forme d'un tracé appelé électro-encéphalogramme. [Makeig et al., 1996] et [Vigario, 1997] sont parmi les premiers qui ont appliqué l'ACI pour les données EEG. Depuis, une explosion de travaux ont vu le jour. Les applications de l'ACI en EEG couvrent principalement la détection et l'élimination des artefacts (débruitage) et l'analyse des réponses cérébrales. Le bruit peut être dû soit à l'instrumentation employée, soit aux activités cérébrales accessoires telles que le mouvement de l'œil et la conversation qui n'ont pas un intérêt pour une étude spécifique.

L'imagerie par résonance magnétique fonctionnelle (FMRI) est une technique d'imagerie qui permet de visualiser quelles sont les régions du cerveau qui sont actives durant une tâche mentale spécifique. La plus populaire technique FMRI mesure le contraste des niveaux d'oxygénation, elle est basée sur le différentiel des propriétés magnétiques des veines sanguines oxygénées et désoxygénées. Les données FMRI sont un des premiers champs explorés par l'ACI [McKeown et al., 1998, McKeown and Sejnowski, 1998].

# 8 Extensions et variantes de l'ACI

## 8.1 ACI non linéaire

Bien que le modèle ACI standard (2) est simple et il a été appliqué avec succès dans plusieurs applications, il demeure trop restreint pour pas mal d'applications réelles. Il est clair que son extension aux mélanges non linéaires parait utile pour décrire des données suivant des distributions plus générales. Le mélange ACI *non linéaire*, *instantané* et *sans bruit* prend la forme suivante :

$$\mathbf{x}(t) = f(\mathbf{s}(t)) \tag{24}$$

Ainsi, chaque variable observée $x_i(t)$ peut s'écrire

$$x_i(t) = f_i(s_1(t), ..., s_n(t)) \tag{25}$$

où $f$ est une fonction multivariable inversible du mélange ACI, $f : R^n \rightarrow R^n$, et supposée inconnue ($f_i$ sont les composantes de $f$).

Une des caractéristiques fondamentales de l'ACI non linéaire est que, les décompositions du vecteur observé $\mathbf{x}$ en un vecteur de CIs $\mathbf{y}$, $\mathbf{y} = h(\mathbf{x})$, généralement existent, mais elles sont fortement non uniques. En conséquence, en plus de l'indépendance, des informations a priori sur les sources et|ou des contraintes sur la structure de la fonction de mélange sont nécessaires pour que l'ACI non linéaire soit en mesure de fournir une solution au modèle BSS de l'équation (24), nous référons à [Comon and Jutten, 2010, Hyvärinen et al., 2001] pour plus de détails sur les propriétés et les approches de mélanges non linéaires. L'ACI post-non linéaire, présentée dans cette section, est un exemple pratique où des contraintes sont supposées sur la fonction du mélange.

### 8.1.1 Mélanges post-non linéaires (PNLs)

Dans les mélanges PNLs, les observations sont des mélanges ayant la forme suivante :

$$x_i(t) = f_i(\sum_{j=1}^{n}[a_{ij}s_j(t)]), \quad i = 1, ..., n) \tag{26}$$

où on suppose que les fonctions de mélanges $f_i$ sont inversibles. Les mélanges ACI type PNL sont identifiables pour des sources dont au plus une est Gaussienne avec les mêmes indéterminations que l'ACI linéaire si la matrice $A$ est de mélange suffisant (*sufficiently mixing*) [Comon and Jutten, 2010]. $A$ est dite de mélange suffisant si on a $a_{ij} \neq 0$,alors $(\exists l \neq i, \exists k \neq j)/(a_{lj} \neq 0, a_{ik} \neq 0)$ [Comon and Jutten, 2010].

### 8.1.2 Mélanges convolutifs

Les mélanges ACI sont souvent considérés instantanés, en d'autres mots, le vecteur observé $\mathbf{x}(t)$ est fonction du vecteur source $\mathbf{s}(t)$ au même instant $t$. Cependant, les sources peuvent contribuent au mélange avec des différents retards de propagation. Comme dans l'exemple de télécommunication, les retards dépendent des distances qui séparent les sources aux capteurs, le milieu de propagation et le phénomène de réflexion. Ces retards peuvent s'exprimer sous forme d'un produit de convolution. Dans un modèle ACI convolutif typique d'ordre $z$, le signal d'observations est déterminé par un modèle de convolution multicanaux comme suit :

$$\mathbf{x}(t) = \sum_{k=0}^{Z} A_k \mathbf{s}(t-k) \tag{27}$$

Toutes les versions du signal sources les plus récentes (de $t - z$ à $t$) sont mélangées ensemble selon les matrices $A_k$.

## 8.2 Analyse en sous-espaces indépendants (ASI)

L'ASI (ISA pour *Independent subspace analysis*) est une version non linéaire de la méthode ACI. Dans cette analyse, les composantes sont rangées dans des groupes et l'indépendance est établie de telle sorte que les groupes soient mutuellement indépendants et les composantes du même groupe soient dépendantes [Hyvaerinen et al., 2009]. Dans le modèle ASI, les variables observées $x_i$ ou les sources $s_i$, $i = 1, ..., n$ sont souvent supposées des variables multidimensionnelles de la même dimension $d$ tel que les variables $s_i$, $i = 1, ..., n$ sont indépendantes et les $d$ entrées d'une variable $s_i$ sont dépendantes. Une mesure de la non Gaussianité à base du kurtosis vectoriel (une généralisation du kurtosis unidimensionnel aux dimensions supérieures) est employée dans [Sharma and Paliwal, 2006] pour résoudre l'ASI, dans lequel une illustration est faite sur des signaux EEG contrôlant les rythmes cardiaques d'un fœtus et de sa mère.

## 8.3 ACI parcimonieuse

Les techniques de décomposition parcimonieuse dans la dernière décennie commencent d'être sensiblement exploitées en BSS comme en ACI. Leur avantage est quelles permettent la séparation lorsque le nombre de sources excède le nombre de variables de mélange observées. En outre, elles contribuent substantiellement dans l'amélioration de la séparation en cas de matrices de mélange carrées. Par la parcimonie (sparsity) nous voulons dire qu'à chaque instant $t$, une source $s_i(t)$ est significativement plus active que les autres sources $|s_i(t)| >> |s_j(t)|$ ,$j \neq i$. Outre l'hypothèse de l'indépendance, la méthode ACI parcimonieuse (SCA pour *Sparse Component Analysis*) tente d'exploiter la sparsity des sources.

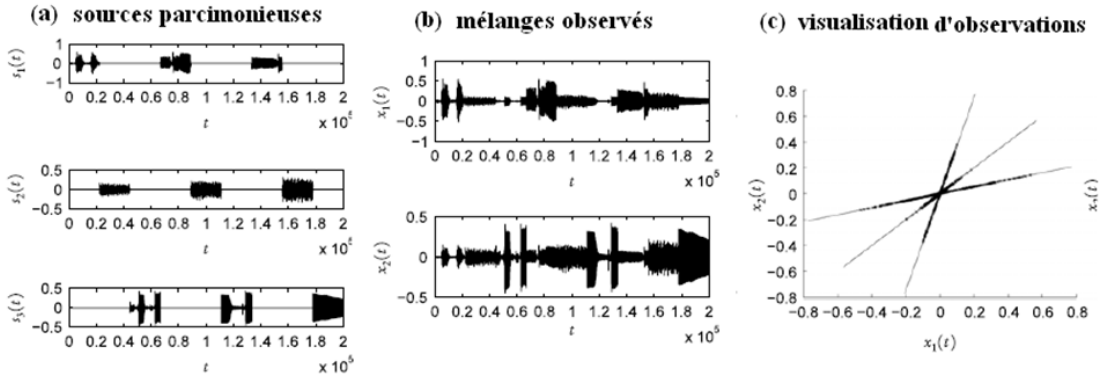



FIGURE 2.3 – Principe géométrique de SCA.

**(a)**: Signaux sources avec sparsity. **(b)**: 02 signaux de mélange. **(c)**: Visualisation des vecteurs de mélange observés, où on constate que les vecteurs sont alignés sur les directions des colonnes de la matrice de mélange à estimer. Exemple pris de [Comon and Jutten, 2010]

Dans l'exemple affiché sur la figure 2.3(c), les vecteurs de données observés forment des

clusters linéaires clairement perceptibles. Par conséquent, il est approprié d'estimer $A$ en se servant d'un algorithme de clustering [Xu and II, 2005]. Ensuite, les directions des clusters constituent la matrice $A$ (le vecteur direction d'un cluster représente une colonne de $A$). Une riche synthèse de la méthode SCA est donnée dans [Comon and Jutten, 2010].

## 8.4 ACI topographique

L'ACI topographique [Hyvärinen et al., 2001] est définie par un modèle génératif qui est une version hiérarchique du modèle ACI standard (2). L'idée consiste à relaxer l'hypothèse de l'indépendance des composantes $s_i, i = 1, ..., n$ de telle sorte que les composantes qui sont proches dans la topographie ne sont pas supposées indépendantes dans le modèle. L'utilité d'un tel modèle est que l'indépendance totale ne peut pas être garantie dans le modèle (2), les composantes ne sont pas toutes mutuellement indépendantes. La structure de dépendance résiduelle est utilisée pour définir un ordre topographique des composantes, cette structure est souvent informative et son estimation est utile. La topographie est généralement modélisée par une grille de proximité, les composantes proches présentent une corrélation d'énergie (un type de dépendance [Hyvärinen et al., 2001]) et les composantes lointaines sont indépendantes. Ce modèle est plus ou moins similaire à l'ASI avec la seule différence qu'il permet d'avoir des groupes de composantes qui se chevauchent. Des applications de l'ACI topographique pour l'analyse des enregistrements magnétoencéphalographiques (MEG) et pour le filtrage des images naturelles sont aussi illustrées dans [Hyvärinen et al., 2001].

## 8.5 Limites de l'ACI

La fonction objectif principale à maximiser par l'ACI est l'indépendance statistique mutuelle globale entre les CIs $y_1, ..., y_n$. L'indépendance statistique entre deux variables aléatoires est un concept facile à comprendre et à définir en théorie (voir sections I.2.1 et II.(3.1), mais sa quantification empirique en se basant sur un échantillon de données n'est guère facile. La mesure de l'indépendance statistique implique l'emploi de toutes les statistiques d'ordre supérieur. A cause de l'augmentation de la complexité, toutes les méthodes ACI développées jusqu'à nos jours se contentent souvent aux statistiques d'ordre quatre (kurtosis, néguentropie, cumulants, etc.) et rarement au statistiques d'ordre six, pour accélérer l'estimation, elles recourent et parfois à des approximations de l'indépendance au détriment de l'efficacité. En conséquence, l'indépendance entre les CIs $y_1, ..., y_n$ est amplifiée relativement à l'indépendance originale entre les variables observées $x_1, ..., x_n$, mais elle n'est pas absolue.

L'estimation ACI est un problème d'apprentissage artificiel semi-paramétrique et non supervisé. Le phénomène de sur-apprentissage s'aggrave lorsque le nombre de paramètres libres à estimer dans le modèle ACI est trop important si comparé avec la taille de l'échantillon disponible, et dans cette circonstance les résultats de l'estimation se détériorent. Une étude appro-

fondie du phénomène de sur-apprentissage peut être trouvée dans [Särelä and Vigário, 2003]. Les solutions proposées dans ce sens incluent, en plus de l'acquisition de plus de données, la réduction de dimensionnalité afin de réduire le degré de liberté du modèle.

Un autre point très important se rapporte à la robustesse des méthodes ACI. En effet, l'indépendance et la sparsity sont fortement liées [Y. Li and Amari, 2004], ce qui peut conduire à une tendance naturelle de n'importe quel algorithme ACI vers des composantes parcimonieuses, plutôt que vers des composantes strictement indépendantes [Comon and Jutten, 2010]. Il convient de noter que la consistance des CIs estimées lorsque l'ACI est utilisée comme un outil de traitement est souvent requise, c'est-à-dire, quand les paramètres de l'algorithme sont changés librement, l'analyse doit aboutir aux mêmes résultats. L'invalidité du modèle ACI pour les données, l'implémentation stochastique, le bruit additif sont parmi les facteurs possibles de la variation des solutions [Comon and Jutten, 2010]. Une des méthodes pour l'analyse de la consistance d'un algorithme ACI consiste au ré-échantillonnage et aux exécutions multiples pour identifier les composantes reproductibles [Duann et al., 2003].

La détermination du nombre de CIs est toujours utile, avant d'exécuter l'ACI, car cela permet de réduire la complexité. Après l'exécution d'un algorithme ACI, l'interprétation des CIs n'est pas habituellement simple. Elle requiert des connaissance a priori, des discussions interactives, et des expérimentations complémentaires par un expert du domaine d'application à l'instar de l'exemple d'exploration de données EEG et FMRI.

# 9 Conclusion

Dans ce chapitre, nous avons exposé brièvement la méthode ACI par la présentation de quelques définitions, principes et propriétés intéressantes de l'ACI. Des méthodes et des approches classiques de l'ACI sont également explorées. Nous avons clôturé par la citation de quelques exemples pratiques d'applications.

Lorsque nous optons pour un modèle ACI spécifique, des questions importantes méritent d'être posés, à savoir : est ce qu'on peut appliquer ACI comme une boite noire ? est ce qu'on peut utiliser n'importe quel algorithme ACI ? Est ce qu'on fait confiance aux résultats obtenus par ACI ?

Typiquement, une méthode d'estimation ACI comporte trois ingrédients : un modèle paramétrique, un critère et un algorithme d'optimisation. Le modèle paramétrique offre une présentation simple et ainsi restreint la solution en un espace de recherche particulier. Par exemple, l'espace de l'ACI standard, modèle (2), est inclus dans l'espace de l'ACI PNL, modèle (26), et ce dernier est inclus dans l'espace de l'ACI non linéaire, modèle (24). Le critère (fonction objectif), généralement l'indépendance statistique entre les composantes, est une mesure de qualité de la solution. L'algorithme d'optimisation décrit la manière de la recherche d'une solution optimale (ou acceptable) dans l'espace de solutions du modèle paramétrique en maximisant (ou

en minimisant) la fonction objectif. En fonction de degré de liberté du modèle paramétrique et le critère considéré, les algorithmes ACI peuvent s'arrêter à des extrema locaux, et ainsi ne pas converger vers la solution optimale.

Bien que l'ACI soit une méthode statistique puissante, elle n'a pas encore atteint un stade de maturité et certains de ses aspects demeurent confus et nécessitent d'être explorés. Pour tirer parti de ses avantages, il faut l'appliquer attentivement et avec responsabilité. D'abord, le choix d'un tel modèle ACI ou d'un tel algorithme ACI demande une connaissance du phénomène étudié ce qui va contribuer à déterminer la nature de mélange. Cela permet de choisir le modèle ACI paramétrique approprié, de restreindre l'espace de recherche et de bien choisir l'algorithme d'optimisation. Des informations a priori sur les distributions de sources telles que la sparsity sont toujours les bienvenues, car elles contribuent dans la simplification et la convergence rapide de l'estimation ACI. En outre, en fonction de l'application voulue (BSS, débruitage, extraction de paramètres, exploration de données, etc.), la fonction objectif est personnalisée et le choix de la méthode ACI en dépend.

Il est bien d'élucider la confusion fréquemment faite entre l'ACI et la BSS. Étant donné un signal observé $\mathbf{x}$, l'ACI réside dans la recherche d'une transformation, $\mathbf{y} = W\mathbf{x}$ ($\mathbf{y} = F(\mathbf{x})$ pour le cas non linéaire) maximisant le critère d'indépendance statistique entre les entrées $y_1, ..., y_n$ de $\mathbf{y}$. Par contre, en BSS le critère d'optimisation mesure la qualité de reconstitution des sources, $\mathbf{y} = W\mathbf{x} = WA\mathbf{s}$, sachant que les sources originales $\mathbf{s}$ (inconnues) sont supposées indépendantes. En effet, l'objectif en BSS est d'optimiser un critère d'identification, de minimiser l'écart $\triangle = \|\mathbf{y} - \mathbf{s}\|$.

Chapitre 3

# Analyse en Composantes Indépendantes, Réseaux Bayésiens et Modèles Causaux

## Sommaire


**1 Introduction . . . . . 58**
**2 RBs et modèles causaux . . . . . 59**
2.1 Le modèle causal est un RB . . . . . 59
2.2 Causalité et dépendance statistique . . . . . 60
2.3 Inférence Bayésienne et inférence causale . . . . . 61
2.4 Causalité probabiliste . . . . . 61
2.5 Connaissances causales et construction des RBs . . . . . 63
**3 ACI et causalité . . . . . 64**
**4 Combinaison des techniques des RBs et de l'ACI . . . . . 65**
4.1 Approche Bayésienne pour l'ACI . . . . . 65
4.2 L'ACI pour la recherche des facteurs cachés . . . . . 66
4.3 Extraction des paramètres par l'ACI : application aux RBs . . . . . 66
**5 Correspondance entre RB et ACI . . . . . 67**
5.1 Relation entre le modèle RB et et l'ACI . . . . . 67
5.2 Similitudes entre RB et ACI . . . . . 69
5.3 Divergences entre RB et ACI . . . . . 69
5.4 Correspondance entre les structure de RB et de l'ACI linéaire . . . . 69
5.4.1 De la structure ACI vers la structure RB . . . . . 69
5.4.2 Exemples . . . . . 71
5.4.3 Le passage de la structure RB vers la structure ACI . . . . 71
**6 ACI, RB et modèles fonctionnels . . . . . 73**
6.1 Relation entre les MESs et les RBs . . . . . 73
6.2 ACI, MESs et RBs . . . . . 74
**7 Conclusion . . . . . 76**

# 1 Introduction

Les deux modèles, RB et ACI, faisant l'objet de cette thèse peuvent être utilisés pour fournir une représentation compactes de données multidimensionnelles. Dans ces deux modèles, surmonter le problème de la haute dimensionnalité est accompli par la recherche d'une factorisation de la fonction de probabilité conjointe $P(X) = P(x_1, ..., x_n)$ en un ensemble de fonctions locales avec un nombre réduit de variables, un produit de probabilités conditionnelles dans les RBs ($P(X) = \prod_{i=1}^{n} P(x_i|Parents_{x_i})$) et un produit de probabilités unidimensionnelles des CIs $P(s_i)$ dans le modèle ACI (par exemple, $P(X) = (\prod_{i=1}^{m} P(s_i))/(Jacobien(X))$ si $m = n$).

Comme points communs entre les RBs et l'ACI, nous soulignons le fait que les deux modèles sont des modèles statistiques fondés autour du concept clé de l'*indépendance statistique* entre variables aléatoires. L'indépendance est une notion utile pour la simplification de modèles et d'outils d'analyse de données, car elle permet une décomposition du problème global en un ensemble de problèmes de petites dimensions (une factorisation de la conjointe, par exemple) où les techniques exactes peuvent être appliquées. En se basant seulement sur un échantillon des données observées, l'indépendance (ou la dépendance) ne peut pas être confirmée d'une manière certaine, mais plutôt estimée en utilisant des mesures empiriques et approximatives susceptibles d'entachées d'erreurs.

En revanche, la principale différence entre les RBs et l'ACI réside dans leurs vocations de base. Les RBs sont fondamentalement conçus pour conduire la propagation de croyances (inférence statistique) d'une manière formelle, alors que l'ACI a été traditionnellement appliquée pour résoudre le problème de séparation aveugle de sources par la considération d'hypothèses faibles entre autres, l'indépendance des sources.

Dans cette thèse, nous visons l'exploration permettant d'exploiter d'éventuelles pistes des résultats obtenus récemment dans le domaine de l'ACI afin d'améliorer certains aspects liés aux RBs. L'amélioration escomptée peut couvrir des différents aspects tels que :

- utilisation des algorithmes d'estimation ACI pour la construction des RBs ;
- proposition d'un nouveau modèle RB, dans lequel les paramètres ACI sont utilisés pour le quantifier;
- extension des RBs à des données non conventionnelles, restées jusqu'à alors inflexibles, à l'instar des données continues.

Il est certain qu'une dépendance statistique observée entre deux variables ou plus n'est pas nécessairement un fait survenu au hasard, mais plutôt le produit d'un mécanisme particulier. Il est évident également que la causalité joue un rôle central dans la création ou l'explication des relations que nous constatons dans le monde réel. A titre d'exemple, la dépendance constatée entre la variable *Etat_du_ciel* {'*dégagé*', '*partiellement dégagé*', '*nuageux*'} et la variable Précipitation {'*beau temps*', '*pluie*', '*neige*'} est une conséquence directe du fait que *Etat_du_ciel*='*nuageux*' combiné avec d'autres facteurs (pression, température,...) cause la pluie ou la neige. Dans un second exemple, l'augmentation simultanée du coût d'une baguette de pain et

du coût d'un ticket de bus dans un marché libre peut être expliquée par l'augmentation du coût d'une cause commune telle que l'énergie, nous avons en effet une dépendance entre les deux variables *Coût_du_pain* et *Coût_du_transport*.

La causalité a été toujours un objet d'intérêt tout au long de l'histoire de la philosophie et les tentatives de définition de la causalité n'ont pas conduit à un consensus. Néanmoins, depuis l'antiquité, la causalité est généralement vue comme une relation de nécessitation où la cause est une *condition suffisante* mais *pas nécessaire* pour avoir l'effet. Cette identification a été étendue vers les années 1960s avec l'apparition des théories probabilistes de la causalité développées autour de la thèse qu'une cause peut ne pas rendre ses effets nécessaires mais plutôt augmente leurs probabilité d'occurrence. Cette thèse probabiliste est justifiée par le fait qu'une cause, peut ne pas agir toute seule, mais combinée avec d'autres causes cachées entraîne l'effet.

Il convient de rappeler que d'une part les RBs et les modèles causaux sont étroitement liés [Pearl, 2000, Drouet, 2007], et d'autre part, l'ACI est intensivement employée pour l'identification des causes cachées (sources) dans le problème BSS [Hyvärinen et al., 2001, Comon and Jutten, 2010]. Dans cette optique, ce chapitre est consacré à la présentation d'une partie de nos contributions théoriques visant à élucider les connexions plausibles entre les modèle RB et ACI autour de la causalité. La mise en évidence de cette correspondance nous permettra par la suite de faire apparaître les opportunités d'application de l'ACI au profit des RBs.

# 2 RBs et modèles causaux

## 2.1 Le modèle causal est un RB

Une méthode simple et triviale de construction d'un RB consiste en la représentation des relations de causalité existant entre les variables d'intérêt. Elle repose sur l'idée que les relations de cause à effet représentent des arcs dans le graphe du RB. Un arc partant du nœud $A$ vers le nœud $B$ est inséré dans le DAG $G$ si et seulement si $A$ *cause directement* $B$. Puis, les probabilités conditionnelles correspondantes $P$ sont estimées. Le RB $(G, P)$ ainsi obtenu est un RB causal qui, en plus d'une représentation des dépendances statistiques, dispose d'une interprétation causale.

Cette méthode exige une connaissance a priori des mécanismes causaux inter-variables et dépend des deux suppositions suivantes. D'abord, le concept de causalité directe est supposé être une relation entre variables. En second lieu, le réseau obtenu $(G, P)$ représente un RB ; c'est-à-dire le graphe $G$ est acyclique et $(G, P)$ satisfait la condition causale de Markov (Théorème 6). La condition de Markov pour les modèles causaux peut être exprimée comme suit :

**Théorème 6 (*Condition causale de Markov*)**

*Chaque modèle causal markovien $M = (G, P)$ induit une distribution $P$ qui satisfait la condition de Markov parentale relative au diagramme causal associé $G$; dans lequel, chaque*

*variable est indépendante de tous ses non descendants (non effets) étant donnés ses parents (ses causes) dans G [Pearl and Verma, 1991].*

Il convient de noter que la construction d'un RB causal n'est possible que si l'hypothèse de *suffisance causale* est vérifiée. Un ensemble de variables $X$ est *suffisant causalement* si et seulement si chaque cause $Y$ commune à plusieurs variables de $X$ appartient aussi à $X$, c'est-à-dire l'absence des causes communes cachées. Cela signifie que les liens causaux entre les variables $X$ sont suffisants pour représenter toutes les relations d'indépendance conditionnelle. En conséquence, la présence des causes cachées amoindrit la capacité des modèles causaux.

## 2.2 Causalité et dépendance statistique

Il est certain, les relations causales jouent un rôle central dans la compréhension des interactions entre les entités du monde physique. Elles sont invariantes et leur détermination nous aide à améliorer la qualité de prédiction. Malheureusement, la causalité est un concept complexe (trop abstrait) dans la réalité qui ne cesse de susciter un bon nombre de divergences et de questions au sein de la communauté des philosophes, des statisticiens et des physiciens. A nos jours, il n'existe pas de définition qui fait l'unanimité de la causalité, nous citons ci-après celle proposée par [Sloman, 2005].

**Définition 11 (*Causalité*)**

*Une relation causale est un mécanisme se déroulant dans le temps qui utilise la cause (et possiblement d'autres choses) et qui produit un effet (et éventuellement d'autres choses).*

En gros, on est tous d'accord que la notion de causalité implique un changement d'état dans le temps que l'intervalle de temps soit court ou long. Une contrainte temporelle générale est que les effets ne peuvent pas précéder leurs causes [Sloman, 2005].

Dans la littérature, la causalité est souvent considérée comme une relation entre événements, en associant des événements effets à d'autres événements causaux. Un événement représente réellement un état (une valeur possible) d'une variable. Ainsi, la définition de la causalité peut être idéalisée pour couvrir les relations entre les variables selon la logique suivante: une variable $A$ cause une variable $B$, si un changement d'état de $A$ aura des effets sur le nouvel état de $B$.

Une connexion possible entre les concepts de la causalité et de la dépendance statistique (voir définition I.2.1) peut tirer sa justification du principe de la cause commune proposé par [Reichenbach, 1956].

***Principe de la Cause Commune (PCC))*** *Si non*$(A \perp B)$ *alors nous avons soit A cause B, soit B cause A, soit il existe un autre ensemble de variables U tel que* $C \in U$ *implique que C cause A et C cause B et* $(A \perp B|U)$, (ici la causalité peut être directe ou indirecte). [2]

2. Dans le contexte du principe PCC, la notion de causalité doit être considérée dans sa forme probabiliste:

Le PCC énonce que toute dépendance statistique $non(A \perp B)$ entre deux variables $A$ et $B$ est expliquée par le fait qu'une des variables cause l'autre ou par l'existence d'un autre ensemble de causes communes $U$. Dans un RB causal, nous avons la proposition suivante.

**Proposition 2 (*Markov, PCC*)**

*La condition causale de Markov implique le principe de la cause commune, [Williamson, 2005].*

## 2.3 Inférence Bayésienne et inférence causale

L'inférence Bayésienne appliquée sur des RBs ou sur des RBs causaux nous permet de savoir ce que nous pouvons s'attendre à observer sur certains objets lorsqu'on dispose de certaines observations sur d'autres objets. Mais, à la différence des RBs ordinaires, les RBs causaux ne se limitent pas à décrire les corrélations entre les observations. En effet, la causalité ne nous informe pas seulement sur ce que nous escomptons avoir comme des observateurs passifs, mais elle nous permet de connaître les effets attendus lorsque nous prenons des actions et nous intervenons comme des agents. La causalité s'intéresse plutôt aux effets des actions prises.

L'inférence Bayésienne, appliquée sur des RBs, après avoir observé la valeur de certaines variables et l'inférence causale, appliquée sur des modèles causaux tels que les RBs causaux, après avoir modifié la valeur de certaines variables (*une intervention*) ne conduisent pas à réviser nos croyances de la même manière. La première est effectuée selon la règle de Bayes, tandis que la deuxième est effectuée selon le mécanisme causal reliant les variables. Pearl [Pearl, 2000] introduit l'opérateur ***do*** pour représenter le concept de l'intervention ; $do(X = a)$ est une manipulation externe réalisée par l'utilisateur pour rendre $X = a$. Des exemples illustrent bien l'inégalité $P(Y|do(X = a)) \neq P(Y|X = a)$ entre l'intervention et l'observation sont proposés dans [Sloman, 2005].

## 2.4 Causalité probabiliste

La causalité est souvent vue comme un mécanisme déterministe dans lequel l'événement $A = a$ cause l'évènement $B = b$ implique que l'effet $B = b$ se produit suite à l'évènement $A = a$ (une nécessitation d'occurrence), dans ce sens, nous avons 'fumer ne cause pas le cancer', car cette implication n'est pas toujours vraie. Par conséquent, il est intéressant parfois de pencher pour la notion de la causalité probabiliste car elle est plus réaliste, dans laquelle l'occurrence de l'évènement $A = a$ augmente la fréquence d'avoir l'évènement $B = b$. L'idée centrale de la causalité probabiliste est que les causes augmentent la probabilité de leurs effets mais sans

l'évènement $A = a$ cause l'évènement $B = b$ veut dire que $A = a$ augmente la probabilité de $B = b$, parce que souvent en réalité, pas seulement la variable $A$, mais $A$ combinée avec d'autres facteurs cachés qui causent $B$, voir la section 2.4

les engendrer nécessairement. L'incorporation des probabilités (l'aspect incertain) est généralement due à la connaissance pas suffisamment parfaite du phénomène à étudier. Cela peut être dû essentiellement à :

- a) L'omission ou l'impossibilité d'inclure d'autres variables cachées contribuant comme des causes dans le phénomène à étudier.(Exemple de la figure 3.1).
- b) La complexité et les erreurs d'estimation de la fonction de causalité. Par exemple, le non déterminisme peut être une conséquence de la discrétisation des variables continues. La discrétisation est souvent imposée soit dans un but de simplification car les RBs ne sont pas bien formulés dans le domaine continu, soit pour l'estimation non paramétrique des relations de dépendance inter-variables car la fonction de causalité est inconnue (Exemple de la figure 3.2).

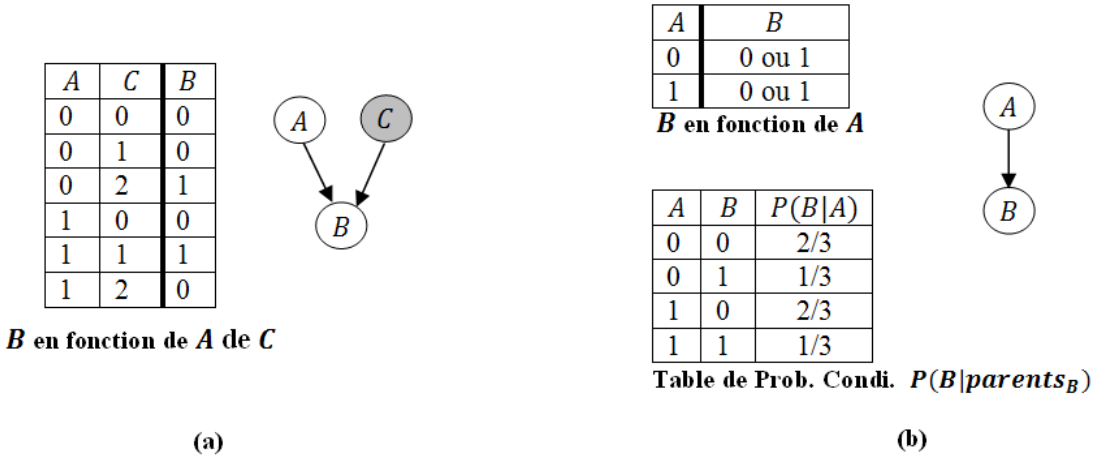



FIGURE 3.1 – Exemple illustrant le rôle des facteurs cachés dans la causalité probabiliste. **(a)**: un RB où la variable cause $C$ est incluse, le RB est causal et déterministe où $B$ prend une seule valeur pour chaque valeur du couple $(A, C)$. **(b)**: Un RB équivalent où la variable $C$ est cachée (pas incluse dans la modélisation), la relation $B$ en fonction de $A$ devient incertaine et elle est quantifiée par une table de probabilité conditionnelle $P(B|A)$.

L'insertion des variables cachées (s'il est possible) permet de simplifier la structure du RB et de converger plus vers un réseau plus déterministe, moins dense et plus représentatif des relations causales. L'introduction des concepts probabilistes dans l'analyse de causalité s'est développée à partir des années 1960s. La plus connue de ces analyses est celle de Mackie : "une cause est une *partie insuffisante mais non redondante* d'une condition *non nécessaire mais suffisante* pour son effet, une *condition* étant ici un ensemble de facteurs" [Mackie, 1974].

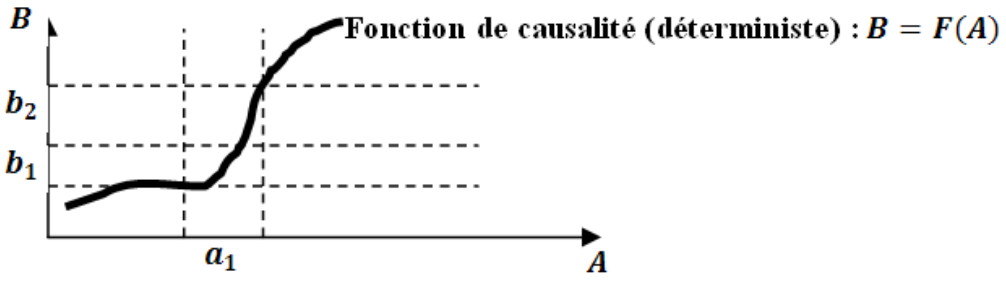



FIGURE 3.2 – Exemple illustrant l'apparition de l'incertitude dans la relation de causalité à cause de la discretisation.
La fonction de causalité $B = F(A)$ est une fonction déterministe et continue, mais après la discrétisation des domaines de $A$ et de $B$, on aura par exemple: si $A$ est instanciée à la valeur discrète $a_1$ ($A = a_1$), alors $B$ prendra soit la valeur $b_1$ soit la valeur $b_2$ avec des probabilités différentes.

## 2.5 Connaissances causales et construction des RBs

Il convient de rappeler que le modèle causal, tel que défini dans la section 2.1, représente un RB, mais le contraire n'est pas toujours vrai. En effet, un RB consiste en un modèle graphique qui représente fidèlement les relations d'indépendance conditionnelle entre les variables d'intérêt et permet ainsi de factoriser la conjointe selon son DAG (Chapitre I). Étant donné que le nombre de DAGs possibles à former à partir d'un ensemble de variables est *super-exponentiel* (explosion combinatoire). Pour un nombre important de variables ($> 10$), la recherche d'une structure (DAG) valide par les algorithmes d'apprentissage automatique est quasi-impossible à cause de l'explosion de la complexité calculatoire. Pour cette raison, des connaissances fournies par l'expert sur les relations causales inter-variables, même si elles sont partielles, sont toujours bénéfiques dans la mesure où elles permettent l'élimination ou l'orientation de certaines arrêtes et là restreignent significativement l'espace de recherche aux arrêtes restantes.

Il est évident que les informations causales établies par l'expert sont très utiles dans l'apprentissage de la structure RB. Malheureusement, en pratique, l'expert ne peut pas toujours identifier toutes les relations de causalité entre les variables d'intérêt en raison de certains facteurs tels que la présence de causes cachées.

L'exemple exposé de la figure 3.3 illustre ce dernier scénario. L'expert intervient en menant des expériences afin de dévoiler les liens causaux sous-jacents entre les 04 variables d'intérêts $A$, $B$, $D$, et $E$. Durant ce processus, il s'aperçoit que la variable $D$ suit les interventions faites sur $E$ d'où la determination du lien $E \rightarrow D$ (figure 3.3(a)). Cependant, les autres interventions sur les autres variables n'aboutissent à aucune conclusion. Ensuite, les algorithmes d'apprentissage automatiques de structure permettent par exemple de déterminer les flèches $B \rightarrow A$ et $B \rightarrow D$ (figure 3.3(b)). Ces deux dernières ne représentent pas réellement des liens causaux, mais plutôt une conséquence des deux causes communes cachées $C_1$ et $C_2$ (figure 3.3(c)).

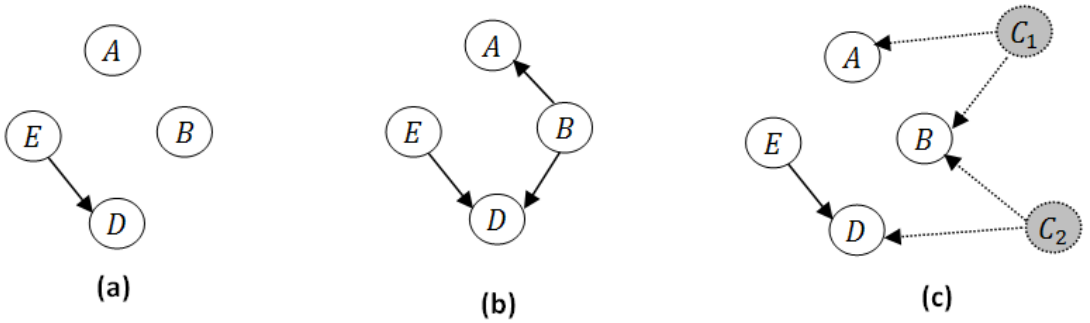



FIGURE 3.3 – Exemple montrant l'insuffisance du diagnostique causal dans la construction de la structure RB à cause des causes communes cachées.

**(a)**: L'arc $E \rightarrow D$ est déterminé par l'expert (*E cause D*). . **(b)**: Un RB correct où $B \rightarrow A$ et $B \rightarrow D$ sont déterminés par les algorithmes d'apprentissage automatiques. **(c)**: Les interventions n'ont pas pu déterminer $B \rightarrow A$ et $B \rightarrow D$ car ils sont les effets des causes communes cachées $C_1$ et $C_2$.

# 3 ACI et causalité

De la même manière que l'ACP, l'ACI peut être vue comme une transformation géométrique de l'espace des variables observées vers un autre espace qui est celui CIs.

Les CIs sont des variables latentes à estimer. Parfois, une variable latente correspond à une *variable physique cachée* que l'on connaît dans le monde réel, mais dans des études empiriques, on ne peut pas la mesurer pour des raisons pratiques. Cependant, une variable latente correspond souvent à une *variable hypothétique* qui n'a pas un aspect physique clair ; par exemple une variable abstraite ou une variable qui remplace l'effet de plusieurs variables physiques cachées.

Le succès des modèles à base de mélanges de variables latentes est argumenté par le fait que ces dernier essayent d'imiter la nature: une variable observée est généralement l'effet d'un ensemble de causes selon une fonction particulière de mélange où un bon nombre de ces causes représentent des variables latentes et la fonction de mélange (fonction de causalité) n' est pas toujours connue a priori.

L'indépendance mutuelle entre les causes est pratiquement plus importante que l'indépendance entre les effets dans un grand nombre de phénomènes réels. Lorsque la forme des fonctions de mélange est connue, la maximisation de l'indépendance (sous certaines hypotheses relatives à la complexité de mélange) a tendance à fournir des variables latentes convergeant vers les variables physiques causales (sources). Dans ce contexte, l'ACI a été appliquée initialement avec succès pour reconstituer les sources dans le problème causal BSS [Comon and Jutten, 2010]. L' exemple célèbre du *'coctail-party'* (II.1) est un modèle causal à variables latentes, utilisé souvent pour introduire les concepts de l'ACI et dans lequel les locuteurs représentent des causes directes des signaux vocaux enregistrés par les microphones.

Lorsque le nombre de sources est inférieur à celui des signaux observés dans le mélange ACI linéaire, la maximisation de l'indépendance statistique mutuelle à elle seule permet de déterminer les CIs avec les deux ambiguïtés suivantes: on ne peut pas déterminer l'énergie et l'ordre [Hyvärinen et al., 2001] des sources; en effet, on aura $y_i = K \times s_j$ $(i, j) : 1, ..., n$ où $y_i$ est une CI estimée et $s_j$ est une source réelle.

Cependant, la maximisation de l'indépendance mutuelle ne suffit pas pour explorer les sources dans les cas de mélanges ACI linéaires sous-déterminés ([Gribonval and Zibulevsky, 2009]) et de mélanges non linéaires. A titre d'exemple, pour le modèle ACI non linéaire $\mathbf{x} = F(\mathbf{s}) + \mathbf{b}$ tel qu'on a autant d'observations que de sources , $F$ est inversible et le bruit $b$ est négligeable, l'estimation des sources sans avoir des connaissances a priori supplémentaires est un problème mal-posé qui pourra avoir plusieurs solutions. Afin d'atteindre la solution appropriée, on aura besoin d'autres hypothèses sur les statistiques des signaux sources, des informations a priori sur la coloration (*coloredness*) et le caractère non stationnaire (Chapitre 14 dans [Comon and Jutten, 2010]). Le problème crucial relatif à l'existence de solution BSS a été traité par Hyvärinen et Pajunen dans une plate-forme générale [Hyvärinen and Pajunen, 1999], et par Taleb et Jutten [Taleb and Jutten, 1999] pour des types particuliers de mélanges ACI.

En résumé, la BSS consiste en un problème causal dans lequel une source représente une cause cachée et l'hypothèse de l'indépendance mutuelle des sources est habituellement considérée. Pour cette raison, l'ACI est principalement appliquée pour résoudre le problème BSS. Il convient de noter que la maximisation de l'indépendance mutuelle entre les sources comme seul critère d'optimisation ne permet de restituer les sources que pour certains types simples de fonctions de mélange. Pour des fonctions complexes et sans informations a priori additionnelles, on risque de converger vers des variables latentes hypothétiques. L'identification des correspondances entre les CIs et les sources est constamment posée en BSS. En revanche, la qualité d'un modèle de représentation de données basé sur l'ACI dépend essentiellement du degré d'indépendance entre les CIs. En effet, dans ce contexte que ces composantes représentent des sources physiques ou des variables latentes hypothétiques n'a pas une grande importance.

# 4 Combinaison des techniques des RBs et de l'ACI

## 4.1 Approche Bayésienne pour l'ACI

Dans tous les problèmes de l'apprentissage machine, l'approche Bayésienne est une des techniques les mieux adaptées pour l'estimation paramétrique pourvu que l'on dispose de connaissances suffisantes sur les distributions a priori. Le schéma général de l'approche Bayésienne pour l'estimation ACI peut se résumer comme suit :

1. Selon les informations disponibles sur le mélange ACI, déterminer les distributions a priori $P(A, S, \Sigma)$ des sources $S$, de la matrice du mélange $A$ et du bruit $\Sigma$.

2. Utiliser le théorème de Bayes pour inférer les distributions a posteriori des paramètres inconnus $P(A, S, \Sigma|X)$ :

$$P(A, S, \Sigma|X) = \frac{P(X|A, S, \Sigma)P(A, S, \Sigma)}{P(X)} \quad (1)$$

qui est proportionnel au numérateur $P(X|A, S, \Sigma)P(A, S, \Sigma)$ et le dénominateur se calcule de cette manière

$$P(X) = \int\int\int P(X|A, S, \Sigma)P(A, S, \Sigma)dS\,dAd\Sigma \quad (2)$$

Cette probabilité a posteriori comporte notre connaissance sur les paramètres inconnus après la prise en compte de mélanges ACI.

3. Utiliser la probabilité postérieure pour inférer les valeurs des paramètres $A$,$S$ et $\Sigma$ maximisant la vraisemblance.

$$(A, S, \Sigma) = argmax_{A,S,\Sigma}P(A, S, \Sigma|X) \quad (3)$$

L'*approximation Bayésienne variationnelle* est souvent utilisée pour simplifier l'estimation des paramètres. A titre d'illustration, voir la méthode *"Bayesian ICA"* proposée par [Winther and Petersen, 2007].

## 4.2 L'ACI pour la recherche des facteurs cachés

L'ACI est une méthode générative qui peut être utile pour révéler les mécanismes cachés du processus génératif de données et permet ainsi de dévoiler, sous certaines conditions, la structure sous-jacente des données observées. Dans [Kiviluoto and Oja, 1998], l'algorithme 'FastICA' a été employé pour explorer les facteurs principaux qui affectent les ventes de plusieurs magasins appartenant à une chaîne de distribution. Dans cette étude, les facteurs (CIs) obtenus ont eu clairement des interprétations différentes : un facteur suit les variations brusques dues aux vacances telles que noël, un deuxième facteur reflète la saisonnalité, un troisième facteur décrit la proximité des concurrents, etc.

L'idée centrale dans une telle analyse est que les variables observées sont le résultat de certains facteurs (causes) cachés qui sont grossièrement indépendants. En appliquant l'ACI, certaines des CIs peuvent correspondre à des facteurs cachés et dans ce cas, les interprétations de l'expert pour identifier, ce que représentent physiquement les CIs obtenues sont, nécessaires.

La recherche des facteurs cachés permet de déceler partiellement la structure causale et de cette manière contribue à la construction de la structure des RBs.

## 4.3 Extraction des paramètres par l'ACI : application aux RBs

Une approche hybride basée sur l'ACI et les RBs a été proposée dans [Kim et al., 2008] pour révéler les variations de la connectivité effective dans les cerveaux atteints de la maladie

de "schizophrenie". Comme les RBs sont sensibles aux données bruitées, l'ACI a été appliquée pour la réduction du bruit dans les images fMRI par la séparation des données brutes afin d'avoir des composantes spatiales indépendantes. Ensuite, les composantes relatives au bruit sont identifiées et éliminées et les composantes retenues ont été prise comme des entrées à un RB dynamique discriminatif (dDBN). Dans cette étude, l'objectif est de déterminer la différence entre les régions cérébrales d'un patient atteint de le schizophrenia et d'un controle sain en se basant sur la mesure de la connectivité effective entre les régions. L'extraction des paramètres par l'ACI est justifié par les deux raisons suivantes:

- Elimination de certaines CIs (variables) non pertinentes pour la prédiction telles que le bruit, ce qui permet de réduire la dimensionnalité.
- Avoir de nouvelles variables maximisant l'indépendance mutuelle, ce qui contribue éventuellement à l'amélioration de la prédiction soit en termes de complexité en temps et en espace, soit en termes de taux de bonne prédiction. Cela est dû à la possibilité de factoriser la conjointe ; à la possibilité de décomposition du problème global en des problèmes unidimensionnels plus simples.

# 5 Correspondance entre RB et ACI

## 5.1 Relation entre le modèle RB et et l'ACI

Un RB consiste en une transcription des relations d'indépendance conditionnelle entre un ensemble fini de variables d'intérêt observées. La notion de la (ind) dépendance statistique peut trouver une justification dans le principe de la cause commune : deux variables $x$ et $y$ sont dépendantes si une variable cause l'autre (directement ou indirectement) ou s'il existe d'autres causes communes à $x$ et $y$. Les relations de causalité sont fondamentalement des fonctions déterministes, le recours aux probabilités dans les RBs, causaux ou non causaux, est en gros justifié par l'incertitude due aux méthodes de quantification non paramétriques (discrétisation des conditionnelles) et notamment à la présence de causes communes cachées.

La relation de causalité $y$ cause $x$ peut être modélisée par une fonction $x = F(y)$ où la variable effet $x$ suit la variable cause $y$ selon la fonction $F$ (figure 3.2).

Selon le PCC, nous avons :

$$non(x \perp y) \equiv \left\{ \begin{array}{ll} x = f_1(y) & ou \\ y = f_2(x) & ou \\ x = f_1(C) \; et \; y = f_2(C) & ou \\ x = f_1(C) \; et \; y = f_2(x, C) & ou \\ x = f_1(y, C) \; et \; y = f_2(C) & \end{array} \right\} \tag{4}$$

avec $C$ des causes communes éventuellement cachées. Les 05 cas de (4) peuvent être réécrits

sous la forme

$$\begin{aligned} non(x \perp y) &\equiv x = F_1(H) \; et \; y = F_2(H) \\ &\equiv (x, y)^T = F(H) \end{aligned} \tag{5}$$

avec $H = C \bigcup \{x\}$ ou $H = C \bigcup \{y\}$ mais pas les deux au même temps. Par exemple, le cas 1 de (4) correspond à $H = \{y\}$ et $F = (F_1 : x = f_1(y), F_2 : y = y)$

Dans les RBs, on quantifie généralement les relations de dépendance statistique entre les variables observées à l'aide des probabilités conditionnelles. De la même manière, l'estimation du modèle ACI généralisé $(x_1, ..., x_n) = \psi(s_1, ..., s_m)$ ($\psi$ : une fonction multivariable de mélange quelconque et inversible) peut être vue comme une quantification semi-paramétrique des dépendances statistiques entre les variables observées. En effet, nous avons $x_i = \psi_i(s_1, ..., s_m)$ et $x_j = \psi_j(s_1, ..., s_m)$ avec $\psi_i$ et $\psi_j$ des composantes de la fonction $\psi$. On peut donc déduire $s_1$ en fonction de $x_j$ : $s_1 = \psi^{-1}_{j,s_1}(x_j, s_2, ..., s_m)$ où $\psi^{-1}_{j,s_1}$ désigne la fonction inverse. Le lien entre $x_i$ et $x_j$: $x_i = \psi_i(\psi^{-1}_{j,s_1}(x_j, s_2..., s_m), s_2, ..., s_m)$ est une relation *probabiliste* définie par la fonction de mélange $\psi$ et les $m - 1$ distributions des variables latentes $P(s_2), ..., P(s_m)$. Une composante indépendante $s_i$ peut correspondre à une cause physique cachée ou tout simplement simuler l'effet d'une ou plusieurs causes cachées.

Les deux modèles RB et ACI peuvent être utilisés pour la représentation de données multivariables. Les paramètres d'un RB comportent la structure DAG et les distributions de probabilités conditionnelles $P(x_i|Parents_{x_i})$ , alors que ceux de l'ACI comportent la fonction de mélange $\psi$ et les distributions de probabilités unidimensionnelles $P(s_i)$.

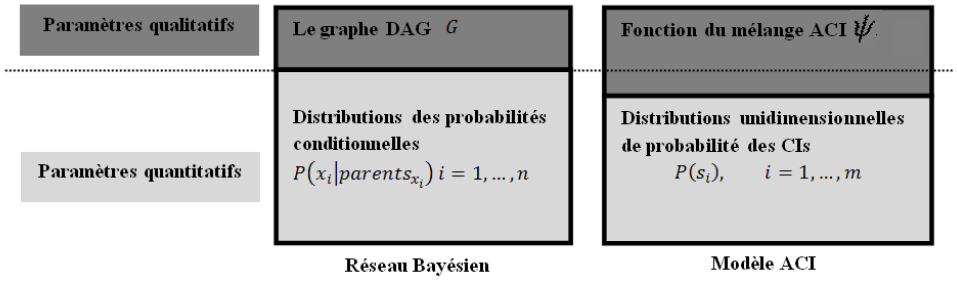



FIGURE 3.4 – Correspondance entre les paramètres qualitatifs-quantitatifs des modèles RB et ACI.

Comme le montre la figure 3.4, la fonction de mélange $\psi$ encode des informations qualitatives et une partie des informations quantitatives, et les densités de probabilités des CIs $P(s_i), i = 1, ..., m$ n'encodent que les informations quantitatives des relations de dépendance entre les variables observées.

## 5.2 Similitudes entre RB et ACI

Nous citons ci-dessous quelques points communs entre les RBs et l'ACI, les deux modèles
- peuvent être utilisés pour la représentation de données;
- visent à décomposer la distribution conjointe à des problèmes de petites dimensions : une factorisation en un produit des probabilités conditionnelles locales par les RBs et une fonction d'un produit des probabilités unidimensionnelles des CIs par l'ACI;
- sont basés sur l'estimation empirique de la dépendance statistique : dépendance statistique conditionnelle dans le cas des RBs et dépendance statistique mutuelle dans le cas de l'ACI;
- leur qualité d'estimation dépend à la fois de la quantité et de la représentativité de l'échantillon de données d'apprentissage et également de l'algorithme d'optimisation;
- ont un rapport fort avec les mécanismes causaux qui ont engendré les données observées;
- tentent d'exploiter les informations a priori. Les informations a priori sur les liens de causalité entre les variables observées, la présence des variables cachées, la nature de la distribution de probabilités des variables observées sont très utiles pour la convergence d'apprentissage des deux modèles.

## 5.3 Divergences entre RB et ACI

Le tableau 3.1 récapitule les principales différences entre les RBs et l'ACI.

## 5.4 Correspondance entre les structure de RB et de l'ACI linéaire

### 5.4.1 De la structure ACI vers la structure RB

Soit un ensemble de variables observées $X = \{x_1, ...x_n\}$. Ici, on considère seulement le cas où les dépendances entre les variables de $X$ peuvent être approximées par un mélange linéaire instantané de type $X = AS$ d'un ensemble de CIs $S = \{s_1, ..., s_m\}$ (Section II.2.1). L'algorithme proposé pour construire la structure DAG d'un RB à partir de la matrice de mélange ACI associé comporte les étapes suivantes :

1. estimation de la matrice du mélange $A$ à partir des données en utilisant un algorithme ACI approprié;
2. binarisation de la matrice de mélange ACI, les éléments nuls (ou avoisinant 0) sont fixés à 0, tandis que les autres éléments sont fixés à 1;
3. construction du graphe RB causal, $D_{XS}$, des variables observées plus les CIs;
4. identification, à partir de $D_{XS}$, les tests d'indépendance conditionnelle entre les variables observées $X$;

Tableau 3.1 – Divergences entre les modèles RB et ACI.

| **Critère de comparaison** | **RB** | **ACI** |
|---|---|---|
| Application principale | – Un RB est un modèle bien adapté pour dérouler la propagation de croyances (inférence statistique), et la classification supervisée | – BSS<br>– Extraction de paramètres<br>– Exploration causale |
| Type de données | – Variables discrètes (catégoriques) en utilisant des TPCs<br>– Variables continues Gaussiennes (modèle CLG) | – Variables continues et non Gaussiennes |
| Principe | – Dépendances conditionnelles entre variables observées | – Dépendance mutuelle |
| Type et degré de généralisation du modèle | – Modèle non paramétrique, ne nécessite pas des informations a priori sur le mécanisme causal (la nature de mélange) | – Modèle semi-paramétrique, l'optimisation est faite dans une famille de fonctions de mélange<br>– Nécessite des informations a priori sur le mélange pour assurer la convergence |
| | – Probabiliste | – Semi-probabiliste (semi-déterministe) en fonction des hypothèses |
| | – Plus général⇒ complexité en temps d'apprentissage et en espace de stockage plus élevée | – Restreint à des familles de mélanges ⇒ complexité en temps d'apprentissage et en espace moins élevée |
| | – Descriptif : comment les données observées sont liées. | – Génératif : comment les données observées sont générées |

5. construction de la structure de DAG équivalente, $D_X$, en utilisant un algorithme d'apprentissage de la structure des RBs à base de tests tels que l'algorithme PC.

La structure du graphe du RB équivalent dépend de la structure de la matrice $A$ qui est définie par le nombre et la position des valeurs nulles (ou presque nulles) ; une matrice $A$ plus parcimonieuse (sparse) correspond à un RB moins dense et vice versa. Les tests d'indépendance conditionnelle entre les variables observées à partir de $D_{XS}$ peuvent être déterminés par

le théorème suivant.

**Théorème 7 (*ACI vers RB*)**

*Soit $D_{XS}$ le DAG du RB causal comportant seulement les arcs $\{s_i \rightarrow x_j$ si $A_{ji} \neq 0\}$. Soient $x_j$ et $x_j$ deux variables observées, $T \subset X$ un sous-ensemble de variables observés et $P_T \subset S$ l'ensemble de parents directs de $T$ qui peuvent être déterminés étant donné $T$. $x_i$ et $x_j$ sont indépendantes conditionnellement à $T$* si et seulement si *tous les chemins entre $x_i$ et $x_j$ passent par $P_T$.*

Dans le premier exemple de la figure 3.5, nous avons $(x_2 \perp x_3|x_1)$. Si la valeur de $x_1$ est donnée, alors $s_1$ peut être déterminée en fonction de $x_1$ : $s_1 = A_{11}x_1$. Une fois la valeur de $s_1$ est connue, la propagation de l'information entre $x_3$ et $x_1$ est bloquée au niveau de $s_1$ car le seul chemin reliant $x_2$ et $x_3$ passe obligatoirement par $s_1$.

Dans le deuxième exemple, si la valeur de $x_1$ est connue alors ni $s_1$ ni $s_2$ peuvent être déterminées car nous avons deux inconnus dans l'équation $x_1 = A_{11}s_1 + A_{12}s_2$, nous avons donc $Not(x_3 \perp x_4|x_1)$ et $Not(x_3 \perp x_4|x_2)$. Cependant, l'instanciation simultanée de $x_1$ et $x_2$ permet de déterminer $s_1$ et $s_2$ : $[s_1, s_2]^T = [A_{11}, A_{12}; A_{21}, A_{22}]^{-1}[x_1, x_2]^T$. Par conséquent, la propagation de l'information sera bloquée entre $x_3$ et $x_4$ au niveau de $s_1$ et $s_2$ si les valeurs de $x_1$ et $x_2$ sont données et cela explique le constat $(x_3 \perp x_4|(x_1, x_2))$.

### 5.4.2 Exemples

La figure 3.5 présente quelques exemples illustrant la correspondance entre la structure d'un graphe d'un RB et la structure de la matrice de mélange ACI linéaire associée sous l'hypothèse que les données observées peuvent être approximées par un modèle ACI.

### 5.4.3 Le passage de la structure RB vers la structure ACI

Réciproquement, les tests d'indépendance conditionnelle peuvent être employés pour l'identification de la structure de la matrice ACI. Pour ce faire, il est possible de déduire la structure de la matrice ACI à partir de la structure du RB en suivant le parcours inverse de l'algorithme présenté dans la Section 5.4.1 :

1. Identification de tous les tests d'indépendance conditionnelle par une lecture directe des relations de d-séparation que le RB comporte (voir chapitre I).
2. Construction d'un graphe RB causal, $D_{XS}$ , défini sur les variables $X \bigcup S$ à partir des tests obtenus à l'étape 1. L'idée centrale ici consiste à déterminer un nombre minimal de variables latentes et indépendantes $S$ ainsi que la structure des arcs sortant de $S$ vers $X$ de telle sorte que le DAG obtenu soit en conformité avec l'ensemble de tests d'indépendance de l'étape 1 (voir les exemples affichés sur la figure 3.5 en les lisant de droite à gauche).

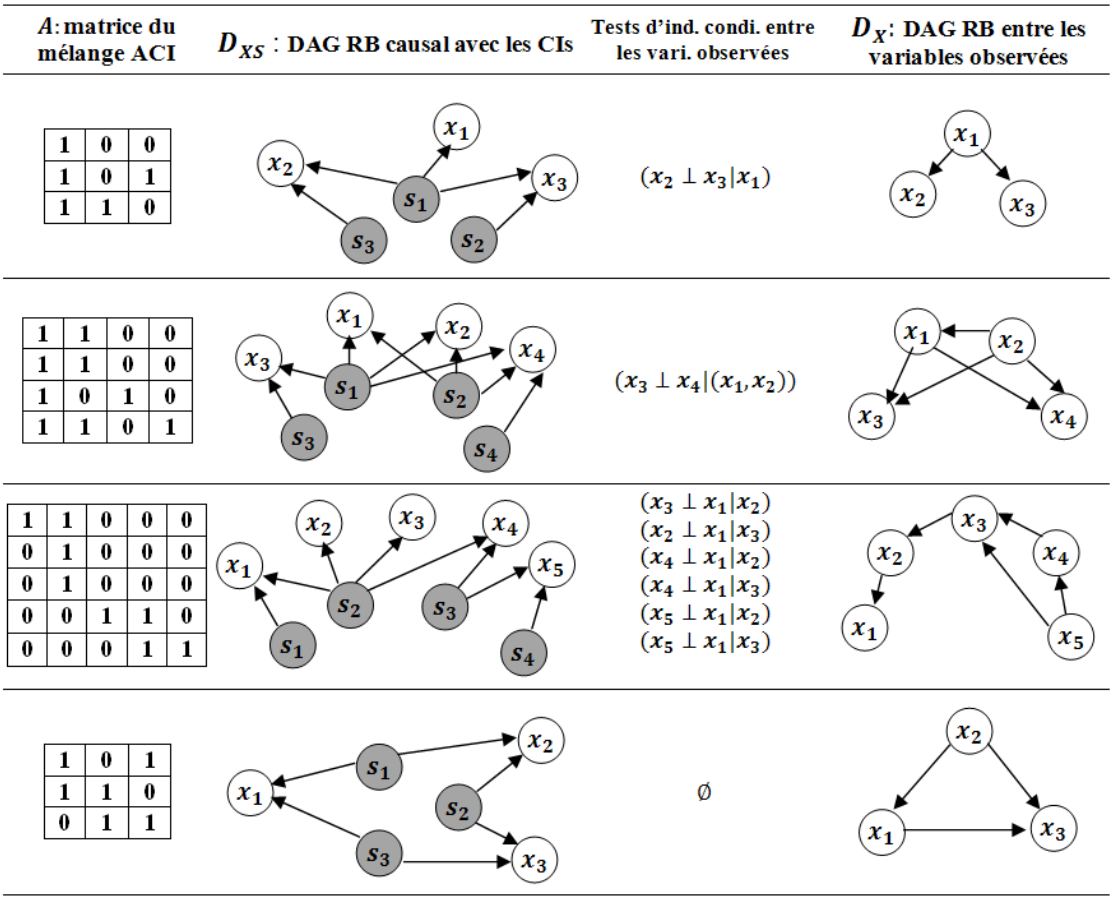



FIGURE 3.5 – Exemples illustrant la procédure de construction d'un DAG d'un RB à partir de la matrice de mélange ACI associée.

3. Déduire la structure de $A$ comme suit, $A(i, j)) = 1$ si le graphe causal contient l'arc $s_i \rightarrow x_j$ et $A(i, j)) = 0$ sinon.

Dans cette section, nous avons montré seulement l'équivalence possible entre la structure de la matrice de mélange ACI linéaire et le DAG d'un RB. Par ailleurs, il y a évidemment des relations à mettre en évidence entre les paramètres quantitatifs, entre la valeur des elements non nuls de la matrice ACI et les distributions de probabilité conditionnelle du RB équivalent. En conclusion, l'établissement d'une telle correspondance entre les paramètres des modèles RB et ACI nous permet de :

1. Se servir des algorithmes d'apprentissage des RBs pour estimer les paramètres du modèle ACI, ce qui nous permettra par exemple d'utiliser les tests d'indépendance conditionnelle pour résoudre le problème BSS.
2. Utiliser, selon le type de la fonction de mélange ACI (linéaire ou une famille particulière de fonctions non linéaires), les paramètres ACI pour quantifier les dépendances entre les variables observées dans un RB. Cela semble intéressant notamment dans le cas des

variables continues et non Gaussiennes car la prise en compte de ce type de données par les RBs reste un problème ouvert.

# 6 ACI, RB et modèles fonctionnels

Un modèle causal fonctionnel, connu sous le nom de 'Modèles d'Equations Structurelles' (MESs), dans sa forme générale consiste en un ensemble d'équations ayant la forme [Pearl, 2000]

$$x_i = f_i(Parents_i, \mu_i), i = 1, ..., n \tag{6}$$

où $Parents_i$ sont des causes directes de $x_i$ et la quantité $\mu_i$ représente un bruit (une erreur) dû aux facteurs oubliés. L'équation (6) est une généralisation non linéaire et non paramétrique des modèles d'équations structurelles linéaires ayant la forme

$$x_i = \sum_j \alpha_j x_j + \mu_i,\ i = 1, ..., n \text{ et } x_j \in Parents_i \tag{7}$$

devenus des outils standard d'analyse en sciences économiques et sociales.

Le but des MESs vise à avoir une représentation plus réduite et plus élégante (une fonction analytique) par la spécification de la manière dont les causes sont combinées pour produire l'effet. Les MESs expriment la nature précise des relations fonctionnelles représentant le mécanisme de production d'un effet, plutôt que d'utiliser les probabilités conditionnelles [Sloman, 2005].

## 6.1 Relation entre les MESs et les RBs

Si le diagramme causal du modèle (6) est acyclique (sans cycles), alors le modèle correspondant est dit *semi-markovien* et les valeurs des variables $X$ peuvent être déterminées seulement par les valeurs prises par les variables $U$. Dans ce cas de figure, la distribution conjointe $P(x_1, ..., x_n)$ est déterminée seulement par la distribution $P(u_1, ..., u_n)$ des variables erreurs. Si, en plus de l'accyclicité du diagramme, les erreurs $u_i$ sont mutuellement indépendantes, le modèle est dit *markovien* [Pearl, 2000]. L'indépendance mutuelle des variables $U$ est une propriété souhaitable et aide à la simplification du modèle (6).

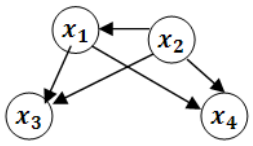
Les MESs sont préférables pour décrire des systèmes *partiellement déterministes*. Dans ce cas, la partie probabiliste est modélisée par les variables erreurs $U$ qui sont des variables latentes à estimer. Dans les MESs, l'estimation de la distribution des variables $U$ est une tâche délicate. A cet effet, des hypothèses portant sur les distributions de $U$ telles que la Gaussianité et leur indépendance mutuelle sont souvent émises pour des besoins de simplification. Les systèmes réels sont souvent complexes et pour lesquels l'effet des variables $U$ est important, les distributions des variables $U$ sont dépendantes et complexes et les relations fonctionnelles entre

les variables $X$ sont inconnues. Dans cette situation, l'augmentation de l'incertitude conduit à favoriser les modèles probabilités tels que les RBs plutôt que les modèles fonctionnels .

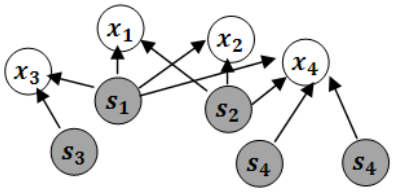
[Druzdzel and Simon, 1993] ont montré que, pour chaque réseau Bayésien $G$ caractérisé par une distribution $P$, il existe un modèle fonctionnel, ayant la forme (6), qui génère une distribution identique à $P$.

## 6.2 ACI, MESs et RBs

Soit $G$ le RB montré sur la figure 3.6(a). Supposons que les relations de dépendance entre les variables $X$ peuvent être approximées par le modèle ACI linéaire $(x_1, ..., x_4)^T = A(s_1, ..., s_5)^T$ dont la structure de la matrice de mélange est montrée sur la figure 3.6(b). Le graphe RB causal correspondant à $G$ avec l'incorporation des CIs est donné sur la figure 3.6(c).




| | | | | |
|---|---|---|---|---|
| $A_{11}$ | $A_{12}$ | 0 | 0 | 0 |
| $A_{21}$ | $A_{22}$ | 0 | 0 | 0 |
| $A_{31}$ | 0 | $A_{33}$ | 0 | 0 |
| $A_{41}$ | $A_{42}$ | 0 | $A_{44}$ | $A_{45}$ |




**(a)** Le RB

**(b)** La matrice $A$, estimée par l'ACI.

**(c)** Le RB causal avec les CIs

FIGURE 3.6 – Exemple d'un RB qui peut être approximé par un mélange ACI.

Dans cet exemple, les variables observées $x_1, ..., x_4$ sont des mélanges linéaires des CIs estimées $s_1, ..., s_5$ :

$$\begin{aligned} x_1 &= A_{11}s_1 + A_{12}s_2 \\ x_2 &= A_{21}s_1 + A_{22}s_2 \\ x_3 &= A_{31}s_1 + A_{33}s_3 \\ x_4 &= A_{41}s_1 + A_{42}s_2 + A_{44}s_4 + A_{45}s_5 \end{aligned} \tag{8}$$

Les relations de dépendances entre chaque variable $x_i$ et ses parents directs dans $G$ sont exprimées comme suit:

$$\begin{aligned} x_1 &= \alpha_1^{(1)}x_2 + \mu_1 && avec \quad \mu_1 = (A_{12} - (A_{11}A_{22})/A_{21})s_2 \\ x_2 &= \mu_2 && avec \quad \mu_2 = A_{21}s_1 + A_{22}s_2 \\ x_3 &= \alpha_1^{(3)}x_1 + \alpha_2^{(3)}x_2 + \mu_3 && avec \quad \mu_3 = A_{33}s_3 \\ x_4 &= \alpha_1^{(4)}x_1 + \alpha_2^{(4)}x_2 + \mu_4 && avec \quad \mu_4 = A_{44}s_4 + A_{45}s_5 \end{aligned} \tag{9}$$

où

$$\begin{matrix} \alpha_1^{(4)} = A_{41}W_{11} + A_{42}W_{21} \\ \alpha_2^{(4)} = A_{41}W_{12} + A_{42}W_{22} \end{matrix} \quad et \quad \begin{matrix} \alpha_1^{(3)} = A_{31}W_{11} \\ \alpha_2^{(3)} = A_{31}W_{12} \end{matrix} \quad et \quad \alpha_1^{(1)} = \tfrac{A_{11}}{A_{21}} \quad et \quad \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1}$$

On constate bien que l'exemple (9) représente un cas typique du MES linéaire spécifié en (7). Ici, nous avons fourni ce développement sur cet exemple dans un but d'illustration. En généralisant cela, il est facile de vérifier qu'un modèle ACI linéaire représente un cas spécial des MESs linéaires à deux niveaux d'abstraction dont la forme correspondante est la suivante

$$\begin{aligned} x_i &= \sum_j \alpha_j^{(i)} x_j + \mu_i \ , \ i = 1, ..., n \ et \ x_j \in Parents_{x_i} \\ \mu_i &= \sum_j \beta_j^{(i)} s_j \ , i = 1, ..., n \ et \ s_j \ (j = 1, ..., m) \ sont \ des \ CIs \end{aligned} \tag{10}$$

Les coefficients $\alpha_j^{(i)}$ et $\beta_j^{(i)}$ sont des constantes réelles. Les variables $U$ sont des mélanges linéaires des CIs et elles peuvent présenter des dépendances mutuelles. La généralisation de (10) au modèle ACI non linéaire est

$$\begin{aligned} x_i &= f_i(Parents_{x_i}, \mu_i) \ , \ i = 1, ..., n \\ \mu_i &= g_i(CIs) \ , \ i = 1, ..., n \end{aligned} \tag{11}$$

Par rapport aux modèles MESs ( linéaire (7) et non linéaire (6)), les modèles basés sur l'ACI (linéaire (10) et non linaire ( 11)) ont l'avantage de ne pas imposer de contraintes restrictives sur les variables $U$ souvent considérées dans les MESs comme la markoviannite (indépendance mutuelle) et la Gaussianité des variables $U$. Le modèle proposé dans (10) et (11) donne plus d'importance à la composante non déterministe de la représentation dans la mesure où le nombre de CIs peut être variable (pas toujours égal à $n$) et les variables erreurs $U$ correspondantes peuvent avoir des distributions quelconques. Il permet aussi de se servir d'un bon nombre d'algorithmes ACI disponibles qui suivent des approches différentes et dédiés chacun à des fonctions de mélange spécifiques.

Le tableau 3.2 montre une comparaison entre les modèles RB, MES et ACI selon différents critères.

Tableau 3.2 – Comparaison entre les modèles RB, MES et ACI.

| **Critère** | **RB** | **MES** | **ACI** |
|---|---|---|---|
| Paramètres | DAG+Probabilités conditionnelles | Fonctions analytiques + $P(U)$ | Fonctions analytiques + $P(s_1), ..., P(s_m)$ |
| Déterminisme du modèle | Fortement Probabiliste | Semi-probabiliste | Semi-probabiliste |
| Taille de représentation | Large | Moyenne | Compacte (petite) |
| Capacité de représentation | Plus générale | Restrictive (hypothèse sur les fonctions de causalité) | Restrictive (hypothèse sur les fonctions de mélange) |
| Type de modèle | Nom-paramétrique | Semi-paramétrique | Semi-paramétrique |
| Exploration causale | Relations entre les variables observées | $U$: des variables latentes souvent *hypothétiques* (non physiques) | $S$: variables latentes mutuellement indépendantes et souvent représentent des *causes physiques* |

# 7 Conclusion

Dans ce chapitre, nous avons essayé de fournir une analyse théorique approfondie dans l'objectif de déceler les aspects communs ainsi que les différences essentiels entre les modèles RB et ACI. Fondamentalement, les RBs sont conçus pour effectuer le raisonnement probabiliste (propagation de croyances) alors que l'ACI s'est montrée utile dans les problèmes de BSS. Néanmoins, les deux modèles peuvent être employés pour la représentation de données multivariables.

Au début de ce chapitre, nous avons accordé une attention particulière à l'interprétation causale des RBs. Les relations d'indépendance conditionnelle qu'exprime un RB sont la partie visible résultante d'un mécanisme causal pas souvent perceptible. La dépendance est réellement une fonction déterministe, et la dépendance probabiliste n'est qu'une conséquence directe des causes communes cachées et, à un degré moindre, du non déterminisme dû aux méthodes non paramétriques (discrétisation) utilisées pour estimer les relations entre des variables continues.

En deuxième lieu, nous avons essayé de faire un rapprochement entre le concept de RBs et celui de l'ACI autour de la notion de causalité. L'équivalence, entre les paramètres qualitatifs et quantitatifs des RBs d'une part et les fonctions de mélange et les densité de probabilité des CIs de l'ACI d'autre part, est mise en évidence. Dans ce contexte, un algorithme de détermination du graphe d'un RB équivalent à partir de la matrice de mélange ACI linéaire est proposé. Dans la dernière section, nous avons montré que le modèle ACI est un cas typique des modèles fonctionnelles MESs. En outre, une comparaison entre les modèles RBs, ACI et MESs est établie selon plusieurs critères, à savoir les propriétés de leurs paramètres, leur capacité de généralisation et leur type d'apprentissage.

L'étude menée dans ce chapitre nous a permis de mettre en lumière quelques nuances et de faire ressortir quelques éléments d'analogie entre les RBs et l'ACI. Notre but est de faire apparaître le mieux possible les éléments de complémentarités possibles faisant l'objet de nos contributions dans les chapitres qui suivent.

Chapitre 4

# Analyse en composantes indépendantes et réseaux Bayésiens pour la représentation de données

## Sommaire


**1 Introduction . . . . . 79**

**2 ACI pour la représentation de données . . . . . 80**

2.1 Estimation de densité par l'ACI linéaire . . . . . 80

2.2 Estimation de densité par l'ACI non linéaire . . . . . 81

2.2.1 ACI non linéaire pour l'estimation de densité . . . . . 81

2.2.2 ACI locale . . . . . 82

2.2.3 Estimation de densité par l'ACI locale . . . . . 82

2.2.4 Validité de clustering . . . . . 84

2.3 Amélioration de la classification des vecteurs MFCCs . . . . . 84

2.3.1 Le clustering pour la reconnaissance du locuteur indépendante de texte . . . . . 84

2.3.2 Le modèle de monde en RAL indépendante de texte . . . . 85

2.3.3 ACI pour l'analyse de la voix . . . . . 86

2.3.4 Notre proposition . . . . . 87

2.4 Reconnaissance du locuteur par la fusion de la voix et du mouvement des lèvres . . . . . 94

2.4.1 Algorithmes de clustering flou . . . . . 94

2.4.2 Système de classification proposé . . . . . 96

2.4.3 Préparation de données . . . . . 98

2.4.4 Résultats et discussions . . . . . 101

**3 ACI locale et RB pour la représentation de données : application en estimation de densité de données de hautes dimensions . . . . . 104**

3.1 ACI locale, hautes dimensions et sparsity . . . . . 104

3.1.1 ACI et hautes dimensions . . . . . 104

3.1.2 Le clustering et les hautes dimensions . . . . . 105
3.1.3 L'ACI et la sparsity . . . . . 106
3.2 Réseaux Bayésiens et ACI pour l'estimation de densité . . . . . 108
3.2.1 Apprentissage de structure . . . . . 108
3.2.2 Arbre de jonction pour la factorisation . . . . . 109
3.2.3 Estimation de densité en utilisant ACI locale et RBs . . . . 110
3.3 Résultats expérimentaux . . . . . 110
3.3.1 Application en estimation de densité . . . . . 110
3.3.2 Application en classification supervisée . . . . . 112
**4 Conclusion . . . . . 116**

# 1 Introduction

Dans le chapitre précédent, nous avons mis en évidence une équivalence envisageable entre les paramètres qualitatifs et quantitatifs des deux modèles étudiés, RB et ACI. Le modèle ACI est un cas typique des modèles d'équations structurelles, qui a à la fois l'avantage d'être caractérisé par des facteurs statistiquement indépendants (des CIs) ce qui représente une propriété convoitée pour des tâches postérieures s'appliquant sur le modèle et aussi l'avantage de disponibilité d'un bon nombre d'algorithmes ACI.

Les RBs sont conçus particulièrement pour accomplir la propagation de croyances par le mécanisme d'inférence Bayésienne, alors que l'ACI est utilisée principalement pour l'estimation aveugle de sources cachées. Toutefois, les deux modèles peuvent être utilisés pour la représentation de données multivariables.

Les algorithmes exacts d'inférence Bayésienne sont donnés seulement pour le cas des RBs discrets où les TPCs sont utilisées comme des paramètres et aussi pour le cas des variables continues Gaussiennes. Un RB discret à base de TPCs est un modèle non paramétrique qui peut être utilisé pour représenter des données multivariables générales qui suivent des distributions quelconques. Les algorithmes d'inférence exacte demeurent pratiques seulement pour les RBs les moins complexes car la complexité en espace mémoire du modèle (volumes des TPCs) et la complexité en espace mémoire et en temps des algorithmes d'inférence augmentent exponentiellement avec la taille du graphe RB (nombre de noeuds et densité du graphe) et avec la résolution de discrétisation (nombre de valeurs discrètes possibles des variables).

En revanche, l'ACI est un modèle (semi-)paramétrique nécessitant que les données suivent les fonctions de mélange considérées. En effet, le modèle ACI fournit une représentation plus petite et plus compacte de données; ses paramètres comportent les fonctions de mélange et les distributions unidimensionnelles des CIs. Il peut être utilisé pour représenter des données continues et non Gaussiennes à condition que les données s'ajustent aux fonctions de mélange.

Les modèles RB et ACI sont complémentaires dans le sens où l'ACI peut réussir sur certains types de données qui représentent un handicap pour les RBs. Dans cette thèse, nous avons préféré de focaliser nos contributions sur l'extension des RBs aux données continues et non Gaussiennes par la proposition de nouveaux algorithmes de propagation de croyances s'exécutant sur les modèles ACI. Pour atteindre cet objectif, il est bien nécessaire de confirmer d'abord le fait que le modèle ACI dispose d'une bonne qualité d'ajustement de certains types de données du monde réel.

Nous avons consacré ce chapitre à l'exploration des performances de méthodes ACI sur des données réelles de natures hétérogènes, à savoir : acoustiques, sismiques, images, etc. Les performances ont été évaluées par le biais de l'erreur empirique de l'estimation de densité et de la classification supervisée.

# 2 ACI pour la représentation de données

## 2.1 Estimation de densité par l'ACI linéaire

Soit $X = (x_1, x_2, ..., x_n)^T$ un vecteur aléatoire multivariable de dimension $n$ dont les variables $x_i (i = 1, ..., n)$ sont continues et non Gaussiennes. Dans un premier temps, nous considérons seulement le cas simple où la variable $X$ peut être approximée par un modèle ACI linéaire et instantané, $X^T = A \times S^T$, avec $A$ une matrice carrée et inversible et $S = (s_1, s_2, ..., s_n)^T$ un vecteur de CIs.

Comme la transformation $X^T = A \times S^T$ est continue, admet des dérivées partielles et définit un mapping de type un-à-un, alors la distribution conjointe de densité de $X$ peut être simplifiée en se servant du théorème célèbre de changement de variables, ce qui donne

$$P_X(X) = P_X(x_1, x_2, ..., x_n) = \frac{1}{|det(Jac(X))|} P_S(A^{-1} \times X) \tag{1}$$

où le Jacobien $det(Jac(X)) = det(Jac(A \times S))$ est le déterminant de la matrice Jacobienne de la transformation, définie comme suit:

$$Jac(X) = \begin{pmatrix} \frac{\partial(A_1 \times S)}{\partial(s_1)} & \cdots & \frac{\partial(A_1 \times S)}{\partial(s_n)} \\ \vdots & \ddots & \vdots \\ \frac{\partial(A_n \times S)}{\partial(s_1)} & \cdots & \frac{\partial(A_n \times S)}{\partial(s_n)} \end{pmatrix} \tag{2}$$

avec $A_i$ la ligne d'indice $i$ de $A$. Cela nous donne

$$P_X(X) = \frac{1}{|det(A)|} P_S(A^{-1} \times X) \tag{3}$$

En exploitant l'hypothèse de l'indépendance mutuelle des CIs $s_1, s_2, ..., s_n$, la conjointe $P_S(S) = P_S(A^{-1} \times X)$ peut être factorisée en un produit. Ainsi, on aura

$$P_X(X) = \frac{1}{|det(A)|} \prod_{i=1}^{n} P_{s_i}(s_i) \tag{4}$$

Il est opportun de noter que dans (4), et après l'estimation de la matrice de mélange $A$, l'estimation de densité $P_X(X)$ est substantiellement simplifiée vue qu'elle nécessitera seulement l'estimation de densité unidimensionnelle de chaque composante $s_i$ séparément par des estimateurs standard tels que les *Fenêtres de Parzen*.

La formule (4) peut être aisément étendue aux cas de mélanges sur-déterminés (bases sous-complètes) lorsque le nombre des CIs est inférieur à celui de variables observées ($n > m$). Si on prend le mélange sur-déterminé $X^T = A \times S^T$ , composé de $N$ vecteurs, suivant :

$$\begin{pmatrix} x_1(1) & \dots & x_1(N) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ x_n(1) & \dots & x_n(N) \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} \times \begin{pmatrix} s_1(1) & \dots & s_1(N) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ s_m(1) & \dots & s_m(N) \end{pmatrix}$$

Puis, nous ajoutons des composantes nulles $s_{m+1}, ..., s_n$ et une matrice $A^{n,n-m}$ de dimension $n \times (n-m)$ à $A$ de telle sorte que la matrice carrée de concaténation $A^{n,n} = [A, A^{n,n-m}]$ soit inversible. Les éléments ajoutés sont montrés en gras sur l'exemple suivant

$$\begin{pmatrix} x_1(1) & \dots & x_1(N) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ x_n(1) & \dots & x_n(N) \end{pmatrix} = \begin{pmatrix} a_{1,1} & \dots & a_{1,m} & \boldsymbol{a}_{1,m+1} & \dots & \boldsymbol{a}_{1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,m} & \boldsymbol{a}_{n,m+1} & \dots & \boldsymbol{a}_{n,n} \end{pmatrix} \times \begin{pmatrix} s_1(1) & \dots & s_1(N) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ s_m(1) & \dots & s_m(N) \\ \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} \end{pmatrix} \quad (5)$$

Il est possible de vérifier que toutes les composantes, y compris les composantes ajoutées $s_{m+1}, ..., s_n$, sont mutuellement indépendantes. De plus, nous avons pour tout $i = m+1, ..., n$, la densité $P_{S_i}(s_i) = 1$ si $s_i = W^i \times X = 0$ et $P_{S_i}(s_i) = 0$ ailleurs ($W^i$ : la ligne d'indice $i$ de $(A^{n,n})^{-1}$). Donc, on peut se servir de (4) pour estimer la conjointe $P_X(X)$ après avoir effectué l'adaptation (5).

Les mélanges déterminés ($m \leq n$) couvrent un nombre important de phénomènes réels. La formule (4) ne peut pas être utilisée pour le cas délicat de mélanges sous-déterminés (lorsque $m > n$). Dans ce cas, la conjointe $P_X(X)$ peut être calculée à l'aide d'un produit de convolution ; géométriquement, cela revient à faire une marginalisation (intégration) de la conjointe $P_S(S)$ sur l'intersection des $n$ hyperplans des variables observées dans l'espace des CIs.

## 2.2 Estimation de densité par l'ACI non linéaire

### 2.2.1 ACI non linéaire pour l'estimation de densité

L'ACI non linéaire est un problème mal posé dans la mesure où le mélange non linéaire peut être n'importe quelle fonction, il admet une infinité de solutions [Hyvärinen and Pajunen, 1999]. Pour cette raison, on trouve en pratique des méthodes ACI non linéaires plus ou moins restrictives, qui sont développés pour des types spécifiques de mélange tels que les mélanges post-non linéaires (PNLs) et les mélanges convolutifs (voir §II. 8), on parle de mélanges paramétriques. Les mélanges ACI non linéaires paramétriques ne conviennent pas à notre objectif car ils sont jugés trop restrictifs; ils ne peuvent pas approximer des distributions générales de

données. En contrepartie, les méthodes ACI non linéaires non paramétriques basées sur le clustering sont préférables grâce à leur capacité de généralisation. La méthode ACI locale proposée par [Karhunen and Malaroiu, 1999] en constitue un exemple typique. Dès lors, nous avons opté pour l'utiliser dans la conception d'un estimateur non linéaire de densité.

### 2.2.2 ACI locale

L'ACI locale est une méthode non linéaire qui cherche une partition des vecteurs de données observées en un ensemble de clusters de telle sorte que les clusters obtenus disposent d'une vraisemblance élevée d'être engendrés par des mélanges ACI linéaires. Dans laquelle, la partie du clustering fournit une représentation globale et non linéaire de données, alors que les modèles ACI linéaires sont utiles pour décrire localement les attributs [Karhunen and Malaroiu, 1999], voir l'exemple de la figure 4.1.

Le clustering est responsable de la prise en compte de l'aspect non linéaire et il est donc crucial pour avoir un bon ajustement. Il convient de souligner le fait que le degré de la non unicité des solutions ACI locale est de la même envergure que les autres approches non linéaires [Karhunen and Malaroiu, 1999]. En effet, le clustering dépend des paramètres de l'initialisation ce qui peut donner une solution différente à chaque exécution.

Le critère de maximisation de la non Gaussianité a été utilisée par l'algorithme populaire 'FastICA' [Hyvärinen et al., 2001] dont le principe s'appuie sur le théorème de la limite centrale : la distribution de chaque variable observée $x_i$ $(i = 1, ..., n)$ est plus proche d'une Gaussienne que les CIs. Cependant, il n y pas une garantie que la conjointe $P(x_1, ..., x_n)$ converge vers une Gaussienne multivariable car les marginales $x_i$ ne sont pas mutuellement indépendantes.

La forme (ou les formes) géométrique exacte que la distribution $P(x_1, ..., x_n)$ tend à avoir est un sujet délicat qui reste à être étudié de près. Réellement, nous voulons connaître quelles sont les formes géométriques des clusters à partir desquels un algorithme ACI linéaire spécifique peut identifier des CIs maximisant l'indépendance mutuelle.

En pratique, les vecteurs de données observés sont compacts et souvent prennent une forme sphérique, elliptique ou linéaire (elliptique étirée). A titre d'exemple, Karhunen et al. ont appliqué l'ACI locale sur des clusters sphériques détectés par l'algorithme K-Means [Karhunen and Malaroiu, 1999] et Honda et al. ont l'appliquée sur des clusters linéaires détectés par l'algorithme *fuzzy c-variétés* [Honda and Ichihashi, 2006].

### 2.2.3 Estimation de densité par l'ACI locale

L'algorithme proposé d'estimation de densité, basé sur l'ACI locale, comporte les étapes suivantes :

1. Grouper les vecteurs de données observés en $L$ clusters homogènes par un algorithme de clustering approprié;

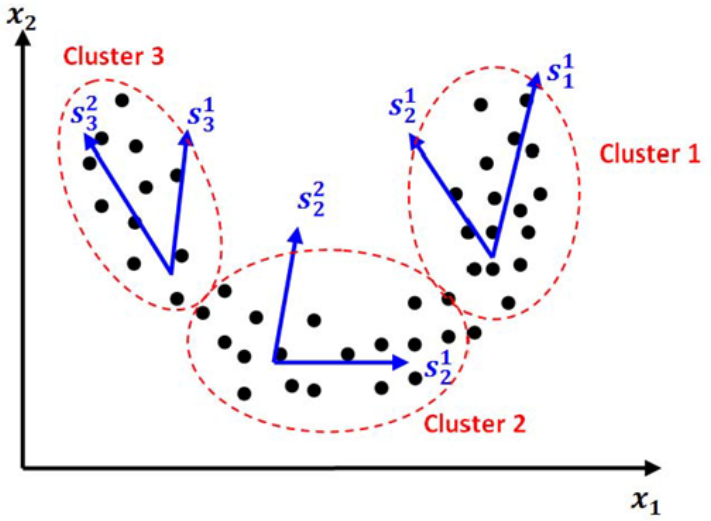



FIGURE 4.1 – Exemple en dimension 2 de l'ACI locale à base de clustering.

2. soustraire le vecteur moyen de chaque cluster des vecteurs du même cluster ;
3. estimer la matrice de mélange $A$ de chaque cluster centré en utilisant un algorithme ACI approprié;
4. estimer la distribution unidimensionnelle $P_{S_i}$ de chaque composante indépendante $s_i$ en utilisant des méthodes standard telles que l'estimateur de Parzen ou les GMMs à base de l'algorithme EM (*Expectation- Maximization*).

L'estimation de la conjointe $P_X(X)$ peut être alors accomplie via une simple marginalisation sur les clusters détectés comme suit :

$$P_X(X) = \sum_{j=1}^{L} P(X, C_j) \tag{6}$$

En utilisant la règle de Bayes, nous pouvons simplifier (6) pour avoir

$$P_X(X) = \sum_{j=1}^{L} P_C(C_j) \times P_X(X|C_j) \tag{7}$$

avec $P_C(C_j)$ $j = 1, ..., L$ la probabilité a priori du cluster $C_j$ et $P_X(X|C_j)$ la vraisemblance de $X$ dans le cluster $C_j$. Pour le cas déterminé ($m \leq n$), la distribution de densité conjointe $P_X(X)$ au niveau du cluster $C_j$ ($j = 1, ..., L$) en utilisant (4) est

$$P_X(X|C_j) = \left(\frac{1}{|\mathtt{det}(A)^{(j)}|}\right) \times \prod_{i=1}^{m} P_{s_i^{(j)}}(s_i^{(j)}) \tag{8}$$

$A^{(j)}$ et $S_i^{(j)} i = 1, ..., m$ représentent la matrice de mélange et les CIs qui sont estimées à partir des vecteurs du cluster $C_j$. En substituant (8) dans (7), nous pouvons déduire $P_X(X)$

$$P_X(X) = \sum_{j=1}^{L} P_C(C_j) \times (\frac{1}{|det(A)^{(j)}|}) \times \prod_{i=1}^{m} P_{s_i^{(j)}}(s_i^{(j)}) \tag{9}$$

### 2.2.4 Validité de clustering

Pour une topologie donnée, les résultats de clustering dépendent de l'initialisation ainsi que du nombre de clusters $L$. L'identification d'une *bonne partition* est une condition nécessaire pour avoir un bon modèle de représentation basé sur l'ACI locale. Trouver une partition acceptable peut être réalisé empiriquement par l'exécution de l'algorithme de clustering plusieurs fois en variant $L$ et les paramètres de l'initialisation, ensuite la meilleure partition selon une mesure de validité adéquate est retenue.

L'estimation du nombre optimal de clusters $L$ est connue d'être un problème difficile, car une telle validation nécessite de faire des comparaisons entre des partitions hétérogènes, composées chacune d'un nombre différents de clusters. La qualité du partitionnement en termes de rapport compacité/séparation s'améliore certainement avec l'augmentation du nombre de clusters. Pour avoir des comparaisons équitables, les mesures de comparaison (de validité) doivent *pénaliser correctement* les partitions composées d'un grand nombre de clusters. Dans notre travail, nous avons adopté le critère Xie et Beni (XB) proposé dans [Xie and Beni, 1991] qui mesure le rapport de la compacité globale des clusters à leur qualité de séparation de la manière suivante

$$XB = \frac{\sum_{i=1}^{L} \sum_{j=1}^{N} u_{ij} \parallel x_j - v_i \parallel^2}{N min_{i \neq j:1...L} \parallel v_j - v_i \parallel^2} \tag{10}$$

où $u_{ij} = 1$ si le vecteur d'indice $j$ appartient au cluster $C_i$ et $u_{ij} = 0$ sinon, $v_i$ est le centre du cluster $C_i$, $N$ est le nombre de vecteurs et $\parallel * \parallel$ est la mesure de distance entre les vecteurs. Ce critère est plutôt invariant de $L$ et permet donc de comparer des partitions qui n'ont pas nécessairement le même nombre de clusters.

## 2.3 Amélioration de la classification des vecteurs MFCCs

### 2.3.1 Le clustering pour la reconnaissance du locuteur indépendante de texte

Le modèle de mélanges Gaussiens (GMM) est incontestablement le modèle le plus utilisé dans la reconnaissance automatique du locuteur (RAL) en mode indépendant de texte (là où il n'y a pas de contraintes sur le contenu prononcé). Un GMM est un modèle probabiliste et génératif qui peut servir d'approximation des distributions multivariables complexes. L'utilisation avec succès des GMMs pour la RAL indépendante de texte est initialement décrite dans

[Reynolds, 1992]. Une extension des systèmes à base de GMMs pour la RAL est donnée dans [Reynolds, 1995]. Le système GMM-UBM basé sur un modèle de monde universel (universal Background Model) pour la modélisation de la représentation alternative du locuteur est illustré dans [Reynolds et al., 2000]. Il est opportun de noter que la majorité des variantes de système GMM-UBM ont donné des meilleures performances dans les évaluations NIST [3] de la RAL.

En RAL indépendante de texte, il n'y a pas de contraintes fortes sur le contenu de la parole d'où les attributs disposent d'une variance intra-classe importante. En effet, les vecteurs cepstraux (utilisés souvent pour caractériser la voix tels que les MFCCs) d'un locuteur sont dispersés sur un ensemble de nuages (clusters) éparpillés. La performance des GMMs dans un tel contexte est justifiée en partie par le fait qu'ils effectuent une sorte du clustering flou ; dans lequel, les vecteurs cepstraux sont regroupés en un ensemble de clusters chevauchés et ayant chacun la forme d'une Gaussienne multivariable.

Le modèle VQ (Vector Quantification) est également l'un des modèles à la fois simple et compétitif en RAL indépendante de texte. Le premier usage de VQ remonte aux années 1980s [Burton, 1987]. Soit $T = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_T})^T$ et $R = (\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_R})^T$ les vecteurs d'attributs (e.g MFCCs) de la phrase de test et les phrases de référence, respectivement. La distorsion (distance) moyenne VQ entre $T$ et $R$ est calculée comme suit :

$$D_{VQ}(T,R) = \sum_{i=1}^{N_T} min_{1 \leq j \leq N_R} d(\mathbf{x}_i, \mathbf{r}_j) \tag{11}$$

où $d(.,.)$ est une mesure de similitude entre vecteurs telle que la distance Euclidienne. Une quantité $D_{VQ}(T,R)$ assez petite est synonyme d'un haut score d'appariement. Comme le nombre de distances élémentaires à évaluer, $N_R \times N_T$, est trop important en pratique, le nombre de vecteurs dans $R$ est souvent réduit à l'aide d'un algorithme de clustering tel que K-Means [Linde et al., 1980]; les distances sont évaluées seulement entre les vecteurs $T$ et l'ensemble de vecteurs centraux des clusters de $R$ connus sous le nom de 'codebooks'.

### 2.3.2 Le modèle de monde en RAL indépendante de texte

La RAL consiste réellement en un problème de la classification supervisée. Établir la décision en utilisant uniquement des modèles génératifs à l'instar de GMM ou VQ n'est pas adéquat, parce que les vraisemblances (ou les scores d'appariement) estimées des vecteurs de la phrase à tester sont souvent biaisées ou non normalisées, ce qui dégrade les performances. A cet effet, les modèles GMM et VQ se sont montrés plus compétitifs lorsqu'ils sont combinés avec une adaptation des vraisemblances à une référence, par exemple par l'ajout d'un modèle de monde (voir le modèle GMM-UBM [Reynolds et al., 2000] et le VQ avec un modèle de monde [Hautamäki et al., 2008]).

3. http://www.nist.gov/itl/iad/mig/sre.cfm

A titre d'exemple, dans le cas GMM-UBM, le score d'un vecteur de test $\mathbf{x}$ est exprimé par le rapport du logarithme de la vraisemblance du modèle du locuteur proclamé $\lambda$ et celle du modèle de monde $\lambda_{UBM}$ :

$$\begin{aligned} score(\mathbf{x}) &= log(P(\mathbf{x}|\lambda)) - log(P(\mathbf{x}|\lambda_{UBM}) \\ &= log(GMM_{\lambda}(\mathbf{x})) - log(GMM_{\lambda_{UBM}}(\mathbf{x})) \end{aligned} \tag{12}$$

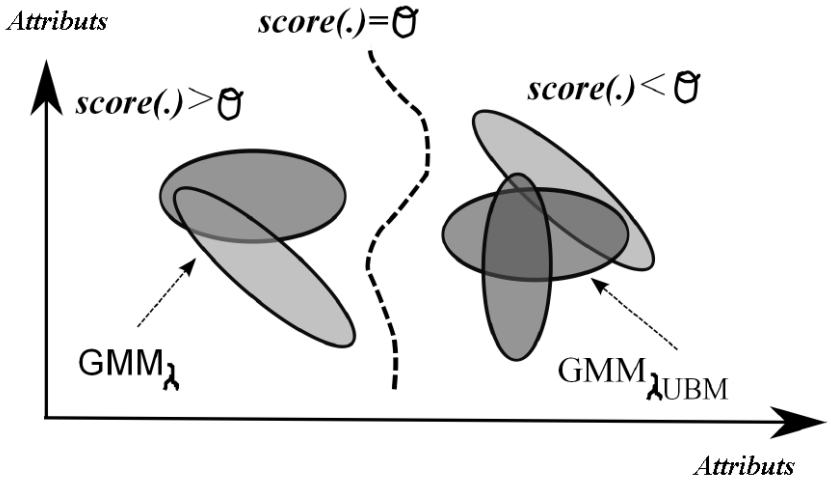



FIGURE 4.2 – Adaptation GMM par l'ajout d'un modèle de monde est vue comme une classification binaire discriminative.

La classe proclamée, $\lambda$, est à gauche et la classe alternative, $'non\ \lambda'$, est à droite de la frontière non linéaire $score(\mathbf{x}) = \theta$ .

Le processus d'adaptation par l'ajout d'un modèle de monde représente réellement une classification binaire supervisée entre deux classes, à savoir : la classe du modèle du locuteur proclamé et la classe du modèle de monde (hypothèse alternative). L'hypothèse que $\mathbf{x}$ soit généré par $\lambda$ et pas par $\lambda_{UBM}$ est acceptée si $score(\mathbf{x})$ est supérieur à un seuil de décision $\theta$ déterminé empiriquement.

$$Classe(\mathbf{x}) = \left\{ \begin{array}{ll} Locuteur\ \lambda & si\ score(\mathbf{x}) \geq \theta \\ NonLocuteur\ \lambda & ailleurs \end{array} \right\} \tag{13}$$

Géométriquement, la frontière séparant les deux classes dans le cas GMM-UBM est non linéaire et son équation est $log(GMM_{\lambda}(\mathbf{x})) - log(GMM_{\lambda_{UBM}}(\mathbf{x}) = \theta$ comme le montre la figure 4.2.

### 2.3.3 ACI pour l'analyse de la voix

Récemment, plusieurs travaux impliquant l'ACI pour la reconnaissance de la parole ou du locuteur ont vu le jour. L'ACI est dans l'ensemble appliquée pour l'extraction des attributs et

pour le filtrage du bruit. En théorie, les classificateurs donnent des bonnes performances avec des attributs indépendants qu'avec ceux qui sont mutuellement corrélés ou dépendants. Dans ce sens, l'algorithme VQ a été appliqué sur des CIs pour la vérification du locuteur [Sentürk and Gürgen, 2006]. Une représentation sur-complète de données (*'overcomplete'*: le nombre des CIs > nombre des variables observées) a été proposée dans [Bai et al., 2007] pour l'extraction des fonctions de base caractérisant le signal vocal pour la vérification du locuteur. Les fonctions de base consistent en colonnes de la matrice du mélange dans le cas de l'ACI linéaire (vecteurs de base) et elles sont utilisées comme des attributs de la classification. Les représentations sur-complètes ont l'avantage d'être moins sensibles au bruit. Dans [Jang, 2001] des fonctions de base estimées localement sur des segments du signal vocal sont utilisées comme des attributs pour la vérification du locuteur.

L'exemple célèbre du '*cocktail-party*' [Hyvärinen et al., 2001] est traditionnellement employé pour introduire la définition de l'ACI. En suivant le même principe, l'ACI est souvent utilisée pour filtrer le signal acoustique corrompu par le bruit environnant où l'idée clé consiste à identifier les CIs relatives au bruit et les CIs non discriminantes afin de les enlever, les CIs pertinentes ou leurs fonctions de base sont retenues pour l'analyse.

### 2.3.4 Notre proposition

Dans cette section, la méthode ACI locale [Karhunen and Malaroiu, 1999] à base de clustering, combinée avec le classificateur SVM a été utilisée pour améliorer la classification des vecteurs MFCCs [Reynolds, 1992] qui sont des paramètres fréquentiels souvent utilisés pour caractériser la voix en RAL indépendante de texte. Dans ce qui suit, nous exposons le modèle proposé, les algorithmes d'apprentissage et de test ainsi que les résultats obtenus.

#### a) Algorithme d'apprentissage

Le pseudo-algorithme d'apprentissage du modèle proposé d'un locuteur $s$ consiste en les étapes suivantes :

1. Regroupement des vecteurs MFCCs du locuteur $s$ en $k^s$ clusters homogènes ;
2. Association d'un modèle ACI linéaire à chaque cluster $j$ : $1...k^s$ de $s$ et estimation de sa matrice de séparation ACI, $W_j^s$, en utilisant un algorithme ACI convenable;
3. Projection des vecteurs d'apprentissage MFCCs du locuteur $s$ et aussi ceux du modèle de monde vers l'espace sous-jacent des CIs, $\mathbf{y} \leftarrow W_j^s \times \mathbf{x}$ ;
4. Attribution de l'étiquette '**1**' aux vecteurs projetés du locuteur $s$ et l'étiquette '**2**' aux vecteurs projetés du modèle de monde;
5. Apprentissage d'un classificateur discriminatif tel que SVM entre les deux classes '**1**' et '**2**'.

Les étapes 2,...,5 sont effectuées pour chaque cluster $j : 1, ..., k^s$ de $s$ et l'algorithme entier, de l'étape 1 à 5, est exécuté pour chaque locuteur enregistré $s : 1, ..., \#Locuteurs$. Un schéma illustrant la phase d'apprentissage est affiché sur la figure 2.3.4.

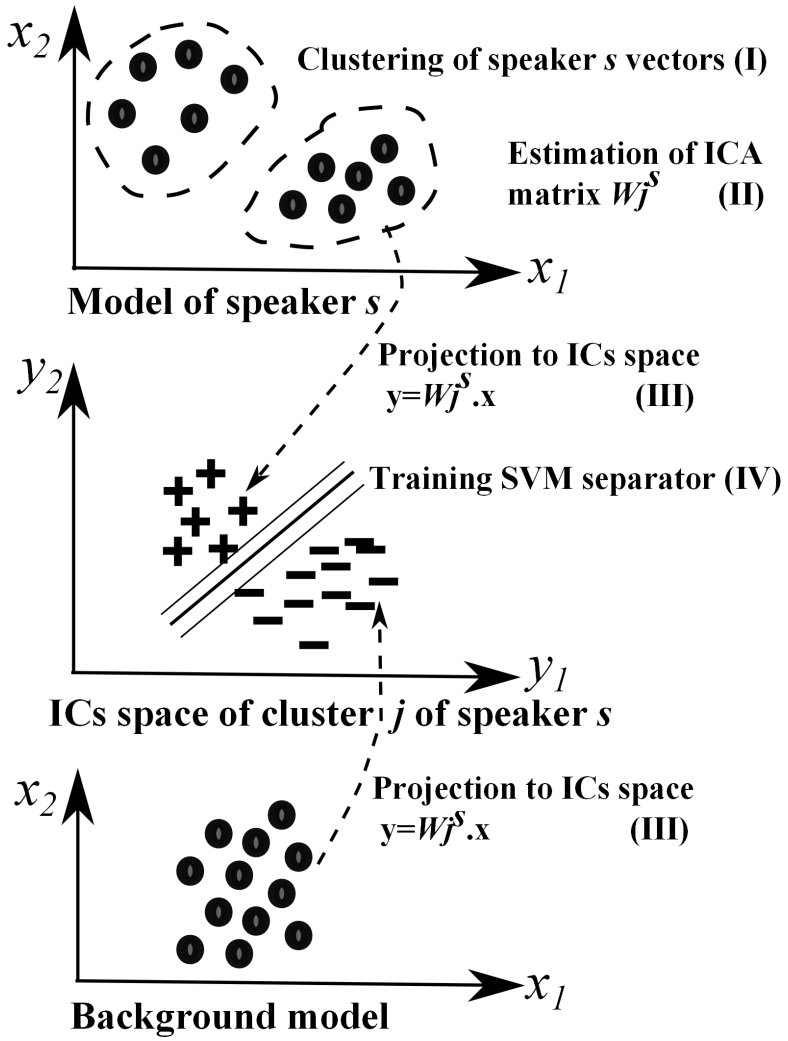



FIGURE 4.3 – Schéma d'apprentissage du modèle de classification en combinant ACI locale et SVM.

Les opérations I,...,IV sont effectuées selon cet ordre chronologique.

Le modèle de monde (MM) doit en principe représenter fidèlement l'hypothèse alternative (tous les locuteurs possibles), une telle condition est difficile à garantir en pratique. Il y a un compromis à faire entre la complexité calculatoire (fonction de la taille du MM) et la qualité de sa représentation, d'où les MMs réduits qui sont composés seulement d'un ensemble raisonnable de vecteurs sélectionnés attentivement sont préférables. Dans nos expériences, le MM est construit par un tirage aléatoire et équiprobable d'un sous-ensemble de vecteurs MFCCs de chaque locuteur enregistré.

**b) Algorithme de prédiction**

Le but de cette phase consiste à prédire la classe (locuteur) d'un vecteur MFCC à tester $\mathbf{x}$ parmi un ensemble fini des locuteurs enregistrés $S = \{s_1, ..., s_{\#Locuteurs}\}$. En réalisant les mêmes projections que celles de la phase d'apprentissage, le pseudo-algorithme de prédiction de la classe la plus vraisemblable est le suivant :

1. Estimation d'une pré-vraisemblance $LL_j^s$ d'appartenance du vecteur $\mathbf{x}$ au cluster $C_j$ du locuteur $s$; Par exemple, $LL_j^s$ est inversement proportionnelle à la distance séparant $\mathbf{x}$ au centre de $C_j$ pour l'algorithme K-Means. $LL_j^s$ peut être considérée comme une probabilité a priori du cluster $C_j$.
2. Calcul de l'image $\mathbf{y}_j^s$ du vecteur $\mathbf{x}$ dans l'espace des CIs correspondant : $\mathbf{y}_j^s \leftarrow W_j^s \times \mathbf{x}$.
3. Calcul de la marge de séparation SVM du vecteur $\mathbf{y}_j^s$ par rapport au classificateur correspondant $SVM_j^s$. Soit $marge_j^s$ la marge calculée, elle est proportionnelle à la vraisemblance $P(\mathbf{x}|C_j)$ au sein de la classe $s$. Répéter les étapes 1, 2 et 3 pour chaque cluster dans chaque locuteur.
4. Calcul de la marge globale de $\mathbf{x}$ à l'intérieur de chaque classe $s$: $marge^s = \sum_{j=1}^{k^s} LL_j^s \times marge_j^s$. La quantité $marge^s$ est proportionnelle à la vraisemblance $\mathbf{x}|P(Classe_s)$
5. Prédiction de la classe de $\mathbf{x}$ par le choix de la classe ayant la marge maximale.

L'algorithme K-Means a été employé dans un but d'illustration. Construire une décision en se basant seulement sur le cluster le plus proche de $\mathbf{x}$ ne semble pas une bonne option, particulièrement lorsque d'autres clusters se trouvent à des distances légèrement supérieures à celle du cluster le plus proche. Pour cette raison, dans l'étape 1, nous proposons d'estimer les pré-vraisemblances $LL_j^s$ de $\mathbf{x}$ dans les différents clusters. Pour le cas de K-Means, $LL_j^s$ peut être calculé comme suit :

$$LL_j^k(\mathbf{x}) = \frac{(\frac{1}{\|\mathbf{x}-v_j\|^2})^m}{\sum_{j=1}^{k^s}(\frac{1}{\|\mathbf{x}-v_j\|^2})^m} \tag{14}$$

Ainsi, on donnera plus d'importance au cluster le plus proche pour des valeurs élevées du paramètre $m$. Dans l'étape 4, la marge SVM calculée, $marge^s$, représente le score d'appariement du vecteur à tester $\mathbf{x}$ avec le modèle du locuteur $s$. $marge^s$ est proportionnelle à la vraisemblance $P(\mathbf{x}|s)$ et elle est également proportionnelle à la probabilité a posteriori $P(s|\mathbf{x})$ du modèle $s$ si les classes ont les mêmes probabilités a priori. La décision finale de la classification dans l'étape 5 est faite selon la règle de maximum a posteriori :

$$Z \leftarrow argmax_{s:1,...,\#locuteurs}(marge^s) \tag{15}$$

Un schéma expliquant le pseudo-algorithme de prédiction est illustré sur la figure 4.4.

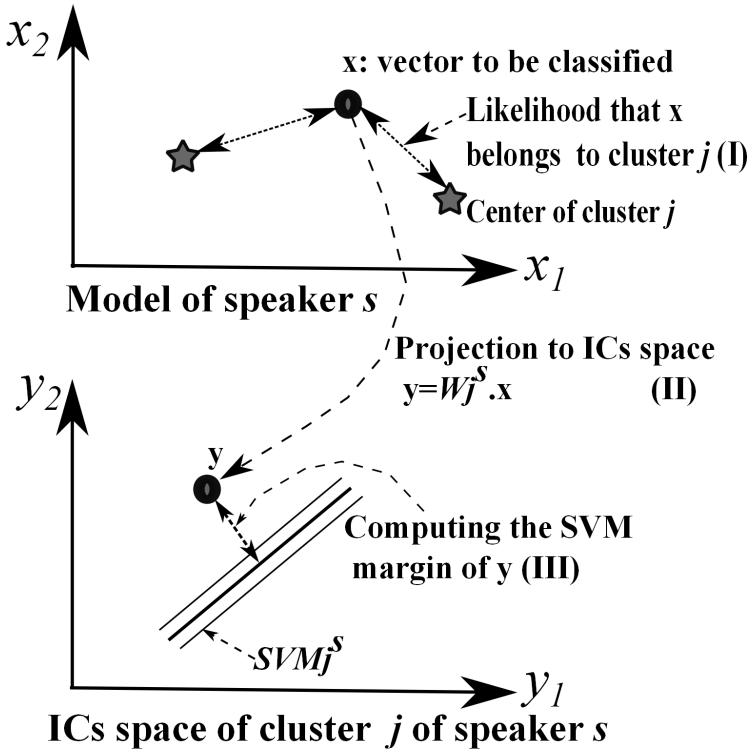



FIGURE 4.4 – Algorithme de prédiction en combinant l'ACI locale et SVM.

Les opérations I,II et III sont effectuées selon cet ordre chronologique.

**c) Modèle proposé**

Le clustering, la classification discriminative, l'adaptation via un modèle de monde et l'ACI sont utiles en RAL indépendante de texte. Dans cette section, nous avons proposé de combiner ces 04 modules afin d'exploiter conjointement leurs avantages dans un objectif d'améliorer le taux de classification des vecteurs MFCCs.

Le modèle d'un locuteur $s$ comporte les paramètres suivants : (i) paramètres de clustering (e.g centres des clusters $v_j$ si K-Means est utilisé), (ii) paramètres de transformation ACI (e.g la matrice de séparation estimée dans le cas linéaire : $W_j^s$), et le séparateur $SVM_j^s$ construit entre les projections de $s$ et celles du modèle de monde.

**d) Préparation de données**

***- Corpus audio-visuel MOBIO***

L'évaluation de performances a été effectuée sur un échantillon tiré du corpus MOBIO [mob, 2013]. Le corpus MOBIO consiste en une base de données biométrique multimodale (audio&vidéo), conçue spécialement pour l'évaluation des systèmes de reconnaissance du lo-

cuteur et de la parole. Ce corpus englobe 12 sessions de 152 personnes et il est collectionné sur 06 sites différents (dialectes différents). La langue d'enregistrement est l'anglais dont certaines personnes l'anglais n'est pas leur langue maternelle. Les séquences audiovisuelles ont été enregistrées à l'aide d'un laptop et d'un téléphone portable de marque NOKIA N93i.

***- Extraction des vecteurs MFCCs***

Les comparaisons ont été menées sur un échantillon de 40 locuteurs tiré aléatoirement de la base de données. Les MFCCs (*mel frequency cepstral coefficients*) ont été utilisés pour caractériser la voix. Les MFCCs sont des coefficients fréquentiels calculés sur des fenêtres temporelles de courte durée, ces derniers sont couramment utilisés en RAL indépendante de texte [Campbell, 1997, Kinnunen and Li, 2010]. Dans nos expériences, la fréquence d'échantillonnage des signaux audio est 16KHz et un vecteur de 12 MFCCs est calculé à chaque 6.25ms sur une fenêtre de Hamming de 25ms.

**e) Résultats de classification**

**e.1) *Comparaison entre les algorithmes ACI***

Les taux moyens de classification de 10 exécutions des 04 algorithmes ACI suivants : FastICA [Hyvärinen et al., 2001], JADE [Cardoso and Souloumiac, 1993], FastKICA [Chen, 2006] et KDICA [Shen et al., 2009] sont affichés dans le tableau 4.1. Ces 04 algorithmes ACI sont appliqués dans le système proposé de la classification sur des clusters sphériques détectés par l'algorithme de clustering K-Means . Nous constatons que leurs taux de classification sont approximativement similaires. Il est à propos de noter qu'il est fort probable que ces algorithmes puissent se comporter différemment avec d'autres algorithmes de clustering. En outre, il est possible que d'autres algorithmes ACI plus appropriés puissent donner des mieux taux que ceux affichés ci-après.

Tableau 4.1 – Comparaison des taux de classification entre quelques algorithmes ACI populaires.

| Algorithme | Nombre de locuteurs (classes) | | | | |
|---|---|---|---|---|---|
| ACI | 2 | 3 | 4 | 5 | 6 |
| FastICA | 0.7643 | 0.5904 | 0.5127 | 0.4701 | 0.4516 |
| Jade | 0.7648 | 0.5925 | **0.5134** | **0.4711** | 0.4470 |
| FastKICA | **0.7670** | **0.5947** | 0.5071 | 0.4634 | **0.4545** |
| KDICA | 0.7599 | 0.5911 | 0.5101 | 0.4675 | 0.4356 |

**e.2) *Comparaison avec d'autres architectures du classificateur SVM***

Dans cette expérience, les performances de la classification de notre solution en termes du taux des vecteurs MFCCs correctement classifiés ainsi que du temps de la prédiction sont comparées avec les architectures SVM suivantes :

- La classification SVM multi-classes conventionnelle *un-contre-un*.
- SVM *locuteur contre modèle de monde* avec une projection ACI : c'est équivalent à notre solution mais sans faire le clustering, c'est-à-dire le nombre de cluster $k^s$ est fixé à 1.
- SVM *locuteur-contre-modèle de monde* avec le clustering : par la prise des matrices de projection $W_j^s$ égales à l'identité $I$ dans notre solution; pas de projection ACI.
- SVM *locuteur-contre-modèle de monde*, sans clustering et sans projection ACI : c'est équivalent à notre solution avec les paramètres $W_j^s = I$ et $k^s = 1$.

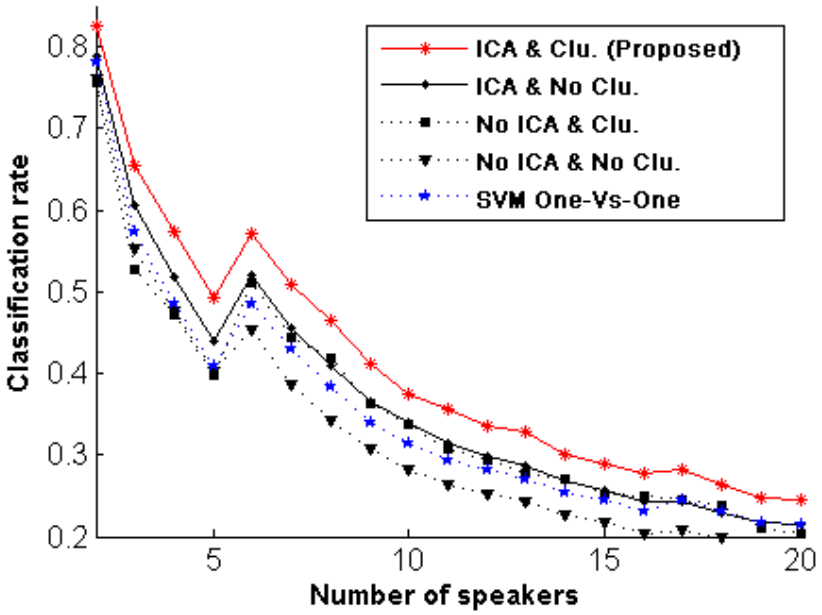



FIGURE 4.5 – Comparaison en termes du taux de la classification de la solution proposée avec quelques classificateurs SVM.

Nous constatons clairement sur la figure 4.5. que, lorsqu'un modèle de monde est ajouté, l'application seulement de l'ACI (courbe : 'ICA & No Clu.') ou seulement du clustering (courbe : 'No ICA & Clu.') dans la solution proposée améliorent significativement les taux de classification par rapport à notre solution sans ACI et sans clustering (courbe : 'No ICA & No Clu.'). En outre, l'application conjointement de l'ACI et du clustering (courbe : 'ICA & Clu.') conduit encore à une amélioration des taux de classification par rapport aux autres solutions. Nous constatons également que la solution proposée, lorsque la projection ACI et le clustering sont effectués, donnent des taux de classification significativement meilleurs que le SVM conventionnel un-contre-un et que l'amélioration eue est particulièrement significative pour un grand

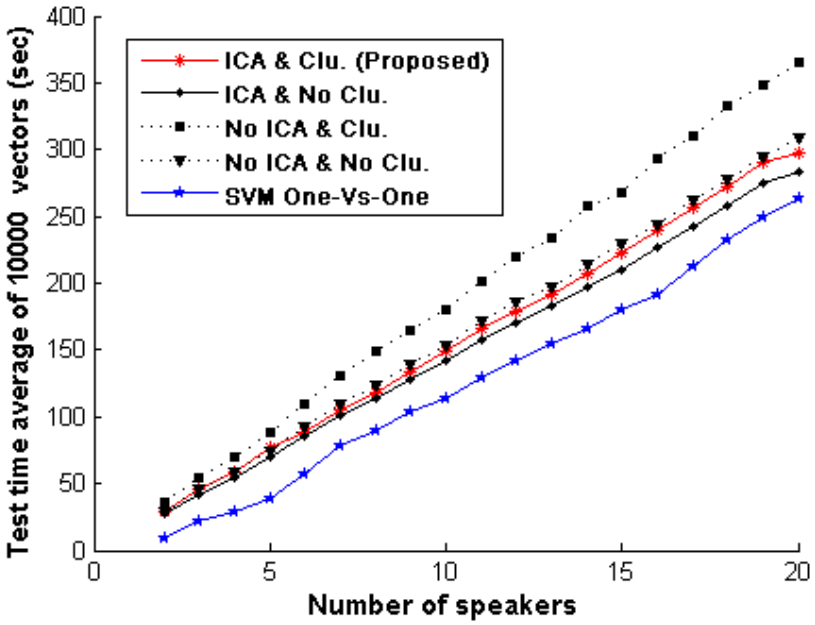



FIGURE 4.6 – Comparaison en termes de temps de la classification de la solution proposée et certains classificateurs SVM.

nombre de classes.

En examinant la figure 4.6, le temps de prédiction des 05 classificateurs semblent avoir une tendance d'accroissement linéaire avec l'augmentation du nombre de locuteurs. Soit $N_{loc}$ le nombre de classes, la complexité du classificateur SVM conventionnel un-contre-un, mesurée en termes du nombre de classifications SVM binaires, est polynomiale $((N_{loc}^2 - N_{loc})/2$ classifications binaires à réaliser). Cependant, la complexité des 04 autres classificateurs (variantes de notre solution) est linéaire où seulement $N_{loc} \times kmax$ classifications SVM binaires doivent être faites (la constante $kmax$ est le nombre maximal de clusters par locuteur). En conséquence, notre proposition sera plus rapide que le SVM un-contre-un pour un grand nombre de classes ($N_{loc}$ important).

En comparant la courbe 'ICA & Clu.' avec la courbe 'No ICA & Clu.' et la courbe 'ICA & No Clu.' avec la courbe 'No ICA & No Clu.', nous constatons clairement que la projection ACI dans notre solution contribue considérablement dans la réduction du temps de prédiction. Cette réduction peut due au fait que des séparateurs SVM plus simples et contenant un nombre moins important des vecteurs de support sont possibles à établir dans l'espace des CIs.

## 2.4 Reconnaissance du locuteur par la fusion de la voix et du mouvement des lèvres

### 2.4.1 Algorithmes de clustering flou

#### a) Définition de clustering flou

Le clustering (regroupement) de données est le processus de partitionnement des vecteurs de données en un ensemble de clusters (groupes) de telle sorte que les vecteurs du même groupe soient similaires (proches) et ceux des différents groupes soient distincts (éloignés) le maximum possible. On distingue deux types de clustering, le clustering déterministe (appelé aussi '*crisp*' ou '*hard*') et le clustering flou (appelé aussi '*fuzzy*', '*soft*'). Un vecteur est affecté à un seul cluster dans le premier type et il peut être affecté à plusieurs clusters avec des degrés d'appartenance différents dans le deuxième type. Dans plusieurs applications du monde réel, les frontières délimitant les clusters naturels peuvent être chevauchées où un vecteur peut appartenir à plusieurs clusters. Par ailleurs, le nombre de vecteurs par cluster doit être suffisamment grand pour avoir un modèle ACI local fiable [Karhunen and Malaroiu, 1999], cet inconvénient est moins grave dans le clustering flou car dans lequel un cluster implique tous les vecteurs à regrouper avec des degrés différents. Pour ces deux raisons, l'emploi du clustering flou semble prometteur dans le modèle ACI local. Dans cette optique, Honda et al. ont proposé la méthode *fuzzy local ICA* par le partitionnement flou de données en un ensemble de clusters linéaires [Honda and Ichihashi, 2006].

**Définition 12 (*espace de clustering flou*)**

*soit* $X = [\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^N]$ *un ensemble fini de* $N$ *vecteurs et soit* $2 \leq L < N$ *un entier représentant le nombre de clusters. L'espace de clustering flou de* $X$ *est l'ensemble*

$$M_{fc} = \left\{ \begin{array}{ll} & u_{ij} \in [0,1] \quad , \forall i, j \; dans \; 1..N, \; 1..L, \; respectivement. \\ U \in R^{N \times L} / & \sum_{j=1}^{L} u_{ij} = 1 \quad , \forall i \; dans \; 1..N \\ & 0 < \sum_{i=1}^{N} u_{ij} \quad , \forall j \; dans \; 1..L \end{array} \right\}$$

La ligne d'indice $i$ de $U$ constitue les probabilités d'appartenance du vecteur $\mathbf{x}^i$ aux différents clusters, ce qui explique que sa somme doit être égale à 1. L'objectif du clustering flou consiste à estimer $U$ en maximisant le rapport compacité-séparation des clusters.

#### b) Algorithme ' Fuzzy c-Means '

L'algorithme Fuzzy c-Means (FCM) est l'une des méthodes du clustering flou les plus populaires. Il a été développé par [Dunn, 1973] et amélioré par [Bezdek, 1981]. Son principe s'appuie sur la minimisation de la fonction objectif suivante.

$$J(X;U;V) = min_{U,V} \sum_{j=1}^{L} \sum_{i=1}^{N} (u_{ij})^m \parallel \mathbf{x}^i - \mathbf{v}^j \parallel_A^2 \tag{16}$$

où $m$ est un nombre réel supérieur à 1 qui définit le degré du caractère flou, $V = \{\mathbf{v}^1, ..., \mathbf{v}^L\}, \mathbf{v}^j \in R^n$ sont les vecteurs centraux des clusters et $\parallel \alpha \parallel_A^2 = \alpha^T A \alpha$ est une norme basée sur la matrice $A$ qui mesure la distance entre les observations et les vecteurs centraux.

$$D_{ijA}^2 = \parallel \mathbf{x}_i - \mathbf{v}_j \parallel_A^2 = (\mathbf{x}^i - \mathbf{v}^j)^T A (\mathbf{x}^i - \mathbf{v}^j) \quad j = 1...L, i = 1...N \tag{17}$$

La forme des clusters recherchés est définie par la matrice $A$, par exemple on aura des clusters sphériques si $A$ est égale à l'identité $I$. Voir une illustration en 2d sur la figure 4.7.

**c) Algorithme ⁄ Gustafson-Kessel ⁄**

L'algorithme de clustering Gustafson-Kessel (GK) est une extension du fameux FCM par l'emploi des normes de distance adaptatives afin de pouvoir détecter des clusters ayant des topologies différentes [Gustafson and Kessel, 1979]. Chaque cluster est défini par sa propre matrice de norme $A_j$, la distance entre un vecteur $\mathbf{x}^i$ et le cluster d'indice $j$ se calcule comme suit :

$$D_{ijA_j}^2 = \parallel \mathbf{x}^i - \mathbf{v}^j \parallel_{A_j}^2 = (\mathbf{x}^i - \mathbf{v}^j)^T A_j (\mathbf{x}^i - \mathbf{v}^j) \qquad j = 1...L, i = 1...N \tag{18}$$

Les matrices $A = \{A_1, ..., A_L\}$ sont considérées comme des paramètres à estimer dans la fonction objectif (19), cela permet ainsi aux clusters d'adapter leurs normes selon la topologie locale de données. La fonction objectif à minimiser par GK est

$$J(X;U;V;A) = min_{U,V,A} \sum_{j=1}^{L} \sum_{i=1}^{N} (u_{ij})^m D_{ijA_j}^2 \tag{19}$$

Dans nos expériences, nous avons utilisé l'implémentation numérique de GK présentée dans [Babuska et al., 2002]. Voir une illustration en d2 sur la figure 4.8.

**d) Algorithme ⁄ Gath Geva ⁄**

L'algorithme Gath Geva (GG) [Gath and Geva, 1989] emploie une norme de distance basée sur les estimateurs de maximum de vraisemblance flous, proposés par [Bezdek and Dunn, 1975], donnée par la formule

$$D_{ij} = \frac{\sqrt{det(F_{wj})}}{\alpha_j} exp(\frac{1}{2}(\mathbf{x}^i - \mathbf{v}^j)^T F_{wj}^{-1} (\mathbf{x}^i - \mathbf{v}^j)) \tag{20}$$

A la différence de GK, le terme exponentiel dans (20) fait diminuer plus rapidement la distance en s'éloignant que le produit scalaire (18). $F_{wj}$ désigne la matrice de covariance floue du cluster d'indice $j$, donnée par la formule

$$F_{wj} = \frac{\sum_{i=1}^{N}(u_{ij})^w(\mathbf{x}^i - \mathbf{v}^j)(\mathbf{x}^i - \mathbf{v}^j)^T}{\sum_{i=1}^{N}(u_{ij})^w} \tag{21}$$

avec $\alpha_j$ la probabilité a priori de selection du cluster d'indice $j$ et $w$ une constante. Voir une illustration en 2d sur la figure 4.9.

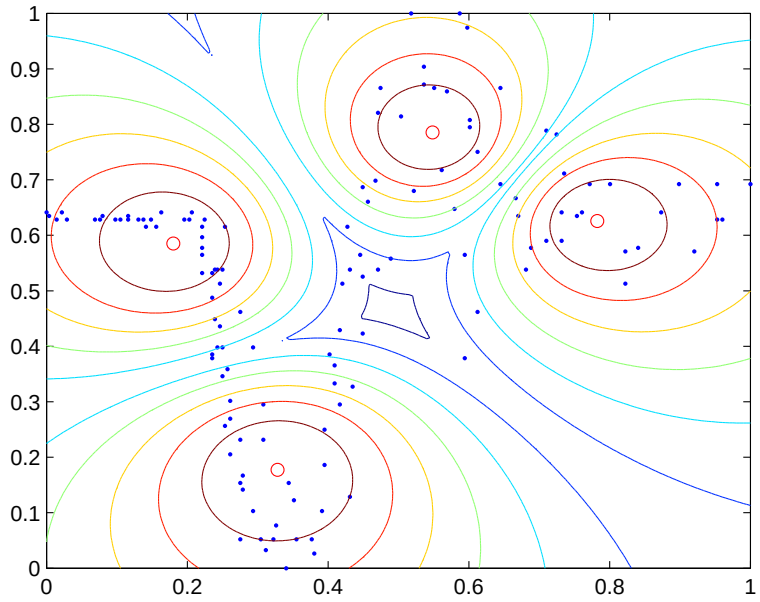



FIGURE 4.7 – Exemple de clustering FCM en deux dimensions avec $L = 4$.

### 2.4.2 Système de classification proposé

Nous utilisons la méthode ACI locale floue pour la construction d'un estimateur de densité multivariable $\pi_i$ pour chaque locuteur $\omega_i, i = 1, ..., N_{loc}$. Les paramètres du modèle $\pi_i$ comportent les paramètres de clustering, les matrices de mélange ACI et les estimateurs de densité unidimensionnelles des CIs.

Le classificateur affecte le vecteur $\mathbf{x}$ à la classe $\omega_i$ s'il maximise la probabilité a posteriori de cette dernière ; formellement, la règle de decision est

$$c = argmax_i(P(\mathbf{x}/\omega_i)P(\omega_i)) \tag{22}$$

où $P(\omega_i)$ est la probabilité a priori du locuteur $\omega_i$.

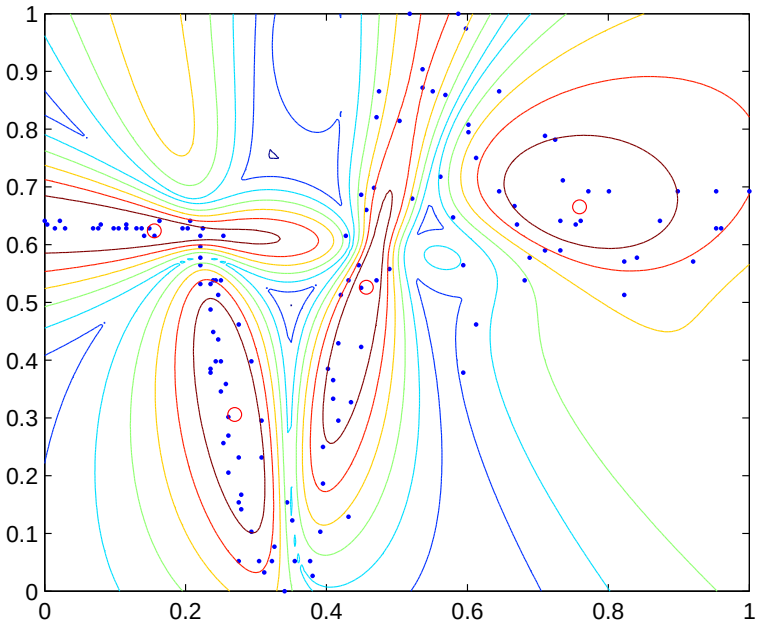



FIGURE 4.8 – Exemple de clustering GK en deux dimensions avec $L = 4$.

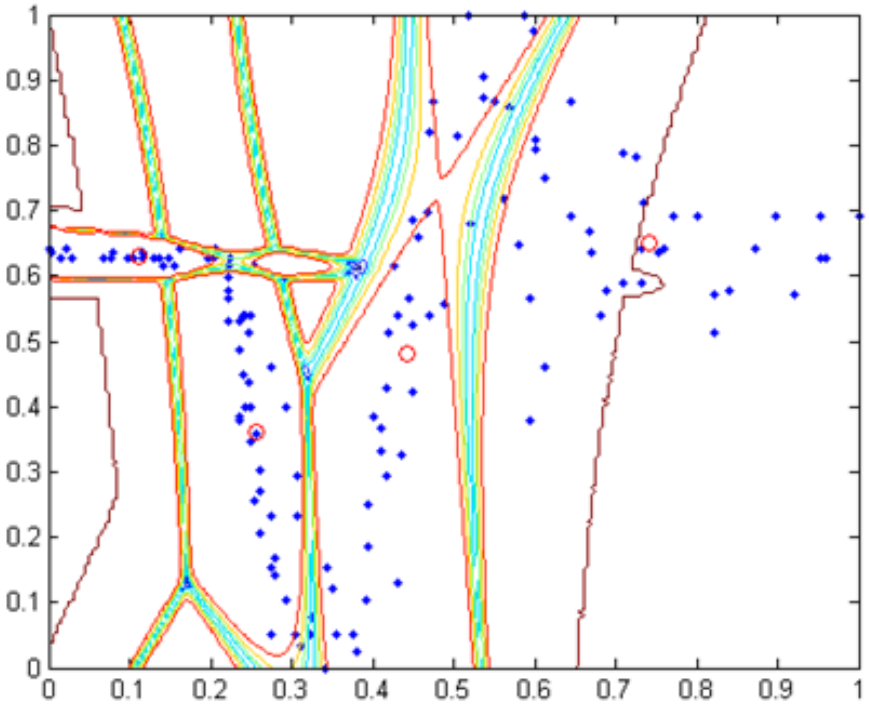



FIGURE 4.9 – Exemple de clustering GG en dimension avec $L = 4$.

Cependant, la fonction de decision (22) n'est pas assez discriminante. En effet, chaque estimateur $\pi_i$ est construit en utilisant un échantillon de données différent tiré de sa classe $\omega_i$. Par conséquent, les estimateurs $\pi_i$ sont biaisés différemment et cela amoindrit les performances même si le biais est faible. Pour cette raison, une étape de poste-traitement telle que la normalisation est y indispensable.

Afin d'atténuer l'effet de bias et aussi pour remédier au problème d'absence de probabilités a priori des classes, nous proposons d'effectuer une classification discriminative sur l'espace des vraisemblances $P(\mathbf{x}/\omega_i)$ des vecteurs MFCCs des différentes classes.

Le pseudo-algorithme d'apprentissage est montré sur le diagramme de la figure 4.10. En outre des estimateurs $\pi_i$, le modèle de classification comporte aussi le séparateur SVM multi-classes. Le pseudo-algorithme de prediction est montré sur la figure 4.11.

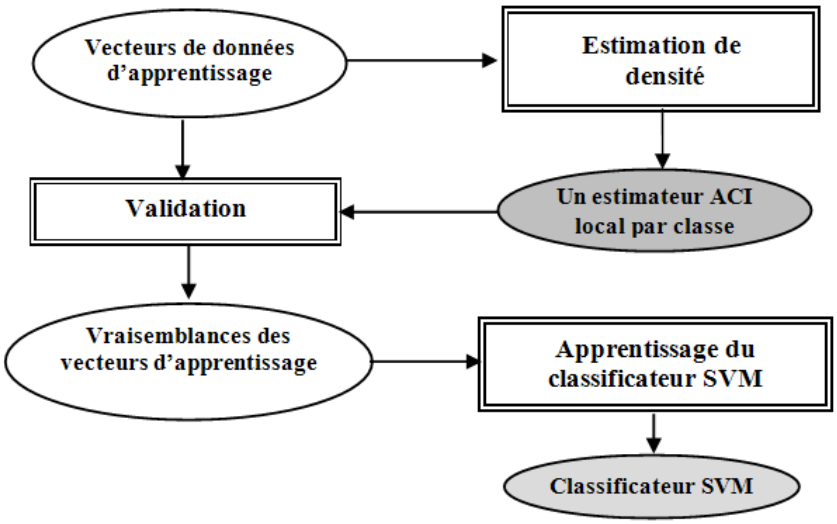



FIGURE 4.10 – Étapes de construction du classificateur ACI-Locale & SVM.

Les paramètres du modèle comportent un estimateur de densité $\pi_i$ basé sur l'ACI locale par locuteur et un classificateur SVM.

### 2.4.3 Préparation de données

Les tests d'évaluation de performances sont réalisés sur des vecteurs biométriques bimodaux caractérisant la voix et le mouvement des lèvres dans un contexte de la RAL indépendante de texte. Les données expérimentées sont tirées du corpus audio-visuel MOBIO [mob, 2013], introduit dans la Section 2.3.4.d). Cette section présente les paramètres labiaux que nous avons considérés ainsi que la méthode de fusion des paramètres audiovisuels.

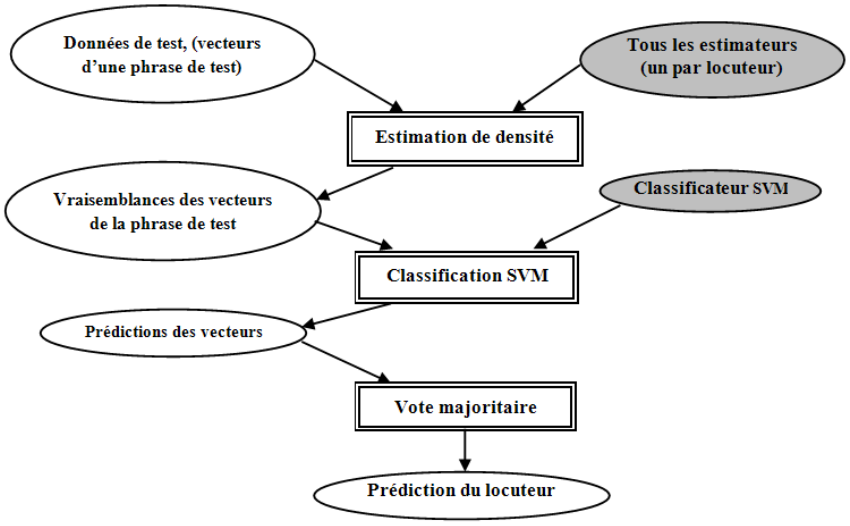



FIGURE 4.11 – Étapes de prédiction du classificateur ACI-Locale & SVM.

La décision de la phrase à tester est obtenue par un vote majoritaire sur les résultats partiels de prédiction des vecteurs MFCCs extraits de la phrase.

### a) Tracking de mouvement des lèvres

Les images sont converties de l'espace de couleurs RGB à l'espace $YCbCr$. L'espace de couleurs $YCbCr$ décompose le RGB en une composante de luminance $Y$ et deux composantes de chrominance $Cb$ et $Cr$. Les composantes de chrominance sont quasi-indépendantes de celle de la luminance et présentent une bonne discrimination entre les régions de la peau [Li et al., 2010]. Nous avons considéré dans notre travail seulement les composantes $Cb$ et $Cr$, plus particulièrement, nous nous sommes intéressés à la détermination de la plus petite fenêtre rectangulaire ($PFR$) contenant entièrement les deux lèvres.

Dans la séquence vidéo, étant donné que la $PFR$ est bien délimitée dans $i^{ème}$ image , la détermination de la $PFR$ de l'image suivante $i+1$ est réalisée comme suit :

1. La recherche d'un nouveau rectangle (région d'intérêt) dans l'image $i+1$ concentrique avec la $PFR$ de l'image $i$ de telle façon que la nouvelle fenêtre choisie soit assez grande pour couvrir les lèvres et pas les autres organes environnant tels que le nez et l'oreille après un éventuel mouvement du visage ente l'instant $i$ et l'instant $i+1$. Les dimensions de la région d'intérêt dans l'image $i+1$ sont prises égales au doubles de la $PFR$ de l'image $i$, (voir le rectangle en ligne discontinue sur la figure 4.12(b)).
2. La recherche de la partie constituant les lèvres dans la fenêtre la région d'intérêt est effectuée par une segmentation en deux régions peau et non-peau en réalisant un seuillage sur l'histogramme de couleurs de pixels. L'usage du *seuillage automatique* est justifié par le fait que l'histogrammes de la région d'intérêt contient deux sommets importants

correspondant aux lèvres et à la peau autour des lèvres, respectivement. Si la bouche est ouverte, le sommet du milieu correspondant aux dents est relativement petit. De plus, nous disposons de l'information a priori sur l'intervalle de couleurs des pixels constituant les lèvres. La segmentation est réalisée selon un seuillage de type sommet-et-vallée (*peak-and-valley thresholding*) [Sezan, 1985] où le seuil est estimé par la recherche de la vallée la plus profonde suivant le sommet des pixels des lèvres, voir un exemple de segmentation sur la figure 4.12(c).

3. L'extraction de la surface blanche connexe la plus large (figure 4.12(c)) et puis la détermination de la nouvelle *PFR* de l'image $i + 1$ (figure 4.12(d)).

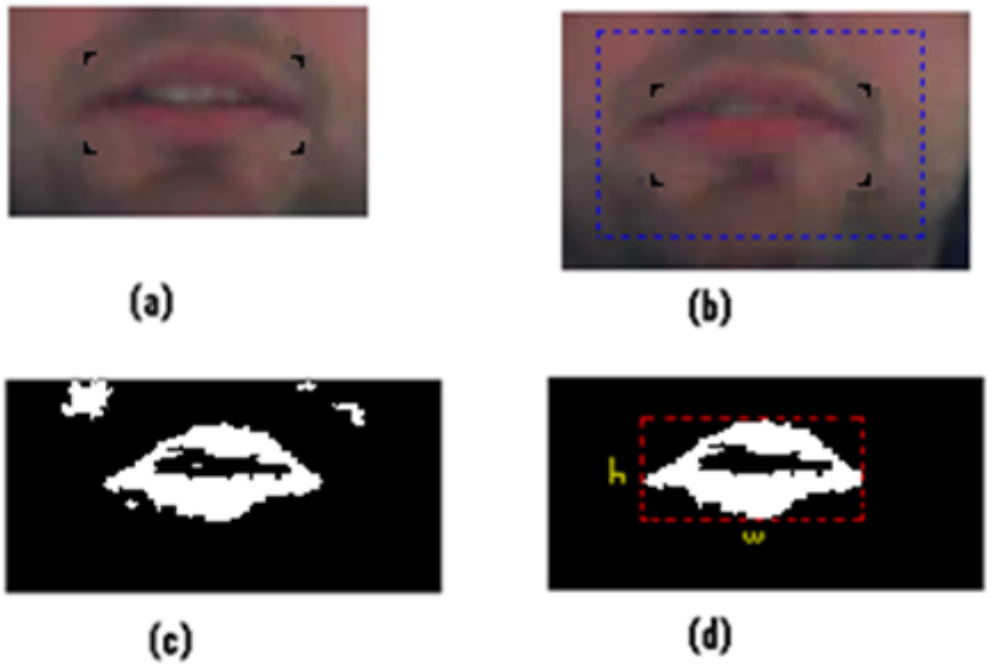



FIGURE 4.12 – Tracking du mouvement des lèvres.

**(a)**: *PFR* de l'image courante $i$. . **(b)**: Région d'intérêt dans l'image suivante $i + 1$, délimitée par le rectangle discontinu. **(c)**: Segmentation. **(d)**: La nouvelle *PFR* détectée dans l'image suivante.

La région d'intérêt contenant la bouche est fixée manuellement dans la première image de la séquence vidéo. Afin de bien caractériser le mouvement des lèvres durant une conversation, 05 attributs globaux décrivant à la fois la géométrie et la texture des lèvres sont considérés. Ces attributs sont : la largeur ($w$) de la *PFR*, la hauteur ($h$) de la *PFR*, la surface (nombre de pixels blancs sur la figure 4.12(d)), la moyenne et l'écart type de la *PFR*. On remarque bien que les attributs sélectionnés varient d'une personne à une autre et varient également en fonction des phonèmes prononcés par la même personne.

**b) Fusion des MFCCs et des attributs labiaux**

Les vecteurs MFCCs [Kinnunen and Li, 2010, Campbell, 1997] sont employés pour la caractérisation de la voix, les paramètres de leur calcul sont spécifiés dans la Section 2.3.4.d. La fusion multimodale est réalisée au niveau des attributs par la concatenation des attributs labiaux et des MFCCs, ce qui donne un vecteur de 17 attributs. La fréquence d'échantillonnage de la voix est 160 vecteurs/sec, cependant celle de la vidéo est 30 vecteur/sec. Pour cette raison, une synchronisation est y nécessaire. Pour ce faire, les attributs labiaux sont interpolés de 30 à 160 vecteurs/sec ; pour chaque vecteur MFCC, le plus proche vecteur labial en termes de l'ordre chronologique est ajouté comme le montre la figure 4.13.

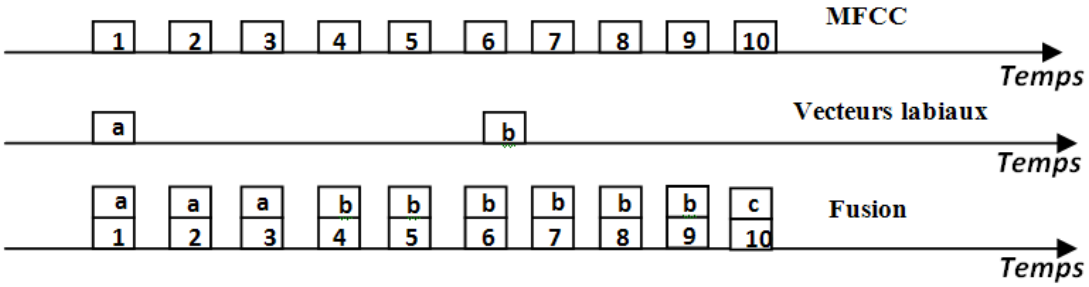



FIGURE 4.13 – Fusion des vecteurs MFCCs et des vecteurs labiaux.

Pour chaque locuteur, nous avons utilisé 05 phrases pour l'apprentissage du classificateur et 05 phrases pour les tests de prédiction.

Le tableau 4.2 montre le nombre par classe des vecteurs d'attributs biométriques impliqués dans l'évaluation.

Tableau 4.2 – Nombre des vecteurs de la classification.

| Classe | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Apprentissage | 3413 | 1884 | 1505 | 1709 | 1010 | 1987 | 1876 | 1628 | 1676 | 1384 |
| Prédiction | 1279 | 1463 | 1495 | 1184 | 944 | 1338 | 1200 | 963 | 1112 | 848 |

### 2.4.4 Résultats et discussions

La courbe *ROC* (*Receiver Operating Characteristics*) et l'indicateur *EER* (*Equal Rate Error*) des phrases sont traditionnellement employés pour évaluer les performances des systèmes de vérification du locuteur. Comme le nombre des phrases employées est relativement petit (50 phrases de test, 5 par locuteur) dans nos expériences, les critères *ROC* et *EER* deviennent obsolètes. En effet, en RAL indépendante de texte, des centaines de vecteurs d'attributs, calculés sur des fenêtres de courte durée (*short time features*), sont extraits de la phrase à tester et la décision

finale de la classe d'une phrase est établie via un vote majoritaire sur les résultats partiels de classification des vecteurs d'attributs. Pour cette raison, nous avons opté pour l'évaluation de performances en termes du taux de classification des vecteurs d'attributs bimodaux proposés.

Le tableau 4.3 montre les meilleurs taux de classification des vecteurs biométriques obtenus par le classificateur SVM standard un-contre-un. Les paramètres optimaux: le coût $\hat{c}$, le sigma du noyau RBF $\hat{\sigma}$, le degré $\hat{d}$ du noyau polynomial sont estimés empiriquement. D'après les résultats, nous constatons que la fusion améliore les taux de classification, ce qui atteste que les deux modalité biométriques comportent des informations distinctes et pertinentes pour la RAL.

Tableau 4.3 – Taux de classification des vecteurs MFCCs par SVM un-contre-un .

| | | Noyau SVM | |
|---|---|---|---|
| Attributs | Gaussien (RBF) | Linéaire | Polynomial |
| Mouvement de lèvres | $\mathbf{0.6783}_{(\hat{c}=1000,\hat{\sigma}=0.0021)}$ | $0.6675_{(\hat{c}=10}$ | $0.6775_{(\hat{c}=100,\hat{d}=2)}$ |
| Voix | $\mathbf{0.2471}_{(\hat{c}=1000,\hat{\sigma}=0.0351)}$ | $0.1859_{(\hat{c}=1000)}$ | $0.2371_{(\hat{c}=10,\hat{d}=3)}$ |
| Mouvement de lèvres + voix | $0.7195_{(\hat{c}=1000,\hat{\sigma}=0.0011)}$ | $0.7233_{(\hat{c}=10)}$ | $\mathbf{0.7245}_{(\hat{c}=10,\hat{d}=3)}$ |

$\hat{c}$ : coefficient du coût d'optimisation SVM, $\hat{\sigma}$ : sigma du RBF, $\hat{d}$ :degré du polynôme.

Les algorithmes ACI linéaires suivants ont été expérimentés dans la plate-forme de la classification proposée :

- FastICA [Hyvärinen et al., 2001], basée sur la maximisation de la non Gaussianité.
- JADE [Cardoso and Souloumiac, 1993], basée sur la diagonalisation conjointe approximée des matrices de cumulants d'ordre 4.
- KDICA [Shen et al., 2009], emploie une fonction noyau de Laplace pour l'estimation de densité.
- FastKernelICA [Chen, 2006], emploie des fonctions à noyaux pour mesurer l'indépendance statistique où une méthode approximée de Newton est utilisée en optimisation, utile pour les problèmes à grande échelle.

Toutes les combinaisons possibles des 04 algorithmes ACI linéaires cités et des 03 algorithmes de clustering auparavant introduits (FCM, GK et GG) ont été testées dans le système de classification proposé. Les taux de classification des attributs labiaux, des attributs vocaux, et des attributs de fusion sont affichés respectivement sur les tableaux 4.4, 4.5 4.6.

Tableau 4.4 – Taux de classification des vecteurs des attributs labiaux.

| | Algorithme de clustering flou | | |
|---|---|---|---|
| Algorithme ACI | GK | FCM | GG |
| FastICA | 0.7068 | 0.7160 | 0.7160 |
| JADE | 0.7029 | 0.7118 | 0.7086 |
| FastKernelICA | 0.7100 | 0.6988 | 0.7157 |
| KDICA | 0.7146 | **0.7170** | 0.7129 |

Tableau 4.5 – Taux de classification des vecteurs des attributs vocaux.

| | Algorithme de clustering flou | | |
|---|---|---|---|
| Algorithme ACI | GK | FCM | GG |
| FastICA | 0.2431 | 0.2391 | 0.2429 |
| JADE1 | 0.2407 | 0.2389 | 0.2438 |
| FastKernelICA | 0.2195 | 0.2244 | 0.2272 |
| KDICA | 0.2418 | 0.2380 | **0.2456** |

Tableau 4.6 – Taux de classification des vecteurs des attributs de fusion.

| | Algorithme de clustering flou | | |
|---|---|---|---|
| Algorithme ACI | GK | FCM | GG |
| FastICA | 0.7623 | 0.7648 | 0.7464 |
| JADE1 | 0.7650 | 0.7658 | 0.7475 |
| FastKernelICA | 0.7435 | 0.7344 | 0.7496 |
| KDICA | **0.7726** | 0.7710 | 0.7275 |

Nous constatons que les taux de classification des 12 combinaisons des algorithmes ACI et de clustering sont globalement similaires. Néanmoins, en examinant de près les résultats obtenus, nous pouvons déceler certains constats intéressants.

Dans la plupart des cas, KDICA a fourni des performances légèrement meilleures que les autres algorithmes ACI.

En examinant les résultats de prédiction des attributs MFCCs affichés sur le Tableau 4.5, nous constatons que l'algorithme GG est légèrement plus performant que GK et FCM lorsqu'ils sont combinés avec les 04 algorithmes ACI. En outre, pour les 03 bases de test, l'algorithme FastKernelICA convient mieux avec GG qu'avec FCM et GK.

Le meilleur taux de fusion est obtenu par l'algorithme de clustering GK, cela peut être expliqué par le fait que les attributs de fusion n'ont pas la même variance ce qui engendre des clusters pas sphériques mais plutôt elliptiques et étirés selon certaines directions.

La figure 4.14 montre le gain eu en taux de classification lorsqu'une classification SVM des vraisemblances des vecteurs est faite afin de corriger le biais d'estimation de densité, les résultats sont comparés avec le classificateur MAP (fonction de décision (22)). Nous constatons que l'amélioration eue est de l'ordre de 10%.

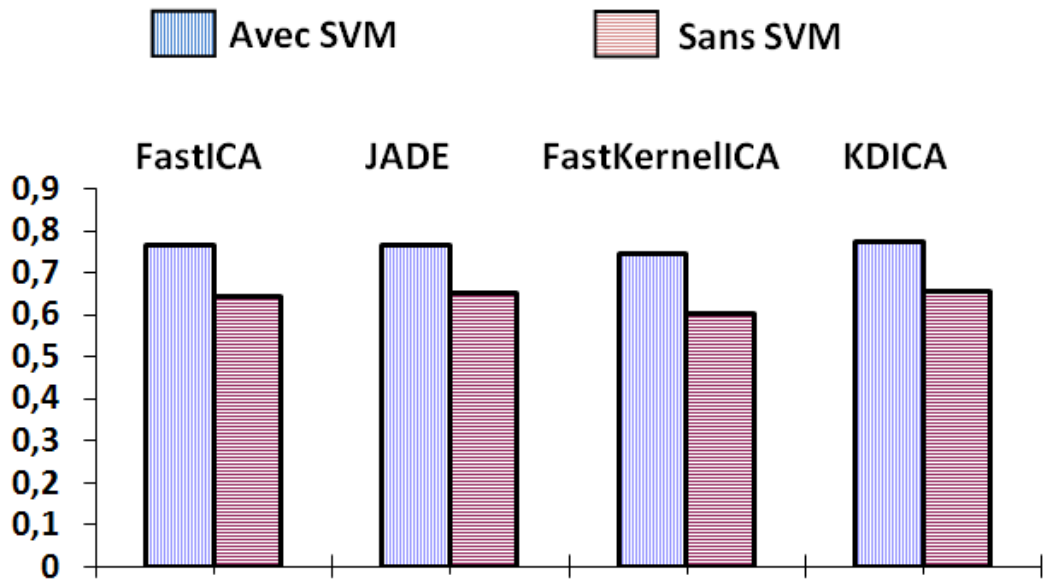



FIGURE 4.14 – Taux de classification: Correction de Biais par SVM.

L'amélioration du taux de classification en ajoutant une poste-classification SVM des vraisemblances des vecteurs d'attributs de fusion (algorithme de clustering employé: GK).

# 3 ACI locale et RB pour la représentation de données : application en estimation de densité de données de hautes dimensions

## 3.1 ACI locale, hautes dimensions et sparsity

### 3.1.1 ACI et hautes dimensions

Les performances des méthodes ACI conventionnelles se détériorent sévèrement lorsque le nombre des variables observées est important avec des cadences différentes en fonction de leurs caractéristiques techniques, de leurs fonctions objectif (approximations de l'indépendance mutuelle) et de leurs algorithmes d'optimisation. Au même temps, la complexité calculatoire spatio-temporelle augmente significativement avec la dimension, elle est souvent exponentielle pour les méthodes ACI compétitives.

Plusieurs travaux récents abordant la dégradation de la capacité des méthodes ACI à reconstituer les signaux sources en hautes dimensions confirment ce constat. Des comparaisons faites, entre quelques algorithmes ACI populaires appliqués sur des sources super Gaussiennes [Giannakopoulos et al., 1999], entre les deux algorithmes Pearson-ICA et JADE sur des sources

suivant la distribution Rayleigh [Karhunen and Koivunen, 2002] et entre les deux algorithmes SHIBBS/SJAD-ICA et FastICA sur des sources uniformes et super Gaussiennes [Liu et al., 2010], témoignent d'une baise rapide du SNR (Signal-to-Noise Ratio) des sources estimées en hautes dimensions.

### 3.1.2 Le clustering et les hautes dimensions

Un bon nombre d'applications relatives à des domaines intéressants tels que la physique, la médecine et la bioinformatique nécessitent une analyse des quantités énormes de données de hautes dimensions. Néanmoins, les données de hautes dimensions posent de nouveaux challenges par rapport aux outils classiques d'analyse de données tels que les algorithmes de clustering. Ces nouvelles difficultés se rapportent en gros à l'augmentation de la complexité calculatoire, au problème de sur-apprentissage et à la non validité de mesures de similarités entre objets.

Les algorithmes standard de clustering deviennent peu fiables lorsque la quantité de données est moins abondante comparativement au nombre important de paramètres de clustering à estimer. Ce phénomène est référé dans la littérature par le terme 'malédiction de la dimensionnalité' (*curse of dimensionality*), il a été mis en évidence par [Bellman, 1957], nous citons l'exemple [Zhao et al., 2011] pour une illustration.

D'autre part, même avec l'abondance de données, le clustering en hautes dimensions souffre d'un autre problème plus sévère qui réside dans la non égalité de pertinence des différentes dimensions par rapport aux différents clusters. En effet, la topologie de clustering dépend de la mesure de similarité adoptée. Toutes les dimensions ne sont pas importantes pour la détermination d'un cluster donné et certaines d'elles ne sont pas pertinentes [Parsons et al., 2004]. Ce point est dû à la spécificité des clusters où des différents clusters sont définis dans des sous-espaces différents. Par conséquent, les mesures de similarité impliquant les dimensions avec la même pertinence deviennent inappropriées [Domeniconi et al., 2007].

Malheureusement, les techniques de sélection des variables et de réduction de dimension ne peuvent pas être appliquées parce qu'elles représentent des transformations globales, dans la mesure où elles déterminent un seul sous-espace des données originaux en considérant l'ensemble complet des vecteurs [Kriegel et al., 2009]. En revanche, les clusters peuvent pratiquement exister dans des sous-espaces distincts. Pour définir des mesures de similitude plus appropriées, une approche récente consistant à trouver des clusters dans des sous-espaces distincts suscite beaucoup d'intérêt [Kriegel et al., 2009, Parsons et al., 2004, Peng and Zhang, 2011].

### 3.1.3 L'ACI et la sparsity

L'exploitation de la sparsity est évidement très utile en ACI comme en BSS ; en effet, elle a été le centre d'intérêt d'un bon nombre de travaux. A titre d'exemple, l'hypothèse de la sparsity des sources a été exploitée pour améliorer les résultats du problème BSS dans la méthode SCA (*Sparse Component Analysis*) [Georgiev et al., 2004]. La sparsity dans le contexte de SCA signifie que la majorité des sources sont nulles à chaque instant (des sources inactives).

Un second type intéressant de la sparsity peut être perçu au niveau de la composition de la matrice de mélange $A$ de l'ACI linéaire (ou la fonction de mélange de l'ACI non linaire), où un grand nombre d'éléments de $A$ sont nuls. Le nombre ainsi que la position des zéros dans la matrice de mélange définissent la structure des relations de dépendance statistique entre les variables observées (voir le chapitre précédent). Dans cette direction, l'ACI a été utilisée pour apprendre la structure de données en adoptant une stratégie d'élagage [Nielsen and Hansen, 2008].

Il est évident que la qualité de séparation de sources par l'ACI s'améliora davantage si la matrice de mélange est sparse. Cela est expliqué par le fait que ce type de sparsity est synonyme d'une faible fusion dans lequel chaque variable observée $x_i$ est un mélange de seulement un petit nombre de sources. En outre, comme c'est expliqué dans la section 3.1.1, les performances d'algorithmes ACI s'améliorent pour un petit nombre de sources.

Les figures 4.15 et 4.16 montrent la variation du SNR des deux algorithmes JADE et FastICA en fonction du nombre de sources et du degré de sparsity de $A$. Nous constatons nettement que le SNR décroît rapidement avec l'augmentation du nombre de CIs et que pour le même nombre de CIs le SNR augmente lorsque la matrice de mélange est plus sparse (e.g la courbe 1/3 est meilleure que les autres courbes: 0/9, 1/9 et 1/6). Quelques limitations d'algorithmes ACI en hautes dimensions sont résumées comme suit:

1. Il est clair que la qualité de séparation des algorithmes ACI existants se dégrade avec l'augmentation du nombre des CIs ; mais il est important de souligner le fait que la dégradation est plus rapide que celle d'un algorithme ACI 'idéal' qui est susceptible d'exploiter toute l'information mutuelle pour identifier les CIs.
2. La sparsity structurelle est omniprésente en hautes dimensions dans le monde réel. Certes, l'hypothèse de la sparsity est un facteur favorisant l'identification des CIs, mais malheureusement, la majorité des algorithmes ACI n'exploitent pas pleinement cette propriété.
3. La complexité calculatoire augmente rapidement avec l'augmentation de la dimension ce qui impose le recours à des approximations au détriment de la qualité de séparation.

Dans la section 2.2.3, nous avons proposé un estimateur de densité basé l'ACI locale où des modèles ACI linéaires sont appliqués localement sur des clusters de données. Afin de surmonter les inconvénients soulevés de l'ACI et de clustering dans les espaces de hautes dimensions, nous suggérons de se servir de la sparsity structurelle qui est omniprésente dans des telles situations,

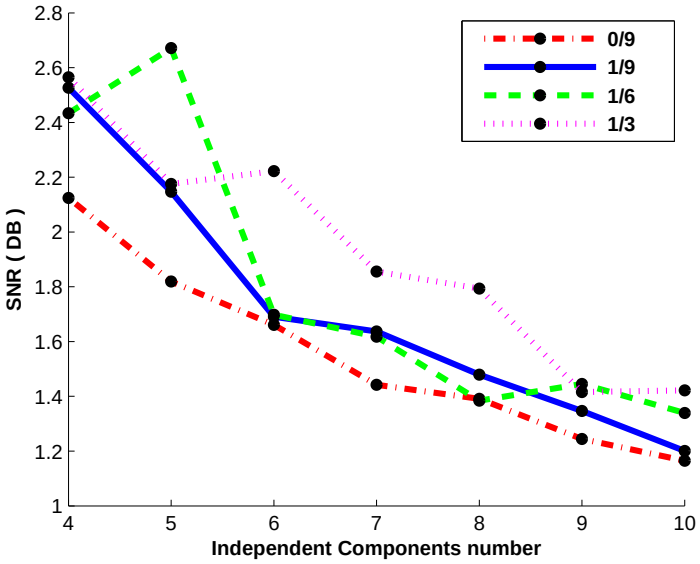



FIGURE 4.15 – Qualité de séparation de sources (mesurée en termes du SNR) par l'algorithme FastICA en fonction du nombre des CIs et de la sparsity.

Les fractions indiquées sur la figure représentent le rapport des éléments nuls de la matrice de mélange $A$.

par l'emploi des méthodes conçues particulièrement pour apprendre la structure de données.

Notre perspective consiste à déterminer la structure de relations de dépendance entre les variables observées en utilisant des algorithmes d'apprentissage dédiés qui sont moins sensibles en hautes dimensions. Puis, la structure trouvée peut être employée pour factoriser la conjointe par la réorganisation des variables observées en un ensemble de groupes composés de peu de variables (de petites dimensions) et inter-connectés entre eux où les variables d'un même groupe sont fortement et mutuellement dépendantes (un cluster de variables). Finalement, l'estimateur proposé dans (9) ou d'autres estimateurs classiques tels que les *fenêtres de Parzen* et le GMM-EM peuvent être appliqués séparément sur chaque groupe.

Concrètement, les RBs fournissent une transcription explicite bien appropriée de la sparsity structurelle sous forme d'un DAG. Les sections suivantes sont consacrées à l'exposition de notre principale contribution qui consiste à utiliser les algorithmes d'apprentissage de la structure RB ainsi que la méthode ACI locale à base de clustering flou pour l'estimation de densité en hautes dimensions.

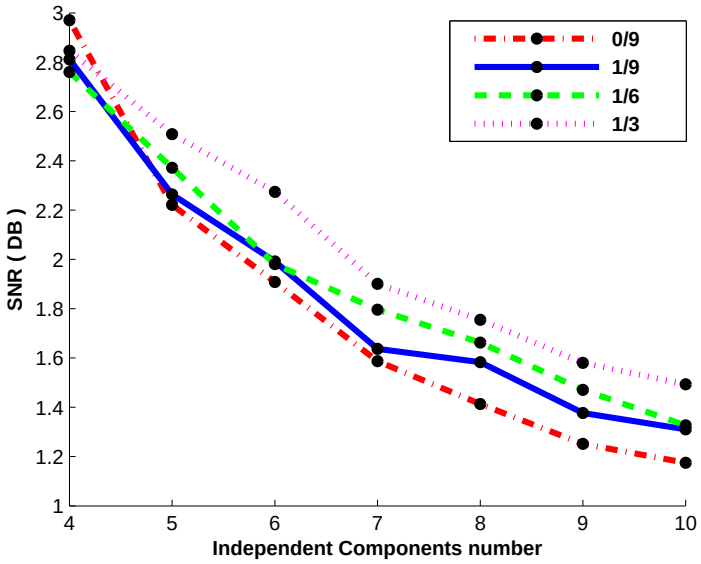



FIGURE 4.16 – Qualité de séparation de sources (mesurée en termes du SNR) par JADE en fonction du nombre des CIs et de la sparsity.

Les fractions indiquées sur la figure représentent le rapport des éléments nuls de la matrice de mélange $A$.

## 3.2 Réseaux Bayésiens et ACI pour l'estimation de densité

### 3.2.1 Apprentissage de structure

L'apprentissage qualitatif de la structure des dépendances statistiques existantes entre les variables d'intérêt constitue la moitie de la solution d'estimation de densité multivariable. Les modèles ACI sont des modèles (semi-)paramétriques et leurs algorithmes d'estimation sont basés souvent sur l'optimisation d'une fonction paramétrique globale en utilisant des approximations de la dépendance statistique (information mutuelle) et leurs performances décroissent sérieusement en hautes dimensions. En outre, les méthodes ACI procèdent par estimer les paramètres qualitatifs (la structure des zéros dans $A$) et les paramètres quantitatifs (valeurs d'éléments de $A$) simultanément, quoique ces deux types de paramètres ne sont pas liés et peuvent être estimés séparément. Ce problème devient plus crucial en hautes dimensions, il est préférable donc de commencer par l'apprentissage de la structure, ensuite l'apprentissage des paramètres quantitatifs. La structure de dépendances statistiques entre les variables observées peut être encodée sous forme d'un graphe RB, pour l'identifier, l'utilisateur peut exploiter les informations fournies par l'expert plus la possibilité de se servir d'un bon nombre d'algorithmes d'apprentissage de la structure des RBs (voir Section I.5).

### 3.2.2 Arbre de jonction pour la factorisation

Un arbre de jonction $T$ est une représentation *équivalente* d'un réseau Bayésien $B = (G, P)$, qui peut être est construit par la triangulation et la moralisation du graphe $G$. Les définitions, les algorithmes de construction ainsi que le principe d'utilisation de $T$ pour la propagation de croyances sont étalées dans le premier chapitre. A titre d'exemple, l'arbre de jonction affiché sur la figure 4.17(c) correspond au RB de la figure 4.17(a). Un arbre de jonction est constitué d'un ensemble de cliques (clusters) de variables qui sont reliées par des séparateurs. Dans cet exemple, l'ensemble de cliques est $C = \{\{X_2, X_4\}, \{X_2, X_3, X_5\}, \{X_1, X_2, X_3\}, \{X_2, X_5, X_6\}\}$ et l'ensemble des séparateurs correspondant est $S = \{\{X_2\}, \{X_2, X_3\}, \{X_2, X_5\}\}$.

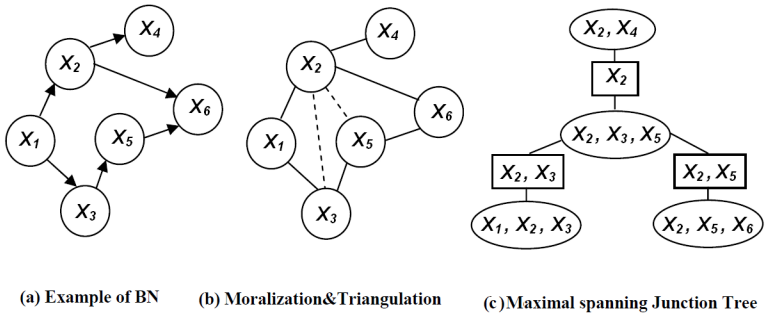



FIGURE 4.17 – Exemple montrant le principe de construction d'un arbre de jonction à partir du graphe d'un RB.

À chaque élément $A$ de $T$, qui peut être soit une clique soit un séparateur, un potentiel de croyances $\Phi_A(A)$ est associé où $\Phi_A(A)$ est une fonction non négative pour chaque sous-ensemble $A$ de $X$.

La distribution de probabilité conjointe $P(X)$ d'un RB correspondant à un arbre de jonction $T(C, S)$ est proportionnelle au potentiel conjoint [Kjaerul, 1995], donnée par la formule :

$$P(X) \propto \Phi_X(X) = \frac{\prod_{c \in C} \Phi_c(c)}{\prod_{s \in S} \Phi_s(s)} \tag{23}$$

Un potentiel $\Phi_A$ est dit normalisé si $\sum_{a \in domaine_A} \Phi_A(a) = 1$ , c'est-à-dire, $\Phi_A$ représente une distribution conjointe de l'ensemble de variables $A$ (la somme devient une intégrale si $A$ est continu ). Si tous les potentiels des cliques et des séparateurs de l'arbre de jonction sont normalisés, alors le potentiel conjoint $\Phi_X$ est normalisé et dans ce cas la distribution conjointe $P(X)$ est égale au potentiel conjoint, alors nous avons

$$P(X) = \frac{\prod_{c \in C} P_c(c)}{\prod_{s \in S} P_s(s)} \tag{24}$$

où $P_c(c)$ et $P_s(s)$ représentent la distribution conjointe de la clique $c$ et du séparateur $s$, respectivement.

Une propriété intéressante de l'arbre de jonction est que les variables qui forment une clique ou un séparateur sont mutuellement et complètement dépendantes. En exploitant la sparsity structurelle par l'apprentissage de la structure du RB associé, la formule (24) nous permet de contourner le problème de hautes dimensions par rendre l'estimation de $P(X)$ équivalant à l'estimation de densité des cliques et des séparateurs qui forment l'arbre de jonction, ces derniers ont des petites dimensions pour des données sparses.

### 3.2.3 Estimation de densité en utilisant ACI locale et RBs

Par la combinaison des résultats obtenus en équation (9) avec la décomposition de la conjointe (24), on aura

$$P_X(X) = \frac{\prod_{q \in Cli} \sum_{j=1}^{Lq} [\frac{P_C(C_{j,q})}{|detA_{j,q}|} \times \prod_{i=1}^{Kq} P_{S_{i,j,q}}(S_{i,j,q})]}{\prod_{p \in Sep} \sum_{j=1}^{Lp} [\frac{P_C(C_{j,p})}{|detA_{j,p}|} \times \prod_{i=1}^{Kp} P_{S_{i,j,p}}(S_{i,j,p})]} \tag{25}$$

où $S_{i,j,q}$ et $A_{j,q}$ sont la $i^{ème}$ composante indépendante et la matrice de mélange du cluster $j$ de la clique $q$, respectivement. $P_C(C_{j,q})$ est la probabilité a priori du cluster $j$ de la clique $q$. $Lq$ est le nombre de cliques, $Kq$ est le nombre de CIs qui est égal à la taille de la clique (mêmes notations pour le séparateur $p \in Sep$ dans le dénominateur).

Le nombre de clusters $Lq|Lp$ est propre à chaque clique ou séparateur. L'estimation du nombre optimal est faite par la minimisation du critère XB [Xie and Beni, 1991], introduit dans §2.2.4.

## 3.3 Résultats expérimentaux

### 3.3.1 Application en estimation de densité

#### a) Méthodologie d'évaluation

Nous avons réalisé un certain nombre de simulations afin de comparer les performances de notre méthode avec le standard GMM-EM où l'algorithme EM (expectation-maximisation) a été utilisé pour estimer les paramètres du modèle GMM.

Pour chaque dimension $dim = 2, 4, ..., 36, 38$, nous avons généré 20 distributions GMM cibles dont 10 comportent 40 composantes Gaussiennes et 10 comportent 80 composantes Gaussiennes. La moyenne et les éléments de la matrice de variance-covariance sont générés uniformément et aléatoirement dans les intervalles [0 10] et [0 5], respectivement. Les poids de contribution des différentes composantes Gaussiennes sont générés également aléatoirement. Ensuite, pour chaque distribution de densité cible, nous avons simulé 8000 observations dont la moitie a été employée pour l'estimation et l'autre moitie a été consacrée aux tests de prédiction.

Nous avons utilisé le critère NMSE (*Normalized Mean Square Error*) pour la comparaison de performances, il est donné par la formule suivante :

$$NMSE = \frac{\sum_{i=1}^{N}(p_{\mathbf{x}_i} - \hat{p_{\mathbf{x}_i}})^2}{N\ var(p_x)} \tag{26}$$

où $p_{\mathbf{x}_i}$ et $\hat{p_{\mathbf{x}_i}}$ sont la densité réelle et la densité estimée du vecteur de test $\mathbf{x}_i$.

**b) Résultats et discussions**

L'estimateur GMM-EM a été comparé avec les deux configurations suivantes de notre méthode : FCM&FastICA et GK&FastICA. Dans les deux configurations, l'algorithme GS (*Grow-Shrink*) [Margaritis, 2003] a été utilisé pour apprendre la structure RB correspondante. Le nombre de composantes du modèle GMM à estimer a été sélectionné par l'emploi du critère BIC (*Bayesian Information Criterion*).

Les figures 4.18 et 4.19 montrent la moyenne NMSE de 10 simulations distinctes. Pour chaque dimension $dim = 2, 4, ..., 36, 38$ la moyenne NMSE est calculée, le NMSE d'une seule simulation est calculé par la formule (26). Nous constatons que sur les deux figures, les NMSEs moyens des estimateurs GMM et de FCM&FastICA ont une tendance à s'accroître plus rapidement que celui de GK&FastICA en augmentant la dimension.

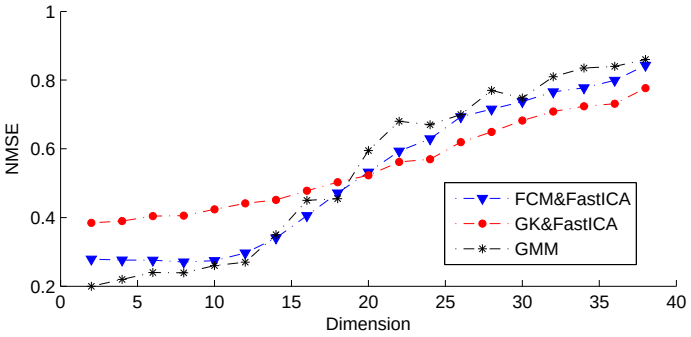



FIGURE 4.18 – Evaluation de performances d'estimation de densité des données moins complexes.

Les lois cibles sont des GMMs plus simples comportant 40 composantes. Pour chaque dimension, le NMSE moyen de 10 simulations est affiché. Dans chaque simulation, 8000 observations sont générées de la loi cible dont 4000 pour l'apprentissage et 4000 pour le test.

Les courbes peuvent être subdivisées en trois régions. Pour les petites dimensions ($dim <$ 14), le NMSE du GMM est clairement plus petit que celui de FCM&FastICA pour des données moins complexes (GMM de 40 composantes sur figure 4.18) et assez proche pour des données plus complexes (GMM de 80 composantes sur la figure 4.19). En revanche, le NMSE de GK&FastICA est significativement élevé (mauvais). Dans la région intermédiaire, ($14 < dim <$ 24) le NMSE du GMM et de FCM&FastICA augmente rapidement et celui de FCM&FastICA

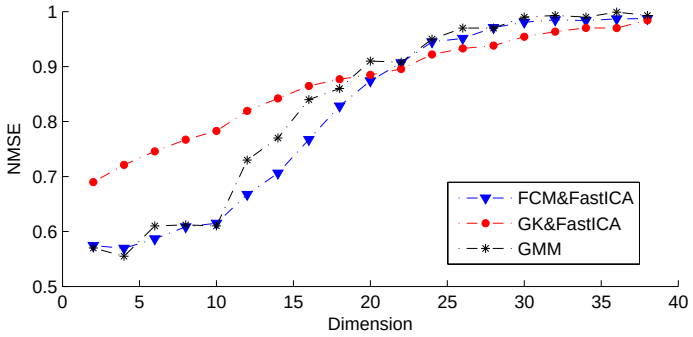



FIGURE 4.19 – Evaluation de performances d'estimation de densité des données plus complexes.

Les lois cibles sont des GMMs plus complexes comportant 80 composantes. Pour chaque dimension, le NMSE moyen de 10 simulations est affiché. Dans chaque simulation, 8000 observations sont générées de la loi cible dont 4000 pour l'apprentissage et 4000 pour le test.

est légèrement meilleur que les deux autres estimateurs, particulièrement pour des données plus complexes (4.19).

Dans la région des hautes dimensions ($dim > 24$), GK&FastICA fournit le NMSE le plus petit. Nous constatons aussi que l'écart entre les 03 courbes est plus clair sur des données complexes (figure 4.19) que sur des données simples (figure 4.18).

En termes de la complexité en temps de calcul, la technique GMM-EM est considérablement plus rapide que notre méthode notamment en hautes dimensions. En effet, l'apprentissage de la structure par l'algorithme *GS* demande beaucoup de ressources calculatoires même si ce dernier ne réalise qu'un sous-ensemble et pas la totalité des tests d'indépendance conditionnelle possibles entre les variables. Les données générées à partir d'un nombre fixe de composantes Gaussiennes (40 ou 80) sont géométriquement plus proches et plus compactes en petites dimensions ce qui explique les bons résultats de FCM&FastICA par rapport à GK&FastICA. Cependant, les données sont plus disperses en hautes dimensions en favorisant l'apparition des clusters elliptiques et hétérogènes ce qui explique les performances de GK&FastICA en hautes dimensions. Les performances de GK&FastICA ont un coût additionnel dans la mesure que l'algorithme de clustering GK a une complexité en temps de calcul plus élevée que FCM.

### 3.3.2 Application en classification supervisée

#### a) Modèle de classification

Dans cette expérience, nous avons opté pour la réalisation des tests de prédiction dans le problème de la classification supervisée car elle permet d'effectuer une comparaison de performances des estimateurs de densité sur des données réelles (non simulées). D'abord, les para-

mètres d'un estimateur de densité $\pi_i$ pour chaque classe $\omega_i (i = 1, ..., Nc)$, basé sur la formule (25), sont estimés à partir des données d'apprentissage.

Les paramètres de $\pi_i$ incluent l'arbre de jonction, les paramètres de clustering, les matrices de mélange ACI locaux et les paramètres d'estimation de densité de chaque composante indépendante. La fonction de décision d'un vecteur $\mathbf{x}$ est basée sur la règle de maximum a posteriori, donnée par

$$Z \leftarrow argmax_i P(\mathbf{x}|\omega_i)P(\omega_i) \tag{27}$$

où $P(\mathbf{x}/\omega_i) = \pi_i(\mathbf{x})$ est la vraisemblance du vecteur $\mathbf{x}$ dans la classe $\omega_i$, calculée en utilisant l'estimateur de densité proposé en (25) et $P(\omega_i)$ est la probabilité a priori de la classe $\omega_i$, généralement prise égale à la fréquence de la classe $\omega_i$ en cas d'absence d'informations a priori.

Nous avons utilisé l'outil 'BNLearn' [Scutari, 2010] pour l'apprentissage de la structure du RB associé. Une brève description des algorithmes employés est fournie dans le tableau 4.7. Pour les algorithmes à base de contraintes, les tests d'indépendance conditionnelle suivants ont été considérés : le test $x^2$ (le test classique de Pearson), le test 'aict' (*Akaike Information Criterion test*), le test 'mi' (*mutual information test*) et le test 'mi-sh'(*mutual information test based on shrinkage estimator*).

Tableau 4.7 – Algorithmes utilisés pour l'apprentissage de la structure RB.

| **Algorithme d'apprentissage** | **Description** | **Type** |
|---|---|---|
| Grow-Shrink (GS) | Basé sur la détection des blankets de Markov en suivant une stratégie de type agrandissement-rétrécissement (*Grow-Shrink*) [Margaritis, 2003]. | A base de contraintes |
| Incremental Association (IAMB) | Basé sur un algorithme de détection des blankets de Markov qui opère en deux phases, une sélection suivie par une tentative de suppression des faux positifs [Tsamardinos et al., 2003]. | A base de contraintes |
| Hill-Climbing (HC) | Une recherche gloutonne de type 'escalade de montagne' (Hill Climbing) dans l'espace des graphes orientés. | A base de score |
| Max-Min Hill-Climbing (MMHC) | Combine l'algorithme Max-Min (Parents et Enfants) pour restreindre l'espace de recherche et l'algorithme 'Hill-Climbing' pour la recherche du graphe optimal dans l'espace restreint. | Hybride |
| Restricted Maximization (RSMAX2) | Une implémentation plus générale de l'algorithme MMHC. | Hybride |

Tableau 4.8 – Description des benchmarks.

| Benchmark | Nombre d'attributs (dimension) | Nombre de classes | Nombre de vecteurs d'apprentissage | Nombre de vecteurs de test |
|---|---|---|---|---|
| SatImage | 36 | 6 | 3104 | 2000 |
| SensIT Vehicle (acoustic) | 50 | 3 | 6000 | 3600 |
| SensIT Vehicle(seismic) | 50 | 3 | 6000 | 3600 |
| Steel fault | 27 | 7 | 1181 | 761 |

**b) Données de test**

Quelques benchmarks standard bien connus dans le domaine de la classification supervisée ont été choisis pour l'évaluation. Les données choisies sont fondamentalement hétérogènes, réelles et définies sur des hautes dimensions ($dim > 20$). Les quatre bases de données impliquées se rapportent à la classification des pixels des images satélitaires, des types de véhicules en utilisant des attributs acoustiques ou des attributs sismiques, et des défauts de tôles d'acier.

En général, la classification des données acoustiques et sismiques des véhicules et la détection des défauts mécaniques sont des problèmes complexes en traitement de signal car les variables observées dépendent de plusieurs facteurs non contrôlés, souvent considérés comme un bruit. Le choix des données acoustiques est motivé par le fait que l'hypothèse de mélange ACI concorde avec la nature plus ou moins additive de sources sonores.

Les trois benchmarks 'SatImage', SensIT Vehicle (acoustic) [Duarte and Hu, 2004] et 'SensIT Vehicle (acoustic)' [Duarte and Hu, 2004] ont été téléchargés de la collection LibSVM disponible sur le lien web [4], alors que le benchmark 'Steel fault' a été téléchargé du site web [5]. voir le tableau 4.8 pour plus de détails.

**b) Résultats expérimentaux**

Nous avons testé a plate-forme proposée toutes les combinaisons possibles des deux algorithmes de clustering flou : GK et FCM, des deux algorithmes ACI : JADE et FastICA, et des algorithmes d'apprentissage de la structure RB présentés dans le tableau 4.7.

Les taux de classification obtenus des 04 bases de données sont montrés dans les tableaux 4.9-4.12, respectivement. Nous remarquons que l'algorithme de clustering GK a donné dans presque toutes les configurations des résultats légèrement meilleurs que l'algorithme FCM. Ce constat est en accord avec notre thèse qui suggère que les clusters elliptiques sont plus convenables à décrire les données dans la méthode ACI locale que les clusters sphériques.

En ce qui concerne le module de séparation ACI, il est évident que les performances ACI

4. http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/
5. http://archive.ics.uci.edu/ml/datasets/Steel+Plates+Fault

Tableau 4.9 – Taux de classification de la base de données 'SatImage'.

| Algo. de clus. flou | Apprentissage de structure | JADE | | | | FastICA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $x^2$ | aict | mi | mi-sh | $x^2$ | aict | mi | mi-sh |
| GK | RSMAX2 | 0.9042 | 0.9042 | 0.9028 | 0.9042 | 0.9056 | 0.9070 | 0.9042 | 0.9035 |
| | MMHC | 0.8881 | 0.9042 | 0.9042 | 0.9042 | 0.8853 | **0.9098** | 0.9070 | 0.9056 |
| | GS | 0.9035 | 0.9035 | 0.9035 | 0.9035 | 0.9049 | 0.9049 | 0.9021 | 0.9063 |
| | IAMB | 0.8972 | 0.9035 | 0.9035 | 0.9035 | 0.8937 | 0.9035 | 0.9091 | 0.9028 |
| | FASTIAMB | 0.8860 | 0.8995 | 0.8916 | 0.8867 | 0.8853 | 0.8951 | 0.8909 | 0.8902 |
| | HC | 0.8986 | | | | 0.8965 | | | |
| FCM | RSMAX2 | 0.9007 | 0.9007 | 0.9007 | 0.9007 | 0.8986 | 0.9000 | 0.8986 | 0.8986 |
| | MMHC | 0.8902 | 0.9007 | 0.9007 | 0.9007 | 0.8923 | 0.8979 | 0.8972 | 0.9000 |
| | GS | 0.8979 | 0.8979 | 0.8979 | 0.8979 | 0.8958 | 0.8972 | 0.8951 | 0.8986 |
| | IAMB | 0.8867 | 0.8979 | 0.8979 | 0.8979 | 0.8860 | 0.8944 | 0.9000 | 0.8937 |
| | FASTIAMB | 0.8874 | 0.8860 | 0.8937 | 0.9000 | 0.8916 | 0.8979 | 0.8916 | **0.9021** |
| | HC | 0.8986 | | | | 0.9042 | | | |

dépendent plus à combien les approximations de la fonction objectif et l'algorithme d'optimisation sont convenables aux caractéristiques statistiques de données. Dans ce sens, nous constatons que FastICA est légèrement meilleur que JADE pour les 04 bases de données.

L'apprentissage de la structure RB affecte sérieusement les performances de la classification. On constate bien que les algorithmes d'apprentissage de la structure ne sont pas stables, chaque algorithme donne des bons taux pour certaines bases de données et des taux modestes pour d'autres. Ce constat peut être expliqué par le fait que tous les algorithmes d'apprentissage de structure RB sont des heuristiques, qui sont appliqués sur une quantité limitée de données discrétisées. En conséquence, il n'y pas une garantie d'avoir une structure RB optimale et les structures obtenues ne sont réellement que des approximations qui expriment partiellement les relations d'indépendance conditionnelle. Par exemple, la structure $S_1$ (où l'arbre de jonction est construit manuellement par le rangement des coefficients fréquentiels voisins dans les mêmes cliques comme suit $Cliques = \{(1,2),(3,4),...,(49,50)\}$ et $S parateur = \{\}$) de la base de données 'SensIT Vehicle acoustic' a donné des taux de classification nettement meilleurs que les algorithmes d'apprentissage automatiques (voir le tableau 4.10).

Nous constatons également que les taux de classification de la base de données 'Steel fault', affichés dans le tableau 4.12, varient fortement dans l'intervalle [0.5243 0.6990] d'une configuration à une autre. Cette variance est due au fait que le nombre de vecteurs par classe (environ 168 vecteurs) est très faible en le comparant à la dimension 20. Le modèle proposé contient plusieurs paramètres et nécessite donc un nombre satisfaisant de données d'apprentissage.

Les meilleurs résultats obtenus par le modèle proposé, ACI&RB, sont comparés à ceux obtenus par les classificateurs populaires SVM et KNN (K-plus proche voisins). Le paramètre K de KNN a été estimé empiriquement en utilisant un échantillon de données d'apprentissage. Les meilleurs résultats de SVM ont été eu avec le noyau Gaussien (RBF) où les paramètres sigma de la fonction noyau et le coût ont été fixés empiriquement par la recherche des valeurs

Tableau 4.10 – Taux de classification de la base de données 'SensIT Vehicle (acoustic)'.

| Algo. de clus. flou | Apprentissage de structure | JADE | | | | FastICA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $x^2$ | aict | mi | mi-sh | $x^2$ | aict | mi | mi-sh |
| GK | RSMAX2 | 0.7200 | 0.6972 | 0.6933 | 0.6988 | 0.7125 | 0.6969 | 0.6966 | 0.6986 |
| | MMHC | 0.7094 | 0.6966 | 0.6944 | 0.6922 | 0.6966 | 0.6836 | 0.6944 | 0.6944 |
| | GS | 0.7163 | 0.7025 | 0.7191 | 0.7172 | 0.7144 | 0.6983 | 0.7205 | 0.7150 |
| | IAMB | 0.6847 | 0.7025 | 0.6916 | 0.6927 | 0.7025 | 0.6925 | 0.6975 | 0.6952 |
| | FASTIAMB | 0.6863 | 0.7138 | 0.7097 | 0.7038 | 0.6766 | 0.7036 | 0.7022 | 0.6980 |
| | HC | 0.7038 | | | | 0.7000 | | | |
| | S1 | **0.7502** | | | | 0.7494 | | | |
| FCM | RSMAX2 | 0.7183 | 0.6969 | 0.6966 | 0.7061 | 0.7186 | 0.6966 | 0.6944 | 0.7000 |
| | MMHC | 0.7016 | 0.7033 | 0.6941 | 0.6947 | 0.6980 | 0.6952 | 0.6913 | 0.6941 |
| | GS | 0.7205 | 0.7116 | 0.7200 | 0.7188 | 0.7233 | 0.7036 | 0.7191 | 0.7141 |
| | IAMB | 0.7036 | 0.7116 | 0.6988 | 0.6805 | 0.7022 | 0.7077 | 0.6961 | 0.6825 |
| | FASTIAMB | 0.7141 | 0.7080 | 0.7113 | 0.6919 | 0.7077 | 0.7030 | 0.7088 | 0.6950 |
| | HC | 0.6922 | | | | 0.6994 | | | |
| | S1 | **0.7333** | | | | 0.7330 | | | |

optimales sur des intervalles larges. Le tableau 4.13 montre un comparatif des taux obtenus. Les taux de ACI&RB sont comparables avec ceux de SVM; meilleurs pour les deux bases de données 'SensIT Vehicle (acoustic)' et 'Steel fault' et légèrement inférieurs pour les bases SensIT Vehicle (seismic) et SatImage.

Nous avons expérimenté seulement les deux méthodes ACI populaires JADE et FastICA dans notre modèle ACI&RB, il est probable que d'autres méthodes ACI puissent être plus adéquates à ces données. En effet, un avantage essentiel de la plate-forme proposée réside dans son aspect générique permettant de se servir des algorithmes ACI différents conçus pour des types de données spécifiques. Par exemple, les méthodes ACI parcimonieuses telle que SCA [Georgiev et al., 2004]conviennent bien aux données parcimonieuses.

# 4 Conclusion

La méthode ACI standard (linéaire) peut être utilisée pour ajuster les données à condition que les l'hypothèse du mélange ACI soit valide et également que l'algorithme ACI (FastICA, JADE, etc) convient aux données. En effet, les méthodes ACI s'appuient en pratique sur des approximations différentes pour la mesure de la dépendance statistique (information mutuelle). Ainsi, l'affinité d'une méthode ACI découle de la validité des approximations faites ainsi que de la technique d'optimisation pour la recherche de la solution optimale. En exploitant un échantillon de données d'apprentissage, l'utilisateur peut comparer des algorithmes ACI divers et empiriquement détermine celui qui marche bien avec les données en mains.

Bien que la méthode ACI standard couvre plusieurs applications du monde réel, elle ne

Tableau 4.11 – Taux de classification de la base de données 'SensIT Vehicle (seismic)'.

| Algo. de clus. flou | Apprentissage de structure | JADE | | | | FastICA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $x^2$ | aict | mi | mi-sh | $x^2$ | aict | mi | mi-sh |
| GK | RSMAX2 | 0.6580 | 0.6627 | 0.6527 | 0.6569 | 0.6619 | 0.6602 | 0.6544 | 0.6566 |
| | MMHC | 0.6627 | 0.6644 | 0.6644 | 0.6552 | 0.6655 | 0.6583 | 0.6558 | 0.6550 |
| | GS | 0.6686 | 0.6527 | 0.6663 | 0.6669 | **0.6702** | 0.6513 | 0.6613 | 0.6658 |
| | IAMB | 0.6588 | 0.6525 | 0.6519 | 0.6563 | 0.6533 | 0.6508 | 0.6586 | 0.6555 |
| | FASTIAMB | 0.6452 | 0.6527 | 0.6502 | 0.6613 | 0.6516 | 0.6513 | 0.6536 | 0.6561 |
| | HC | 0.6591 | | | | 0.6594 | | | |
| FCM | RSMAX2 | 0.6600 | 0.6613 | 0.6591 | 0.6561 | 0.6666 | 0.6597 | 0.6572 | 0.6605 |
| | MMHC | 0.6661 | 0.6630 | 0.6630 | 0.6594 | 0.6613 | 0.6572 | 0.6594 | 0.6516 |
| | GS | 0.6608 | 0.6508 | 0.6611 | 0.6619 | 0.6613 | **0.6669** | 0.6591 | 0.6555 |
| | IAMB | 0.6527 | 0.6475 | 0.6586 | 0.6555 | 0.6611 | 0.6577 | 0.6583 | 0.6555 |
| | FASTIAMB | 0.6516 | 0.6591 | 0.6536 | 0.6561 | 0.6591 | 0.6575 | 0.6536 | 0.6561 |
| | HC | 0.6594 | | | | 0.6594 | | | |

peut pas être utilisée pour décrire des données provenant des distributions quelconques. En revanche, les méthodes ACI non linéaires ont théoriquement le potentiel d'ajuster des données plus générales, mais leur espace de solutions est tellement étendu qu'on risque de converger vers des optima locaux, vers des solutions non admissibles.

Parmi les méthodes ACI non linéaires, la méthode ACI locale [Karhunen and Malaroiu, 1999] semble être une bonne option pour la représentation de données. Dans laquelle, le module de clustering est un composant nom-paramétrique qui est responsable de la prise en compte de l'aspect non linéaire et à la prestation d'une représentation globale, alors que les modèles ACI linéaires locaux décrivent localement les données en plus de détails.

Les RBs et l'ACI locale peuvent être utilisés pour la représentation de données multivariables, chacun d'eux a ses points forts, ses limites et ses restrictions. Les RBs ont l'avantage d'avoir une séparation explicite entre les paramètres qualitatifs (DAG) et les paramètres quantitatifs (distributions conditionnelles). En plus des informations fournies par l'expert, plusieurs algorithmes d'apprentissage de la structure DAG ont été développés récemment. En domaine discret, l'apprentissage des TPCs souffre du problème de sur-apprentissage [Hawkins, 2004] si la résolution de discrétisation est importante. En domaine continu, il n'existe pas des méthodes pratiques et générales pour l'apprentissage des distributions conditionnelles, et dans ce cas de figure, le recours à la discrétisation n'est pas toujours la bonne option car il en découle une perte d'information plus le risque d'avoir des TPCs trop volumineuses.

En contrepartie, les méthodes ACI ont l'avantage de pouvoir analyser des données continues et non Gaussiennes. En outre, les modèles ACI sont des modèles semi-paramétriques; des algorithmes différents ont été développés pour des types de données spécifiques. La sparsity structurelle, qui manifeste sous forme d'indépendance conditionnelle entre les variables observées, est souvent présente dans les données réelles. Malheureusement, la majorité des méthodes ACI procèdent par l'estimation des paramètres qualitatifs (structurels) et quantitatifs au même

Tableau 4.12 – Taux de classification de la base de données 'Steel fault'.

| Algo. de clus. flou | Apprentissage de structure | JADE $x^2$ | JADE aict | JADE mi | JADE mi-sh | FastICA $x^2$ | FastICA aict | FastICA mi | FastICA mi-sh |
|---|---|---|---|---|---|---|---|---|---|
| GK | RSMAX2 | 0.6333 | 0.6452 | 0.6662 | 0.6465 | 0.6294 | 0.6425 | 0.6386 | 0.6741 |
| | MMHC | 0.6373 | 0.6333 | 0.6346 | 0.6517 | 0.6517 | 0.6452 | 0.6530 | 0.6530 |
| | GS | 0.6333 | 0.6636 | 0.6688 | 0.6136 | 0.6452 | 0.6438 | 0.6675 | 0.6057 |
| | IAMB | 0.6425 | 0.6465 | 0.6346 | 0.6386 | 0.6491 | 0.6636 | 0.5808 | 0.6241 |
| | FASTIAMB | 0.6688 | 0.6215 | 0.6110 | 0.6609 | **0.6990** | 0.6517 | 0.5939 | 0.6859 |
| | HC | 0.6373 | | | | 0.6688 | | | |
| FCM | RSMAX2 | 0.6031 | 0.6307 | 0.6123 | 0.6097 | 0.5978 | 0.6189 | 0.6254 | 0.6189 |
| | MMHC | 0.6491 | 0.6452 | 0.6320 | 0.6189 | 0.6044 | 0.6294 | 0.6386 | 0.6268 |
| | GS | 0.6320 | 0.6281 | 0.6425 | 0.6097 | 0.6189 | 0.6018 | 0.6517 | 0.5992 |
| | IAMB | 0.6452 | 0.6360 | 0.5900 | 0.6228 | 0.6044 | 0.6136 | 0.6084 | 0.6346 |
| | FASTIAMB | 0.6412 | 0.6254 | 0.5243 | 0.6596 | 0.6070 | 0.6517 | 0.5952 | **0.6675** |
| | HC | 0.6123 | | | | 0.6162 | | | |

&

Tableau 4.13 – Comparaison avec les classificateurs SVM et KNN.

| Dataset | SVM | KNN | ACI&RB |
|---|---|---|---|
| SatImage | **0.9099** | 0.8903 | 0.9098 |
| SensIT Vehicle (acoustic) | 0.7401 | 0.7192 | **0.7502** |
| SensIT Vehicle (seismic) | **0.6844** | 0.6358 | 0.6702 |
| Steel fault | 0.6662 | 0.6045 | **0.6990** |

temps et n'exploitent pas pleinement ce type de sparsity. Ce problème s'amplifie notamment en hautes dimensions.

De ce point de vue, les modèles RB et ACI sont complémentaires. Dans ce chapitre, nous avons essayé de proposer des modèles combinant les RBs et l'ACI où les paramètres du modèle ACI jouent le rôle des distributions conditionnelles des RBs. L'objectif de ce chapitre est sonder la qualité de représentation de données réelles par les méthodes ACI, en particulier la méthode ACI locale. Des tests de la classification supervisée réalisés sur des applications différentes (reconnaissance du locuteur par la voix, par la voix et le mouvement des lèvres, reconnaissance des types de véhicules par le son, etc.) ont fourni des résultats acceptables.

Notre objectif dans les deux chapitres suivants consiste à proposer des nouveaux algorithmes de propagation de croyances s'exécutant sur les modèles ACIs. En d'autres termes, nous visons la proposition de nouveaux algorithmes d'inférence Bayésienne sur des données continues et non Gaussiennes.

Chapitre 5

# Inférence sur le modèle ACI linéaire (application dans la fusion de scores multi-biométriques

## Sommaire


**1 Introduction . . . . . 120**
**2 Énonce du problème . . . . . 121**
**3 Solution par les réseaux Bayésiens discrets . . . . . 122**
**4 Algorithme d'inférence sur le modèle ACI . . . . . 122**
4.1 Modèle proposé (ACI-RB) . . . . . 122
4.2 Apprentissage des paramètres . . . . . 124
4.3 Algorithme d'inférence . . . . . 125
4.4 Expérimentation . . . . . 129
4.4.1 Méthodologie . . . . . 129
4.4.2 Datasets . . . . . 130
4.4.3 Résultats . . . . . 131
4.4.4 Analyse de complexité . . . . . 132
**5 Application pour la fusion des scores multi-biométriques . . . . . 133**
5.1 Fusion des scores multi-biométriques . . . . . 133
5.2 Scores absents ou de mauvaise qualité . . . . . 134
5.3 Approche proposée . . . . . 134

# 1 Introduction

Le défi principal dans les problèmes d'apprentissage artificiel est de prédire avec le moins possible d'erreurs les valeurs vraisemblables d'une (ou plusieurs) variable(s) d'intérêt que l'on ne peut pas mesurer directement étant données les valeurs d'autres variables sous-jacentes. Parmi les outils conçus particulièrement pour répondre d'une manière formelle à ce problème d'inférence figurent les réseaux Bayésiens [Pearl, 1988] dont le fonctionnement s'appuie sur la fameuse règle de Bayes.

Lorsqu'un RB ne comporte que des valeurs discrètes, les TPCs sont utilisées pour la quantification du modèle. Dans la littérature, plusieurs algorithmes d'inférence exacte sont publiés [Shenoy and Shafer, 1990, Lauritzen and Spiegelhalter, 1988] dont la majorité adoptent une approche du calcul local (voir chapitre I). Les méthodes exactes ne sont pratiques que pour le cas des petits RBs ; elles ne peuvent pas être utilisées si le graphe est dense (contenant un grand nombre d'arcs) et/ou grand (contenant un grand nombre de nœuds) ou si la résolution de discrétisation (nombre de valeurs par variable) est importante à cause de l'explosion de la complexité spatio-temporelle.

Toutefois, la prise en compte des variables continues jusqu'à nos jours représente une limitation majeure des RBs. La solution triviale consistant à discrétiser les domaines des variables continues [Kozlov and Koller, 1997, Friedman and Goldszmidt, 1996] n'est pas malheureusement toujours convenable. Dans certaines applications, une haute résolution de discrétisation est demandée et ainsi une gestion délicate du compromis complexité-précision s'impose. Comme alternative, peu de solutions sans discrétisation sont simplement proposées pour des types spécifiques de RBs continus ou hybrides.

Le modèle Gaussien linéaire conditionnel (CLG pour "*Conditional linear Gaussian*") [Lauritzen, 1992] est un exemple typique des RBs hybrides contenant des variables discrètes et continues, avec les deux restrictions suivantes : (i) la distribution conjointe des variables continues étant donnée la valeur prise par les variables discrètes est supposée être une Gaussienne multivariable. (ii) dans le graphe DAG associé, les nœuds discrets ne doivent pas avoir de nœuds parents continus. Des algorithmes d'inférence exacte pour les modèles CLGs sont proposés par [Lauritzen and Jensen, 2001] et par [Cowell, 2005].

Une autre classe importante des RBs hybrides utilise un mélange d'exponentielles tronquées (MTE pour "*mixture of truncated exponentials*") [Moral et al., 2001] pour modéliser les lois conditionnelles. Les modèles MTEs peuvent être vus comme une généralisation de la discrétisation où le domaine des variables continues est partitionné en un ensemble d'hypercubes et puis un mélange d'exponentielles est utilisé pour approximer la conjointe à l'intérieur de chaque hypercube, cela représente en d'autres termes une sorte de compromis entre la discrétisation en hypercubes et une fonction de densité analytique locale au niveau de chaque hypercube. Un bon modèle MTE est celui qui comporte peu d'hypercubes et peu d'exponentielles et qui approxime le mieux possible la conjointe. Cependant l'estimation des approximations MTEs n'est

pas toujours une mince tâche, la recherche de la meilleure approximation MTE dans un graphe important est extrêmement difficile dans la mesure où cela nécessite la résolution des problèmes d'optimisation non linéaires en hautes dimensions au sein de chaque hypercube, ainsi l'algorithme d'optimisation peut converger vers des optima locaux [Romero et al., 2006].

Même si le modèle ACI linéaire [Hyvärinen et al., 2001] apparaît dans sa forme assez simple, il offre une représentation compacte d'une large classe de données multivariables, ses paramètres englobent seulement la matrice de mélange et les distributions de densité des CIs. Dans le chapitre précédent, nous avons fait surgir que les modèles ACI peuvent être à la base de plusieurs représentations acceptables de données réelles.

Parfois, il est intéressant de réaliser des tâches d'inférence Bayésienne sur les variables ACI observées à l'instar des RBs où la matrice de mélange et les densités des CIs correspondent au graphe DAG et aux TPCs du RB équivalent, respectivement. Dans ce chapitre, par l'exploitation de la propriété d'indépendance mutuelle des variables latentes du modèle ACI, nous concevons un algorithme d'inférence Bayésienne pour le modèle ACI linéaire que nous désignons par le sigle 'ACI-RB' dans ce qui suit.

introduction

# 2 Énonce du problème

Soit $X = \{x_1, ..., x_n\}$ un ensemble de $n$ variables aléatoires continues en supposant que certaines variables de $X$ sont statistiquement dépendantes, il existe donc plusieurs paires de variables $(x_i, x_j / i \neq j : 1, ..., n)$ tel que $P(x_i, x_j) \neq P(x_i) \times P(x_j)$. Dans ce cas d'étude, nous considérons seulement le cas où les relations de dépendance statistique à l'intérieur de $X$ peuvent être approximées par une combinaison linéaire d'un ensemble de CIs $S = \{s_1, ..., s_m\}$ ; autrement dit, par un modèle ACI linéaire et instantané : $\mathbf{x} = (x_1, ..., x_n)^T = A \times (s_1, ..., s_m)^T = A\mathbf{s}$. Dans un but de simplification, nous supposons que le nombre des composantes $m$ est égal au nombre des variables observées $n$.

Notre objectif consiste à calculer la probabilité a posteriori d'une variable d'intérêt $y \in X$ sachant que les valeurs d'un sous-ensemble de variables $Z \subset (X - \{y\})$ sont connues ($Z$ sont communément appelées des évidences). Du fait que les variables observées sont continues, il est important en pratique que les valeurs prises par la variable cible $y$ ou par les évidences $Z$, soient tolérées d'être des intervalles quelconques dans leurs domaines. Formellement, nous voulons estimer

$$P(y \in \Delta_y | z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k}) \tag{1}$$

qui est la probabilité que la variable $y$ soit instanciée dans l'intervalle $\Delta_y$ sachant que la variable de l'évidence $z_e (e = 1, ..., k)$, après être mesurée, est susceptible de prendre une valeur à l'intérieur de l'intervalle $\Delta_{z_e}$. Il convient de noter que les variables considérées, $X$ , ne sont pas

Gaussiennes, elles peuvent avoir des distributions relativement proches de la Gaussienne (cette restriction se rapporte aux hypothèses du modèle ACI).

# 3 Solution par les réseaux Bayésiens discrets

Les dépendances entre les variables de $X$ peuvent être modélisées par un RB, dans lequel il n y a pas un arc entre deux nœuds disjoints $x_i$ et $x_j$ seulement si la variable $x_i$ est conditionnellement indépendante de $x_j$ étant données les autres variables $X - \{x_i, x_j\}$. En d'autres mots, $x_i$ et $x_j$ ne sont pas connectés si $P(x_i|(X - \{x_i, x_j\})) = P(x_i|(X - \{x_i\}))$; c'est-à-dire si $x_j$ ne comporte pas bien plus d'informations que l'ensemble $X - \{x_i, x_j\}$ sur la variable $x_i$.

Si nous optons pour les RBs, nous serons contraints de discrétiser les variables continues car les RBs ne sont pas bien formulés en domaine continu. En revanche, les modèles CLGs ne sont pas appropriés pour notre cas car les variables considérées dans notre cas n'ont pas des distributions Gaussiennes.

Une fois les variables discrétisées, les TPCs sont souvent utilisées pour quantifier les RBs. Avec une faible résolution de discrétisation, nous perdons une partie de l'information et nous perdons aussi dans la flexibilité de définir des intervalles assez personnalisés dans la requête d'inférence (1). En revanche, une haute résolution entraîne une explosion exponentielle de complexité en espace mémoire nécessaire pour le stockage des TPCs et en temps d'inférence, et également aggrave le problème de sur-apprentissage [Hawkins, 2004] par le fait que le nombre de cases des TPCs à estimer devient démesuré comparé avec la quantité de données d'apprentissage.

Comme alternative aux RBs discrets à base de TPCs, un algorithme d'inférence Bayésienne exacte s'exécutant sur l'espace sous-jacent des CIs est proposé dans la section 4.

# 4 Algorithme d'inférence sur le modèle ACI

## 4.1 Modèle proposé (ACI-RB)

Soit $X = \{x_1, ..., x_n\}$ un ensemble de $n$ variables aléatoires continues et qui peuvent être approximées par un modèle ACI linéaire et instantané, $\mathbf{x}(t) = A\mathbf{s}(t)$. Notre objectif consiste à développer un algorithme de propagation de croyances sur les variables observées. En effet, les variables de $X$ sont caractérisées par un ensemble de relations mutuelles de dépendances statistiques, où les CIs $S$ sont perçues comme des sources ou des causes qui expliquent les liens existants au sein de $X$. En faibles dimensions, la matrice de mélange $A$, estimée par l'ACI, ne contient pas généralement de valeurs nulles, ou contient peu de valeurs proches de zéro. La faible sparsity de la matrice $A$ implique que la plupart des variables de $X$ sont condi-

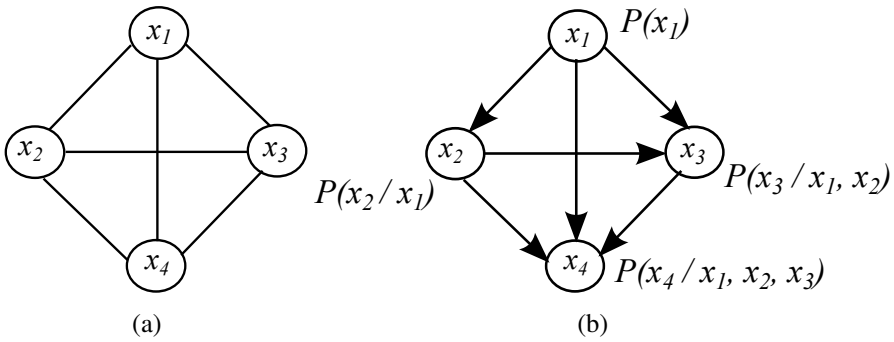



FIGURE 5.1 – Exemple d'un mélange ACI en dimension quatre.

(**a**) Les dépendances entre les variables ACI observées sous forme d'un cluster. (**b**) Un réseau Bayésien pour modéliser les variables ACI observées.

tionnellement dépendantes entre elles ; en d'autres termes, pour la majorité des paires $(x_i, x_j)$ $i \neq j : 1, ..., n$, une part de l'information concernant $x_j$ est contenue dans $x_i$ et n'est pas redondante dans le reste des variables $X - \{x_i, x_j\}$.

L'indépendance statistique est par définition un concept non orienté (symétrique), ainsi les relations entre les variables de $X$ peuvent être modélisées par un cluster complet comme le montre la figure 5.1(a). Ces relations peuvent être aussi modélisées par un RB tel que pour chaque couple $i < j$ $(i, j : 1, ..., n)$, il existe un arc du nœud $x_i$ vers le nœud $x_j$ et une fonction de distribution conditionnelle $P(x_i, Parents_{x_i})$ associée à chaque nœud $x_i$. Voir l'exemple de la figure 5.1(b).

Selon les hypothèses du modèle ACI linéaire, et après l'introduction des CIs résultantes de l'estimation ACI, toutes les variables $X \cup S$ peuvent être modélisées par un RB dans lequel:

1. Il y a un arc allant de chaque variable latente $s_i \in S$ vers chaque variable observée $x_j \in X$.

2. Il n y pas d'arcs entre les variables de $S$ parce qu'elles sont supposées être mutuellement et statistiquement indépendantes.

3. Il n y pas d'arcs entre les variables de $X$ parce que chaque paire variables $x_i \neq x_j$ sont conditionnellement indépendantes lorsque les valeurs de leurs parents communs $S$ sont données, $x_i \perp x_j | S (i \neq j : 1, ..., n)$.

L'exemple affiché sur la figure 5.2 illustre le modèle proposé ACI-RB. Les paramètres du modèle comportent la matrice estimée $A$ du mélange ACI ainsi que les distributions de densité de probabilité unidimensionnelle $P(s_i)$ de chaque CI $s_i \in S$.

Dans la suite de ce chapitre, nous utilisons le terme "variables observées" pour désigner $X$, "variables mesurées" pour désigner les évidences $Z \subset X$, et "sources", "causes" ou "composantes indépendantes" pour désigner $S$.

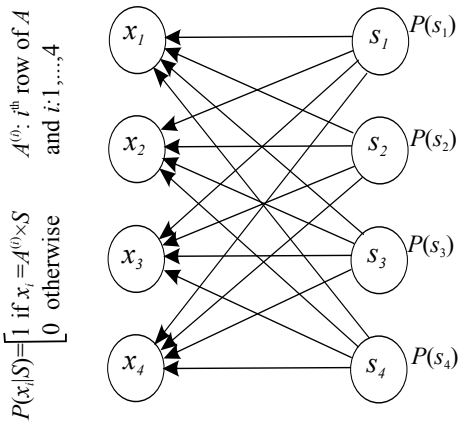



FIGURE 5.2 – Modèle proposé ACI-RB. Exemple de quatres variables.

$X$ : variables observées. $S$ : composantes indépendantes estimées par l'ACI.

## 4.2 Apprentissage des paramètres

Nous disposons d'un échantillon de données d'apprentissage comportant $T$ vecteurs rangés sous forme d'une matrice $X_{app}$ de taille $n \times T$. La tâche d'apprentissage des paramètres du modèle proposé consiste en les étapes suivantes :

1. Estimation de la matrice $A$ de mélange ACI en utilisant un algorithme ACI approprié.
2. Calcul des CIs, $S_{app} = A^{-1} \times X_{app}$.
3. Estimation de la densité de probabilité $P(s_i)$ de chaque composante $s_i \in S$ en utilisant les données $S_{app}$.

Notre objectif ne réside pas dans la reconstitution de sources comme dans la BSS [Amari and Yang, 1996], mais plutôt nous voulons identifier des composantes qui soient indépendantes le maximum possible. Même si cette relaxation rend l'estimation ACI plus facile que le problème BSS, il n'y a pas un algorithme universel qui soit toujours performant; en effet, certains algorithmes ACI sont plus performants que d'autres pour chaque type de données. La sélection d'un algorithme approprié doit être soigneusement menée pour avoir un bon ajustement ACI.

Il est bien de mentionner que la difficulté et la complexité de l'algorithme d'inférence dépend évidement de la fonction adoptée pour approximer la densité de probabilité de chaque composante indépendante. Dans un premier temps et pour raison de simplicité, nous avons adopté la technique non paramétrique d'estimation en utilisant des histogrammes. Le domaine de chaque CI est subdivisé en un ensemble fini d'intervalles et la densité de probabilité à l'intérieur de chaque intervalle est approximée par une distribution uniforme. La technique MAP peut être appliqué pour estimer les valeurs d'histogrammes. L'algorithme d'inférence qui correspond à ce choix est détaillé dans la section suivante (§4.3).

## 4.3 Algorithme d'inférence

Nous rappelons que notre objectif est de calculer d'une manière efficace la probabilité a posteriori d'une variable cible non mesurée $y \in X$ sachant que les valeurs prises par un sous-ensemble disjoint des variables $Z = \{z_1, ..., z_k\} \subset (X - \{y\})$ sont connues. Nous traitons le cas où les valeurs de la variable cible et des évidences peuvent être des intervalles quelconques dans leurs domaines correspondants. La probabilité a posteriori $P(y \in \Delta_y | z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k})$ peut être évaluée en utilisant la règle d'inférence de Bayes comme suit :

$$P(y \in \Delta_y | z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k}) = \frac{P(y \in \Delta_y, z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k})}{P(z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k})} \tag{2}$$

Au numérateur comme au dénominateur de (2), nous évaluons des termes ayant la même forme : $\alpha = P(v_1 \in \Delta_{v_1}, ..., v_l \in \Delta_{v_l})$. L'évaluation d'un tel terme peut être réalisée par une intégration multiple de la distribution conjointe $P(v_1, ..., v_l)$ sur les intervalles fixés $\Delta_{v_i}, i : 1, ..., l$ comme suit :

$$\alpha = P(v_1 \in \Delta_{v_1}, ..., v_l \in \Delta_{v_l}) = \int_{\Delta_{v_1}} ... \int_{\Delta_{v_l}} P(v_1, ..., v_l) dv_1 ... dv_l \tag{3}$$

En étendant le calcul à l'ensemble de CIs, $\alpha$ peut être évalué par la marginalisation de la distribution conjointe $P(v_1, ..., v_l, s_1, ..., s_n)$ comme suit :

$$\alpha = \int_{s_1=-\infty}^{s_1=+\infty} ... \int_{s_n=-\infty}^{s_n=+\infty} \int_{\Delta_{v_1}} ... \int_{\Delta_{v_l}} P(v_1, ..., v_l, s_1, ..., s_n) dv_1 ... dv_l ds_1 ... ds_n \tag{4}$$

La conjointe $P(v_1, ..., v_l, s_1, ..., s_n) = 0$ si au moins une variable $s_i \in S$ est en dehors de son domaine, ainsi les limites de l'intégrale (4) peuvent être simplifiées

$$\alpha = \int_{\phi_{s_1}} ... \int_{\phi_{s_n}} \int_{\Delta_{v_1}} ... \int_{\Delta_{v_l}} P(v_1, ..., v_l, s_1, ..., s_n) dv_1 ... dv_l ds_1 ... ds_n \tag{5}$$

où $\phi_{s_i}$ est le domaine de composante $s_i \in S$.

$V = \{v_1, ..., v_l\}$ est un sous-ensemble de variables observées $X$. Selon l'hypothèse du modèle ACI ($\mathbf{x} = A\mathbf{s}$), les variables de $V$ sont des fonctions linéaires des variables $S$ et l'évaluation de $\alpha$ revient à intégrer la conjointe $P(s_1, .., s_n)$ dans l'espace multidimensionnel sous-jacent des CIs sur l'hyper-intervalle $\Delta_v$ défini comme étant l'intersection des intervalles fixés des variables de $V$, $\Delta_v = \Delta_{v_1} \bigcap ... \bigcap \Delta_{v_l}$. Nous aurons donc

$$\alpha = \int \int_{\Delta_v} P(s_1, ..., s_n) ds_1 ... ds_n \tag{6}$$

$$\alpha = \int_{\phi_{s_1}} ... \int_{\phi_{s_n}} P(s_1, ..., s_n) Q_{vs} ds_1 ... ds_n \tag{7}$$

avec $Q_{vs} = 1$ si le point $(s_1, ..., s_n)$ se situe à l'intérieur de l'hyper-intervalle $\Delta_v$ et $Q_{vs}$=0 ailleurs.

Il est encore possible de réduire les limites inférieures et supérieures de chaque composante $s_i \in S$ dans (7). En effet, nous avons $s_i = w_{i,1}x_1 + ... + w_{i,n}x_n$ où $W = A^{-1}$ est la matrice estimée de séparation ACI de taille $n \times n$.

En intégrant, chaque variable $x_j \in V$ impliquée dans $\alpha$ varie à l'intérieur de son intervalle $\Delta_{x_j}$, alors qu'une variable libre $x_j \in (X - V)$ peut varier le long de son domaine $\phi_{x_j}$. En conséquence, les limites inférieure $a_i$ et supérieure $b_i$ dans l'intégrale (7) de la composante $s_i$ peuvent être rétrécies en utilisant l'algorithme montré sur le tableau 5.1 .

Tableau 5.1 – Calcul des limites d'intégration de la composante $s_i$

**Input**: $V$: variables instanciées dans $\alpha$
$\Delta_v$: intervalles des évidences
$\phi_x$: domaines des variables observées
$W$: matrice estimée $n \times n$ de séparation ACI
**Output**: Limite inférieure $a_i$ et supérieure $b_i$ de la variable $s_i$

$a_i \leftarrow 0$; $b_i \leftarrow 0$;
**for** chaque variable observée $x_j$ $(j = 1, ..., n)$ do
  **if** $(W_{i,j} > 0)$ **then**
    **if** $x_j \in V$ **else**
      $a_i \leftarrow a_i + W_{i,j} \times min(\Delta_{x_j})$; $b_i \leftarrow b_i + W_{i,j} \times max(\Delta_{x_j})$;
      //$min(\Delta)$ and $max(\Delta)$ sont les limites inférieure et supérieure de l'intervalle $\Delta$
    **else**
      $a_i \leftarrow a_i + W_{i,j} \times min(\phi_{x_j})$; $b_i \leftarrow b_i + W_{i,j} \times max(\phi_{x_j})$;
    **end**
  **else**
    **if** $x_j \in V$ **then**
      $a_i \leftarrow a_i + W_{i,j} \times max(\Delta_{x_j})$; $b_i \leftarrow b_i + W_{i,j} \times min(\Delta_{x_j})$;
    **else**
      $a_i \leftarrow a_i + W_{i,j} \times max(\phi_{x_j})$; $b_i \leftarrow b_i + W_{i,j} \times min(\phi_{x_j})$;
    **end**
  **end**
**end**

$Q_{vs} = 0$ si au moins une composante $s_i$ est en dehors du nouvel intervalle calculé $[a_i, b_i]$. L'intégrale (7) est alors équivalente à

$$\alpha = \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} P(s_1, ..., s_n) Q_{vs} ds_1 ... ds_n \tag{8}$$

Nous avons utilisé des histogrammes pour l'approximation des densités des composantes $S$. Supposons que les limites $a_i$ et $b_i$ de la variable $s_i$ se trouvent dans les boites d'indices $Ia_i$ et

$Ib_i$ de l'histogramme $P(s_i)$, respectivement. Donc, l'intégrale (8) peut être décomposée en une sommation d'intégrales partielles sur les boites d'histogrammes comme suit :

$$\alpha = \sum_{j_1=Ia_1}^{Ib_1} ... \sum_{j_n=Ia_n}^{Ib_n} (\int_{h^1_{j_1}}^{h^2_{j_1}} ... \int_{h^1_{j_n}}^{h^2_{j_n}} P(s_1, ..., s_n) Q_{vs} ds_1 ... ds_n) \tag{9}$$

avec $h^1_{j_i}$ et $h^2_{j_i}$ les limites inférieure et supérieure, respectivement, de la boite de l'histogramme d'indice $j_i$ de la composante $s_i$.

Soit $\alpha_j = \int_{h^1_{j_1}}^{h^2_{j_1}} ... \int_{h^1_{j_n}}^{h^2_{j_n}} P(s_1, ..., s_n) Q_{vs} ds_1 ... ds_n$ et $H = (j_1, ..., j_n, \Delta_{v_1}, ..., \Delta_{v_l})$ le polytôpe convexe de dimension $n$ défini dans l'espace de $S$ par l'intersection des demi-hyperplans comme suit :

$$H = \begin{cases} s_i \geq h^1_{j_i} & \textit{pour chaque composante indépendante } s_i,\ i = 1, ..., n \\ s_i \leq h^2_{j_i} & // \\ A^k \times S \geq \Delta^1_{v_k} & \textit{pour chaque variable instanciée } v_k,\ k = 1, ..., l \\ A^k \times S \leq \Delta^2_{v_k} & // \end{cases} \tag{10}$$

où $\Delta^1_{v_k}$ et $\Delta^2_{v_k}$ sont les limites inférieure et supérieure, respectivement, de l'intervalle $\Delta_{v_k}$. $A^k$ est le vecteur ligne d'indice $k$ de la matrice $A$ ($A^k \times \mathbf{s} = A_{k,1} \times s_1 + ... + A_{k,n} \times s_n$). $\alpha_j$ devient alors égal à l'intégrale multiple de la conjointe $P(S)$ sur le polytôpe $H$

$$\alpha_j = \int_H P(s_1, ..., s_n) ds_1 ... ds_n \tag{11}$$

Soit $B$ l'hypercube de dimension $n$ (par hypercube nous voulons dire aussi hyper-rectangle) borné par les boites ayant les indices $(j_1, ..., j_n)$ des histogrammes des composantes $s_1, ..., s_n$, respectivement. Formellement, $B$ est défini comme suit :

$$B = \begin{cases} s_i \geq h^1_{j_i} & \textit{pour chaque composante indépendante } s_i,\ i = 1, ..., n \\ s_i \leq h^2_{j_i} & // \end{cases} \tag{12}$$

Sur l'intervalle $[h^1_{j_i}, h^2_{j_i}]$ de la boite d'indice $j_i$ de l'histogramme de la composante $s_i$, la densité de probabilité est approximée par une distribution uniforme. Et comme toutes les variables $s_i$ $(i : 1, ..., n)$ sont mutuellement indépendantes et chacune d'elles a une distribution uniforme sur son intervalle correspondant $[h^1_{j_i}, h^2_{j_i}]$, alors la conjointe $P(S)$ est également uniforme sur l'hypercube $B$. Géométriquement, $H$ est inclus dans $B$ et de plus il découle de l'uniformité locale de $P(S)$ sur $B$ que

$$\alpha_j = \frac{Volume(H)}{Volume(B)} \times P(B) \tag{13}$$

$P(B)$ est la probabilité de $S$ sur l'hypercube $B$, c'est-à-dire $P(B) = P(s \in B)$. Puisque toutes les variables $s_i (i : 1, ..., n)$ sont estimées en maximisant leurs indépendances statistiques, $P(B)$ peut être évalué par la multiplication des valeurs discrètes des histogrammes des $s_i$ $(i : 1, ..., n)$.

$$P(B) = \prod_{i=1}^{n} hist_i(j_i) \tag{14}$$

où $hist_i(j_i)$ est la probabilité de la boite d'indice $j_i$ de la variable $s_i$. D'ailleurs, $Volume(B)$ peut être calculé d'une manière simple

$$Volume(B) = \prod_{i=1}^{n} Largeur_i(j_i) \tag{15}$$

avec $Largeur_i(j_i)$ la largeur de la boite d'indice $j_i$ de l'histogramme de la variable $s_i$.

Cependant, le calcul de volume du polytôpe $Volume(H)$ est relativement complexe car il n'a pas une forme régulière. Pour le faire, nous pouvons d'abord trouver la représentation minimale par l'élimination des contraintes redondantes. Ensuite, on peut déterminons les points extrêmes (sommets) de $H$. Il est démontré dans la littérature que le calcul de volumes de polytôpes quelconques est un problème np-difficile [Dyer and Frieze, 1988, Bäräny and Füredi, 1986], mais il reste abordable en un temps raisonnable pour des dimensions peu élevées (en général pour $n \leq 7$). Dans notre cas d'étude, ce problème est significativement relaxé parce que $H$ est défini au plus par $4 \times n$ contraintes : deux contraintes pour chaque composante indépendante $s_i \in S$ et deux contraintes pour chaque variable observée instanciée $v_i \in (V \subset X)$, voir la définition de $H$ en (10). En outre, plusieurs contraintes sont redondantes et elles sont supposées être éliminées dans une étape antérieure. Cela, aboutit généralement à des polytôpes $H$ qui sont définis par un nombre relativement petit de sommets dans l'espace de dimension $n$.

Le volume exact de $H$ peut être déterminé analytiquement pour les dimensions 2 et 3. Mais pour les dimensions élevées ($n > 3$), la tâche devient de plus en plus complexe et les méthodes basées sur la triangulation sont souvent employées. De telles méthodes commencent par subdiviser $H$ en un ensemble de simplexes, un simplexe est un polytôpe convexe défini par $n + 1$ sommets dans l'espace de dimension $n$ et dont le volume est non nul: par exemple un triangle en dimension 2 et un tétraèdre en dimension 3. Le volume total de $H$ est ainsi obtenu par la sommation des volumes de ses simplexes. Le volume d'un simplexe en dimension $n$ défini par les sommets $\{M_0, M_1, ..., M_n\}$ est donné par la formule suivante

$$Volume(Simplexe) = \frac{1}{n!}|det(\begin{matrix} 1 & 1 & ... & 1 \\ V_0 & V_1 & ... & V_n \end{matrix})| \tag{16}$$

où $V_k$ est le vecteur colonne définissant les cordonnées du $k^{ème}$ sommet et $|*|$ dénote l'opérateur de la valeur absolue.

Dans nos expérimentations, nous avons utilisé l'outil 'Qhull' qui implémente l'algorithme *Quickhull* [Barber et al., 1996] pour le calcul de volume de $H$. Qhull a été utilisé dans de nombreux travaux, particulièrement en apprentissage artificiel où la solution est basée sur la géométrie calculatoire.

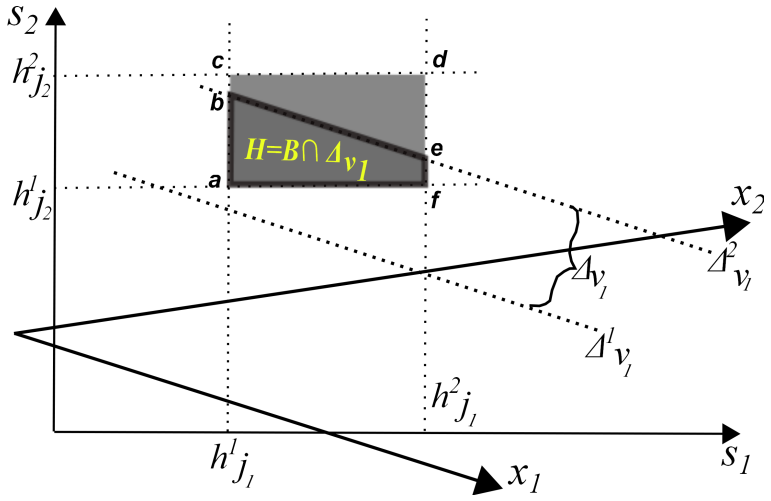



FIGURE 5.3 – Exemple géométrique en 2D illustre la technique d'inférence sur l'espace des composantes indépendantes.

L'exemple affiché sur la figure 5.3 montre une illustration en 2D de la technique d'inférence proposée. Nous avons deux variables observées $X = \{x_1, x_2\}$, deux CIs $S = \{s_1, s_2\}$, et une variable instanciée $V = \{x_2\}$. Nous voulons calculer $\alpha_j = P(x_2 \in \Delta_{x_2})$ sur les deux boites $j_1$ et $j_2$ des histogrammes de $s_1$ et $s_2$, respectivement. L'hypercube $B$ est le rectangle gris "acdf" défini par l'intersection de $s_1 \in [h^1_{j_1}, h^2_{j_1}]$ et $s_2 \in [h^1_{j_2}, h^2_{j_2}]$, $P(B) = hist_1(j_1) \times hist_2(j_2)$ et $volume(B) = (h^2_{j_1} - h^1_{j_1}) \times (h^2_{j_2} - h^1_{j_2})$. Le polytôpe $H$ est le polygone "abef". Ainsi, nous avons tous les éléments nécessaires pour calculer $\alpha_j$ en utilisant la formule (13).

## 4.4 Expérimentation

### 4.4.1 Méthodologie

Le but de l'algorithme d'inférence consiste à calculer la vraisemblance des variables non mesurées dans les mélanges ACI. Les performances de notre algorithme ("ACI-RB") sont comparées avec la solution conventionnelle basée sur les RBs discrets où les TPCs sont employées pour leur quantification. Nous avons utilisé l'outil ACI FastICA [Hyvärinen, 1999] pour l'estimation de la matrice $A$ de mélange et l'estimateur MAP pour apprendre les histogrammes de probabilités des CIs.

Nous comparons les deux algorithmes, ACI-RB et TPCs( RBs discrets), par le biais de l'évaluation de la vraisemblance conditionnelle de la valeur prise par la variable cible étant données des évidences, la vraisemblance d'un seul vecteur de test $\mathbf{x} = (y, z_1, ..., z_k)$ est la probabilité conditionnelle $LLx = P(y \in \Delta_y / z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k})$, les valeurs sont exprimées sous forme d'intervalles.

En supposant que $y$ prend la valeur réelle $r$ dans le vecteur de test (noter que les variables

considérées dans notre modèle sont continues), nous avons fixé $\Delta_y$ égal à la valeur discrète de la solution TPCs tel que $r$ se trouve à l'intérieur de $\Delta_y$ (même chose pour les évidences $Z$). Pour un vecteur de test $\mathbf{x} = (y, z_1, ..., z_k)$, une valeur élevée de $LL_x$ signifie une bonne performance de prédiction.

Nous avons employé un échantillon de test composé de $N = 100$ vecteurs tirés aléatoirement du benchmark dans chaque expérience. Étant donné que les vecteurs de test sont supposés identiquement et indépendamment distribués, alors la vraisemblance de l'échantillon entier est égale au produit des vraisemblances des vecteurs. Le logarithme de la vraisemblance moyenne (mean-likelihood) de l'échantillon de test $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ est

$$LogLL = \frac{\sum_{i=1}^{N} log(LL_{\mathbf{x}_i})}{N} \tag{17}$$

Les résultats sont comparés en termes de vraisemblance moyenne $LL = exp(LogLL_N)$.

### 4.4.2 Datasets

Dans le tableau 5.2, l'origine, la taille et une brève description des datasets utilisés dans les expériences sont donnés.

Tableau 5.2 – Description des datasets.

| Dataset | Origine | Taille (# vect.) | Description |
|---|---|---|---|
| Uniform | Synthétique | 30000 | Des signaux générés selon la loi uniforme. |
| Halo | site web[a] | 6000 | Des signaux de parole qui sont fortement corrélés, les locuteurs prononcent les mêmes phrases. |
| ACsparse | site web[a] | 5000 | Des sources qui sont approximativement indépendantes. |
| Speech | site web[a] | 1000 | Des sources sonores de musique et de parole. |

[a]Benchmarks proposés dans le projet ICALAB [Cichocki et al., 2012].

Les datasets utilisés comportent des échantillons de signaux sources hétérogènes issus de différentes applications qui sont plus ou moins indépendants. Afin de simuler des données qui satisfont l'hypothèse du modèle ACI linéaire, des mélanges linéaires sont constitués en suivant les étapes suivantes :

1. Générer une matrice carrée et non singulière (inversible) $A$ d'une manière aléatoire.
2. Réaliser les mélanges ACI, $X = A \times S$.
3. Supprimer $A$ et $S$. L'algorithme d'inférence proposé a en entrée seulement la matrice de données $X$ des variables observées. 100 vecteurs de $X$ sont utilisés pour le test tandis que le reste des vecteurs est utilisé pour l'apprentissage des paramètres de modèle.

### 4.4.3 Résultats

Les résultats de prédiction obtenus des quartes datasets sont affichés dans les tableaux 5.3-5.6. Nous dénotons par $LL = P(y|z_1, ..., z_k)$ la vraisemblance moyenne de l'échantillon de test de la variable $y$ étant données des évidences $\{z_1, ..., z_k\}$. La résolution dans la première colonne de chaque tableau indique le nombre de valeurs discrètes de domaine de chaque variable lorsqu'un RB à base des TPCs est utilisé, il est fixé à 10 ou à 20. Les intervalles $\Delta_y$ et $\Delta_{z_i}$ $(i : 1, ..., k)$ de la requête d'inférence sont pris égaux aux valeurs discrètes contenant les valeurs réelles prises par $y$ et les $z_i$, leurs largeurs sont donc égales au domaine/résolution. Nous avons utilisé la même résolution de discrétisation pour les histogrammes des CIs dans notre solution ACI-RB que celle de la solution TPC.

Tableau 5.3 – Résultats de prédiction du dataset " Uniform ", un exemple de 04 variables.

| | $LL(x_4)$ | | $LL(x_4\|x_1)$ | | $LL(x_4\|x_1, x_2)$ | | $LL(x_4\|x_1, x_2, x_3)$ | |
|---|---|---|---|---|---|---|---|---|
| Résolution | TPC | ACI-RB | TPC | ACI-RB | TPC | ACI-RB | TPC | ACI-RB |
| 10 | 0.1549 | 0.1551 | 0.3493 | 0.4981 | 0.4958 | 0.4981 | 0.5216 | 0.5229 |
| 20 | 0.0799 | 0.0802 | 0.2094 | 0.3477 | 0.3468 | 0.3478 | 0.3379 | 0.3756 |

Tableau 5.4 – Résultats de prédiction du dataset "Halo", un exemple de 04 variables.

| | $LL(x_4)$ | | $LL(x_4\|x_1)$ | | $LL(x_4\|x_1, x_2)$ | | $LL(x_4\|x_1, x_2, x_3)$ | |
|---|---|---|---|---|---|---|---|---|
| Résolution | TPC | ACI-RB | TPC | ACI-RB | TPC | ACI-RB | TPC | ACI-RB |
| 10 | 0.3104 | 0.3024 | 0.3691 | 0.3755 | 0.3696 | 0.3755 | 0.3939 | 0.3919 |
| 20 | 0.1599 | 0.1588 | 0.2052 | 0.2382 | 0.2362 | 0.2382 | 0.0899 | 0.2809 |

Tableau 5.5 – Résultats de prédiction du dataset "ACsparse", un exemple de 04 variables.

| | $LL(x_4)$ | | $LL(x_4\|x_1)$ | | $LL(x_4\|x_1, x_2)$ | | $LL(x_4\|x_1, x_2, x_3)$ | |
|---|---|---|---|---|---|---|---|---|
| Résolution | TPC | ACI-RB | TPC | ACI-RB | TPC | ACI-RB | TPC | ACI-RB |
| 10 | 0.1084 | 0.1007 | 0.4561 | 0.3890 | 0.4937 | 0.3837 | 0.6007 | 0.4157 |
| 20 | 0.0556 | 0.0502 | 0.3366 | 0.2527 | 0.5195 | 0.2488 | 0.6873 | 0.2649 |

Tableau 5.6 – Résultats de prédiction du dataset "Speech", un exemple de 04 variables.

| | $LL(x_4)$ | | $LL(x_4\|x_1)$ | | $LL(x_4\|x_1, x_2)$ | | $LL(x_4\|x_1, x_2, x_3)$ | |
|---|---|---|---|---|---|---|---|---|
| Résolution | TPC | ACI-RB | TPC | ACI-RB | TPC | ACI-RB | TPC | ACI-RB |
| 10 | 0.1832 | 0.1828 | 0.1939 | 0.2313 | 0.1048 | 0.2310 | 0.0693 | 0.2781 |
| 20 | 0.0952 | 0.0943 | 0.0863 | 0.1368 | 0.0348 | 0.1366 | 0.0057 | 0.1663 |

nous constatons dans le tableau 5.3 que les résultats de prédiction, exprimés sous forme de vraisemblance de l'échantillon de test, de notre solution (ACI-RB) sont proches de ceux de la

solution TPC pour une faible résolution (Res.=10) et ils sont légèrement meilleurs pour une résolution importante (Res.=20). Cette observation peut être expliquée par le fait que d'une part les sources utilisées pour générer cette base sont uniformes et fortement indépendantes et cela représente un avantage pour l'estimation ACI dans notre approche. D'une autre part, l'abondance des données d'apprentissage (29900 vecteurs) atténue le problème de sur-apprentissage et contribue dans l'amélioration de la solution TPC.

Pour les deux datasets, "Halo" et "Speech", les résultats de notre algorithme sont nettement meilleurs que ceux de la solution TPC. En effet, l'amélioration est plus significative lorsque la résolution de discrétisation est haute (Res.=20). Dans une telle situation, les tables TPCs du RB discret sont parcimonieuses parce qu'elles sont trop larges (contenant jusqu'à $20^4$ paramètres par table) comparées avec la quantité modeste de vecteurs d'apprentissage, 900 vecteurs pour "Speech" et 5900 pour "Halo". Nous constatons que notre solution est moins sensible au problème du sur-apprentissage.

En revanche, la solution TPC s'est montrée plus performante que notre algorithme pour le dataset "ACsparse". La dégradation des résultats de l'algorithme ACI-RB peut être due à l'inefficacité ou à la non appropriation de l'outil "FastICA" pour ce dataset. En effet, dans le contexte de la BSS, "ACsparse" est connu pour être un dataset difficile, il contient des signaux sparses et plusieurs méthodes ACI standard sont défaillantes pour la séparation de telles sources [Cichocki et al., 2012].

Dans l'objectif de faire des comparaisons équitables, les intervalles de la cible et des évidences sont fixés de telle sorte qu'ils soient égaux aux valeurs discrètes de la solution CPT. Il convient de noter que dans notre algorithme, nous pouvons personnaliser des intervalles plus flexibles et cela représente un avantage pour notre proposition. En outre, comme il est constaté, ACI-RB souffre moins du problème de sur-apprentissage. Par ailleurs, la séparation ACI permet également d'utiliser des histogrammes de probabilités de haute résolution pour les CIs, et par conséquent, de gagner plus dans la flexibilité de fixer des intervalles très petits dans la requête d'inférence sans diminuer les performances.

### 4.4.4 Analyse de complexité

En supposant que pour les $n$ variables observées, le nombre moyen des valeurs discrètes par variable est $D$ quand les CPTs sont utilisées, et que le nombre moyen des boites d'histogrammes par composante indépendante est également $D$ quand le modèle ACI-RB est utilisé.

En termes d'espace mémoire nécessaire pour le stockage des paramètres des modèles, la complexité de la solution CPT est de l'ordre de $O(D^n)$, elle augmente exponentiellement en fonction de $n$. Cependant la complexité de notre proposition est linéaire dans la mesure où seulement $n^2$ éléments de la matrice $A$ du mélange ACI et $n \times D$ paramètres d'histogrammes des CIs sont à sauvegarder. En termes de temps d'inférence, la complexité de la marginalisation de la conjointe sur $K$ variables de la solution CPT est de l'ordre de $O(D^K)$ additions ($n - K$

variables sont instanciées). Pour l'algorithme proposé, la complexité est de l'ordre de $O(D^K) \times O(Calcul\ de\ volume\ d'un\ polytpe\ en\ dimension\ n)$ ce qui est difficile à évaluer avec précision, pour cela nous nous contentons d'en donner une approximation.

Les polytôpes obtenus proviennent en effet de l'intersection des valeurs (sous forme d'intervalles) prises par les variables instanciées et les boites d'histogrammes des CIs de telle sorte que chaque polytôpe est défini par $4n$ demi-plans au maximum, voir la définition de l'équation(10). Les polytôpes obtenus ont approximativement le même nombre de sommets que l'hypercube de dimension $n$. Pour calculer les volumes des polytôpes nous avons utilisé l'outil Qhull [Barber et al., 1996] qui procède par la triangulation en simplexes des polytôpes. La majorité des techniques publiées de triangulation d'un hypercube en dimension $n$ aboutissent à moins de $n!$ simplexes [Sallee, 1982]. Par conséquent, la complexité d'inférence est approximativement $O(D^K) \times n! \times O(volume\ d'un\ simplexe\ en\ dimension\ n$. Due à la présence du terme factoriel, la complexité augmente exponentiellement avec la dimension $n$. En effet, la complexité est pratiquement abordable seulement pour les petites dimensions ($n \leq 7$). Les polytôpes obtenus ne sont pas complètement arbitraires, ils se distinguent par une régularité relative vu qu'ils sont définis comme étant l'intersection d'un ensemble de paires de demi-hyperplans parallèles. Donc, il reste envisageable de trouver de nouvelles méthodes plus efficaces du calcul de volumes en contournant la triangulation par l'exploitation de la pseudo-régularité des polytôpes obtenus.

En récapitulant, nous avons conclu que l'algorithme proposé convient aux problèmes de dimensions inférieures qui exigent une précision importante (des petits intervalles dans la requête d'inférence) et peu d'espace mémoire.

# 5 Application pour la fusion des scores multi-biométriques

## 5.1 Fusion des scores multi-biométriques

L'emploi des systèmes biométriques multimodaux [Ross and Jain, 2003] pour l'identification d'individus par la considération de plusieurs caractéristiques biométriques afin d'améliorer la robustesse, la précision et l'universalité d'usage a gagné plus d'intérêt ces dernières années. Les techniques de fusion de scores des différents sous-systèmes biométriques sont les plus répandues car elles sont faciles à implémenter et donnent de bonnes performances [Poh et al., 2010, Maurer and Baker, 2008, Ding and Ross, 2012]. La fusion de scores peut être perçue comme un problème de classification binaire supervisée dans l'espace des vecteurs de scores dans lequel nous avons deux classes, la classe des utilisateurs légitimes et celle des imposteurs et dans lequel les attributs correspondent aux scores des modalités adoptées, voir l'exemple de la figure 5.4. Par conséquent, nous pouvons nous servir d'un large répertoire des classificateurs disponibles (SVM, Réseaux de neurones, GMM, etc.).

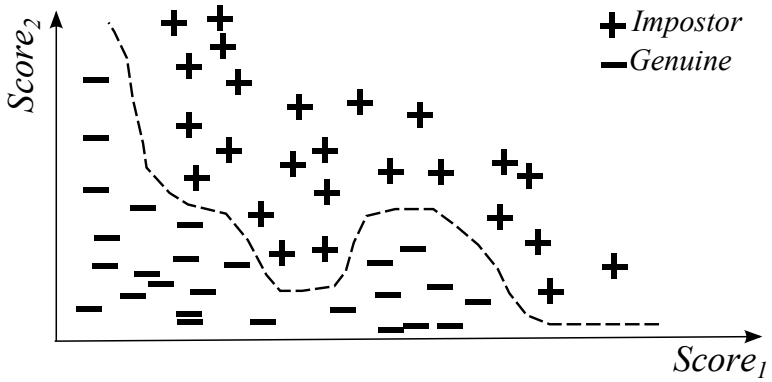



FIGURE 5.4 – Fusion de scores vue comme un problème de classification binaire supervisée, exemple de deux scores.

## 5.2 Scores absents ou de mauvaise qualité

Il est évident que la fusion multi-biométrique entraîne un coût supplémentaire; nous optons pour elle particulièrement pour remédier aux problèmes de qualité insuffisante ou de l'absence de certains scores. Généralement, les attributs absents dans un vecteur de scores sont remplis par des valeurs prédites (des valeurs maximisant la vraisemblance) comme illustré dans [Ding and Ross, 2012], ensuite les techniques standard de fusion , conçues pour des données complètes, sont appliquées. La qualité des scores peut être aussi traitée par l'ajout de mesures de qualité sous forme d'attributs au classificateur comme le montre l'exemple [Maurer and Baker, 2008].

Les scores absents peuvent être également traités parfaitement par les RBs où les nœuds du graphe représentent les variables de scores. En effet, les RBs comme concept sont conçus spécialement pour la prise de décision d'une manière efficace en se basant seulement sur les variables (scores) instanciées dont la valeur est connue (évidences). Dans la section suivante (§5.3), motivée par l'expansion d'emploi des RBs, nous allons montrer comment l'algorithme d'inférence Bayésienne décrit en Section 4 est appliqué pour résoudre simultanément les problèmes de scores absents et de scores de mauvaise qualité.

## 5.3 Approche proposée

Les RBs ne sont pas bien formulés en domaine continu. En contrepartie, les scores de classification biométriques sont des variables continues et leur manipulation requiert une haute précision. Comme solution à ce compromis, nous proposons d'appliquer l'algorithme décrit dans la section 4 pour accomplir la fusion de scores. Au lieu de considérer les mesures de qualité de scores comme des attributs de classification séparés, nous proposons de les représenter sous forme d'intervalles de confiance; autant l'intervalle de confiance est large, autant la qualité est médiocre. Par exemple, si la valeur estimée du score $S$ est $v$ et l'intervalle de confiance correspondant est $\Delta$, alors la valeur réelle du score $S$ doit se trouver à l'intérieur de l'intervalle

$[v - \frac{\Delta}{2}, v + \frac{\Delta}{2}]$.

Les vecteurs de scores sont partitionnés en deux groupes distincts : celui des imposteurs et celui des utilisateurs légitimes (comme le montre la figure 5.4). Cela rend l'association d'un modèle ACI linéaire à tous les vecteurs de scores une approximation non appropriée. En s'inspirant du travail réalisé par [Karhunen and Malaroiu, 1999] où un ensemble de modèles ACI linéaires locaux appliqués sur des clusters de données ont été utilisés pour produire une représentation de données multivariables qui peuvent avoir des distributions générales, nous suggérons de considérer deux modèles ACI linéaires associés aux vecteurs de scores des utilisateurs imposteurs et légitimes, respectivement.

Dans les deux modèles ACI suggérés, les scores des différentes modalités biométriques $S = \{s_1, ..., s_n\}$ plus le score de fusion $f$ sont considérées comme des variables observées du modèle ACI. Le score de fusion $f$ est la variable à prédire ; pour les données d'apprentissage, nous avons simulé ses valeurs de telle sorte qu'elle varie d'une manière considérables entre les vecteurs imposteurs et les vecteurs légitimes.

Soit $fs = \sum_{i=1}^{n} s_i$ la variable *somme de scores*. Dans ce travail, pour les vecteurs légitimes, nous avons ajouté à la variable somme $fs$ une valeur constante $y_{leg}$ tel que la moyenne de la variable de fusion, défini ainsi $f = fs + y_{leg}$, des vecteurs d'apprentissage sera égale à $+\theta$. De la même façon, la variable $fs$ des vecteurs d'apprentissage des imposteurs est décalée par une valeur constante $y_{imp}$, $f = fs + y_{imp}$, tel que la moyenne sera égale à $-\theta$. La constante positive $\theta$ est un paramètre en entrée fixé par l'utilisateur, tandis que les constantes de décalage $y_{leg}$ et $y_{imp}$ sont déterminées en fonction de $\theta$, nous avons $y_{leg} = +\theta - moyenne(fs_{lgitime})$ et $y_{imp} = -\theta - moyenne(fs_{imposteur})$, voir l'exemple de deux modalités affiché sur la figure 5.5.

Comme la variable de fusion $f$ est définie de telle sorte qu'elle prend des valeurs positives pour la quasi-totalité des utilisateurs légitimes et des valeurs négatives pour la quasi-totalité des utilisateurs imposteurs, la probabilité a posteriori pour qu'un vecteur de scores à tester $\mathbf{t} = (s_i \in \Delta_{s_i}, i = 1, ..., n)$ soit d'un utilisateur légitime est

$$P_{gen}(\mathbf{t}) = P(f \in [0, max_f] | \{s_i \in \Delta_{s_i}, i = 1, ..., n\}) \tag{18}$$

Du fait que l'espace des vecteurs de scores est scindé en deux clusters, on peut déduire $P_{leg}$ par la marginalisation sur les probabilités conditionnelles des clusters. Soit $P_1(t)$ et $P_2(t)$ les probabilités postérieures que $f \in [0, max_f]$ sachant que $\mathbf{t}$ est légitime ou imposteur, respectivement. $P_1(\mathbf{t})$ et $P_2(\mathbf{t})$ sont définies comme suit :

$$\begin{aligned} P_1(\mathbf{t}) &= P(f \in [0, max_f] | \{s_i \in \Delta_{s_i}, i = 1, ..., n\}, \mathbf{t} \text{ est dans le cluster légitime}) \\ P_2(\mathbf{t}) &= P(f \in [0, max_f] | \{s_i \in \Delta_{s_i}, i = 1, ..., n\}, \mathbf{t} \text{ est dans le cluster imposteur}) \end{aligned} \tag{19}$$

ainsi, on aura

$$P_{gen}(t) = P_g^{'} \times P_1(\mathbf{t}) + P_i^{'} \times P_2(\mathbf{t}) \tag{20}$$

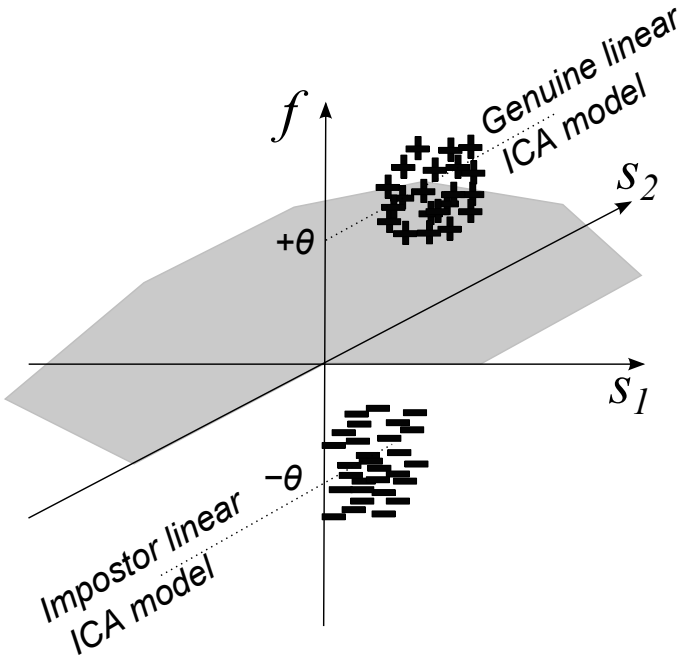



FIGURE 5.5 – Modèle de fusion proposé.

$f$ est la variable de fusion des scores $s_1$ et $s_2$, elle est simulée pour les vecteurs d'apprentissage et elle est la variable à prédire pour les vecteurs de test.

où $P_g^{'}$ et $P_i^{'}$ sont les probabilités a priori des clusters des vecteurs de scores légitimes et imposteurs, respectivement. $P_g^{'}$ et $P_i^{'}$ sont souvent prises égales à 0.5 en cas d'absence des a priori. Pour l'estimation de $P_1(\mathbf{t})$ et $P_2(\mathbf{t})$, nous avons utilisé l'algorithme d'inférence proposé dans la section 4. La fonction de décision est comme suit: un vecteur de scores $\mathbf{t}$ est assigné à la classe des utilisateurs légitimes si $P_{leg} > 0.5$ et à la classe des imposteurs sinon.

Chapitre 6

# Inférence Bayésienne sur les variables continues en utilisant la méthode ACI locale

## Sommaire


**1 Introduction . . . . . 139**
**2 Limites du modèle ACI-RB . . . . . 140**
**3 Propositions . . . . . 140**
**4 Ajustement d'histogrammes par des séquences polynomiales . . . . . 141**
4.1 Approximation de la densité par des fonctions polynomiales . . . . . 141
4.2 Approximation d'une fonction par une spline . . . . . 143
4.2.1 Définitions . . . . . 143
4.2.2 Régression par des splines . . . . . 143
4.2.3 Choix des nœuds de jonction . . . . . 144
4.2.4 Contraintes sur les extrémités . . . . . 144
4.2.5 Régression polynomiale (un seul morceau) . . . . . 145
4.2.6 Algorithme de régression développé . . . . . 147
**5 Algorithme d'inférence . . . . . 149**
5.1 Apprentissage des paramètres . . . . . 149
5.2 Inférence . . . . . 151
**6 Intégration de polynômes sur des polytôpes convexes . . . . . 153**
6.1 Introduction . . . . . 153
6.2 Énonce du problème . . . . . 154
6.3 Intégration d'un polynôme sur un polytôpe par la triangulation . . . . 155
6.4 Intégration d'un polynôme multivariable sur le simplexe standard (simplexe unitaire) . . . . . 155
6.5 Intégration d'un polynôme multivariable sur un simplexe (cas général) 156
6.6 Amélioration proposée pour accélérer l'intégration . . . . . 158
6.6.1 Objectif . . . . . 158

6.6.2 Une constatation intéressante . . . . . 159
6.6.3 Proposition . . . . . 159
6.7 Conclusion . . . . . 162
**7 Expérimentation . . . . . 163**
7.1 Description de données de test . . . . . 163
7.2 Méthodologie de comparaison . . . . . 165
7.3 Résultats obtenus . . . . . 166
**8 Conclusion . . . . . 168**

# 1 Introduction

Dans le modèle ACI standard, un modèle linéaire est employé pour fournir une description globale de données. Même si le modèle ACI linéaire s'est montré satisfaisant dans de nombreuses applications [Comon and Jutten, 2010, Hyvärinen and Oja, 2000, Hyvärinen et al., 2001], il permet seulement de procurer une approximation triviale et parfois non appropriée à des données quelconques ayant une distribution non linéaire. Le modèle d'inférence Bayésienne ACI-RB, introduit dans le chapitre précédent, est restreint aux mélanges ACI linéaires. Par conséquent, son extension à des données générées selon des distributions générales est souhaitable.

Par ailleurs, la méthode ACI non linéaire est connue d'être un problème mal-posé dans la mesure où l'espace de recherche des fonctions de mélange ACI est infini [Hyvärinen and Pajunen, 1999, Achard and Jutten, 2005]. Cela explique pourquoi le peu d'algorithmes ACI non linéaires développés à nos jours, soit limitent la recherche au sein d'une famille de fonctions paramétriques telles que les modèles ACI poste-linéaires [Achard and Jutten, 2005], soit se servent des techniques d'estimation non paramétrique afin de traiter la non linéarité.

Dans la première approche, la fonction $F$ du mélange ACI ($\mathbf{x} = F(\mathbf{s})$ est souvent approximée par des techniques d'optimisation neuronales. En essayant de généraliser notre algorithme ACI-RB aux modèles ACI non linaires, l'inférence Bayésienne reviendra donc à intégrer la conjointe des CIs sur des formes géométriques non linéaires ce qui est théoriquement très compliqué et voire impossible.

De nombreuses techniques d'intégration des fonctions multivariables ordinaires telles que les polynômes et les exponentielles sur des formes géométriques simples telles que les polytôpes sont publiées dans la littérature. Afin d'exploiter les résultats théoriques obtenus dans le chapitre précédent en ce qui concerne le modèle ACI-RB, le développement d'un algorithme d'inférence Bayésienne pour les modèles ACI non linéaires semi-paramétriques s'avère intéressant. En effet, l'option pour des modèles ACI semi-paramétriques par la considération d'un ensemble de modèles ACI linéaires appliqués localement est expliquée par leur simplicité d'estimation, par leur bonne capacité de représentation de données et surtout par le fait que l'inférence sur ces modèles revient à intégrer sur des formes géométriques simples (polytopes convexes).

Dans le chapitre IV, nous avons introduit la méthode ACI locale basée sur le clustering [Karhunen and Malaroiu, 1999] où la combinaison d'une méthode ACI linéaire (FastICA, JADE, etc) avec un algorithme de clustering (K-means, GMM, etc) convenable peut constituer une base de représentation de données multivariables. Dans un objectif d'illustration, nous avons l'employée pour l'estimation de densité et pour la classification de données réelles. Dans ce contexte, nous allons décrire dans ce chapitre comment le modèle ACI-RB a été étendu au modèle ACI local. En outre,des améliorations proposées impliquant l'utilisation de fonctions analytiques et continues pour l'estimation de densité des CIs au lieu de se servir d'histogrammes seront également entamées.

## 2 Limites du modèle ACI-RB

Les deux principales limites du modèle ACI-RB que nous avons proposé dans le chapitre précédent sont :

1. Restriction à des données qui peuvent être ajustées par un modèle ACI linéaire.

2. Emploi d'histogrammes de probabilité pour l'approximation de la distribution de densité des CIs. En effet, ces dernières se caractérisent par des distributions continues. Ainsi, il résulte des approximations discrètes telles que les histogrammes une perte d'information. Cependant, même l'augmentation de la résolution de discrétisation pour diminuer l'erreur de l'estimation n'est pas toujours désirable, car cela va entraîner une accentuation du problème de sur-apprentissage [Hawkins, 2004] et une explosion du temps d'inférence.

## 3 Propositions

Afin de remédier aux inconvénients précités, nous avons apporté au modèle ACI-RB les deux améliorations suivantes :

1. Généralisation du modèle ACI-RB à des données générées selon des distributions non linéaires (cas général) par l'emploi de la méthode ACI locale.

2. Utilisation de fonctions analytiques ou semi-analytiques pour approximer la distribution de densité des CIs plutôt que la discrétisation (histogrammes). Cette alternative est suggérée dans le but d'atténuer le rapport de la qualité d'approximation de la densité des CIs au temps d'inférence. Les fonctions potentielles doivent d'une part avoir un grand degré de liberté pour pouvoir approximer des densités complexes et doivent d'autre part disposer de certaines propriétés analytiques facilitant l'intégration de la conjointe des CIs sur des polytôpes en hautes dimensions. Pour ces deux raisons, et dans un premier lieu, nous avons opté pour les fonctions polynomiales définies par morceaux, connues sous le nom des "fonctions splines" [STROUD, 1971, Ichida and Kiyono, 1977, Plass and Stone, 1983]. Évidemment, il est possible qu'il y ait d'autres fonctions plus adéquates à cette fin que les fonctions splines, et cela va certainement ouvrir une brèche prometteuse pour améliorer l'algorithme d'inférence proposé.

Dans le reste de ce chapitre, le modèle d'inférence Bayésienne proposé 'ACIL-RB' basé sur la méthode ACI locale et les fonctions splines est présenté.

# 4 Ajustement d'histogrammes par des séquences polynomiales

## 4.1 Approximation de la densité par des fonctions polynomiales

Une des deux améliorations proposées consiste à estimer la densité d'une CI à partir d'un échantillon de données en utilisant des fonctions analytiques. Notre choix s'est porté sur les fonctions polynomiales grace à leurs caractéristiques analytiques d'intégration potentiellement avantageuses pour l'inférence.

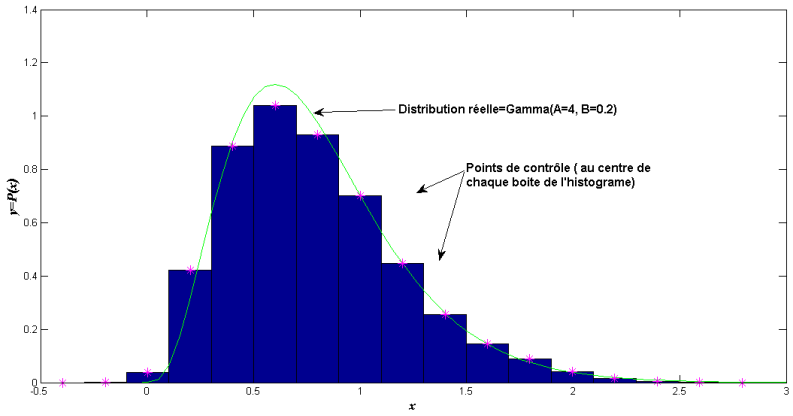
Pour un ensemble $P = \{(x_1, y_1), ..., (x_N, y_N)\}$ de $N$ points sur le plan, il existe un seul polynôme de degré $N - 1$ qui passe par tous les points de $P$. L'inconvénient d'ajuster $P$ par un seul polynôme est que la détermination des coefficients d'un tel polynôme nécessite la résolution d'un système de $N$ équations linéaires. De plus, les fonctions polynomiales de degrés importants ne sont pas souhaitables dans notre contexte car la complexité spatio-temporelle de l'algorithme d'inférence augmente exponentiellement avec le degré.

Pour palier à ce problème, en pratique, des séquences de polynômes de degrés inférieurs ($d \leq 3$) sont généralement utilisées pour l'interpolation ou pour la régression, car même avec peu de pièces polynomiales de degrés inférieurs nous pouvons atteindre des résultats comparables et parfois meilleurs que ceux des polynômes de hauts degrés. Les fonctions polynomiales par morceaux permettent également d'éviter le phénomène de Runge [Boyd, 1992] qui se manifeste lorsque le degré de polynôme est important.

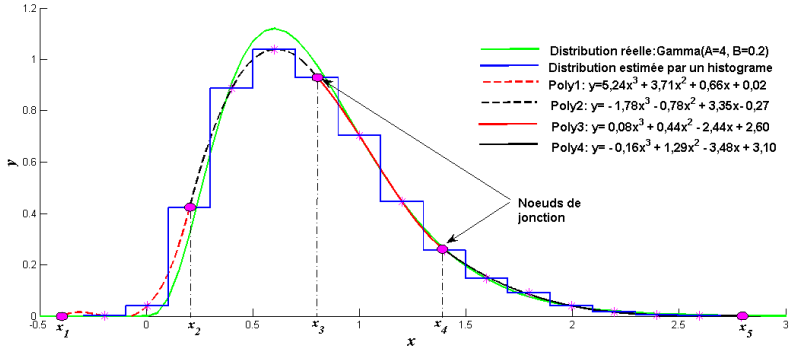
La distribution de densité d'une CI est inconnue, pour l'estimer à partir d'un échantillon de données nous pouvons procéder comme suit :

1. Estimation non paramétrique d'un histogramme de probabilité.
2. Sélection d'un ensemble de points de contrôle à partir de l'histogramme.
3. Régression des points de contrôle sélectionnés par une fonction polynomiale définie par morceaux (fonctions splines).

Dans l'exemple affiché sur la figure 6.1, un échantillon de 1000 valeurs réelles est généré à partir de la distribution gamma avec les paramètres $A = 4$ et $B = 0.2$ dont la courbe est affichée en vert. Les points de contrôle sont fixés aux centres des intervalles de l'histogramme comme le montre la figure 6.1(a). Au sens des moindres carrés, les points de contrôle sont ajustés par une séquence de 04 polynômes cubiques dont les nœuds de jointure sont $\{x_2, x_3, x_4\}$. La qualité de l'estimation est évaluée en fonction de la somme des erreurs carrées ($SSE$ pour '*sum of squares due to error*'), nous constatons que l'ajustement par une spline cubique a permis de diminuer considérablement l'erreur. L'échantillon de données a été généré à partir d'une distribution connue (e.g. gamma) seulement pour pouvoir calculer l'erreur d'estimation par rapport à la distribution originale. Il faut bien souligner qu'en pratique la distribution originale est souvent inconnue. Cet exemple a été choisi pour montrer que des fonctions polynomiales




(a)




(b)

FIGURE 6.1 – Exemple de l'estimation de densité en utilisant une spline cubique.

Courbe verte : densité originale utilisée pour la génération d'un échantillon de données. (a) : Sélection des points de contrôle. (b) : fonctions estimés. Courbe bleue : histogramme estimé en se basant sur un échantillon de 1000 réalisations. $P(x) = \{poly_1\ six \in [x_1, x_2];\ poly_2\ si\ x \in [x_2, x_3];\ poly_3\ si\ x \in [x_3, x_4];\ poly_4\ si\ x \in [x_4, x_5]\}$ est la densité ajustée par une spline cubique. La somme des erreurs carrées de la courbe bleue est $SSE = 0.0165$, tandis que celui de la spline est $SSE = 0.0027$.

définies par morceaux sont en principe mieux adaptées pour approximer la densité des variables continues que les histogrammes et les fonctions polynomiales.

Il faut bien préciser que notre finalité n'est pas de faire l'interpolation qui consiste à trouver une fonction passant exactement par les points sélectionnés, mais plutôt la régression par une spline constituée de peu de pièces polynomiales et minimisant l'erreur d'ajustement.

## 4.2 Approximation d'une fonction par une spline

### 4.2.1 Définitions

**Définition 13 (*spline*)**

*une spline est une fonction polynomiale définie par morceaux et suffisamment lisse aux endroits de jointure des morceaux polynomiaux (appelés des nœuds ou des knots).*

**Définition 14 (*spline*)**

*une courbe spline* $S : [a, b] \rightarrow R$ *est une fonction polynomiale réelle par morceaux définie sur un intervalle* $[a, b]$ *composé de* $k$ *sous-intervalles ordonnés et disjoints* $[t_{i-1}, t_i]$ *avec* $a = t_0 < t_1 < ... < t_{k-1} < t_k = b$*, tel que :*

$$\begin{array}{lcll} S(t) & = & P_1(t), & si\ t_0 \leq t \leq t_1 \\ S(t) & = & P_2(t), & si\ t_1 \leq t \leq t_2 \\ \vdots & \vdots & \vdots & \vdots \\ S(t) & = & P_k(t), & si\ t_{k-1} \leq t \leq t_k \end{array} \quad avec\ P_i : [t_{i-1}, t_i] \rightarrow R\ un\ polynme.$$

Le degré de la spline est défini comme étant celui du polynôme $P_i(t)$ de plus haut degré. Sachant que les polynômes sont infiniment continus, la continuité de la spline dépend donc de la continuité aux endroits de jointure des polynômes. Une spline est dite de continuité $n$ si pour tout $0 \leq j \leq n$, et au niveau de chaque jointure $t_i$ $(1 \leq i \leq k-1)$ nous avons :

$$P_i^{(j)}(t) = P_{i+1}^{(j)}(t) \tag{1}$$

### 4.2.2 Régression par des splines

Que signifie-t-elle une bonne régression par une courbe spline d'un ensemble de points bruités et générés selon une fonction (inconnue) telle que l'exemple affiché sur la figure 6.1 ? La courbe recherchée doit évidement se trouver le plus proche possible des points, mais elle doit être aussi efficace en se constituant *de peu de polynômes de degrés inférieurs*. Cela représente un compromis entre le resserrement de l'ajustement et l'efficacité de la représentation. Dans d'autres applications telles qu'en géométrie ou en conception mécanique, on cherche souvent une courbe qui soit suffisamment continue et lisse. La continuité veut dire que deux polynômes consécutifs séparés par le nœud $h$ sont continus au niveau de ce point : $poly_i(h) = poly_{i+1}(h)$. La courbe est dite lisse si la direction de la tangente est préservée, en d'autres mots nous avons la continuité de la dérivée : $poly'_i(h) = poly'_{i+1}(h)$ et ce concept peut être généralisé à la continuité d'ordre $n$ (Equ. (1)).

Dans notre cas d'étude, les contraintes sur la continuité et le lissage sont souhaitables étant donnée l'hypothèse que les distributions de densité des CIs que nous voulons à estimer sont

généralement continues et lisses. Cependant, surtout que les splines recherchées sont constituées de peu de jointures, nous donnons plus d'importance à la qualité globale de l'ajustement (mesurée par le critère SSE) qu'à la garantie des contraintes de continuité des jointures.

### 4.2.3 Choix des nœuds de jonction

Nous considérons le problème de détermination d'une approximation par des segments polynômiaux d'un ensemble fini de points $E = [t_1, t_2, ..., t_N]$ qui sont issus d'une fonction sous-jacente. En d'autres termes, l'algorithme à développer a en entrée un ensemble de points de contrôle et dérive une séquence de polynômes qui soit la plus proche possible des points à ajuster. L'approximation est faite en minimisant une métrique d'erreur basée sur la somme des erreurs carrées (SSE). La recherche d'une telle approximation est une tâche délicate parce que les nœuds de jonction sont des paramètres inconnus et nécessitent eux mêmes d'être estimés, nous ne connaissons pas ni leur nombre ni leurs emplacements optimaux.

Le nombre total de combinaisons des nœuds de jonctions possibles est énorme. Par exemple, sur un ensemble de 100 points de contrôle et pour des fonctions splines contenant 2, 3, 4 et 5 polynômes ajustant chacun au minimum 04 points, nous avons 94, 4186, 117480, 2331890 combinassions possibles des endroits des nœuds, respectivement. Les méthodes exactes examinent toutes les combinaisons et par conséquent garantissent l'obtention d'une solution optimale. Tandis que les méthodes heuristiques qui sont plus rapides, telles que celles présentées dans [Ichida and Kiyono, 1977, Plass and Stone, 1983], scrutent seulement un sous-ensemble de combinaisons et aboutissent à des résultats acceptables.

Dans le cas où la régression d'un ensemble de points $E_{ij} = [t_i, t_{i+1}, ..., t_j]$ ne dépend pas des points qui sont dehors l'intervalle $E_{ij}$ ($\{E - E_{ij}\} = \{[t_1, ..., t_{j-1}] \cup [t_{j+1}, ..., t_N]\}$), par exemple cela est vérifié si on n'impose pas des contraintes de continuité d'ordre $n \geq 1$ sur les jointures), alors la régression $(P, SSE)_{i,j} = Reg(E_{ij})$ peut être effectuée une seule fois et ses résultats sauvegardés pour être utilisés lors de la recherche, ainsi on peut faire au pire des cas $N \times (N-1)/2$ régressions polynomiales pour explorer l'espace entier des combinaisons possibles des nœuds.

### 4.2.4 Contraintes sur les extrémités

Des contraintes sur la continuité au niveau des points de jonction sont souvent imposées. Une spline est continue si dans laquelle chaque deux polynômes consécutifs $P_i$ et $P_{i+1}$ sont connectés au point de jonction $t_i$, $P_i(t_i) = P_{i+1}(t_i)$. Elle est de continuité $n$ si toutes les dérivées de l'ordre $j$ ($0 \leq j \leq n$) prennent les mêmes valeurs aux points de jonction, $P_i^{(j)}(t_i) = P_{i+1}^{(j)}(t_i)$. Les contraintes sur la continuité rendent l'estimation des coefficients de la spline plus complexe car la régression d'un polynôme $P_i$ dépend des résultats de la régression des autres polynômes $P_k (k \neq i)$. La regression $(P, SSE)_{i,j} = Reg(E_{ij})$ est faite autant des combinaisons des noeuds sont possibles et cela explique pourquoi il est nécessaire d'avoir recours aux heuristiques [Ichida and Kiyono, 1977, Plass and Stone, 1983] au lieu de chercher la solution optimale.

Dans notre cas d'étude, nous préférons des solutions minimisant la somme des erreurs carrées. La seule contrainte sur les extrémités que nous avons imposée dans quelques variantes de nos implémentations est le passage de chaque polynôme par le premier et le dernier point de contrôle de son intervalle. Formellement, le modèle de régression considéré est comme suit :

$$\begin{array}{lcll} S(t) & = & P_1(t) & t_0 \le t \le t_1 \\ S(t) & = & P_2(t) & t_1 \le t \le t_2 \\ \vdots & = & \vdots) & \vdots \\ S(t) & = & P_k(t) & t_{k-1} \le t \le t_k \end{array} \quad \textit{sous contraintes} \quad \begin{array}{lllll} P_1(t_0) & = & h(t_0) & & \\ P_1(t_1) & = & P_2(t_1) & = & h(t_1) \\ P_2(t_2) & = & P_3(t_2) & = & h(t_2) \\ \vdots & = & \vdots & & \\ P_{k-1}(t_{k-1}) & = & P_k(t_{k-1}) & = & h(t_{k-1}) \\ P_k(t_k) & = & h(t_k) & & \end{array} \tag{2}$$

$h(t)$ est la coordonnée du point $t$, dans notre cas $h(t)$ est la probabilité du point $t$. En procédant ainsi, la continuité de l'ordre 0 est établie et la régression d'un ensemble de points $E_{ij} = \{t_i, ..., t_j\}$ par un polynôme ne dépend pas des résultats de régression des autres points restants. Par conséquent, la régression $(P, SSE)_{ij} = Reg(E_{ij})$ peut être faite une seule fois et ses résultats sauvegardés pour la recherche des autres combinaisons, ce qui permet de trouver la solution optimale en un temps de calcul raisonnable. Nous avons accordé avec ce choix plus d'importance à la minimisation de l'erreur de régression qu'au lissage de la spline.

### 4.2.5 Régression polynomiale (un seul morceau)

Considérons un ensemble de $N$ points empiriques $\{(x_i, y_i)\ i = 1, ..., N\}$. Nous voulons ajuster cet ensemble par un polynôme $P(x) = \sum_{c=0}^{d} \alpha_c x^c$ de degré $d$ et de coefficients $\alpha_0, \alpha_1, ..., \alpha_d$ (pour réaliser cela il faut que $N > d$). L'objectif de l'ajustement consiste à minimiser l'écart entre les valeurs prises par la fonction de régression $P$ aux points $x_i$ et les valeurs empiriques $y_i$. Au sens des moindres carrés, on minimise la somme des carrés des écarts entre les $y_i$ et les $P(x_i)$.

$$SSE(\alpha_0, \alpha_1, ..., \alpha_d) = \sum_{i=1}^{N} \xi_i = \sum_{i=1}^{N} (y_i - \sum_{c=0}^{d} \alpha_c x_i^c)^2 \tag{3}$$

**• Régression linéaire**

Les deux variables $y$ et $x$ sont liées par l'équation linéaire : $y = \alpha_0 + \alpha_1 x + \xi$. Les valeurs $\hat{\alpha_0}$ et $\hat{\alpha_1}$ qui minimisent la fonction objectif $SSE(\alpha_0, \alpha_1) = \sum_{i=1}^{N} \xi_i$ sont

$$\hat{\alpha_1} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}, \hat{\alpha_0} = \bar{y} - \hat{\alpha_1}\bar{x} \tag{4}$$

avec $\bar{x}$ , $\bar{y}$ les moyennes des variables $y$ et $x$, respectivement.

• **Régression polynomiale**

C'est les modèles quadratique ($d = 2$) et cubique ($d = 3$) qui sont généralement les plus utilisés. En généralisant la régression au modèle polynomial général $P(x) = \sum_{j=0}^{d} \alpha_j x^j$ de degré $d$, l'erreur résiduelle est donnée par

$$R = SSE(\alpha_0, \alpha_1, ..., \alpha_d) = \sum_{i=1}^{N} [y_i - \sum_{j=0}^{d} \alpha_j x_i^j]^2 \tag{5}$$

En annulant les dérivées partielles de $R$

$$\begin{array}{rcl} \frac{\partial(R)}{\partial(\alpha_0)} & = & -2\sum_{i=1}^{N}[y_i - (\alpha_0 + \alpha_1 x_i + ... + \alpha_d x_i^d)] = 0 \\ \frac{\partial(R)}{\partial(\alpha_1)} & = & -2\sum_{i=1}^{N}[y_i - (\alpha_0 + \alpha_1 x_i + ... + \alpha_d x_i^d)]x = 0 \\ \vdots & = & \vdots = 0 \\ \frac{\partial(R)}{\partial(\alpha_d)} & = & -2\sum_{i=1}^{N}[y_i - (\alpha_0 + \alpha_1 x_i + ... + \alpha_d x_i^d)]x^d = 0 \end{array} \tag{6}$$

Nous obtenons les équations :

$$\begin{array}{lcccccl} \alpha_0 N & + & \alpha_1 \sum_{i=1}^{N} x_i & + & ... & + & \alpha_d \sum_{i=1}^{N} x_i^d = \sum_{i=1}^{N} y_i \\ \alpha_0 \sum_{i=1}^{N} x_i & + & \alpha_1 \sum_{i=1}^{N} x_i^2 & + & ... & + & \alpha_d \sum_{i=1}^{N} x_i^{d+1} = \sum_{i=1}^{N} x_i y_i \\ \vdots & + & \vdots & + & ... & + & \vdots = \vdots \\ \alpha_1 \sum_{i=1}^{N} x_i^d & + & \alpha_1 \sum_{i=1}^{N} x_i^{d+1} & + & ... & + & \alpha_d \sum_{i=1}^{N} x_i^{2d} = \sum_{i=1}^{N} x_i^d y_i \end{array} \tag{7}$$

En forme matricielle, 7 est équivalent à

$$\begin{pmatrix} N & \sum_{i=1}^{N} x_i & \dots & \sum_{i=1}^{N} x_i^d \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 & \dots & \sum_{i=1}^{N} x_i^{d+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_i^d & \sum_{i=1}^{N} x_i^{d+1} & \dots & \sum_{i=1}^{N} x_i^{2d} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_d \end{pmatrix} = \begin{matrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \\ \vdots \\ \sum_{i=1}^{N} x_i^d y_i \end{matrix} \tag{8}$$

Cela permet d'avoir

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \dots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^d \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_d \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \tag{9}$$

En notation matricielle, on peut écrire le système d'equations obtenu sous la forme $\mathbf{X}\alpha = \mathbf{y}$. La solution optimale au sens des moindres carrés est

$$\alpha = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{10}$$

Cette solution est unique si les colonnes de **X** sont linéairement indépendantes. Comme **X** est une matrice de '*Vandermonde*', ce critère est vérifié si au moins $d + 1$ des $x_i$ sont distincts.

### 4.2.6 Algorithme de régression développé

Afin d'avoir une bonne approximation de la distribution de densité d'une CI, nous avons développé un algorithme de recherche des paramètres (le nombre et la position des nœuds, et les coefficients des polynômes) de la fonction spline optimale.

Certes, la seule contrainte considérée que chaque polynôme doit passer par le premier et le dernier point du sous-intervalle à ajuster diminue la qualité du lissage au voisinage des nœuds, mais cela ne représente pas un inconvénient essentiel dans la mesure où notre objectif est de trouver une spline contenant peu de nœuds et minimisant l'erreur. L'utilité d'une telle résolution est de permettre la décomposition du problème d'estimation en sous-problèmes de tailles inférieures $Reg(E_{ik}) = min_j(Reg(E_{ij}) + Reg(E_{jk}))$ ,$i < j < k$; cette décomposition est une conséquence directe du fait que la régression d'un intervalle de points est réalisée sans intervenir les résultats de régression des autres polynômes de la spline.

Ainsi, l'optimisation peut être faite d'une manière récursive. La méthode proposée est détaillée par les deux algorithmes 1 et 2, affichés sur les tableaux 6.1 et 6.2. La fonction '*RegresserSpline*' permet la recherche de la position optimale des $K$ nœuds de la spline de degré $d$ sur le vecteur de points de contrôle $(x, y)$. L'initialisation des variables globales est effectuée par l'Algorithme 1 et la recherche récursive est faite par la fonction '*AjusterRecursif*'.

La régression polynomiale '*RegresserPol*'; de l'intervalle de points de $E_{ij} = [(x(i), y(i))$ ,..., $(x(j), y(j))]$ est exécutée au maximum une seule fois et si oui son résultat est sauvegardé dans la case $H(i, j)$, initialisé au préalable à -1.

L'espace de recherche peut être vu comme un arbre de profondeur $K$ dont la racine est le nœud '1' et dont chaque nœud $i$ est le père des nœuds $(i + d)$ à $N - d \times (K - k)$.

La fonction '*AjusterRecursif*' explore progressivement l'espace de combinaisons en profondeur d'abord. Le chemin menant de la racine vers une feuille quelconque représente une solution acceptable. Durant l'exploration, la meilleure solution trouvée jusqu'à un moment donné est toujours enregistrée et la solution obtenue à la fin représente la solution optimale. Pour accélérer l'exploration, les arborescences dont l'erreur cumulée au niveau de leurs sous-racines est supérieure à l'erreur da la meilleure solution obtenue sont élaguées. Le chemin parcouru d'un exemple est montré sur la figure 6.2.

Les solutions intermédiaires obtenues de cet exemple sont montrées sur la figure 6.3, ces solutions représentent la trace du couple ($CombinaisonOptimale$ , $SSEOptimal$). Les résultats s'améliorent d'une solution intermédiaire à une autre et la solution optimale est la dernière solution (Solu .8 dans cet exemple).

La réutilisation des résultats de régression polynomiale et l'élagage opéré sur l'arbre de

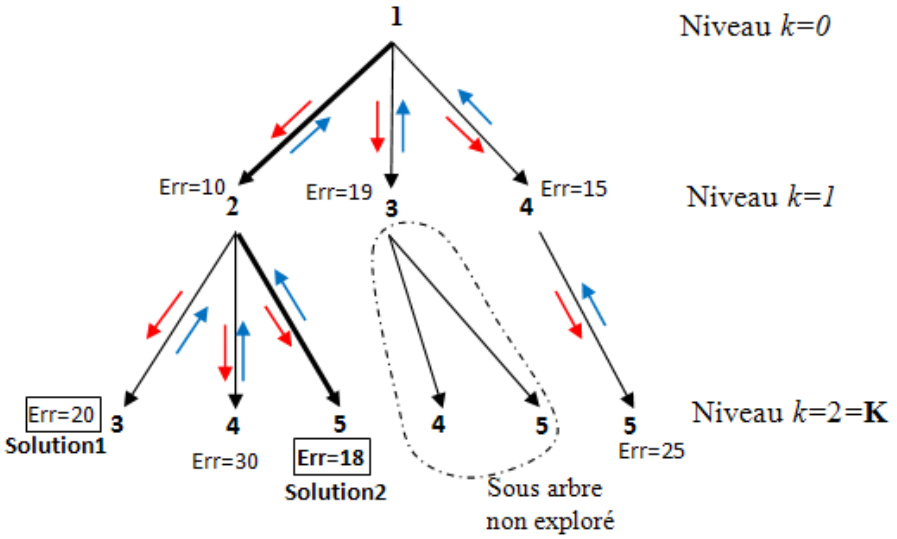



FIGURE 6.2 – Exemple de la trace de recherche de 06 points de contrôle.

K=2 (3 pieces), dégrée d=1. Le parcours suivi est indiqué en flèches rouges et bleues. La première solution trouvée est [1 2 3 6], la deuxième amélioration est [1 2 5 6]. La solution optimale est [1 2 5 6]. L'arborescence entourée est élagué (19 > 18). Ici, $Err(i)$ représente l'erreur cumulée ($SSEtotal$) de la partie de la spline estimée sur l'intervalle [1,...,i].

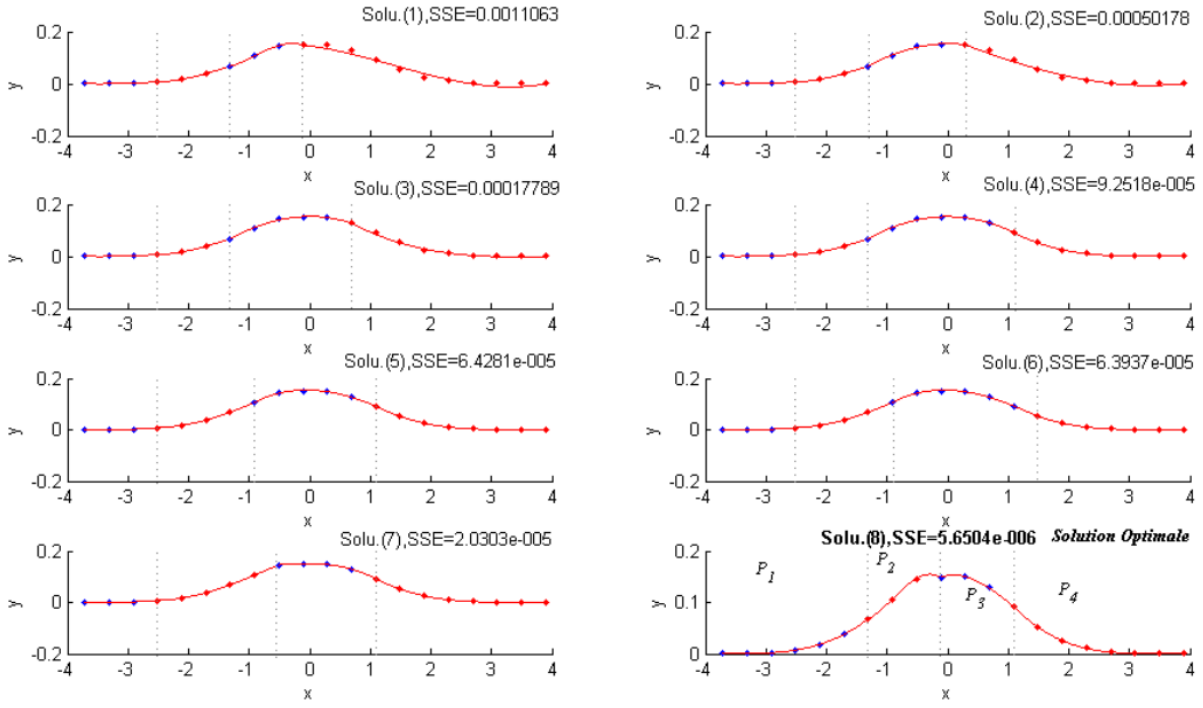



FIGURE 6.3 – Trace de recherche de la spline de l'exemple de la figure 6.2.

Tableau 6.1 – Algorithme 1. *RegresserSpline*($\mathbf{x}, \mathbf{y}, K, d$)

**Entrées**

$\mathbf{x}$,$\mathbf{y}$: vecteurs des points à ajuster ( $\mathbf{x}$ : abscisses et $\mathbf{y}$ : cordonnées)
$K$: le nombre de nœuds ⇒ nombre de polynômes est $K + 1$
$d$: degré de la spline

**Sorties**

*CombinaisonOptimale*: vecteur qui va contenir les indices des nœuds de la solution optimale
*SSEOptimal*: la somme des erreurs carrées de la solution optimale

**Début**

(1) $N \leftarrow taille(x) = taille(y)$ ; \ nombre de points à ajuster
(2) $SSEtotal \leftarrow 0$;
(3) $SSEOptimal \leftarrow 100000$; \ diminue au fur et à mesure de l'optimisation, initialisé à une valeur très grande.
(4) $Combinaison \leftarrow []$; \ vecteur vide
(5) $H$ : matrice $N \times N$ initialisé à -1
(6) *AjusterRecursif* (1,0); \ Appel de la fonction récursive d'optimisation

**Fin**

l'exploration permettent de réduire significativement le temps de régression.

# 5 Algorithme d'inférence

## 5.1 Apprentissage des paramètres

L'algorithme d'apprentissage du modèle d'inférence proposé (ACIL-RB) consiste en les étapes suivantes :

1. Regroupement des vecteurs de données d'apprentissage $X$ en un ensemble de $Nclu$ clusters 'homogènes', avec $X$ une matrice $n \times T$ composée de $T$ vecteurs de dimension $n$.
2. Estimation d'une matrice $W_j$ de séparation ACI pour chaque cluster $j$ ($j : 1, ..., Nclu$) en utilisant un algorithme ACI approprié, et puis le calcul des CIs locales $S_j = W_j \times X_j$ où $X_j$ est la matrice des vecteurs appartenant au cluster $j$.
3. Pour chaque composante indépendante locale $s_j^{(i)}$ ($i : 1, ..., n$) de chaque cluster $j$ ($j : 1, ..., Nclu$), $s_j^{(i)}$ est la $i^{ème}$ ligne de $S_j$, on effectue :
   - a) Estimation non paramétrique d'un histogramme de probabilité du vecteur $s_j^{(i)}$. Soit $H_j^{(i)}$ l'histogramme obtenu, constitué de $Nh$ intervalles.
   - b) Identification d'un ensemble de points de contrôle $PC_j^{(i)}$=[($s_j^{(i)}(1)$,$H_j^{(i)}(1)$),...,($s_j^{(i)}(Nh)$,$H_j^{(i)}(Nh)$)] avec $s_j^{(i)}(k)$ le centre du $k^{ème}$ intervalle et $H_j^{(i)}(k)$ sa probabilité estimée (voir l'exemple de la figure 6.1(a)).

Tableau 6.2 – Algorithme 2. $AjusterRecursif(i,k)$

**Entrées et Sorties**

Toutes les variables de la fonction *RegresserSpline*, à savoir : $x$, $y$, $K$, $d$, $N$ , $H$, $SSEtotal$, $SSEOptimal$, $Combinaison$ et $CombinaisonOptimale$ sont déclarées comme des variables globales et par conséquent elles sont accessibles en lecture et en écriture par cette fonction. La combinaison optimale des nœuds sera rangée dans $CombinaisonOptimale$ à la fin de l'execution.

$k$: on recherche le $k^{ème}$ nœud

$i$: indice du point du début de l'intervalle à explorer [i...N] sachant que $k-1$ nœuds sont trouvés sur l'intervalle [1...$i$]

**Début**

(1) **Si** $(k = K)$ \ dernier segment
(2) **Si** $(H(i,N) = -1)$
(3) $SSE \leftarrow RegresserPol(i, N, \mathbf{x}, \mathbf{y}, d)$; \ appel de la fonction de régression polynomiale de l'intervalle des points de $i$ à $N$ par un polynôme de degré $d$
(4) $H(i,N) \leftarrow SSE$;
(5) **Sinon**
(6) $SSE \leftarrow H(i,N)$;
(7) **Fin Si**
(8) $SSEtotal \leftarrow SSEtotal + SSE$;
(9) **Si** $(SSEtotal < SSEOptimal)$;
(10) $SSEOptimal \leftarrow SSEtotal$;
(11) $CombinaisonOptimale \leftarrow Combinaison$;
(12) **Fin Si**
(13) $SSEtotal \leftarrow SSEtotal - SSE$;
(14) **Sinon**
(15) **Pour** $j = (i + d)$ à $(N - d) \times (K - k)$
(16) $Combinaison(k + 1) \leftarrow j$;
(17) **Si** $(H(i, j) = -1)$
(18) $SSE \leftarrow RegresserPol(i, j, \mathbf{x}, \mathbf{y}, d)$;
(19) $H(i, j) \leftarrow SSE$;
(20) **Sinon**
(21) $SSE \leftarrow H(i, j)$;
(22) **Sinon**
(23) $SSEtotal \leftarrow SSEtotal + SSE$;
(24) **Si** $(SSEtotal > SSEOptimal)$;
(25) $SSEtotal \leftarrow SSEtotal - SSE$;
(26) **continue**; \ Itération suivante : $j \leftarrow j + 1$ et **Aller** à 15;
(27) **Fin Si**
(28) $AjusterRecursif(j, k + 1)$;
(29) $SSEtotal \leftarrow SSEtotal - SSE$;
(30) **Fin Pour**
(31) **Fin Si**

**Fin**

– c) Régression des points $PC_j^{(i)}$ par une spline en utilisant l'algorithme proposé dans la section 4.2.6.

Les paramètres du modèle proposé comportent les paramètres de clustering, les matrices de séparation ACI locale et les paramètres des splines utilisées pour l'approximation de la distribution de densité des CIs.

## 5.2 Inférence

Il convient de rappeler que notre objectif consiste à calculer la probabilité a posteriori d'une variable d'intérêt $y$ étant données des évidences $Z$ (une généralisation de l'algorithme présenté dans V.§2 au cas non linéaire), en d'autres mots nous voulons estimer

$$P(y|Z) = P(y \in \Delta_y | z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k}) \tag{11}$$

Par la marginalisation de la conjointe sur les clusters du modèle ACIL-RB, nous avons

$$P(y|Z) = \sum_{j=1}^{Nclu} P(y|Z, Clu_j) \tag{12}$$

et qui peut être factorisée ainsi

$$P(y|Z) = \sum_{j=1}^{Nclu} [P(y|(Z, Clu_j)) \times P(Clu_j)] \tag{13}$$

avec $P(y|(Z, Clu_j)) = P(y \in \Delta_y | (z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k}\ et\ [y, Z]^T \in Clu_j))$ la probabilité conditionnelle $P(y|Z)$ au niveau du cluster $j$ et $P(Clu_j)$ la probabilité a priori du cluster $j$.

En suivant une approche d'évaluation similaire à celle décrite dans V.§4.3, nous avons

$$P(y|(Z, Clu_j)) = \frac{P(y \in \Delta_y, z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k} | Clu_j)}{P(z_1 \in \Delta_{z_1}, ..., z_k \in \Delta_{z_k} | Clu_j)} \tag{14}$$

$$P(y|(Z, Clu_j)) = \frac{\int\int_{\Delta_{yZ}} P(s_j^{(1)}, ..., s_j^{(n)}) ds_j^{(1)} ... s_j^{(n)}}{\int\int_{\Delta_Z} P(s_j^{(1)}, ..., s_j^{(n)}) ds_j^{(1)} ... s_j^{(n)}} \tag{15}$$

où $\Delta_{yz} = \Delta_y \cap \Delta_{z_1} \cap ... \cap \Delta_{z_k}$ et $\Delta_z = \Delta_{z_1} \cap ... \cap \Delta_{z_k}$ sont des hyper-espaces dans l'espace des CIs locales associé au cluster $j$, définis comme étant l'intersection des valeurs instanciées des variables observées dans la requête d'inférence. Formellement, $\Delta_v = \Delta_{v_1} \cap ... \cap_{v_k}$ est défini comme suit :

$$\Delta_v = \left\{ (s_j^{(1)}, ..., s_j^{(n)}) \quad / \quad \Delta_{v_{ia}} \leq A_j^{(i)} \times (s_j^{(1)}, ..., s_j^{(n)})^T \leq \Delta_{v_{ib}} \quad i = 1, ..., k \right\} \tag{16}$$

avec $\Delta_{v_{ia}}$ et $\Delta_{v_{ib}}$ les limites inférieure et supérieure de l'intervalle $\Delta_{v_i}$, et $A_j^{(i)}$ la $i^{ème}$ ligne de la matrice de mélange ACI du cluster $j$.

La distribution de densité de chaque CI est approximée par une fonction polynomiale définie par morceaux. Du fait que les CIs sont déterminées en maximisant leurs indépendances statistiques mutuelles, la conjointe $P(s_j^{(1)}, ..., s_j^{(n)})$ peut être factorisée en produit, elle est en effet définie par morceaux comme suit :

$$Ps = P(s_j^{(1)}, ..., s_j^{(n)}) = \prod_{i=1}^{n} Poly_{k_i,i,j}(s_j^{(i)}) \;\; S\,i \;\; s_j^{(i)} \in I(Poly_{k_i,i,j})(\forall i : 1...n) \tag{17}$$

avec $I(Poly_{k_i,i,j})$ le domaine de définition du polynôme d'indice $k_i$ de la $i^{ème}$ CI du cluster $j$. Le domaine des CIs peut être vu comme un hypercube en dimension $n$. Avec la modélisation basée sur les fonctions splines que nous avons adoptée, la conjointe $P(s)$ est définie par un polynôme multivariable dans chaque sous-hypercube $H = \{(s_j^{(1)}, ..., s_j^{(n)}) \;/\; s_j^{(i)} \in I(Poly_{k_i,i,j}) \; (\forall i : 1...n)\}$.

L'évaluation de l'intégrale multiple du numérateur et aussi bien du dénominateur de (15) peut être effectuée de la même manière. A titre d'exemple, l'évaluation du numérateur peut être décomposée en une sommation d'intégrales des polynômes multivariables :

$$\int\int_{\Delta_Z} Ps \; ds_j^{(1)}...ds_j^{(n)} = \sum_{i=1}^{N\Phi} \int\int_{\Phi_i} Ps \; ds_j^{(1)}...ds_j^{(n)} \tag{18}$$

où $\Phi_i(i : 1...N_\Phi)$ sont des polytôpes en dimension $n$ qui sont issus de l'intersection de l'hyperespace des variable observées instanciées $\Delta_Z$ (16) avec les sous-hypercubes du domaine de définition de la conjointe des CIs $H = \{(s_j^{(1)}, ..., s_j^{(n)}) \;/\; s_j^{(i)} \in I(Poly_{k_i,i,j}) \; (\forall i : 1...n)\}$. Nous avons proposé l'exemple de la figure 6.4 pour bien illustrer la technique proposée.

Dans le modèle ACI local associé au cluster d'indice $j$ de l'exemple affiché sur la figure 6.4, nous avons $(x_1, x_2)^T = A_j \times (s_j^{(1)}, s_j^{(2)})^T$. La densité de la composante $s_j^{(1)}$ est approximée par la spline $[Poly_{1,1,j}, Poly_{2,1,j}, Poly_{3,1,j}$, et celle de $s_j^{(2)}$ est approximée par la spline $[Poly_{1,2,j}, Poly_{2,2,j}]$. Comme le montre la figure 6.4, $\Delta_Z = \Delta_{(Z_1)} \cap \Delta_{(Z_2)}$ est l'hyper-espace défini comme suit :

$$\Delta = \begin{cases} A_j^{(1)} \times (s_j^{(1)}, s_j^{(2)})^T \geq \Delta_{Z_{1a}} & \backslash A_j^{(1)} = [A_j(1,1), A_j(1,2)] \\ A_j^{(1)} \times (s_j^{(1)}, s_j^{(2)})^T \geq \Delta_{Z_{1b}} & \backslash A_j^{(1)} = [A_j(1,1), A_j(1,2)] \\ A_j^{(2)} \times (s_j^{(1)}, s_j^{(2)})^T \geq \Delta_{Z_{2a}} & \backslash A_j^{(2)} = [A_j(2,1), A_j(2,2)] \\ A_j^{(2)} \times (s_j^{(1)}, s_j^{(2)})^T \geq \Delta_{Z_{2b}} & \backslash A_j^{(2)} = [A_j(2,1), A_j(2,2)] \end{cases}$$

L'intersection de $\Delta_Z$ avec le domaine des CIs est le polygone 'abcde'. La conjointe $Ps$ est approximée par les polynômes multivariables $Poly_{1,1,j} \times Poly_{1,2,j}$, $Poly_{2,1,j} \times Poly_{1,2,j}$ et $Poly_{2,1,j} \times Poly_{2,2,j}$ définis sur les polygones $\Phi_1$, $\Phi_2$ et $\Phi_3$, respectivement. Ainsi , nous avons $\int\int_{\Delta_Z} Ps \; ds_j^{(1)} ds_j^{(2)} = \sum_{i=1}^{3} \int\int_{\Phi_i} Ps \; ds_j^{(1)} ds_j^{(2)}$.

En résumé, l'inférence dans le modèle proposé revient à évaluer des intégrales multiples de fonctions polynomiales multivariables sur des polytôpes convexes issus de l'intersection des valeurs observées instanciées des variables observées avec les intervalles des polynômes

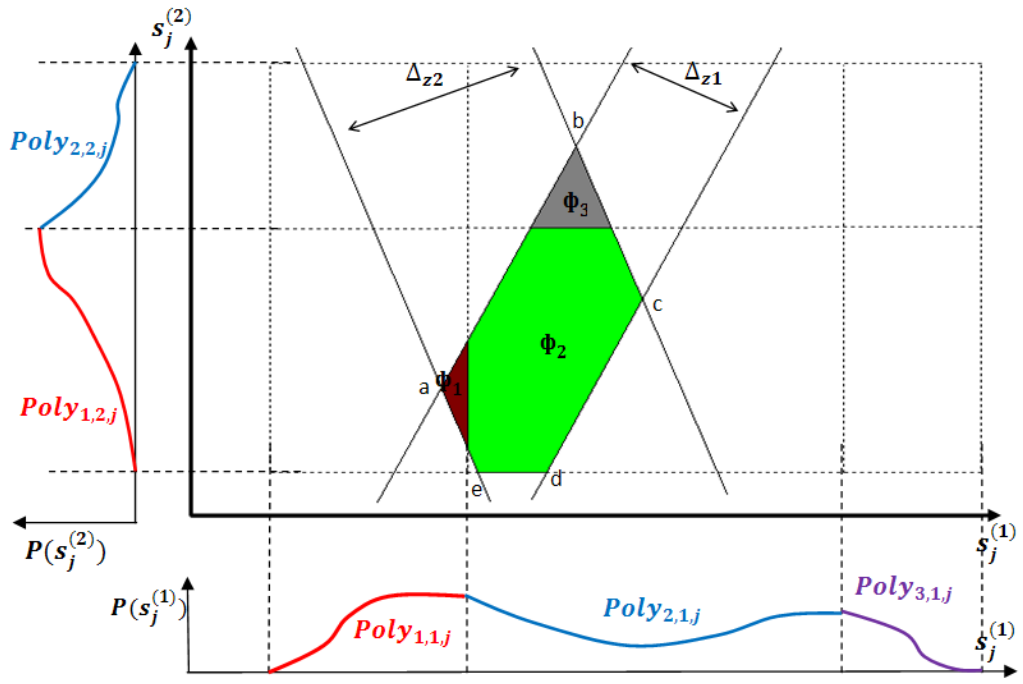



FIGURE 6.4 – Exemple en dimension 2 d'inférence géométrique dans le modèle ACIL-RB.

des splines utilisées pour approximer la densité des CIs. En conséquence, l'efficacité d'une telle approche dépend des performances d'intégration des polynômes multivariables sur des polytôpes en dimensions élevées. Dans cette optique, la section suivante (VI.§6) sera consacrée à la présentation de l'algorithme d'intégration des polynômes que nous avons proposé. Il est opportun de mentionner que le modèle ACI-RB proposé dans le chapitre précédent (en utilisant des histogrammes : spline de degré 0 et comportant un seul cluster) est un cas spécial du modèle ACIL-RB (utilisant des splines de degré $d$ et comportant plusieurs clusters).

# 6 Intégration de polynômes sur des polytôpes convexes

## 6.1 Introduction

L'évaluation d'intégrales multiples de polynômes multivariables sur des polytôpes convexes est une tâche indispensable dans de nombreuses applications. L'intégration sur des polytôpes est communément utilisée par les méthodes des éléments finis [Zienkiewicz and Taylor, 1988], ou bien dans le calcul de la masse et du centre de gravité en mécanique. Les intégrales multiples sur les polytôpes jouent aussi un rôle important en statistique où on s'intéresse à intégrer la distribution conjointe d'une variable aléatoire sur un domaine convexe. Il convient de souligner sur le fait que l'emploi des fonctions polynomiales dont les splines pour l'approximation de

distributions a été bien traité, l'application des splines dans l'estimation non paramétrique de densité est décrite dans [Wright, 1977].

En général, l'intégration de polynômes sur des polytôpes est une tâche difficile. Il est démontré que le calcul des volumes de polytôpes (le volume correspond au polynôme de degré 0) de dimensions quelconques est un problème NP-Complet (voir [Gritzmann and Klee, 1994] et les références citées dedans), et même l'approximation des volumes de polytôpes est un problème crucial [Elekes, 1986].

Une des approches couramment adoptées pour ce faire consiste en la décomposition du polytôpe en simplexes (pour une certaine dimension, un simplexe est le polytôpe ayant le moins nombre de sommets et dont le volume est non nul, par exemple un triangle en dimension 2). Un état de l'art riche ainsi que quelques algorithmes d'intégration de polynômes sur un simplexe sont fournis dans [Baldoni et al., 2011].

En examinant la littérature des méthodes d'intégration de polynômes sur des polytôpes, les techniques publiées sont souvent soit difficiles à mettre en œuvre, soit restreintes à des classes réduites de polynômes et de polytôpes, soit basées sur des approximations qui sont faites au détriment de la précision. Dans cette section, nous présentons la méthode d'intégration générale basée sur la triangulation ainsi que notre contribution, faite pour accélérer l'intégration.

## 6.2 Énonce du problème

Soit $\Phi$ un polytôpe convexe dans l'espace de dimension $n$ (un polytôpe est une généralisation du concept du polygone ($n = 2$) à des dimensions quelconques ($n = 2, 3, ...$)) défini par un ensemble de sommets, et soit $f(x_1, ..., x_n)$ un polynôme multivariable de degré $d$ défini sur cet espace. Le polynôme $f$ peut être alors écrit sous la forme simple d'une sommation de ses $K$ monômes comme suit :

$$f(x_1, ..., x_n) = \sum_{j=1}^{K} \gamma_j x_1^{\alpha_{1j}} x_2^{\alpha_{2j}} ... x_n^{\alpha_{nj}} \tag{19}$$

où les puissances entières $\alpha_{ij}$ ($i = 0, ..., n$) et le réel $\gamma_j$ représentent respectivement les degrés et le coefficient du $j^{eme}$ monôme du $f$, et nous avons pour tout $j = 1, ..., k$ la contrainte

$$\alpha_{1j} + \alpha_{2j} + ... + \alpha_{nj} \leq d$$

$f$ est dit de degré $d$ si le degré de chaque monôme est inférieur ou égal à $d$. Par exemple, le polynôme $f(x_1, x_2) = 2x_1^2 x_2 - 7x_1 x_2 + x_2^3 + 3$ est un polynôme multivariable de dégrée 3 ($d = 3$) dans l'espace de dimension 2 ( $n = 2$).

Notre objectif consiste à évaluer numériquement la valeur précise de l'intégrale multiple suivante définie ainsi

$$If = \int_{\Phi} f(x_1, ..., x_n) dx_1 ... dx_n \tag{20}$$

## 6.3 Intégration d'un polynôme sur un polytôpe par la triangulation

La méthode générale et également la plus utilisée pour l'évaluation de $If$ consiste en la triangulation simpliciale du polytôpe $\Phi$. Pour ce faire, l'intégration est réalisée en suivant les 03 étapes :

1. Décomposition du polytôpe $\Phi$ en un ensemble de simplexes;
2. Evaluation des intégrales de $f$ sur chaque simplexe;
3. Evaluation de $If$ ; $If$ est déduit par la sommation des intégrales des simplexes obtenues à l'étape 2).

Les performances de cette approche dépendent en gros de celles de l'intégration de $f$ sur les simplexes en dimensions élevées. Il convient de noter que les simplexes obtenus sont souvent non homogènes ; en effet, leurs formes ainsi que leurs positions dans l'espace sont pseudo-aléatoires.

## 6.4 Intégration d'un polynôme multivariable sur le simplexe standard (simplexe unitaire)

Dans l'espace de dimension $n$, le simplexe unitaire $\Delta_u$ est défini par les $n + 1$ sommets suivants $V_0 = (0, 0, ..., 0)$, $V_1 = (1, 0, ..., 0)$, $V_2 = (0, 1, ..., 0)$,... ,$V_n = (0, 0, ..., 1)$ En d'autres mots, un point $x(x_1, x_2, ..., x_n)$ appartient au simplexe $\Delta_u$ seulement si $x_1 + x_2 + ... + x_n \leq 1$ et pour tout $i = 1, ..., n$, nous avons la contrainte $x_i \geq 0$.

On dénote par $If_{\Delta_u}$ l'intégrale du polynôme $f$ sur le simplexe unitaire $\Delta_u$, définie ainsi

$$If_{\Delta_u} = \int_{x_1=0}^{x_1=1} \int_{x_2=0}^{x_2=1-x_1} ... \int_{x_n=0}^{x_n=1-x_1...-x_{n-1}} f(x_1, ..., x_n) dx_1 ... dx_n \tag{21}$$

L'intégrale d'un seul monôme $x^{\alpha} = x_1^{\alpha_1} x_2^{\alpha_2} ... x_n^{\alpha_n}$ de $f$ sur $\Delta_u$ est donnée par la formule [STROUD, 1971]:

$$\begin{array}{rcl} Ix^{\alpha}_{\Delta_u} & = & \int_{\Delta_u} f(x_1, ..., x_n) dx_1 ... dx_n \\ & = & \frac{\alpha_1! \alpha_2! ... \alpha_n!}{(n+\alpha_1+\alpha_2+...+\alpha_n)!} \end{array} \tag{22}$$

Par conséquent, nous pouvons déduire facilement la valeur de $If_{\Delta_u}$ par l'addition des valeurs d'intégrales de ses monômes $x^{\alpha j}$ ($j = 1, ..., K$) et qui peuvent être calculées par la formule (22), nous avons donc

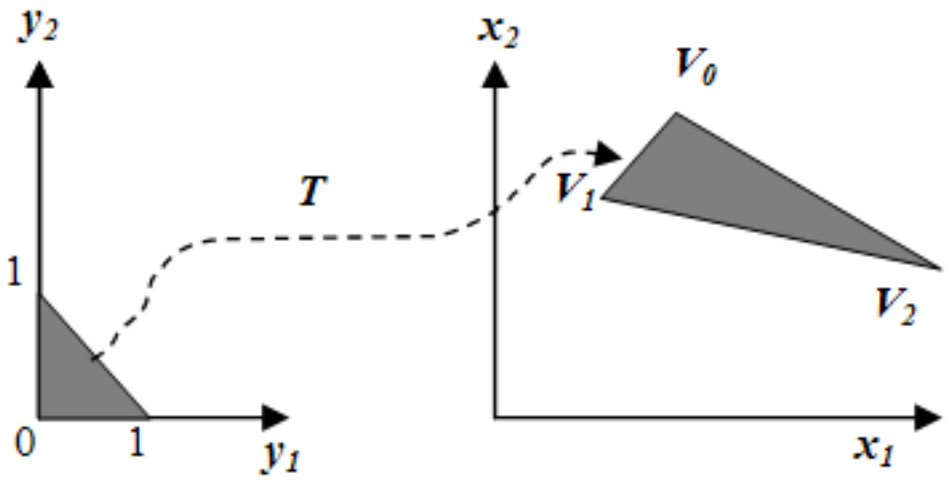



FIGURE 6.5 – Exemple de transformation affine du simplexe standard vers un simplexe général, cas de $n = 2$.

$$If_{\Delta_u} = \sum_{j=1}^{K} \gamma_j \, x_{\Delta_u}^{\alpha j} \tag{23}$$

## 6.5 Intégration d'un polynôme multivariable sur un simplexe (cas général)

Les simplexes provenant de la triangulation du polytôpe $\Phi$ sont souvent non unitaires. L'évaluation de $If$ sur un simplexe général $\Delta$ défini par les $n+1$ sommets $V_0 = (v_1^0, v_2^0, ..., v_n^0)$,$V_1 = (v_1^1, v_2^1, ..., v_n^1)$,...,$V_n(v_1^n, v_2^n, ..., v_n^n)$ est relativement complexe. Afin de pouvoir se servir de la formule (22), il semble intéressant d'effectuer un changement de variables pour aller du simplexe unitaire $\Delta_u$ vers le simplexe général $\Delta$, et puis d'en déduire la valeur de $If$. Pour ce faire, nous pouvons déterminer une transformation affine $T : \Delta_u \rightarrow \Delta$ qui mappe le simplexe unitaire $\Delta_u$ vers le simplexe $\Delta$. Un des problèmes qu'on doit inspecter avec soin est: est-ce que cette transformation $T$ existe toujours? Pour répondre à cette question, on va démontrer l'existence d'une telle transformation en donnant ses paramètres pour n'importe quel simplexe $\Delta$. On suppose que $T$ mappe chaque sommet du simplexe standard $\Delta_u$ dans l'espace $Y$ de dimension $n$ vers un sommet distinct du simplexe de l'intégrale $\Delta$ dans l'espace $X$ de dimension $n$ (pour l'illustration, voir l'exemple de 02 variables sur la figure 6.5).

Formellement, nous avons $X = A \times Y + B$ avec $A$ une matrice carrée de taille $n \times n$ et B un vecteur de translation de taille $n$. La transformation des sommets du simplexe standard $\Delta_u$ vers les sommets du simplexe général $\Delta$ nous donne le système d'équations suivant :

$$\begin{pmatrix} v_1^0 & \dots & v_1^n \\ \vdots & \ddots & \vdots \\ v_n^0 & \dots & v_1^n \end{pmatrix} = \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{n,1} & \dots & A_{n,n} \end{pmatrix} \times \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} + \begin{pmatrix} B_1 \\ \vdots \\ B_n \end{pmatrix} \tag{24}$$

L'ordre des sommets n'est pas important, il suffit de choisir arbitrairement un mapping bijectif entre les sommets de $\Delta_u$ et les sommets de $\Delta$. Si le volume de $\Delta$ est non nul, ce système d'équations linéaires admet toujours une seule solution exacte qui est

$$\begin{cases} B = V_0 & (c'est - á - dire\ B_i = v_i^0\ pour\ i : 1..n \\ A_{i,j} = v_i^j - B_i & pour\ i, j : 1..n \end{cases} \tag{25}$$

La transformation proposée, $T : X = A \times Y + B$ mappe les sommets du simplexe unitaire $\Delta_u$ vers le simplexe $\Delta$. Pour ce type de transformation (transformation affine), nous avons la propriété de la conservation de la convexité ; si un point $y$ se situe à l'intérieur de $\Delta_u$, alors son image, $x = T(y)$ est forcement à l'intérieur de $\Delta$. En outre, si le volume de $\Delta$ est non nul, cette transformation est inversible (déterminant de $A \neq 0$) et son inverse, $T^{-1} : Y = A^{-1} \times X - A^{-1}B$, est également une transformation affine. Pour l'évaluation de $If_\Delta$

$$If_\Delta = \int_\Delta f(x_1, ..., x_n)dx_1...dx_n \tag{26}$$

Nous pouvons procéder par le changement de variables $X = A \times Y + B$, on aura donc

$$\begin{aligned} x_i &= A_{i1}y_1 + ... + A_{in}y_n + Bi \\ &= \textstyle\sum_{j=1}^n A_{ij}y_j + B_i \quad pour\ i = 1, ..., n \end{aligned} \tag{27}$$

Cela donne l'intégrale suivante

$$if_\Delta = \textstyle\int_{\Delta_u} f(\sum_{j=1}^n A_{1j}y_j + B_1, ..., \sum_{j=1}^n A_{nj}y_j + B_n)|det(Jac_{T(y_1,...,y_n)})|dy_1...dy_n \tag{28}$$

avec $Jac_{T(y_1,...,y_n)} = A$ la matrice jacobienne de la transformation $T$. Après la simplification on aura

$$if_\Delta = \int_{\Delta_u} f(\sum_{j=1}^n A_{1j}y_j + B_1, ..., \sum_{j=1}^n A_{nj}y_j + B_n)|det(A)|dy_1...dy_n \tag{29}$$

Pour le cas simple où le polynôme $f$ comporte un seul monôme $x^\alpha = x^{\alpha_1}...x^{\alpha_n}$, par la substitution de $x^\alpha$ dans (29), $Ix_\Delta^\alpha$ peut se calculer comme suit :

$$Ix_\Delta^\alpha = \int_{\Delta_u} (\sum_{j=1}^n A_{1j}y_j + B_1)^{\alpha_1}...(\sum_{j=1}^n A_{nj}y_j + B_n)^{\alpha_n}|det(A)|dy_1...dy_n \tag{30}$$

Soit $h$ le polynôme de degré $\alpha$ défini comme suit :

$$h(y) = (\sum_{j=1}^{n} A_{1j}y_j + B_1)^{\alpha_1}...(\sum_{j=1}^{n} A_{nj}y_j + B_n)^{\alpha_n}|det(A)| \tag{31}$$

Donc $Ix_{\Delta}^{\alpha}$ peut se réécrire

$$Ix_{\Delta}^{\alpha} = \int_{\Delta_u} h(y)dy_1...dy_n \tag{32}$$

La simplification de $h(y)$ par la distribution donne un polynôme du même degré $d = \alpha = \alpha_1 + ... + \alpha_n$ dans le nouvel espace $Y$. Par conséquent, l'évaluation de $Ix_{\Delta}^{\alpha}$ peut se faire par l'intégration du polynôme $h(y)$ sur le simplexe unitaire $\Delta_u$ et qui peut être accomplie à l'aide des formules (22) et (23).

Pour le cas général où $f$ est constitué de plusieurs monômes, l'évaluation de $If_{\Delta}$ peut être faite en suivant les étapes :

1. Calcul du polynôme $h(y)$ issu du changement de variables correspondant à chaque monôme de $f$ ;
2. Addition des polynômes $h(y)$ obtenus en étape 1 tout en simplifiant les monômes ayant les mêmes degrés. Par exemple, $3y_1^2y_2^1$ et $4y_1^2y_2^1$ sont remplacés par $7y_1^2y_2^1$ ;
3. Simplification des monômes équivalents de $h(y)$, deux monômes sont équivalents s'ils donnent la même valeur d'intégrale sur le simplexe unitaire. En analysant la formule (22), ils sont équivalents s'il y a une permutation entre les ensembles de leurs puissances. Par exemple, $2y_1^2y_2^1$ et $3y_1^1y_1^2$ sont simplifiées par $5y_1^2y_2^1$, car $(1, 2)$ est une permutation de $(2, 1)$. A la fin, on aura un seul polynôme multivariable simplifié dans l'espace Y ;
4. Evaluation de l'intégrale du polynôme simplifié obtenu en étape 3 en utilisant les formules (22) et (23).

## 6.6 Amélioration proposée pour accélérer l'intégration

### 6.6.1 Objectif

Dans la solution proposée précédemment, l'intégration d'un polynôme de degré $d$ sur un polytôpe de dimension $n$ est faite par la décomposition du polytôpe $\Phi$ en simplexes et ensuite par le changement de variables de l'équation (27). Pour un seul simplexe, après avoir effectué le changement de variables conduisant du simplexe d'intégration $\Delta$ vers le simplexe unitaire $\Delta_u$, le calcul de $h(y)$ dans (31) est réalisé par une série d'opérations de multiplication et d'addition de polynômes multivariables intermédiaires. Si la multiplication de deux polynômes est faite par la distribution de leurs monômes, la simplification des monômes résultants ayant les mêmes degrés est nécessaire. Malheureusement, la simplification requit énormément du temps et par

conséquent elle nécessite d'être optimisée. On rappelle que sans simplification, les polynômes intermédiaires et le polynôme final $h(y)$, seront trop volumineux (une explosion combinatoire) et la formule (22) sera exécutée autant de monômes que $h(y)$ a.

### 6.6.2 Une constatation intéressante

Nous avons constaté qu'après avoir effectué le changement de variables de l'équation (27), toutes les opérations de simplification de $h(y)$ en (31) comportant entre autre les additions et les multiplications sont faites sur des polynômes multivariables complets ; par le mot '*complet*', nous voulons dire qu'un polynôme de degré $d$ est complet, s'il contient tous les monômes de degré $\leq d$. Cette remarque est intéressante dans la mesure où tous les polynômes intermédiaires obtenus durant le changement de variables ne sont pas creux.

La multiplication des polynômes est très gourmande en terme du temps d'exécution, la multiplication de deux polynômes revient à multiplier les monômes qui les composent deux a deux par la multiplication de leurs coefficients et l'addition de leurs puissances. A la fin de distribution, la simplification des monômes ayant les mêmes degrés est nécessaire. Durant la multiplication de deux polynômes, la simplification peut être également faite au fur et à mesure de la distribution par le rangement des monômes résultat dans le bon endroit dans le polynôme produit.

Soient $m_1$ et $m_2$ deux monômes et leur multiplication le monôme $z = m_1 \times m_2$. Plusieurs couples $(p_1, p_2)$ peuvent donner des monômes ayant les mêmes degrés que ceux de $z$. Par exemple, $x_1^2 x_2^1 \times x_1^0 x_2^3$ et $x_1^1 x_1^2 \times x_1^1 x_1^2$ donnent le mémé monôme $x_1^2 x_2^4$.

Étant donnés deux monômes $m_1$ et $m_2$ à multiplier, c'est la recherche de l'emplacement d'un monôme $w$ (s'il existe) ayant les mêmes degrés que $z = m_1 \times m_2$ dans le polynôme produit qui consomme beaucoup de temps. Dans l'exemple précité, en supposant qu'on va multiplier $2x_1^2 x_2^1$ par $5x_1^0 x_2^3$ en premier lieu, le résultat $w = 10x_1^2 x_2^4$ est sauvegardé en mémoire quelque part dans le polynôme produit. Ensuite, lorsque on multiplie $2x_1^1 x_1^2$ par $x_1^1 x_1^2$, le résultat est $z = 2x_1^2 x_2^4$, on cherche d'abord si un monôme ayant les degrés $x_1^2 x_2^4$ est déjà créé dans le polynôme produit, dans cet exemple on trouve $w = 10x_1^2 x_2^4$, on lui ajoute $2x_1^2 x_2^4$, ce qui donne $12x_1^2 x_2^4$.

### 6.6.3 Proposition

La mémoire physique est un espace continu et n'importe quel polynôme, final ou intermédiaire obtenu lors de l'intégration, est sauvegardé effectivement sur cet espace. Les multiplications et les additions des polynômes sont faites selon les puissances des variables des monômes qui les constituent. L'addition de deux polynômes consiste seulement à additionner des coefficients des monômes ayant les mêmes puissances. Cependant, la multiplication est réalisée par la distribution en multipliant les coefficients des monômes et en additionnant leurs puissances, puis par la simplification des monômes obtenus qui ont les mêmes degrés dans le polynôme

produit.

Notre idée consiste à proposer une structure de données compacte pour la représentation de polynômes de telle sorte que les monômes sont accessibles directement par leurs puissances, ensuite de trouver une fonction du mapping rapide des éléments de cette structure vers la mémoire physique et la fonction du mapping inverse.

Puisque tous les polynômes manipulés sont complets, nous avons proposé de représenter un polynôme de degré $d$ et impliquant $n$ variables par *un simplex entier et droit* dans l'espace de dimension $n$. Un exemple illustrant la structure proposée où les coefficients d'un polynôme de $n = 3$ et de $d = 3$ sont représentés sur une structure simpliciale est montré sur la figure 6.6. Pour chaque monôme $\gamma x_1^{V_1} x_2^{V_n} ... x_n^{V_n}$, on sauvegarde seulement son coefficient $\gamma$ dans la case $(V_1, ..., V_n)$ du simplexe entier, les puissances des variables $V_1, ..., V_n$ ne sont pas sauvegardées explicitement, mais elles sont liées à l'emplacement du monôme dans le simplexe entier.

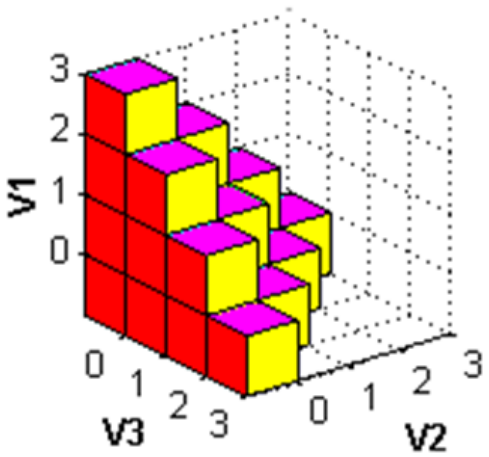



FIGURE 6.6 – Exemple d'un polynôme complet de degré $d = 3$ en dimension $n = 3$ sauvegardé dans un simplexe entier.

Les coefficients sont stockés dans une structure simpliciale, la case $(V_1, V_2, V_2)$ du simplexe contient le coefficient du monôme $x_1^{V_1} x_2^{V_2} x_3^{V_3}$.

Sur l'exemple affiché sur la figure 6.7, le monôme $0.5 x_1^2 x_2^1 x_3^0$ a les puissances $(2, 1, 0)$, son coefficient 0.5 est rangé dans la case $(2, 1, 0)$ du simplexe entier et qui correspond à la case mémoire d'indice 18. Les numéros affichés dans les cubes du simplexe entier de la figure 6.7 représentent les emplacements physiques des monômes en mémoire physique.

Comme le montre l'exemple de la figure 6.7, nous avons proposé l'ordre du rangement des monômes en mémoire par l'incrémentation des coordonnées les plus à droite en premier; c'est-à-dire, on incrémente $V_i$ seulement si les $V_j$ , $j = i + 1, ..., n$ sont tous saturés.

Les fonctions associées du mapping des coordonnées d'un monôme dans le simplexe entier

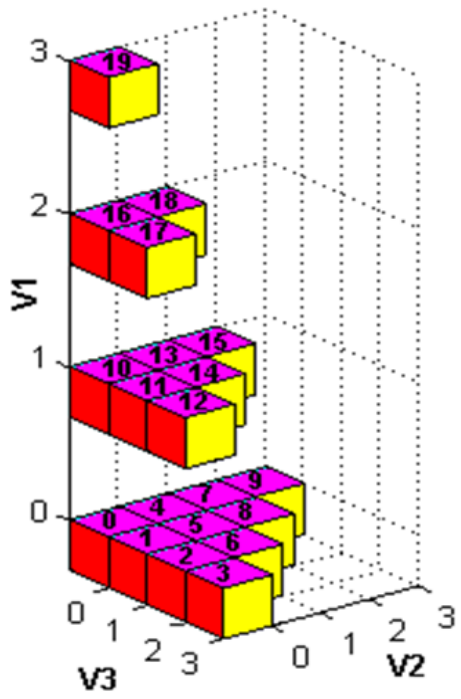



FIGURE 6.7 – Ordre des coefficients des monômes en mémoire physique. Un exemple d'un polynôme complet de degré $d = 3$ en dimension $n = 3$.

vers l'emplacement physique et l'opération inverse sont données respectivement par Algorithme 3 et Algorithme 4 montrés ci-dessous. Le calcul de la base de l'incrémentation des coordonnées des monômes dépend de degré $d$ et de dimension $n$, et elle est en effet contrôlée par le '*carré de Pascal*' affiché sur la figure 6.8(a).

La trace d'exécution d'un exemple de ces deux algorithmes de conversion est montrée sur la figure 6.8(b) et sur la figure 6.8(c), respectivement. Le nombre d'itérations de ces algorithmes est au pire de cas égal à $d + n$, l'accès au mémoire est linéaire.

Le temps de multiplication et d'addition de polynômes dépend particulièrement de la rapidité de la fonction du mapping des coordonnées du simplex entier (*Cord*: puissances du monôme) vers l'endroit physique (*Ind*) où le coefficient est sauvegardé et de la fonction du mapping inverse. Les deux algorithmes du mapping que nous avons proposés nécessitent au maximum $n + d$ itérations, ainsi l'accès est optimisé (l'accès est linéaire). Avec cette nouvelle technique, le temps d'intégration est réduit d'environ 60% par rapport à implémentation de l'algorithme d'intégration détaillé en Section 5.

Tableau 6.3 – Algorithme 3. Conversion des puissances d'un monôme vers son indice en mémoire physique.

**Sorties**

Ind: indice en mémoire physique du monôme ayant les puissances $Cord$.

**Entrées**

$Cord$: un vecteur $1 \times n$ des cordonnées (puissances) du monôme dans le simplexe.
$n$: dimension du polynôme.
$d$:degré du polynôme.
$Pascal$: matrice constante telle qu'elle est définie sur la figure 6.8(a).

**Début**

(1) $Ind \leftarrow 0$;
(2) $var \leftarrow n - 1$;
(3) $i \leftarrow d$; \ la case de départ dans dans $Pascal[var, i]$.
(4) $c \leftarrow 0$; \ la cordonné la plus à gauche dans $Cord$
(5) **Tant que** $(Cord \neq 0)$ \ toutes les composantes de $Cord \neq 0$
(6) **Si** $(Cord[c] > 0)$
(7) $Ind \leftarrow Ind + Pascal[var, i]$
(8) $i \leftarrow i - 1$; \ décaler vers la case voisine haute dans le carré de Pascal
(9) $Cord[c] \leftarrow Cord[c] - 1$;
(10) **Sinon**
(11) $var \leftarrow var - 1$; \ décaler vers la case voisine gauche dans le carré de Pascal ;
(12) $c \leftarrow c + 1$; \ décaler à droite vers la coordonnée suivante.
(13) **Fin Si**
(14) **Fin Tant que**
(15) Retourner $Ind$;

**Fin**

## 6.7 Conclusion

Dans cette section, nous avons proposé une méthode d'intégration de polynômes multivariables de degrés quelconques sur un simplexe. Nous avons procédé par la recherche d'une transformation affine entre le simplexe d'intégration et le simplexe unitaire, ce qui nous a permis de se servir de la formule d'intégration donnée par [STROUD, 1971] pour le simplexe unitaire. Notre contribution majeure est de proposer une structure compacte pour la représentation des polynômes multivariables afin d'optimiser les opérations d'addition et de multiplication des polynômes complets et par conséquent d'accélérer l'intégration.

Pour bien valider l'approche proposée, une analyse approfondie de la complexité est nécessaire. Nous envisageons dans les travaux futurs de conduire des comparaisons de performances avec d'autres méthodes récentes en termes du temps, de l'espace mémoire, et de précision sur des benchmarks dédiés tels que les collections utilisées dans [De Loera et al., 2012].

Tableau 6.4 – Algorithme 4. Conversion de l'indice mémoire d'un monôme vers ses cordonnées dans le simplex.

**Sorties**

$Cord$: $1 \times n$ vecteur de coordonnées d'un monôme dans le simplex = leur puissances.

**Entrées**

$Ind$: indice du monôme dans la mémoire.
$n$: dimension du polynôme.
$d$:degré du polynôme.
$Pascal$: matrice constante telle qu'elle est définie sur la figure 6.8(a).

**Début**

(1) $Cord \leftarrow 0$; \ mettre tous les elements de $Cord$ à 0
(2) $Var \leftarrow n - 1$;
(3) $i \leftarrow d$; \ la case de départ dans $(var, i)$ dans le carré de Pascal
(4) $c \leftarrow 0$; \ la cordonnée gauche dans vecteur $Cord$
(5) **Tant que** $(Ind \neq 0)$
(6) **Si** $(Ind \geq Pascal[var, i])$
(7) $Cord[c] \leftarrow Cord[c] + 1$;
(8) $Ind \leftarrow Ind - Pascal[var, i]$;
(9) $i \leftarrow i - 1$;
(10) **Sinon**
(11) $var \leftarrow var - 1$; \ décaler vers la case voisine à gauche dans $Pascal$
(12) $c \leftarrow c + 1$; \ décaler à droite dans le vecteur $Cord$
(13) **Fin Si**
(14) **Fin Tant que**
(15) Retourner $Cord$;

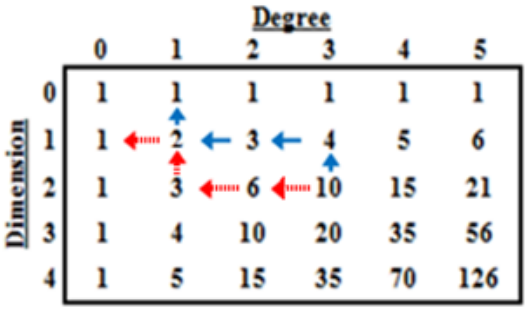
**Fin**

# 7 Expérimentation

## 7.1 Description de données de test

L'algorithme d'inférence sur le modèle ACIL-RB est conçu pour réaliser l'inférence Bayésienne sur des variables continues et non Gaussiennes. Dans le but d'illustrer ses performances, nous avons effectué des tests de prédiction sur le benchmark 'abalone' [Blake and Merz, 1998]. Les données proviennent d'une étude originale portée sur les oreille-de-mer (un mollusque appelé aussi '*ormier*', ou '*abalone*' en Anglais), collectées dans le cadre d'un projet de recherche au niveau de l'université de Tasmania (Australie). Ce benchmark comporte 4177 instances décrites par le biais de 9 attributs, leurs détails sont rapportés dans le tableau 7.1.

Ce benchmark a été utilisé principalement en la classification supervisée, en particulier la prédiction de l'âge de l'oreille-de-mer en fonction des attributs physiques mesurables. En effet, l'âge est déterminé par le découpage du coquillage, le maculage de la coupe et puis le comptage




(a)

| Ind | var | i | c | Cord | Ite. |
|---|---|---|---|---|---|
| 18 | 2 | 3 | 0 | 0 0 0 | Ini. |
| 08 | 2 | 2 | 0 | 1 0 0 | (1) |
| 02 | 2 | 1 | 0 | 2 0 0 | (2) |
| 02 | 1 | 1 | 1 | 2 0 0 | (3) |
| 00 | 1 | 0 | 1 | 2 1 0 | (4) |

(b)

| Ind | var | i | c | Cord | Ite. |
|---|---|---|---|---|---|
| 0 | 2 | 3 | 0 | 0 2 1 | Ini. |
| 0 | 1 | 3 | 1 | 0 2 1 | (1) |
| 4 | 1 | 2 | 1 | 0 1 1 | (2) |
| 7 | 1 | 1 | 1 | 0 0 1 | (3) |
| 7 | 0 | 1 | 1 | 0 0 1 | (4) |
| 8 | 0 | 0 | 1 | 0 0 0 | (5) |

(c)

FIGURE 6.8 – Exemple de conversion de puissances entières de|vers emplacement mémoire d'un monôme.

**(a)**: Triangle de Pascal. **(b)**: La trace des iterations de conversion de $Ind = 18$ vers $Cord = (210)$. Le chemin suivi par (b) est indiqué en rouge (flèches discontinues) dans (a). **(c)**: La trace de conversion de $Cord = (021)$ vers $Ind = 8$. Le chemin suivi par (c) est indiqué en bleu (flèches continues) dans (a).

des anneaux à l'aide d'un microscope ; une tache fastidieuse et trop lente.

Le choix de ce benchmark est justifié par le fait qu'il est accessible, relevé des données réelles, ses attributs sont continues et présentent des dépendances mutuelles complexes. Nous avons réalisé les tests de prédiction sur ce benchmark seulement pour illustrer l'efficacité de l'algorithme proposé en comparaison avec la solution standard des RBs discrets en utilisant des TPCs. Certainement, il pourra avoir d'autres domaines où l'application de cet algorithme sera plus prometteuse. Nous avons restreint l'étude aux cinq premières variables, plus précisément l'estimation de $P(v_5|(v_1, v_2, v_3, v_4))$; la prédiction du poids comestible du mollusque $v_5$ en fonction des variables $v_1, ..., v_4$ qui sont facilement mesurables.

Tableau 6.5 – Description des attributs du benchmark 'abalone' [Blake and Merz, 1998].

| Variable | Nom | Type | Unité | Description |
|---|---|---|---|---|
| $v_1$ | Longueur | continu | mm | La mesure la plus longue du coquillage. |
| $v_2$ | Diamètre | continu | mm | Perpendiculaire à la longueur |
| $v_3$ | Hauteur | continu | mm | Avec la chair dans le coquillage |
| $v_4$ | Poids complet | continu | gramme | poids de l'oreille-de-mer entier |
| $v_5$ | Poids écaillé | continu | gramme | Seulement le poids de la chair comestible |
| $v_6$ | Sexe | continu | M\|F | Mâle ou femelle |
| $v_7$ | Poids du coquillage | continu | gramme | Poids après séchage |
| $v_8$ | Poids des viscères | continu | gramme | Après dissection |
| $v_9$ | Nombre de stries | Entier | | +1.5 donne l'âge en années |

## 7.2 Méthodologie de comparaison

Dans le modèle d'inférence proposé ACIL-RB, nous avons expérimenté toutes les combinaisons possibles des algorithmes de clustering à savoir: K-means, Fuzzy C-Means (FCM) [Bezdek, 1981, Dunn, 1973], fuzzy Gustafson Kessel (GK) [Gustafson and Kessel, 1979] et Gath Geva (GG) [Gath and Geva, 1989] avec les algorithmes ACI : Fast ICA [Hyvärinen et al., 2001], JADE [Cardoso and Souloumiac, 1993], Kernel ICA [Bach and Jordan, 2002], Fast kernel ICA [Shen et al., 2009], KDICA [Chen, 2006] et MILKA [Stogbauer et al., 2004].

Notre algorithme a été comparé avec l'algorithme d'inférence exacte en utilisant les RBs discrets à base des TPCs pour quantifier les dépendances conditionnelles entre variables. Dans la solution TPCs, les variables $v_i$ ($i$ : 1...5) sont d'abord discrétisées, le domaine $\Phi_i$ de chaque variable $v_i$ est subdivisé en 15 valeurs discrètes ayant la même largeur ($Pas_i = \Phi_i/15$). Le réseau Bayésien dont le graphe DAG est complet, constitué des arcs $E = \{v_1 \rightarrow v_2, v_1 \rightarrow v_3, v_1 \rightarrow v_4, v_1 \rightarrow v_5, v_2 \rightarrow v_3, v_2 \rightarrow v_4, v_2 \rightarrow v_5, v_3 \rightarrow v_4, v_3 \rightarrow v_5, v_4 \rightarrow v_5\}$, est suffisant pour représenter toutes les éventuelles dépendances possibles.

Soit $\mathbf{v} = (t_1, t_2, t_3, t_4, t_5)$ un vecteur de test. La comparaison est faite en termes de la vraisemblance conditionnelle $LL(\mathbf{v}) = P(v_5 \in \Delta_{v_5} | (v_1 \in \Delta_{v_1}, v_2 \in \Delta_{v_2}, v_3 \in \Delta_{v_3}, v_4 \in \Delta_{v_4}))$. Les valeurs prises par la cible $v_5$ et les évidences $v_{1..4}$ sont prises égales aux intervalles centriques $\Delta_{v_i} = [t_i - \frac{Pas_i}{20}, t_i + \frac{Pas_i}{20}]$, voir la figure 6.9.

Supposons que les intervalles $\Delta_{v_i}$ se trouvent à l'intérieur des valeurs discrètes $d_i$, respectivement. Dans la solution basée sur les TPCs, l'introduction d'une évidence plus précise que le pas de discrétisation ($\Delta_i \subset d_i$) ne peut pas être exploitée pleinement car la distribution de densité de $v_i$ est approximée par une distribution uniforme sur l'intervalle $d_i$. Par conséquent, nous avons

$$LL_{TPC}(\mathbf{v}) = \frac{P(v_5 \in d_5 | (v_1 \in d_1, v_2 \in d_2, v_3 \in d_3, v_4 \in d_4))}{10} \tag{33}$$

Contrairement aux TPCs, notre modèle ACIL-RB peut avoir en entrée des intervalles quelconques dans la requête d'inférence.

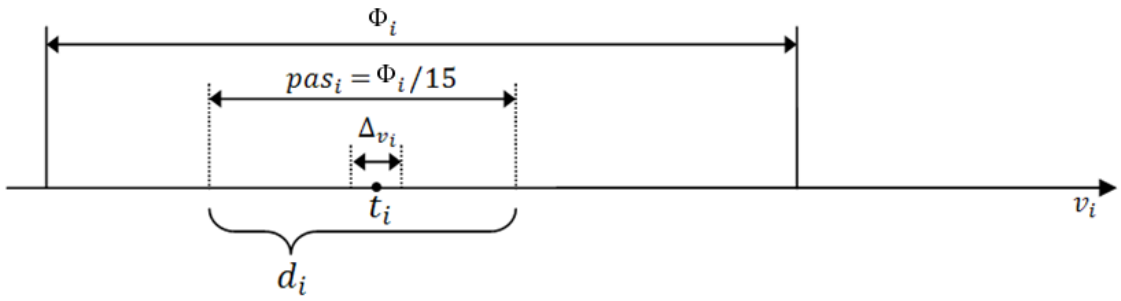



FIGURE 6.9 – Intervalle instanciée de la variable $v_i$.

Les performances ont été comparées sur la base d'un échantillon S composé de $N = 100$ vecteurs tirés aléatoirement (4177-100 =4077 vecteurs pour l'apprentissage). Étant donné que les vecteurs de test sont i.i.d, le critère de comparaison est la vraisemblance moyenne de l'échantillon entier, calculé comme suit :

$$LL(S) = (\prod_{i=1}^{N} LL(\mathbf{v}_i))^{\frac{1}{N}} = exp[\frac{\sum_{i=1}^{N} log(LL(\mathbf{v}_i))}{N}] \tag{34}$$

Dans ce sens, plus grande est la valeur $LL(S)$, meilleure est la prédiction.

## 7.3 Résultats obtenus

Les résultats de prédiction du modèle ACIL-RB et ceux des RBs discrets à base des TPCs sont comparés selon le critère introduit dans (34). On rappelle que le nombre de clusters considéré dans le modèle ACIL-RB est 03, la largeur des valeurs discrètes dans la solution TPCs est $pas_i = \Phi_i/15$ (où $\Phi_i$ est le domaine de la variable $x_i$) et la largeur des valeurs instanciées dans la requête d'inférence (11) est $\Delta_i = (\frac{pas_i}{10} = \frac{\Phi_i}{150})$.

Dans toutes les expériences, le modèle ACLL-RB ou les tables TPCs sont estimés en utilisant seulement les données d'apprentissage tirées du benchmark 'abalone'. Par contre, les tests sont scindés en deux groupes. Dans le premier groupe, les vecteurs de test sont tirés du benchmark 'abalone', tandis que les vecteurs de test du deuxième groupe sont générés aléatoirement à partir des distributions uniforme est Gaussienne.

Les résultats du premier groupe sont montrés sur le tableau 6.6 et ceux du deuxième groupe sont montrés sur le tableau 6.7. En concevant les tests de cette manière, une vraisemblance élevée dans le tableau 6.6 ou une faible vraisemblance dans le tableau 6.7 signifie une bonne performance d'inférence parce que les vecteurs de test du tableau 6.7 n'appartiennent pas au benchmark 'abalone' utilisé en apprentissage.

Par l'analyse des résultats d'inférence du premier groupe, nous constatons que, à l'exception de la méthode ACI 'Fast Kernel ICA', les résultats des autres différentes combinaisons (spline, méthode ACI, algorithme du clustering) dans le modèle ACIL-RB sont supérieurs à 0.0500 ($LL_{ACIL-RB}(S) \geq 0.0500$) qui sont nettement meilleurs que le résultat de la solution TPCs ($LL_{TPC}(S) = 0.0319$). Le manque de performance de 'Fast Kernel ICA' est justifié par le fait qu'elle n'a pas pu identifier des composantes suffisamment indépendantes sur le benchmark 'abalone', ce qui a rendu l'approximation $Ps = P(s_j^{(1)}, ..., s_j^{(n)}) = \prod_{i=1}^{n} Poly_{k,i,j}(s_j^{(i)})$ faite dans (17) non appropriée.

Nous constatons également que toutes les combinaisons excepté celles de 'Fast Kernel ICA' ont abouti à des vraisemblances très similaires, variant entre 0.0552 et 0.0621, sans qu'on puisse trancher qu'un algorithme ACI ou un algorithme de clustering est plus performant que les autres. Néanmoins, des conclusions peuvent être dégagées, à savoir :

1. Les splines sans contraintes sur la continuité des polynômes consécutifs (sous-tableaux (a), (b) et (c)) sont légèrement plus performants que les splines avec contraintes (sous tableau (d)). Cela confirme notre hypothèse disant que la qualité de l'ajustement des splines comportant peu de polynômes ($d \leq 10$) a plus d'impact sur les résultats de prédiction que la garantie des contraintes de continuité aux jointures.
2. Des splines de degrés inférieurs ($d = 1$) avec 6 nœuds ont données des résultats proches des splines cubiques ($d = 3$) avec 03 nœuds.
3. En général, l'algorithme MILKA et l'algorithme FCM ont donné des résultats légèrement meilleurs que les autres algorithmes.
4. 'Fast Kernel ICA' se combine bien avec K-Means que les autres algorithmes de clustering.

Les vraisemblances des vecteurs de test du deuxième groupe sont affichées dans le tableau 6.7. La composante $t_i (i : 1, ..., 5)$ du vecteur de test $v = (t_1, t_2, t_3, t_4, t_5)$ est générée selon la loi uniforme sur le domaine $\Phi_i$ dans le premier ensemble et elle est générée selon la distribution normale de moyenne $\mu = centre\ de\ \Phi_i$ et d'écart type sigma $\sigma = \| \frac{\Phi_i}{4} \|$ dans le deuxième ensemble. La quasi-totalité des résultats d'inférence de notre solution sont nettement meilleurs que ceux de RBs discrets à base des TPCs (on note que ici une petite vraisemblance signifie une bonne performance). Nous apercevons que les vraisemblances du premier ensemble (uniforme) sont inférieures que celles obtenues pour le deuxième ensemble (Gaussienne). La diminution de performances des données de test Gaussiennes est attendue parce que la distribution des données d'apprentissage du benchmark 'abalone' est plus proche d'une Gaussienne que de la distribution uniforme.

Nous constatons également que 'K-means' dans cette expérience est plus compétitif que les autres algorithmes de clustering dans notre approche. A l'instar des résultats du premier groupe (Tableau 6.6), l'algorithme 'Fast Kernel ICA' s'est montré moins performant que les autres algorithmes ACI.

Par la comparaison des vraisemblances du premier groupe (vecteurs positifs) et du deuxième

Tableau 6.6 – Résultats d'inférence obtenus sur le benchmark 'abalone' [Blake and Merz, 1998].

| Spline | Méthode ACI | Algorithmes de clustering | | | |
|---|---|---|---|---|---|
| | | K-means | FCM | GK | GG |
| | FastICA | 0.0588 | **0.0596** | 0.0589 | 0.0585 |
| (**a**) Degré : $d = 1$. Nombre | JADE | 0.0591 | **0.0621** | 0.0593 | 0.0590 |
| de polynômes : $K = 6$. Pas | Kernel ICA | 0.0593 | 0.0587 | 0.0593 | 0.0585 |
| de contraintes sur les | FastKernelICA | 0.0384 | 0.0284 | 0.0304 | 0.0202 |
| extrémités des polynômes | KDICA | 0.0589 | **0.0614** | 0.0589 | 0.0592 |
| | MILKA | **0.0612** | **0.0611** | **0.0606** | **0.0605** |
| | Fast ICA | 0.0589 | 0.0591 | 0.0589 | 0.0587 |
| (**b**) Degré : $d = 2$. Nombre | JADE | 0.0592 | 0.0582 | 0.0588 | 0.0589 |
| de polynômes : $K = 5$. Pas | Kernel ICA | **0.0605** | 0.0582 | 0.0594 | 0.0597 |
| de contraintes sur les | FastKernelICA | 0.0432 | 0.0203 | 0.0401 | 0.0236 |
| extrémités des polynômes | KDICA | 0.0594 | **0.0609** | 0.0599 | 0.0597 |
| | MILKA | 0.0590 | **0.0616** | **0.0609** | **0.0605** |
| | Fast ICA | 0.0575 | 0.0545 | **0.0609** | 0.0552 |
| (**c**) Degré : $d = 3$. Nombre | JADE | 0.0577 | **0.0608** | **0.0612** | 0.0558 |
| de polynômes : K =3. Pas | Kernel ICA | 0.0589 | **0.0612** | 0.0566 | 0.0586 |
| de contraintes sur les | FastKernelICA | 0.0349 | 0.0297 | 0.0275 | 0.0274 |
| extrémités des polynômes | KDICA | 0.0589 | 0.0574 | **0.0606** | 0.0563 |
| | MILKA | 0.0594 | 0.0589 | 0.0592 | 0.0586 |
| (**d**) Degré : $d = 3$. Nombre | Fast ICA | 0.0572 | 0.0557 | 0.0582 | 0.0552 |
| de polynômes : $K = 3$. | JADE | 0.0576 | 0.0587 | 0.0583 | 0.0583 |
| Deux polynômes | Kernel ICA | 0.0583 | 0.0586 | 0.0583 | 0.0596 |
| consécutifs $P_i$ et $P_{i+1}$ sont | FastKernelICA | 0.0331 | 0.0324 | 0.0185 | 0.0229 |
| connectés à leur point de | KDICA | 0.0578 | 0.0581 | 0.0589 | **0.0614** |
| jonction $t_i$ : $P_i(t_i) = p_{i+1}(t_i)$ | MILKA | 0.0576 | **0.0609** | **0.0602** | **0.0612** |

La vraisemblance conditionnelle de la solution TPCs est $LL_{TPC}(S) = 0.0319$.

groupe (vecteurs négatifs), nous pouvons conclure que l'algorithme d'inférence proposé est globalement non biaisé dans la mesure où les vraisemblances des vecteurs positifs sont élevées et celles des vecteurs négatifs sont très faibles par rapport aux RBs discrets.

# 8 Conclusion

Dans ce chapitre, nous avons proposé un modèle de représentation de données multivariables en utilisant la méthode ACI locale basée sur les clustering. Le modèle proposé, désigné par le sigle ACIL-RB, utilise le clustering pour quantifier la non linéarité globale entre les variables et utilise des modèles ACI linéaires pour fournir une description locale de données. Les paramètres du modèle ACIL-RB regroupent les paramètres de clustering, les matrices de mélange ACI et les densités de probabilité des CIs.

Au lieu d'utiliser des histogrammes de probabilité, nous avons opté pour les fonctions splines car avec une séquence constituée de peu de polynômes de degrés inférieurs on peut

Tableau 6.7 – Résultats d'inférence obtenus sur des données de test hors abalone' [Blake and Merz, 1998].

| Apprentissage | TPCs | Méthode ACI | Algorithmes de clustering | | | |
|---|---|---|---|---|---|---|
| | | | K-means | FCM | GK | GG |
| Données d'app. : 'abalone'. Données de test: 'Uniforme'. Spline=(Degré: $d = 1$. Nombre de polynômes: $K = 6$). Pas de contraintes sur les extrémités des polynômes | 0.00607 | FastICA | **0.00006** | 0.00151 | 0.00081 | 0.00055 |
| | | JADE | **0.00006** | 0.00092 | 0.00158 | 0.00061 |
| | | Kernel ICA | 0.00006 | 0.00073 | 0.00038 | 0.00102 |
| | | FastKernelICA | 0.00111 | 0.00705 | 0.00492 | 0.00574 |
| | | KDICA | **0.00006** | 0.00113 | 0.00068 | 0.00410 |
| | | MILKA | **0.00008** | 0.00081 | 0.00101 | 0.00232 |
| Données d'app. : 'abalone'. Données de test: 'Gaussiennes'. Spline=(Degré: $d = 1$. Nombre de polynômes: $K = 6$). Pas de contraintes sur les extrémités des polynômes | 0.00666 | FastICA | **0.00009** | 0.00862 | 0.00061 | 0.00106 |
| | | JADE | **0.00009** | 0.00590 | 0.01018 | 0.00129 |
| | | Kernel ICA | 0.00015 | 0.00209 | 0.00404 | 0.00160 |
| | | FastKernelICA | 0.00348 | 0.00563 | 0.00494 | 0.00874 |
| | | KDICA | **0.00010** | 0.00295 | 0.00185 | 0.00135 |
| | | MILKA | **0.00007** | 0.01682 | 0.00206 | 0.00239 |

Les vecteurs de test sont générés artificiellement (n'appartiennent pas au benchmark 'abalone'). Les vraisemblances de ces vecteurs ont été calculées sur un modèle construit à partir des donnes du benchmark 'abalone'.

avoir une bonne approximation de densité des CIs. La deuxième raison qui a motivé ce choix c'est les caractéristiques analytiques d'intégration des polynômes facilitant l'inférence Bayésienne. Lors de l'approximation des histogrammes par des splines (particulièrement les splines de degrés supérieurs), la densité de probabilité peut être rarement négative mais faible sur certaines zones étroites. Ce problème est bien connu, nous l'avons traité simplement par l'addition d'une quantité positive aux polynômes présentant ce problème, ainsi la surface de l'aire de probabilité sous la courbe de la spline peut être légèrement supérieure à 1. Conséquemment, cet élément nécessite d'être amélioré dans des travaux futurs.

Nous avons également proposé un algorithme d'inférence Bayésienne exacte s'exécutant sur le modèle ACIL-RB. Cet algorithme s'appuie principalement sur un module d'intégration de fonctions polynomiales multivariables sur des polytôpes. L'intégration des polynômes sur les polytôpes est connue d'être un problème difficile et de complexité NP-dur. Une des nos contributions consiste en la proposition d'un algorithme d'intégration afin d'accélérer l'inférence.

L'algorithme d'inférence est exact et toute l'erreur de prédiction provient du pouvoir de représentation de données par le modèle ACIL-RB. En effet, l'erreur de prédiction est due essentiellement à la défaillance de la méthode ACI, appliquée sur chaque cluster, d'extraire des composantes suffisamment indépendantes. Avec un degré moins, une partie de l'erreur a son origine au niveau de l'estimation de densité des CIs, notamment quand nous avons peu d'observations.

Le modèle ACIL-RB ainsi que l'algorithme d'inférence associé permettent d'effectuer l'inférence Bayésienne sur des variables continues et non Gaussiennes en accordant à l'utilisateur la faculté de spécifier des intervalles quelconques, même bien petits, dans la requête d'inférence.

Les comparaisons de ACIL-RB avec les RBs discrets à base des TPCs, faites sur le benchmark 'abalone' pris du dépôt UCI [Blake and Merz, 1998], ont montré un gain significatif en performance de notre solution.

Il convient de préciser que la complexité en temps d'inférence, si nous optons pour l'intégration par la triangulation des polytôpes, reste trop élevée pour des dimensions importantes. En outre, les performances des algorithmes de clustering généralement se détériorent avec l'augmentation de la dimension. Conséquemment, l'utilité de notre solution se montre mieux pour les petites des dimensions ($n \leq 7$)).

# Conclusion générale et perspectives

## Conclusion générale

Dans ce travail, nous avons proposé des nouveaux algorithmes d'inférence Bayésienne numériques sur des données qui peuvent être ajustées par un modèle ACI linéaire ou par un modèle ACI local à base de clustering. En effet, les RBs ne sont pas encore formulés pour le cas des variables continues non Gaussiennes et, pour palier à ce problème,un des objectifs principaux fixés au préalable dans cette thèse consiste à développer des algorithmes de propagation de croyances sur ce type de données. Les algorithmes d'inférence implémentés ont été appliqués pour effectuer la fusion multimodale de scores biométriques où des estimateurs de qualité des scores biométriques ont été exploités sous forme d'intervalles de confiance afin d'améliorer la fonction de décision. Une qualité médiocre d'un score est modélisée par un intervalle de confiance plus large en introduisant ainsi un degré d'incertitude important et vice-versa.

Notre choix du modèle ACI local est justifié par le fait qu'il est d'une part semi-paramétrique et permet d'approximer une famille assez étendue de distributions de données, et d'autre part parce que la généralisation des résultats obtenus pour le modèle ACI linéaire au modèle ACI local est relativement simple.

Les algorithmes d'inférence proposés sont exacts ; en conséquence, tout le manque de performances de prédiction dépend en effet de la qualité d'ajustement de données par les modèles proposés ACI-RB et ACIL-RB, et plus précisément de combien les CIs estimées par l'ACI sont statistiquement indépendantes, la quantité d'erreurs est véritablement proportionnelle à la divergence entre la conjointe des CIs $P(s_i, ..., s_N)$ et le produit des marginales $\prod_{i=1}^{N} P(s_i)$.

Le clustering qui joue un rôle primordial dans le modèle ACIL est un problème mal-posé dans la mesure qu'il dépend de la distance de similarité adoptée et des contraintes faites sur le nombre de clusters. A la base des expérimentations réalisées, nous avons vu que généralement des clusters elliptiques (obtenus par l'algorithme GK [Gustafson and Kessel, 1979]) sont légèrement plus adéquats que des clusters sphériques (obtenus par K-means ou FCM [Bezdek, 1981, Dunn, 1973]). Des études approfondies décidant qu'un tel algorithme ACI linéaire peut

estimer des CIs assez indépendantes sur une telle forme de clusters sont y nécessaires; c'est-à-dire nous avons besoin d'avoir une comparaison rigoureuse de performances des différentes méthodes ACI linéaires existantes en fonction de distributions des variables observées.

Nous avons développé seulement le cas simple où le nombre des CIs est égal au nombre des variables observées ($n = m$). Les algorithmes d'inférence ACI-RB et ACIL-RB proposés dans les deux derniers chapitres restent valides aux cas des représentations sur-complètes ($n < m$) et sous-complètes ($n > m$). Évidement, les algorithmes ACI ont plus de difficulté à estimer des représentations sur-complètes ($n < m$) en s'appuyant sur des hypothèses faibles telle que l'indépendance, mais ils peuvent donner des solutions convenables sous certaines hypothèses additionnelles. Dans ce sens, un de nos objectifs était de proposer des algorithmes d'inférence sur des représentations ACI acceptables.

## perspectives

Parmi les nouvelles directions de recherche possibles pour étendre ce travail, nous envisageons de :

1. Trouver des applications d'intérêt de l'algorithme d'inférence ACIL-RB.
2. L'algorithme d'intégration de fonctions polynomiales sur des polytôpes permet d'effectuer l'inférence Bayésienne sur le modèle ACIL en un temps raisonnable seulement pour des petites dimensions ($n <= 7$). L'optimisation de ce module d'intégration permet d'accélérer le calcul et d'étendre son emploi à des dimensions supérieures.
3. Proposer des nouvelles fonctions analytiques pour approximer les distributions des CIs comme une alternatives aux fonctions splines, les fonctions recherchées doivent avoir un haut degré de liberté pour pouvoir approximer des distributions complexes et au même temps disposer de certaines propriétés analytiques facilitant leur intégration sur des polytôpes en hautes dimensions.
4. ACIL-RB réalise l'inférence Bayésienne sur un seul cluster (une clique) de variables mutuellement dépendantes, faire un apprentissage en amont de la structure RB et puis l'extension de cet algorithme sur l'arbre de jonction correspondant représente une perspective intéressante pour sa généralisation aux données de hautes dimensions.

# Bibliographie


[mob, 2013] (2013). http://www.mobioproject.org/.

[Achard and Jutten, 2005] Achard, S. and Jutten, C. (2005). Identifiability of post nonlinear mixtures. *IEEE Signal Processing Letters*, 12:423–426.

[Albera et al., 2003] Albera, L., Ferréol, A., Comon, P., and Chevalier, P. (2003). Sixth order blind identification of underdetermined mixtures (birth) of sources. In *In International Symposium of Independent Component Analysis and Blind Sources Separation, ICA*, pages 909–914, Nara, Japan.

[Amari and Yang, 1996] Amari, S. and Yang, A. C. Y. L. H. (1996). A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8:752–763.

[Anandkumar et al., 2013] Anandkumar, A., Hsu, D., Javanmard, A., and Kakade, S. M. (2013). Learning linear bayesian networks with latent variables. In *30th International Conference on Machine Learning, Atlanta, Georgia, USA, . JMLR: W&CP*, volume 28.

[Asadi et al., 2008] Asadi, N. B., Meng, T. H., and Wong, W. H. (2008). Reconfigurable computing for learning bayesian networks 2008. In *FPGA08*, pages 24–26, Monterey, California, USA.

[Babuska et al., 2002] Babuska, R., van der Veen, P. J., and Kaymak, U. (2002). Improved covariance es timation for gustafson-kessel clustering. In *In IEEE International Conference on Fuzzy Systems*, pages 1081–1085, Honolulu, Hawaii.

[Bach and Jordan, 2002] Bach, F. and Jordan, M. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.

[Bai et al., 2007] Bai, S., Liu, J., Sun, G., and Zhang, W. (2007). Text independent speaker verification based on mixing ica overcomplete basis functions. In *3th Inter. Conf. on IIHMSP*, volume 1, pages 533–536.

[Baldoni et al., 2011] Baldoni, V., Berline, N., Loera, J. A. D., Köppe, M., and Vergne, M. (2011). How to integrate a polynomial over a simplex. *Journal of Math. Comp.*, 80:297–325.

[Bäräny and Füredi, 1986] Bäräny, I. and Füredi, Z. (1986). Computing the volume is difficult. In *18th Annual ACM Symposium on Theory of Computing*, pages 442–447, New York.

[Barber et al., 1996] Barber, C., Dobkin, D., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions. on Mathematical Software*, 22(4):469–483.

[Bell and Sejnowski, 1995] Bell, A. J. and Sejnowski, J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159.

[Bellman, 1957] Bellman, R. (1957). *Dynamic programming*. Princeton University Press, Princeton, NJ.

[Belouchrani et al., 1997] Belouchrani, A., Meraim, K. A., Cardoso, J. F., and Moulines, E. (1997). A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–443.

[Bezdek, 1981] Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York.

[Bezdek and Dunn, 1975] Bezdek, J. and Dunn, J. (1975). Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal dustrubutions. *IEEE Transactions on Computers*, 00:835–838.

[Binder et al., 1997] Binder, J., Koller, D., Russell, S., and Kanazawa, K. (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244.

[Blake and Merz, 1998] Blake, C. and Merz, C. (1998). Uci repository of machine learning databases.

[Boyd, 1992] Boyd, J. (1992). Defeating the runge phenomenon for equispaced polynomial interpolation via tikhonov regularizationregularization. *Applied mathematics letters*, 05:57–59.

[Burton, 1987] Burton, D. (1987). Text-dependent speaker verification using vector quantization source coding. *IEEE Trans. Acoustics, Speech, Signal Process.*, 35(02):133–143.

[Campbell, 1997] Campbell, J. P. (1997). Speaker recognition: A tutorial. In *of the IEEE*, volume 85.

[Cardoso and Souloumiac, 1993] Cardoso, F. and Souloumiac, A. (1993). Blind beamforming for non gaussian signals. *in IEE Proceedings-F*, 140(06):362–370.

[Cardoso, 1989] Cardoso, J. (1989). Source separation using higher order moments. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP89)*, pages 2109–2112, Glasgow, UK.

[Cardoso, 2003] Cardoso, J. (2003). Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203.

[Chen, 2006] Chen, A. (2006). Fast kernel density independent component analysis. *Lecture Notes in Computer Science*, 3889:24–31.

[Cheng et al., 1997a] Cheng, J., Bell, D., and Liu, W. (1997a). An algorithm for bayesian network construction from data. In *6th International Workshop on Artificial Intelligence and Statistics AI&STAT97*, pages 83–90.

[Cheng et al., 1997b] Cheng, J., Bell, D., and Liu, W. (1997b). Learning belief networks from data : An information theory based approach. In *the sixth ACM International Conference on Information and Knowledge Management CIKM*, pages 325–331.

[Chickering, 1996] Chickering, D. M. (1996). Learning bayesian networks is np-complete. *Learning from Data: Artificial Intelligence and Statistics*, 0:121–130.

[Chickering et al., 1995] Chickering, D. M., Geiger, D., and Heckerman, D. (1995). Learning bayesian networks: Search methods and experimental results. In *In Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 112–128.

[Cichocki et al., 2012] Cichocki, A., Amari, S., Siwek, K., Tanaka, T., and al. (2012). Icalab - matlab toolbox ver. 3 for signal processing.

[Comon, 1994] Comon, P. (1994). Independent component analysis-a new concept? *Signal Processing*, 36:287–314.

[Comon and Jutten, 2010] Comon, P. and Jutten, C. (2010). *handbook of blind source separation: independent component analysis and its applications*. Elsevier.

[Cooper and Herskovits, 1992] Cooper, F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–348.

[Cooper, 1990] Cooper, G. (1990). Computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42:393–405.

[Cover and Thomas, 1991] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

[Cowell, 2005] Cowell, R. (2005). Local propagation in conditional gaussian bayesian networks. *Journal of Machine Learning Research*, 6:1517–1550.

[Cowell et al., 1999] Cowell, R., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert systems*. Springer-Verlag, New York.

[Dagum and Luby, 1993] Dagum, P. and Luby, M. (1993). Approximantely probablistic reasonning in bayesian belief networks is np-hard. *Artificial Intelligence*, 60:141–153.

[de Campos et al., 2003] de Campos, L. M., Fernndez-Luna, J. M., and Puerta, J. M. (2003). An iterated local search algorithm for learning bayesian networks with restarts based on conditional independence tests. *International Journal of Intelligent Systems*, 18:221–235.

[De Loera et al., 2012] De Loera, J. A., Dutra, B., Köppe, M., Moreinis, S., Pinto, G., and Wu, J. (2012). Software for exact integration of polynomials over polyhedra. *ACM Commun. Comput. Algebra*, 45(3/4):169–172.

[Ding and Ross, 2012] Ding, Y. and Ross, A. (2012). A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognition*, 45(3):919–933.

[Domeniconi et al., 2007] Domeniconi, C., Gunopulos, D., Ma, B. Y. S., Al-Razgan, M., and Papadopoulos, D. (2007). Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery*, 4(1):63–97.

[Drouet, 2007] Drouet, I. (2007). *Causalité et probabilités : réseaux Bayésiens, propensionnisme*. PhD thesis, Université Paris I.

[Druzdzel and Simon, 1993] Druzdzel, M. J. and Simon, H. A. (1993). Causality in bayesian belief networks. In Heckerman, I. D. and (Eds.), A. . M., editors, *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, pages 3–11, San Mateo, CA. Morgan Kaufmann.

[Duann et al., 2003] Duann, J., Jung, T., Makeig, S., and Sejnowski, T. (2003). Consistency of infomax ica decomposition of functional brain imaging data. In *4th Int. Symposium on Independent Component Analysis and Blind Signal Separation, ICA 2003*, pages 289–294, Nara, Japan.

[Duarte and Hu, 2004] Duarte, M. and Hu, Y. (2004). Vehicle classi_cation in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(07):826–838.

[Dunn, 1973] Dunn, J. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 03(03):32–57.

[Dyer and Frieze, 1988] Dyer, M. and Frieze, A. (1988). On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17(5):967–974.

[Elekes, 1986] Elekes, G. (1986). A geometric inequality and the complexity of computing volume. *Discrete Comput. Geom.*, 01(04):289–292.

[Friedman and Goldszmidt, 1996] Friedman, N. and Goldszmidt, M. (1996). Discretizing continuous attributes while learning bayesian networks. In *Proceedings of the* $13^{th}$ *international conference on machine learning. San Francisco*, pages 157–165. Morgan Kaufmann.

[Gath and Geva, 1989] Gath, I. and Geva, A. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 07:773–781.

[Georgiev et al., 2004] Georgiev, P., Theis, F., and Cichocki, A. (2004). Blind source separation and sparse component analysis for over-complete mixtures. In *Blind source separation and sparse component analysis for over-complete mixtures*, pages 493–496, Montreal, Canada.

[Giannakopoulos et al., 1999] Giannakopoulos, X., Karhunen, J., and Oja, E. (1999). Experimental comparison of neural algorithms for independent component analysis and blind separation. *Int. Journal of Neural Systems*, 9(2):651–656.

[Gilks et al., 1996] Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics.

[Gribonval and Zibulevsky, 2009] Gribonval, R. and Zibulevsky, M. (2009). *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, chapter Sparse Component Analysis. academic press.

[Gritzmann and Klee, 1994] Gritzmann, P. and Klee, V. (1994). On the complexity of some basic problems in computational convexity: Ii. volume and mixed volumes. Technical report, Universität Trier.

[Gustafson and Kessel, 1979] Gustafson, D. E. and Kessel, W. C. (1979). Fuzzy clustering with fuzzy covariance matrix. In *the IEEE CDC*, pages 761–766, San Diego.

[Hautamäki et al., 2008] Hautamäki, V., Kinnunen, T., Kärkkäinen, I., Tuononen, M., Saastamoinen, J., and Frnti, P. (2008). Maximum a posteriori estimation of the centroid model for speaker verification. *IEEE Signal Process. Lett.*, 15:162–165.

[Hawkins, 2004] Hawkins, D. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243.

[Henrion, 1988] Henrion, M. (1988). Propagating uncertainty in bayesian networks by probabilistic logic sampling. *Uncertainty in Artificial Intelligence*, 2:149–163.

[Honda and Ichihashi, 2006] Honda, K. and Ichihashi, H. (2006). Fuzzy local independent component analysis with external criteria and its application to knowledge discovery in databases. *International Journal of Approximate Reasoning*, 42:159–173.

[Hyvaerinen et al., 2009] Hyvaerinen, A., Hurri, J., and O.Hoyer, P. (2009). *Natural Image statistics, A probabilistic approach to early computational vision*. Springer.

[Hyvärinen, 1999] Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

[Hyvärinen et al., 2001] Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons, New York, USA.

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430.

[Hyvärinen and Pajunen, 1999] Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12:429–439.

[Hyvärinen et al., 2001] Hyvärinen, A., Hoyer, P. O., and Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558.

[Ichida and Kiyono, 1977] Ichida, K. and Kiyono, T. (1977). Curve fitting by a one-pass method with a piecewise cubic polynomial. *ACM Transactions on Mathematical Software*, 03(02):164–174.

[Jang, 2001] Jang, G. (2001). Learning statistically efficient features for speaker recognition. In *ICASSP‘ 01*, volume 01, pages 437–440.

[Jensen et al., 1990] Jensen, F. V., Lauritzen, S. L., and Olesen, K. (1990). Bayesian, updating in causal probabilistic networks by local computations. *Computational Statistics Quaterly*, 4:269–282.

[J.F.Cardoso, 1998] J.F.Cardoso (1998). Blind signal separation: Statistical principles.,. In *Proc. of IEEE*, volume 9, pages 2009–2025.

[Jordan and Bach, 2002] Jordan, M. I. and Bach, F. R. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.

[Karhunen and Koivunen, 2002] Karhunen, J. and Koivunen, V. (2002). Blind separation methods based on pearson system and its extensions. *Signal Processing*, 84(04):663–673.

[Karhunen and Malaroiu, 1999] Karhunen, J. and Malaroiu, S. (1999). Independent component analysis using clustering. In *First Int. Workshop on ICA and Signal Separation*, pages 43–48.

[Kim et al., 2008] Kim, D., Burge, J., Lane, T., Pearlson, G., Kiehl, K., and Calhoun, V. (2008). Hybrid ica–bayesian network approach reveals distinct effective connectivity differences in schizophrenia. *NeuroImage*, 42(04):1560–1568.

[Kim and Pearl, 1983] Kim, J. and Pearl, J. (1983). A computational model for combined causal and diagnostic reasoning in inference systems. In *the Eighth International Joint Conference on Artificial Intelligence*, pages 190–193, IJCAI-83: Karlsruhe, Germany.

[Kinnunen and Li, 2010] Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52:12–40.

[Kiviluoto and Oja, 1998] Kiviluoto, K. and Oja, E. (1998). Independent component analysis for parallel financial time series. In *In Proc. Int. Conf. on Neural Information Processing (ICONIP9́8)*, volume 2, pages 895–898, Tokyo, Japan.

[Kjaerul, 1995] Kjaerul, U. (1995). dhugin a computational system for dynamic time-sliced bayesian networks. *International Journal of Forecasting*, 11(1):89–111.

[Koller et al., 1999] Koller, D., Lerner, U., and Anguelov, D. (1999). A general algorithm for approximate inference and its application to hybrid bayes nets. In Kaufmann, M. ., editor, *the 15th Conference on Uncertainty in Artificial Intelligence,*, pages 324–333.

[Kozlov and Koller, 1997] Kozlov, A. and Koller, D. (1997). Nonuniform dynamic discretization in hybrid networks. In *Proceedings of the 13th conference on uncertainty in artificial intelligence, 314-25*, pages 314–325, San Francisco, US. Morgan Kaufmann.

[Kriegel et al., 2009] Kriegel, H., Krger, P., and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 03(01):1–58.

[Kruskal, 1956] Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. In *the American Mathematical Society*, volume 7, pages 48–50.

[Kwak et al., 2001] Kwak, N., Choi, C., , and Choi, J. (2001). Feature extraction using ica. In *ICANN 2001, LNCS 2130*, pages 568–573. Springer.

[Kwona and Lee, 2004] Kwona, O. and Lee, T. (2004). Phoneme recognition using ica-based feature extraction and transformation. *Signal Processing*, 84:1005 –1019.

[Lauritzen, 1992] Lauritzen, S. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108.

[Lauritzen and Jensen, 2001] Lauritzen, S. and Jensen, F. (2001). Stable local computation with conditional gaussian distributions. *Statistics and Computing*, 11(2):191–203.

[Lauritzen and Spiegelhalter, 1988] Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, Series B 50(2):157–224.

[Li et al., 2010] Li, K., Wang, M., Liu, M., and Zhao, A. (2010). Improved level set method for lip contour detection. In *IEEE 17th International Conference on Image Processing*, pages 26–29.

[Linde et al., 1980] Linde, Y., Buzo, A., and Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Trans. Comm.*, 28:84–95.

[Liu et al., 2010] Liu, X., Cardoso, J., and R.Randall (2010). Very fast blind source separation by signal to noise ratio based stopping threshold for the shibbs/sjad algorithm. *Mechanical Systems and Signal processing*, 24(7):2096–2103.

[Mackie, 1974] Mackie, J. (1974). *The cement of the universe. A study of causation.* Oxford university press.

[Madsen and Jensen, 1999] Madsen, A. L. and Jensen, F. (1999). Lazy propagation: A junction tree inference algorithm based on lazy evaluation. *Artificial Intelligence*, 113(1-2):203–245.

[Makeig et al., 1996] Makeig, S., Bell, A., Jung, T.-P., and Sejnowski, T. (1996). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 8:145–151.

[Margaritis, 2003] Margaritis, D. (2003). Learning bayesian network model structure from data. Available as technical report cmu-cs-03-153, Phd thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh.

[Martiriggiano et al., 2005] Martiriggiano, T., Leo, M., P.Spagnolo, and Orazio, T. D. (2005). Facial feature extraction by kernel independent component analysis. In *AVSS*, pages 270–275. IEEE Computer Society.

[Maurer and Baker, 2008] Maurer, D. and Baker, J. (2008). Fusing multimodal biometrics with quality estimates via a bayesian belief network. *Pattern Recognition*, 41(3):821–832.

[McKeown et al., 1998] McKeown, J., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A., and Sejnowski, T. (1998). Analysis of fmri data by blind separation into independent spatial components. *Human Brain Mapping*, 6:160–188.

[McKeown and Sejnowski, 1998] McKeown, M. and Sejnowski, T. (1998). Independent component analysis of fmri data: Examining the assumptions. *Human Brain Mapping*, 6:368–372.

[Mittal and Kassim, 2007] Mittal, A. and Kassim, A. (2007). *Bayesian Network Technologies: Applications and Graphical Models*. IGI Publishing, Hershey, New York.

[Moral et al., 2001] Moral, S., Rumí, R., and Salmeró, A. (2001). Mixtures of truncated exponentials in hybrid bayesian networks, in: P. besnard, s. benferhart (eds.), symbolic and quantitative approaches to reasoning under uncertainty. *Lecture Notes in Artificial Intelligence*, 2143:156–167.

[Naim et al., 2004] Naim, P., Wuillemin, P., Leray, P., Pourret, O., and Becker, A. (2004). *Rseaux baysiens*. Eyrolles, Paris.

[Neal, 1993] Neal, R. (1993). Probabilistic inference using markov chain monte carlo methods. Technical report crg-tr-93-1, Department of Computer Science, University of Toronto.

[Niculescu et al., 2006] Niculescu, R. S., Mitchell, T. M., and Rao, R. B. (2006). Bayesian network learning with parameter constraints. *Journal of Machine Learning Research*, 7:1357–1383.

[Nielsen and Hansen, 2008] Nielsen, A. and Hansen, L. (2008). Structure learning by pruning in independent component analysis. *Neurocomputing*, 71(10-12):2281–2290.

[Parsons et al., 2004] Parsons, L., Haque, E., and Liu., H. (2004). Subspace clustering for high dimensional data: A review. *Explorations*, 06(01):90–105.

[Pearl, 1982] Pearl, J. (1982). Reverend bayes on inference engines: A distributed hierarchical approach. In *the Second National Conference on Artificial Intelligence*, pages 133–136, AAAI-82: Pittsburgh, PA. Menlo Park, California: AAAI Press.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, USA.

[Pearl, 2000] Pearl, J. (2000). *Causality : Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, England.

[Pearl and Verma, 1991] Pearl, J. and Verma, T. (1991). A theory of inferred causation. In In J. A. Allen, R. F. and (Eds.), E. . S., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, pages 441–52, San Mateo, CA. Morgan Kaufmann.

[Peng and Zhang, 2011] Peng, L. and Zhang, J. (2011). An entropy weighting mixture model for subspace clustering of high-dimensional data. *Pattern Recognition Letters*, 32(08):1154–1161.

[Plass and Stone, 1983] Plass, M. and Stone, M. (1983). Curve-fitting with piecewise parametric cubics. *computer graphics*, 17(03):229–239.

[Poh et al., 2010] Poh, N., Bourlai, T., and Kittler, J. (2010). A multimodal biometric test bed for qualitydependent, cost-sensitive and client-specific score-level fusion algorithms. *Pattern Recognition*, 43(03):1094–1105.

[Reichenbach, 1956] Reichenbach, H. (1956). *The Direction of Time*. Berkeley, University of Los Angeles Press.

[Reynolds, 1992] Reynolds, D. A. (1992). *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*. PhD thesis, Georgia Institute of Technology.

[Reynolds, 1995] Reynolds, D. A. (1995). Speaker identification and verification using gaussian mixture speaker models. *Speech Communnication*, 17:91–108.

[Reynolds et al., 2000] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Process.*, 10:19–41.

[Riggelsen, 2008] Riggelsen, C. (2008). *Approximation Methods for Efficient Learning of Bayesian Networks*. IOS Press.

[Robinson, 1973] Robinson, R. (1973). *New Directions in the Theory of Graphs*, chapter Counting labeled acyclic digraphs, pages 239–273. Academic Press.

[Romero et al., 2006] Romero, V., Rumí, R., and Salmeró, A. (2006). Learning hybrid bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 42(1-2):54–68.

[Ross and Jain, 2003] Ross, A. and Jain, A. (2003). Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125.

[Rouigueb et al., 2012] Rouigueb, A., Chitroub, S., and Bouridane, A. (2012). Fuzzy local ica for speaker recognition using voice and lip motion. In of the World Congress on Engineering, P., editor, *International Conference of Data Mining and Knowledge Engineering,*, pages 420–425, London (UK).

[Rouigueb et al., 2013a] Rouigueb, A., Chitroub, S., and Bouridane, A. (2013a). Intégration des fonctions polynomiales sur des polytôpes de dimensions quelconques. In *8 eme Conférence sur le Génie Electrique CGE08*. EMP.

[Rouigueb et al., 2013b] Rouigueb, A., Chitroub, S., and Bouridane, A. (2013b). Text-independent mfccs vectors classification improvement using local ica. In *IEEE International Workshop On Machine Learning For Signal Processing*, Southampton, (UK).

[Rouigueb et al., 2014a] Rouigueb, A., Chitroub, S., and Bouridane, A. (2014a). Bayesian inference over ica models: Application to multibiometric score fusion with quality estimates. *Journal of Applied Statistics*, 41:2123–2140. DOI:10.1080/02664763.2014.909780.

[Rouigueb et al., 2014b] Rouigueb, A., Chitroub, S., and Bouridane, A. (2014b). Density estimation of high dimensional data using ica and bayesian networks. *Intelligent Data Analysis*, 18(02):157–179.

[Sallee, 1982] Sallee, J. (1982). A triangulation of the n-cube. *Discrete Mathematics*, 40(01):81–86.

[Särelä and Vigário, 2003] Särelä, J. and Vigário, R. (2003). Overlearning in marginal distribution-based ica: analysis and solutions. *Journal of Machine Learning Research*, 04:1447–1469.

[Scutari, 2010] Scutari, M. (2010). Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(03):1–22.

[Sentürk and Gürgen, 2006] Sentürk, A. and Gürgen, F. S. (2006). Feature selection by independent component analysis for robust speaker verification. *Inter. Journal of CSNS*, 06(06):229–239.

[Sezan, 1985] Sezan, M. I. (1985). A peak detection algorithm and its application to histogram-based image data reduction. *Graph. Models Image Processing*, 29:47–59.

[Shafer and Shenoy, 1990] Shafer, G. R. and Shenoy, P. (1990). Axioms for probability and belief function propagation. *Uncertainty in Artificial Intelligence*, 4:169–198.

[Sharma and Paliwal, 2006] Sharma, A. and Paliwal, K. (2006). Subspace independent component analysis using vector kurtosis. *Pattern Recognition*, 39:2227–2232.

[Shen et al., 2009] Shen, H., Jegelka, S., and Gretton, A. (2009). Fast kernel-based independent component analysis. *IEEE Trans. on Signal Processing*, 57(09):3498–3511.

[Shenoy and Shafer, 1990] Shenoy, P. and Shafer, G. (1990). Axioms for probability and belief function propagation. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence, 169-198*. Morgan Kaufmann.

[Sloman, 2005] Sloman, S. (2005). *Causal Models, How People Think about the World and Its Alternatives*. Oxford University Press.

[Smail, 2004] Smail, L. (2004). *Algorithmique pour les Réseaux Bayésiens et leurs Extensions*. PhD thesis, Université de Marne-la-Vallée.

[Spirtes et al., 1993] Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction, and search*. Springer-Verlag.

[Stogbauer et al., 2004] Stogbauer, H., Kraskov, A., Astakhov, A. S., and Grassberger, P. (2004). Least dependent component analysis based on mutual information. *Physical Review E*, 70(066123):066123.

[STROUD, 1971] STROUD, A. (1971). *Approximate Calculation of Multiple Integrals*. Prentice-Hall, NJ.

[Taleb and Jutten, 1999] Taleb, A. and Jutten, C. (1999). Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47:2807–2820.

[Tsamardinos et al., 2003] Tsamardinos, I., Aliferis, C., and Statnikov, A. (2003). Algorithms for large scale markov blanket discovery. In *the 16th International Florida Artificial Intelligence Research Society Conference*, pages 376–381.

[Verma and Pearl, 1991] Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. In *the Sixth Conference on uncertainty in Artificial Intelligence*, pages 220–227, San Francisco. Morgan Kaufmann.

[Vigario, 1997] Vigario, R. (1997). Extraction of ocular artefacts from eeg using independent component analysis. *Electroencephalography Clinical Neurophysiology*, 103:395–404.

[Wen, 1991] Wen, W. (1991). Optimal decomposition of belief networks. *Uncertainty in Artificial Intelligence*, 6:209–224.

[Williamson, 2005] Williamson, J. (2005). *Bayesian nets and causality*. Oxford University Press, New York.

[Winther and Petersen, 2007] Winther, O. and Petersen, K. (2007). Bayesian independent component analysis: Variational methods and non-negative decompositions. *Digital Signal Processing*, 17:858–872.

[Wright, 1977] Wright, I. (1977). Splines in statistics. Technical report, Department of Statistics, University of North Carolina, USA.

[Xie and Beni, 1991] Xie, X. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Trans. on PAMI*, 13(08):841–847.

[Xu and II, 2005] Xu, R. and II, D. W. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16:645–678.

[Y. Li and Amari, 2004] Y. Li, A. C. and Amari, S. (2004). Analysis of sparse representation and blind source separation. *Neural Computation*, 16:1193–1204.

[York, 1992] York, J. (1992). Use of the gibbs sampler in expert systems. *Artificial Intelligence*, 56:115–130.

[Zhao et al., 2011] Zhao, Y., Cao, J., Zhang, C., and Zhang, S. (2011). Enhancing grid-density based clustering for high dimensional data. *The Journal of Systems and Software*, 84(09):1524–1539.

[Zienkiewicz and Taylor, 1988] Zienkiewicz, O. and Taylor, R. (1988). *The finite element method*. McGraw-Hill, London.

# Table des figures


1.1 Exemple d'un réseau Bayésien à variables discrètes. . . . . 12
1.2 Exemple de manipulations de base sur les potentiels. . . . . 20
1.3 Exemple de construction de l'arbre de jonction. . . . . 22
1.4 Exemple d'inférence sur l'arbre de jonction en utilisant l'algorithme Hugin. . . 25
1.5 Exemple d'introduction d'une variable latente. . . . . 29

2.1 Interprétation géométrique de l'ACI et l'ACP. . . . . 36
2.2 Exemple de centrage et de blanchiment. . . . . 46
2.3 Principe géométrique de SCA. . . . . 52

3.1 Exemple illustrant le rôle des facteurs cachés dans la causalité probabiliste. . . 62
3.2 Exemple illustrant l'apparition de l'incertitude dans la relation de causalité à cause de la discretisation. . . . . 63
3.3 Exemple montrant l'insuffisance du diagnostique causal dans la construction de la structure RB à cause des causes communes cachées. . . . . 64
3.4 Correspondance entre les paramètres qualitatifs-quantitatifs des modèles RB et ACI. . . . . 68
3.5 Exemples illustrant la procédure de construction d'un DAG d'un RB à partir de la matrice de mélange ACI associée. . . . . 72
3.6 Exemple d'un RB qui peut être approximé par un mélange ACI. . . . . 74

4.1 Exemple en dimension 2 de l'ACI locale à base de clustering. . . . . 83
4.2 Adaptation GMM par l'ajout d'un modèle de monde est vue comme une classification binaire discriminative. . . . . 86

4.3 Schéma d'apprentissage du modèle de classification en combinant ACI locale et SVM. . . . . 88
4.4 Algorithme de prédiction en combinant l'ACI locale et SVM. . . . . 90
4.5 Comparaison en termes du taux de la classification de la solution proposée avec quelques classificateurs SVM. . . . . 92
4.6 Comparaison en termes de temps de la classification de la solution proposée et certains classificateurs SVM. . . . . 93
4.7 Exemple de clustering FCM en deux dimensions avec $L = 4$. . . . . 96
4.8 Exemple de clustering GK en deux dimensions avec $L = 4$. . . . . 97
4.9 Exemple de clustering GG en dimension avec $L = 4$. . . . . 97
4.10 Étapes de construction du classificateur ACI-Locale & SVM. . . . . 98
4.11 Étapes de prédiction du classificateur ACI-Locale & SVM. . . . . 99
4.12 Tracking du mouvement des lèvres. . . . . 100
4.13 Fusion des vecteurs MFCCs et des vecteurs labiaux. . . . . 101
4.14 Taux de classification: Correction de Biais par SVM. . . . . 104
4.15 Qualité de séparation de sources (mesurée en termes du SNR) par l'algorithme FastICA en fonction du nombre des CIs et de la sparsity. . . . . 107
4.16 Qualité de séparation de sources (mesurée en termes du SNR) par JADE en fonction du nombre des CIs et de la sparsity. . . . . 108
4.17 Exemple montrant le principe de construction d'un arbre de jonction à partir du graphe d'un RB. . . . . 109
4.18 Evaluation de performances d'estimation de densité des données moins complexes. . . . . 111
4.19 Evaluation de performances d'estimation de densité des données plus complexes. . . . . 112

5.1 Exemple d'un mélange ACI en dimension quatre. . . . . 123
5.2 Modèle proposé ACI-RB. Exemple de quatres variables. . . . . 124
5.3 Exemple géométrique en 2D illustre la technique d'inférence sur l'espace des composantes indépendantes. . . . . 129
5.4 Fusion de scores vue comme un problème de classification binaire supervisée, exemple de deux scores. . . . . 134
5.5 Modèle de fusion proposé. . . . . 136

6.1 Exemple de l'estimation de densité en utilisant une spline cubique. . . . . 142
6.2 Exemple de la trace de recherche de 06 points de contrôle. . . . . 148
6.3 Trace de recherche de la spline de l'exemple de la figure 6.2. . . . . 148
6.4 Exemple en dimension 2 d'inférence géométrique dans le modèle ACIL-RB. . 153
6.5 Exemple de transformation affine du simplexe standard vers un simplexe général, cas de $n = 2$. . . . . 156
6.6 Exemple d'un polynôme complet de degré $d = 3$ en dimension $n = 3$ sauvegardé dans un simplexe entier. . . . . 160
6.7 Ordre des coefficients des monômes en mémoire physique. Un exemple d'un polynôme complet de degré $d = 3$ en dimension $n = 3$. . . . . 161
6.8 Exemple de conversion de puissances entières de|vers emplacement mémoire d'un monôme. . . . . 164
6.9 Intervalle instanciée de la variable $v_i$. . . . . 166

# Liste des tableaux


2.1 Comparaison entre quelques méthodes ACI populaires. . . . . 48

3.1 Divergences entre les modèles RB et ACI. . . . . 70
3.2 Comparaison entre les modèles RB, MES et ACI. . . . . 75

4.1 Comparaison des taux de classification entre quelques algorithmes ACI populaires. . . . . 91
4.2 Nombre des vecteurs de la classification. . . . . 101
4.3 Taux de classification des vecteurs MFCCs par SVM un-contre-un . . . . . 102
4.4 Taux de classification des vecteurs des attributs labiaux. . . . . 102
4.5 Taux de classification des vecteurs des attributs vocaux. . . . . 103
4.6 Taux de classification des vecteurs des attributs de fusion. . . . . 103
4.7 Algorithmes utilisés pour l'apprentissage de la structure RB. . . . . 113
4.8 Description des benchmarks. . . . . 114
4.9 Taux de classification de la base de données 'SatImage'. . . . . 115
4.10 Taux de classification de la base de données 'SensIT Vehicle (acoustic)'. . . . . 116
4.11 Taux de classification de la base de données 'SensIT Vehicle (seismic)'. . . . . 117
4.12 Taux de classification de la base de données 'Steel fault'. . . . . 118
4.13 Comparaison avec les classificateurs SVM et KNN. . . . . 118

5.1 Calcul des limites d'intégration de la composante $s_i$ . . . . . 126
5.2 Description des datasets. . . . . 130
5.3 Résultats de prédiction du dataset " Uniform ", un exemple de 04 variables. . . 131

5.4 Résultats de prédiction du dataset "Halo", un exemple de 04 variables. . . . . 131
5.5 Résultats de prédiction du dataset "ACsparse", un exemple de 04 variables. . . 131
5.6 Résultats de prédiction du dataset "Speech", un exemple de 04 variables. . . . . 131

6.1 Algorithme 1. $RegresserSpline(\mathbf{x}, \mathbf{y}, K, d)$ . . . . . 149
6.2 Algorithme 2. $AjusterRecursif(i, k)$ . . . . . 150
6.3 Algorithme 3. Conversion des puissances d'un monôme vers son indice en mémoire physique. . . . . 162
6.4 Algorithme 4. Conversion de l'indice mémoire d'un monôme vers ses cordonnées dans le simplex. . . . . 163
6.5 Description des attributs du benchmark 'abalone' [Blake and Merz, 1998]. . . . 165
6.6 Résultats d'inférence obtenus sur le benchmark 'abalone' [Blake and Merz, 1998]. 168
6.7 Résultats d'inférence obtenus sur des données de test hors abalone' [Blake and Merz, 1998]. . . . . 169

# Glossaire

**ACI** : Analyse en composantes indépendantes
**ACI-RB** : Analyse en composantes indépendantes et réseau Bayésien
**ACIL-RB**: Analyse en composantes indépendantes locale et réseau Bayésien
**ACP** : Analyse en composantes principales
**ASI** : Analyse en sous-espaces indépendants
**BSS** : blind source separation
**CI** : Composante Indépendante
**CLG** : Modèle conditionnel Gaussien
**DAG** : Directed acyclic graph
**FCM** : Fuzzy c-means
**JADE** : Joint approximate diagonalization of eigen-matrices
**GG** : Gath Geva
**GK** : Gustafson-Kessel
**GMM** : Gaussian mixture model
**EEG** : électro-encéphalographie
**EM** : Maximization-expectation
**FMRI** : Functional magnetic resonance image
**KDICA** :Fast kernel-based independent component analysis
**KLD** : Kullback-Leibler divergence
**MAP** : Maximum a posteriori
**MFCC** : Mel frequency cepstral coefficients
**MTE** : mélange d'exponentielles tronquées
**NMSE** : Normalized mean square error
**PNL** : post-non linéaires
**RAL** : Reconnaissance automatique du locuteur
**RB** : Réseau Bayésien
**SCA** : Sparse Component Analysis
**TPC** : Table de Probabilités Conditionnelles
**VQ** : Vector quantification

## Abstract

Uncertainty is omnipresent in many real-world applications. Bayesian Networks (BN) are probabilistic graphical models which allow to representing data with uncertainty and to reasoning over these representations. Independent component analysis (ICA) is a high-order statistical method which aims to finding components not only decorrelated in the sense of ACP, but fully independent. BNs are mainly used for inference while ICA is used for blind source separation problems look-alike. Even their principal vocations are different, both models identify, measure, and model relationships of statistical dependency among observed variables. Our aim in this thesis is to exploit recent results of ICA to improve algorithms of learning and inference of BNs, or to extend the BN use to non standard data.

This thesis focus on two related aspects; the use of ICA for the representation of multivariable data with uncertainty and probabilistic reasoning over these representations. In particular, we envisage to propose an extension of BN on continuous non Gaussian data using ICA basis and then to propose inference algorithms over these extensions.

Multimodal biometric systems allow person verification by considering many biometrics in order to strengthen the authentication process and also to widen the universality of use. The data fusion can be achieved at the features or the matching scores levels. In this sense, local ICA based on clustering is used to classify the fusion features of voice and lips movement because they are highly correlated. Moreover, scores fusion may be seen as a binary classification between the geniuses' and impostor's score vectors. We have modeled each of both classes by a local ICA model. The proposed Bayesian inference algorithm over continuous variables is used to achieve score fusion with quality estimate; the quality is introduced as a confidence interval.



**KeyWords** : Bayesian Networks, Independent component analysis, inference, probabilistic reasoning, continuous non Gaussian variables, biometry, multimodal fusion, supervised classification.

# Résumé

L'incertitude est omniprésente dans la plupart des applications du monde réel. Les réseaux Bayésiens (RBs) et l'analyse en composantes indépendantes (ACI) sont deux modèles statistiques qui ont vu leur essor vers les années 90s. Les RBs permettent une représentation graphique de données avec incertitude et le raisonnement probabiliste sur ces représentations, tandis que l'ACI vise à identifier des composantes non seulement décorrélées au sens de l'ACP, mais indépendantes au sens plein. Les RBs sont essentiellement employés pour l'inférence, alors que l'ACI pour la séparation aveugle de sources. Bien que les vocations de base des RBs et d'ACI soient différentes, les deux modèles identifient, mesurent et modélisent les dépendances statistiques entre les variables observées. L'objectif principal des travaux effectués dans cette thèse est d'exploiter les aboutissements des techniques ACI, obtenus récemment, pour enrichir quelques aspects relatifs aux RBs, soit par l'amélioration des algorithmes d'apprentissage ou d'inférence des RBs, soit par l'extension des RBs à des données non standard.

Cette thèse est donc articulée en deux grandes parties. Une partie porte sur l'utilisation de la méthode ACI ou l'une de ses variantes pour la représentation de données multivariables. Dans la deuxième partie, nous proposons des algorithmes d'inférence Bayésienne sur les représentations ACI considérées. Pour conclure, nous aimons proposer une extension des RBs aux données continues et non Gaussiennes où le modèle ACI sera employé pour qualifier et quantifier les relations entre les variables d'intérêt.

Les systèmes biométriques multimodaux permettent de reconnaître une personne en utilisant plusieurs modalités biométriques afin de consolider l'authentification et d'étendre l'universalité d'emploi. La fusion de données peut être faite au niveau d'attributs ou au niveau de scores d'appariement. L'ACI locale à base de clustering a été appliquée pour la classification des attributs de la voix et de mouvement de lèvres, la fusion de données est faite niveau d'attributs car il y a une forte dépendance entre les MFCCs et les attributs labiaux. En outre, la fusion de scores biométriques est une classification binaire supervisée entre la classe des vecteurs légitimes et imposteurs. Chacune des deux classes est modélisée par un modèle ACI linéaire. L'algorithme d'inférence Bayésienne proposé a été appliqué pour réaliser la fusion de scores où la qualité $q$ de score $s$ est insérée sous forme d'un intervalle de confiance $\Delta$ dont la largeur est inversement proportionnelle à $q$ ; $s$ est à l'intérieur de $\Delta$ mais on ne connaît pas sa valeur exacte, ce qui représente une forme d'incertitude.

**Mots-clés :** Réseaux Bayésiens, analyse en composantes indépendantes, inférence, raisonnement probabiliste, variables continues non Gaussiennes, biométrie, fusion multimodale, classification supervisée.