

UNIVERSITÉ CLERMONT AUVERGNE

HABILITATION À DIRIGER DES RECHERCHES

Présentée par :

Julien AH-PINE

Sur le sujet :

CONTRIBUTIONS EN SCIENCE DES DONNÉES

**FUSION D'INFORMATIONS, FONCTIONS D'AGRÉGATION, MESURES
EN CLUSTERING, VARIÉTÉS NON LINÉAIRES ET DONNÉES
FONCTIONNELLES**

Soutenue publiquement le 24 juin 2024, devant le jury composé de :

Stéphane Chrétien	Professeur à l'Université Lumière Lyon 2	(Rapporteur)
Éric Gaussier	Professeur à l'Université Grenoble Alpes	(Rapporteur)
Nathalie Vialaneix	Directrice de Recherche INRAe Toulouse	(Rapporteuse)
Michel Grabisch	Professeur à l'Université Paris 1 Panthéon-Sorbonne	(Examineur)
Pierre Latouche	Professeur à l'Université Clermont Auvergne	(Examineur)
Vincent Barra	Professeur à l'Université Clermont Auvergne	(Tuteur)

Table des matières

Remerciements	v
0 Introduction générale	1
1 Fusion d'informations multimodales texte-image en RI basée sur le contenu	8
1.1 Introduction	8
1.1.1 Contexte	8
1.1.2 Travaux antérieurs	10
1.2 Contributions	13
1.2.1 <i>Reranking, clustering</i> et promotion de la diversité en recherche d'information	13
1.2.2 Apprentissage actif et <i>multimedia information seeking</i>	14
1.2.3 Marche aléatoire sur des graphes et unification de méthodes de fusion transmodale	19
1.3 Discussions et perspectives	24
2 Fusion d'informations linguistiques et statistiques en TALN	29
2.1 Introduction	29
2.1.1 Contexte	29
2.1.2 Travaux antérieurs	33
2.2 Contributions	36
2.2.1 Méthodes de fusion et enrichissement d'hypergraphes linguistiques	36
2.2.2 <i>Clique Based Clustering</i> et désambiguïsation d'entités nommées	42
2.3 Discussions et perspectives	47
3 Fonctions d'agrégation et explicabilité en apprentissage supervisé	52
3.1 Introduction	52
3.1.1 Contexte	52

3.1.2	Travaux antérieurs	55
3.2	Contributions	61
3.2.1	Intégrale de Choquet bipolaire et 2-additive	61
3.2.2	Une nouvelle famille de fonctions d'agrégation	70
3.3	Discussions et perspectives	76
4	Mesures de proximité et critères de partitionnement en <i>clustering</i>	79
4.1	Introduction	79
4.1.1	Contexte	79
4.1.2	Travaux antérieurs	81
4.2	Contributions	89
4.2.1	Maximisation de mesures d'association en <i>clustering</i>	89
4.2.2	Normalisation de mesures de similarité	96
4.3	Discussions et perspectives	104
5	Variétés non linéaires en apprentissage non supervisé	109
5.1	Introduction	109
5.1.1	Contexte	109
5.1.2	Travaux antérieurs	112
5.2	Contributions	119
5.2.1	Plongement et partitionnement spectral de graphes	119
5.2.2	Un nouveau modèle générique de classification ascendante hiérarchique	125
5.3	Discussions et perspectives	133
6	Méthodes d'apprentissage en analyse de données fonctionnelles	136
6.1	Introduction	136
6.1.1	Contexte	136
6.1.2	Travaux antérieurs	139
6.2	Contributions	143
6.2.1	Représentation de données fonctionnelles avec dérivées	143
6.2.2	<i>k-means</i> à noyaux multiples pour données fonctionnelles avec dérivées	145
6.2.3	SVM à noyaux multiples pour données fonctionnelles avec dérivées . .	150
6.3	Discussions et perspectives	154
7	Conclusion générale, travaux actuels et futurs	159
7.1	Conclusion générale	159
7.2	Travaux actuels et futurs	160
	Notations	163
	Table des figures	170
	Liste des tableaux	173

Remerciements

Je remercie vivement Vincent Barra de m'avoir accompagné dans mon projet en remplissant avec grande efficacité son rôle de tuteur. Je remercie chaleureusement Stéphane Chrétien de m'avoir encouragé à préparer mon habilitation et d'avoir accepté d'être rapporteur. Ses grandes qualités scientifiques et humaines sont une source d'inspiration. Nathalie Vialaneix et Éric Gaussier me font également l'honneur d'évaluer mon manuscrit malgré leurs emplois du temps chargés. Je leur suis très reconnaissant pour cette précieuse contribution. Je remercie sincèrement Michel Grabisch et Pierre Latouche d'avoir accepté de faire partie de mon jury. Leur expertise et leur engagement dans mon projet sont des atouts inestimables.

Au cours de ma carrière j'ai côtoyé de nombreuses personnes qui m'ont enrichi scientifiquement et humainement et qui ont ainsi contribué, d'une façon ou d'une autre, à la réalisation de mon projet. Je suis profondément reconnaissant à celles et ceux avec qui j'ai partagé un "bout de chemin" à Grenoble au *Xerox Research Centre Europe* (devenu *Naver Labs*), et à Lyon au Laboratoire ERIC et à l'Université Lumière Lyon 2. Depuis septembre 2023, mon chemin se poursuit plus spécifiquement à Clermont-Ferrand. Je souhaite remercier vivement pour leur accueil, mes collègues avec qui je collabore au sein du LIMOS, de SIGMA Clermont, du LMBP et du CERDI de l'Université Clermont Auvergne.

Je suis particulièrement reconnaissant à mes coauteurs avec qui j'ai travaillé de façon étroite. Les travaux que nous avons menés sont exposés dans ce mémoire et ils sont donc des contributeurs de premier plan. Merci à Jean-Michel Renders, Gabriela Csurka, Guillaume Jaquet, Stéphane Clinchant, Antoine Rolland, Brice Mayag, Jérôme Darmont, Sabine Loudcher, Xinyu Wang, Pavel Soriano, Anne-Françoise Yao et Noé Lebreton. Je souhaite également remercier sincèrement Jean-François Marcotorchino, Rodolphe Le Riche et Mourad Baïou pour le soutien précieux qu'ils m'ont apporté à des moments importants de mon parcours ; et Anne-Françoise Yao, Engelbert Mephu Nguifo et Jean-Marc Bourinet pour leur aide bienveillante dans mon processus actuel d'intégration en m'associant à différents projets.

Je remercie de tout mon coeur Elda pour son soutien inconditionnel, sa grande patience et son amour profond. Je lui dédie ce mémoire ainsi qu'à Abélia et Méloé, nos deux merveilles.



Introduction générale

Ce manuscrit synthétise les activités scientifiques que j'ai menées de 2008 à 2024. Au cours de ces seize années en tant que chercheur, j'ai eu l'occasion de travailler dans divers environnements et de contribuer à différents domaines de recherche.

Mais avant de lister ces différentes thématiques qui annoncent le contenu de ce manuscrit, je souhaite écrire quelques mots sur ma thèse de doctorat en préambule. En effet, c'est au cours de cette période que j'ai découvert la recherche et c'est donc à ce moment là que j'ai principalement forgé mon identité de chercheur.

J'ai soutenu mon doctorat en mathématiques en octobre 2007 à l'Université Pierre et Marie Curie Paris 6 (UP6). J'ai effectué une thèse dans le cadre d'un contrat CIFRE au sein du groupe Thales à Paris. Je devais travailler sur l'analyse relationnelle (AR) qui est l'approche développée par mon directeur de thèse Jean-François Marcotorchino qui fût également mon professeur en DEA de Statistiques à UP6. L'AR est une approche à l'intersection des graphes, des statistiques et de l'optimisation. Elle s'intéresse à la représentation, l'association et l'agrégation de relations binaires (RB). Elle trouve de nombreuses applications notamment en analyse de données, en statistiques des variables catégorielles, en aide multicritère à la décision et, dans une certaine mesure, en théorie des tresses. J'ai donc été initié à la recherche dans un cadre où plusieurs disciplines mathématiques et plusieurs domaines de recherche s'entremêlaient.

Pendant mon doctorat, j'ai bénéficié d'une grande liberté dans le choix de mes sujets de recherche. Curieux de nature, je m'étais alors intéressé à l'ensemble des sujets susmentionnés et en particulier à la classification automatique, à la théorie du choix social et à la théorie des tresses. Le point commun à ces trois sujets est donc la mise en perspective des structures relationnelles sous-jacentes aux objets mathématiques impliqués. La classification automatique ou *clustering* revient à agréger ou à rechercher une relation d'équivalence. La théorie du choix social et l'aide multicritère à la décision impliquent l'agrégation de relations de préférences. La théorie des tresses peut être abordée du point de vue du groupe des permutations dont elle est une généralisation, et les permutations sont en bijection avec les relations d'ordre total.

J'ai éprouvé beaucoup d'épanouissement à étudier, réfléchir et tenter d'innover dans le cadre de tous ces domaines. Ce plaisir que j'ai eu à explorer des contrées qui m'étaient inconnues mais qui me fascinaient ne m'a plus quitter et explique en grande partie la diversité des thèmes que j'ai abordés au cours de mes expériences passées et dont je vais livrer une synthèse ici.

Mon parcours professionnel est lui aussi emprunt d'expériences riches et variées. Après ma thèse à Thales, un groupe français, j'ai poursuivi ma carrière dans la recherche industrielle en rejoignant Xerox, un groupe américain. Plus précisément, dans le contexte du projet de recherche Infom@gic auquel je participais pendant ma thèse, on m'a proposé un postdoctorat au *Xerox Research Centre Europe* (XRCE) à Grenoble. Au cours de cette période de trois années, entre 2007 et 2010, j'ai abordé de nouveaux domaines scientifiques : la recherche d'information (RI), la fusion d'informations multimédias et le traitement automatique du langage naturel (TALN).

Puis, en 2010, j'ai rejoint l'enseignement et la recherche publique. J'ai été recruté en tant que MCF par l'Université Lumière Lyon 2 (UL2). Il s'agit d'une université en sciences humaines et sociales. J'ai d'abord effectué mes enseignements au sein de la faculté de sciences économiques et de gestion, puis au sein de l'institut de la communication. J'ai effectué ma recherche au sein du laboratoire ERIC qui est composé d'enseignants-chercheurs en section CNU 27 et 26. J'ai passé treize années sur ce poste où j'ai pu étudier différents thèmes.

J'ai eu l'occasion de travailler plus en profondeur en aide multicritère à la décision en découvrant les intégrales de Choquet et plus généralement les fonctions d'agrégation. J'ai continué des recherches entamées à XRCE en TALN et en fusion d'informations hétérogènes. J'ai également développé des idées issues de ma thèse sur les similarités et les critères de partitionnement. Je me suis ensuite investi plus fortement dans les domaines du *machine learning* et du *big data*. La pluralité des enseignements que j'ai dû monter et donner¹ au cours de ces treize années à l'UL2, explique également la variété des objets et des outils que j'ai étudiés et exploités en recherche. Dans le cas de l'apprentissage automatique, je me suis intéressé à la détection des variétés non linéaires, à la classification ascendante hiérarchique et aux aspects en lien avec les données massives et le passage à l'échelle. Plus récemment, je me suis plongé dans l'analyse de données fonctionnelles à la suite d'un accueil en délégation CNRS de six mois en 2019, au Laboratoire Mathématiques de Blaise Pascal (LMBP) de l'Université Clermont Auvergne (UCA).

Le présent mémoire présente une synthèse de ces activités de recherche. Il est organisé en six Chapitres dont voici les intitulés :

1. Fusion d'informations multimodales texte-image en RI basée sur le contenu.

1. Les enseignements que j'ai donnés couvrent les thématiques suivantes : algèbre linéaire, analyse, logiciels de calcul numérique, prédiction de séries temporelles, recherche opérationnelle, théorie des graphes, modélisation linéaire, analyse de données, optimisation numérique, agrégation des préférences, *spatial statistics*, *data mining*, *clustering*, *unsupervised learning*, *supervised learning*, *ensemble methods*, *deep learning*.

-
2. Fusion d'informations linguistiques et statistiques en TALN.
 3. Fonctions d'agrégation et explicabilité en apprentissage supervisé.
 4. Mesures de proximité et critères de partitionnement en *clustering*.
 5. Variétés non linéaires en apprentissage non supervisé.
 6. Méthodes d'apprentissage en analyse de données fonctionnelles.

A la fin de l'introduction de chaque Chapitre, j'indique les publications qui y sont associées. En **gras et rouge** sont les articles de journaux ou de conférence de premier plan².

L'ordonnement des Chapitres est proche de l'ordre chronologique de développement de ces sujets. J'étudie ou ai étudié chacune de ces thématiques sur une période plus ou moins longue. Il existe ainsi des recouvrements de nature temporelle entre elles. Par ailleurs, il existe plusieurs dénominateurs en commun entre ces six domaines variés qui permettent d'expliquer une certaine cohérence dans mon parcours. J'explique ci-dessous trois axes qui structurent mes motivations pour l'étude de ces thèmes de recherche.

Au coeur de cette structuration se trouve **la donnée** que je considère conceptuellement comme **objet numérique** d'une part et **objet mathématique** d'autre part. Chacune des deux dimensions conduit à des instanciations diverses de la donnée et des caractéristiques intrinsèques variées qu'elle peut posséder.

Le **premier axe concerne la diversité des types de données numériques** sur lesquels j'ai travaillé. Chaque sorte de données donne lieu à des propriétés, des challenges et des applications diverses. J'ai eu beaucoup d'intérêt à appréhender, modéliser et analyser les types de données suivants :

- Les **données qualitatives et quantitatives**. Il s'agit des tables de données classiques. Il existe toutefois une pluralité de sortes de données parmi lesquelles les données qualitatives nominales, ordinales ou de rangs ; et les données quantitatives discrètes, continues. On qualifie souvent ces types de *data* de données structurées ou tabulaires, dans la mesure où on peut les stocker dans des tables de données classiques au sens des bases de données. Les données qualitatives ordinales ou quantitatives sont abordées dans le Chapitre 3 dans le cadre de l'aide multicritère à la décision et de l'apprentissage automatique. Les données qualitatives nominales sont en lien avec le problème d'agrégation de relations d'équivalence et les critères de partitionnement que j'étudie dans le Chapitre 4. Au sein de ce Chapitre, je propose également une extension des critères de partitionnement à des données quantitatives continues.
- Les **données relationnelles ou de graphe**. Mon doctorat abordait l'étude des relations binaires et j'ai poursuivi un certain nombre de travaux impliquant ces objets. Ceux-ci concernent en particulier le Chapitre 4. J'ai également étendu mon intérêt à l'étude des graphes quelconques et de terrain. En effet, les données de graphe sont puissantes

2. Selon les classements divers suivants : Scimago, Core Portal et Qualis

car elles permettent de représenter de façon flexible certains types d'information que les données tabulaires ne permettent pas. Par ailleurs, l'avènement du réseau internet et des réseaux sociaux a engendré un vaste intérêt pour l'étude des graphes. L'algorithme *PageRank* de Google est d'ailleurs fondé sur une mesure de centralité dans un graphe. Dans ce manuscrit, les graphes de similarités ou d'affinités sont omniprésents à l'exception du Chapitre 3. Le Chapitre 2 aborde par ailleurs les hypergraphes.

- Les **données multimédias**. Les données textes, images, voix et sons, se situent au centre des applications en intelligence artificielle dans la mesure où elles représentent les principaux modes de communication de l'humain. Ainsi, ces données sont essentiellement composés d'éléments sémantiques. Par exemple, une image est un ensemble de pixels mais elle représente des objets, des personnes, des situations, ... Dès lors que l'on cherche à analyser ce type de données par des algorithmes, un premier verrou concerne la représentation numérique des éléments sémantiques. Dans le Chapitre 1, je traite ce sujet dans le contexte de la recherche d'information basée sur le contenu dans le cas de données image-texte (base de photos annotées par exemple). Ensuite, le Chapitre 2 traite exclusivement de textes et s'inscrit dans le domaine du TALN. Un challenge inhérent à ce type de données que j'étudie, est le problème de *data sparsity*.
- Les **données fonctionnelles**. Dans le cadre de l'internet des objets, des appareils de mesures ou capteurs sont capables de collecter des données en temps réel. Ces données horodatées et/ou géoréférencées permettent de surveiller l'évolution de systèmes environnementaux, physiques ou sociétaux. Ces données se présentent sous forme de flux, elles sont volumineuses et peuvent être variées. Elles entrent naturellement dans le cadre des enjeux du *big data*. J'aborde ce type d'objet du point de vue de l'analyse des données fonctionnelles. Prenons l'exemple type d'une série temporelle. Au lieu d'analyser le vecteur des mesures horodatées, on va d'abord chercher à approximer la série par une fonction continue. Privilégier la forme continue de la donnée permet ici une analyse plus riche. J'étudie exclusivement ce type de données dans le Chapitre 6 de ce rapport.

Le deuxième axe est relatif aux caractéristiques sémantiques et mathématiques des données dont on doit tenir compte dans les modèles afin de mieux capter les informations pertinentes selon la tâche à traiter. J'identifie dans ce contexte les éléments suivants :

- La **représentation numérique des informations d'une donnée**. Cette notion questionne l'adéquation entre d'une part, la donnée brute et d'autre part, son encodage numérique en vue de résoudre une tâche. Prenons un premier exemple simple qui est celui d'une variable ordinale comme l'âge. Cette variable infère naturellement une relation d'ordre sur les individus mais si on considère des tranches d'âge elle implique aussi une relation d'équivalence. Par ailleurs, il peut avoir plusieurs façons de représenter ces deux informations : des listes de valeurs ou des graphes encodant les relations binaires sous-jacentes. Chaque format possède ses propres propriétés et peut être appréhendé par différentes techniques d'analyse. J'aborde ce sujet dans les Chapitres 3 et 4.

Le deuxième exemple correspond au concept de fossé sémantique (*semantic gap*) qui s'applique en particulier à une donnée multimédia. Prenons l'exemple d'une image qui est un ensemble de pixels organisés sur une grille. Cette donnée brute est dite de bas niveaux. Elle n'est pas facilement manipulable à des fins d'analyse. L'extraction de *features* vise donc à représenter numériquement l'image par des concepts de plus haut niveau. Le fossé sémantique est alors l'écart existant entre la représentation de l'image par les *features* et le contenu sémantique de l'image initiale. Je m'intéresse à cette question dans les Chapitres 1 et 2.

Ensuite, j'étudie dans le Chapitre 6 le cas des données fonctionnelles. Reprenons l'exemple type d'une série temporelle. Celle-ci est l'observation partielle d'un phénomène qui est continu. Si on approxime la suite de nombres par une fonction continue alors nous pouvons enrichir l'information initiale par des éléments supplémentaires comme la vitesse et l'accélération du phénomène en question. Ces informations peuvent être obtenues à partir des dérivées premières et secondes de la fonction initiale.

- **L'hypothèse de variété (*manifold hypothesis*)**. L'encodage numérique évoqué plus haut correspond du point de vue mathématique, à la représentation de la donnée dans un espace qui est souvent de très grande dimension. Dans ce contexte, la "malédiction de la dimensionalité" (*curse of dimensionality*) indique que la difficulté d'apprentissage croît de façon exponentielle en fonction de la dimensionalité de l'espace de description. Cependant, en pratique, on constate très souvent que la donnée appartient en fait à un sous-espace de plus petite dimension. En effet, même s'il est nécessaire de représenter numériquement la donnée dans un espace de grande dimension, cela ne veut pas dire que toutes les données du phénomène que l'on peut mesurer, vont remplir l'espace de façon uniforme. En d'autres termes, l'espace de grande dimension est principalement constitué de vecteurs qui ne font pas sens pour le phénomène à l'étude et seule une infime partie de cet espace est pertinent. De ce point de vue, il est donc important, pour l'analyse, d'être en mesure de détecter ces sous-espaces significatifs ce qui est le but des techniques de réduction de dimension. Par ailleurs, dans le cas de données complexes, les sous-espaces pertinents sont très souvent non linéaires. Cette caractéristique présente des challenges et je présente dans les Chapitres 5 et 6 des contributions qui permettent de les relever. Dans le premier Chapitre cité, je propose des algorithmes de *clustering* basés sur le graphe des plus proches voisins qui facilite la détection des variétés non linéaires. Dans le second Chapitre qui est consacré aux données fonctionnelles, j'emploie des méthodes à noyaux qui permettent de capter la géométrie intrinsèque des données à l'instar du noyau Gaussien. Les critères de partitionnement que je développe dans la Chapitre 4 pour la décomposition d'un graphe en composantes connexes et la normalisation de la matrice Laplacienne, sont également en lien avec cette problématique.
- **L'interdépendance des données**. Je fais la distinction entre l'interdépendance entre individus et l'interdépendance entre variables. Le premier cas est directement lié aux données relationnelles et aux graphes. J'étudie dans les Chapitres 5 et 4, le partition-

nement d'un graphe en cliques ce qui permet de déterminer des groupes d'individus fortement interdépendants à l'instar de la détection de communautés dans un réseau social. Dans le cas du TALN et du Chapitre 2, je m'intéresse aux relations de proximité sémantique entre termes ce qui permet de créer des *synset* c'est à dire des ensembles de mots qui sont synonymes.

L'interdépendance entre variables quant à elle, fait écho au point précédent puisqu'un sous-espace non linéaire peut être défini par des dépendances fonctionnelles entre variables. Une autre façon d'appréhender cet aspect est par le biais des fonctions noyaux utilisées dans les *kernel machines*. Une fonction noyau peut être associée à une application dite *feature map*, qui projette les individus dans un espace étendu et dont les variables sont des transformations non linéaires des variables de l'espace de départ. Ces transformations représentent des interactions entre les variables initiales. Les fonctions noyaux sont des outils que j'utilise beaucoup et ils interviennent fortement dans les Chapitres 6 et 5.

Toutefois, les fonctions noyaux sont de nature géométrique et les interactions entre variables qu'elles encodent implicitement ne s'interprètent pas en général. A l'inverse de ces approches de type "boîte noire", j'étudie dans le Chapitre 3 des objets mathématiques en aide multicritère à la décision dont le but est de modéliser les préférences d'un agent rationnel. Il s'agit de l'intégrale de Choquet et des fonctions d'agrégation. Ces outils visent à dépasser les limites de fonctions classiques telles que la moyenne pondérée qui suppose implicitement l'indépendance entre variables. J'examine l'intégrale de Choquet bipolaire et je définis une fonction d'agrégation dans le Chapitre 3. Ces outils permettent de modéliser et d'interpréter les interactions entre variables pour la prise de décision.

Les deux axes précédents présentaient des éléments caractérisant la donnée en tant qu'objet. J'ajoute une [troisième dimension qui définit plus spécifiquement les types de tâches qui m'ont intéressé lorsque l'on dispose de plusieurs données pouvant être hétérogènes](#). Ceci positionne le propos dans un contexte d'**analyse de données complexes** avec une prise en compte des **aspects multi-vues ou multi-sources**, le cas échéant.

- [Le problème de la mesure de similarité entre données de nature différente ou décrites par des descripteurs hétérogènes](#). Les mesures de similarité ou d'affinité ou d'association, visent à évaluer un lien entre deux objets et sont centrales pour les tâches en analyse de données complexes et en apprentissage automatique. Dans le Chapitre 1, je présente des techniques de fusion permettant de quantifier des relations de proximité entre objets multimédias image-texte.

Puis, dans le Chapitre 4, je définis des mesures de similarité pour deux relations binaires quelconques et hétérogènes.

J'introduis le concept de similarité pénalisée dans le Chapitre 5.

Enfin, concernant les données fonctionnelles avec dérivées qui font l'objet du Chapitre 6, je présente une procédure permettant de définir un noyau multiple approprié pour

mieux comparer deux fonctions dans un contexte non supervisé ou supervisé.

- Le [problème de l'agrégation de plusieurs données](#). Ce mémoire fait la part belle aux méthodes de fusions d'informations. Plusieurs contextes sont étudiés. Les approches de fusion précoce, tardive et transmodale utilisées dans les Chapitres 1 et 2 permettent d'agréger des données et informations de natures différentes. Dans le premier cas, la fusion s'opère entre des informations provenant de médias distincts, texte et image; tandis que dans le deuxième cas, ce sont des informations de types divers, linguistique et statistique, qui sont fusionnées.

Ensuite, les fonctions d'agrégation examinées dans le Chapitre 3, servent à combiner des nombres réels en un score synthétique en tenant compte des interactions entre variables. La nature de ces interactions peut être révélée à partir de l'observation d'un échantillon et en ayant recours à des modèles d'apprentissage supervisé.

Enfin, les différents types de critères de partitionnement que j'examine dans le Chapitre 4 peuvent être interprétés comme l'agrégation de variables qualitatives nominales ou l'agrégation de variables quantitatives continues.

- Le [problème de la structuration d'un ensemble de données](#). Des tâches classiques en science des données consistent soit à ordonner un ensemble de données, soit à segmenter un ensemble de données. Les fonctions d'agrégation du Chapitre 3 évoquées précédemment, donnent un score globale à différentes alternatives afin de les ordonner et de prendre des décisions.

La recherche d'information qui est la thématique traitée au Chapitre 1 vise à ordonner les éléments d'une base étant donnée une requête.

Dans les Chapitres 3 et 6, il est question de catégorisation en apprentissage supervisé et le but est d'inférer une fonction qui affecte un individu à une classe. Cette fonction revient à attribuer une distribution de scores sur l'ensemble des classes étant donné un individu. Elle sert donc à ordonner les classes.

Enfin, le *clustering* est une tâche récurrente dans ce mémoire. La partition est la structure de segmentation employée par les méthodes décrites dans les Chapitres 4, 5 et 6. Je traite des structures hiérarchiques dans le Chapitre 5.

La grande majorité des éléments constituant ces trois axes sont commun à au moins trois des six Chapitres de ce rapport. J'espère qu'ils fourniront au lecteur des fils conducteurs pour naviguer avec intérêt au sein de la variété des six étapes du parcours non linéaire qui va suivre.

Fusion d'informations multimodales texte-image en RI basée sur le contenu

Sommaire du chapitre

1.1	Introduction	8
1.1.1	Contexte	8
1.1.2	Travaux antérieurs	10
1.2	Contributions	13
1.2.1	<i>Reranking, clustering</i> et promotion de la diversité en recherche d'information	13
1.2.2	Apprentissage actif et <i>multimedia information seeking</i>	14
1.2.3	Marche aléatoire sur des graphes et unification de méthodes de fusion transmodale	19
1.3	Discussions et perspectives	24

1.1 Introduction

1.1.1 Contexte

Ce premier Chapitre expose mes contributions en fusion d'informations hétérogènes. Je me suis investi dans ce sujet de recherche au cours de mon postdoctorat de trois années au sein de *Xerox Research Centre Europe* (XRCE) entre 2007 et 2010. Ces travaux ont contribué au projet Infom@gic émanant du pôle de compétitivité francilien Cap Digital dont Thales, le groupe au sein duquel j'ai effectué ma thèse CIFRE, était le coordinateur et XRCE un des membres du consortium qui était composé de plus de 20 partenaires académiques et industriels.

Ces activités se situent vers la fin des années 2000 et en amont du développement des méthodes de *deep learning*. Le projet Infom@gic avait pour but de réunir des acteurs de l'ingénierie de la connaissance afin de développer des synergies dans le domaine de l'indexation, de la recherche et de l'analyse de l'information multimédia. Au sein de XRCE, je travaillais

alors dans l'équipe *Textual and Visual Pattern Analysis* (TVPA), sur le problème de la **fusion d'informations texte-image pour la recherche d'information multimédia basée sur le contenu** (*Content-Based Multimedia Information Retrieval -CBMIR-*).

Les objets multimédias que j'étudiais, étaient composés d'une image et de tags ou d'un texte descriptif. Un exemple typique dans ce cas est la plateforme de partage de photos *Flickr*. La tâche consistait à retrouver les objets multimédias correspondants à une requête qui, elle aussi, pouvait être composée d'une image et de tags/textes. Il s'agit toutefois d'une recherche basée sur le contenu, c'est à dire que les résultats attendus devaient être sémantiquement pertinents vis-à-vis de la requête. Prenons l'exemple d'une requête image montrant un dauphin dans l'océan, si la top liste de l'algorithme retourne des images de poissons dans l'eau qui sont visuellement très proches de la requête mais, qui ne sont pas des dauphins alors celles-ci ne sont pas pertinentes. Les méthodes initiées par les collègues de XRCE et auxquelles j'ai contribué à enrichir par la suite, sont fondées sur des **opérations de diffusion (ou propagation) transmodale (ou cross-modale) de similarités** et se sont distinguées à la fois par leur originalité et par les performances qu'elles permettaient à cette période. Nous avons ainsi remporté plusieurs challenges de *photo retrieval* et de *multimedia retrieval* lors des campagnes d'évaluation internationale ImageCLEF en 2008 et 2009.

Je présente ici les différents travaux auxquels j'ai contribué afin d'enrichir cette approche dans les divers contextes suivants : promotion de la diversité dans la top liste, *information seeking*, boucle de pertinence et liens avec les modèles en recherche d'information (RI) fondés sur les chaînes de Markov.

Les publications dans des revues ou conférences avec comités de lecture qui sont concernées par ce Chapitre sont les suivantes :

- **J. Ah-Pine**, Gabriela Csurka, Stéphane Clinchant. 2015. Unsupervised Visual and Textual Information Fusion in CBMIR Using Graph-Based Methods. *ACM Transactions on Information Systems*. 33(2). [Lien vers le journal, <http://dl.acm.org/citation.cfm?id=2699668>].
- S. Clinchant, **J. Ah-Pine**, G. Csurka. 2011. Semantic Combination of Textual and Visual Information in Multimedia Retrieval. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR 2011)* [Taux d'acceptation < 35%]. [Lien vers la conférence, <http://www.icmr2011.org/>].
- **J. Ah-Pine**, S. Clinchant, G. Csurka, F. Perronnin, J.M. Renders. 2010. Leveraging text, image and cross-media similarities for diversity-focused multimedia retrieval. *ImageCLEF*. Springer. The Information Retrieval series. Vol 32.
- **J. Ah-Pine**, S. Clinchant, G. Csurka. 2010. Comparison of several combinations of multimodal and diversity seeking methods for multimedia retrieval. *Multilingual Information Access Evaluation Vol. II Multimedia Experiments, 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009), Lecture Notes in Computer Science (LNCS 6242)*. [Lien vers le journal, <http://www.springerlink.com/content/u517184286474j81/>]

- **J. Ah-Pine**, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, J.M. Renders. 2009. Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications Journal*. 42(1) :31-56. [Lien vers le journal, <http://www.springerlink.com/content/5777145v45407864/>].
- **J. Ah-Pine**, J.M. Renders, M.L. Viaud. 2009. A Continuum between Serendipitous Browsing and Query-based Search for Multimedia Information Access. *Proceedings of the Adaptive Multimedia Retrieval (AMR 2009)* [Workshop]. [Lien vers la conférence, <http://cabrillo.lsi.uned.es/nlp/amr2009/program>].
- **J. Ah-Pine**, G. Csurka, J.M. Renders. 2009. Evaluation of diversity-focused strategies for Multimedia Retrieval. *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008)*, *Lecture Notes in Computer Science (LNCS 5706)*. [Lien vers le journal, <http://www.springerlink.com/content/h3123p4555v79112/>].

1.1.2 Travaux antérieurs

Dans le contexte de la fusion d'informations multimédias, deux approches classiques sont la fusion précoce ou *early* et la fusion tardive ou *late*. Prenons le cas d'objets multimédias composés d'une image et d'un texte (ou de tags) la décrivant. Nous supposons que pour chacun des deux médias, des pré-traitements spécifiques d'extraction de *features* ont été appliqués, si bien que nous disposons pour un objet multimédia X , deux représentation vectorielles \mathbf{x}^v et \mathbf{x}^t .

La **fusion précoce**, consiste à prendre comme représentation numérique d'un objet X , un vecteur qui concatène les deux vues, image et texte. Chaque objet $X_i; i = 1, \dots, n$ de la base est représenté par un vecteur \mathbf{x}_i qui concatène \mathbf{x}_i^v et \mathbf{x}_i^t . Les pré-traitements image et texte se font de façon indépendante, la fusion consiste uniquement à enrichir la représentation numérique par des vecteurs composites. Les vecteurs \mathbf{x}_i de la base sont de même taille. On peut alors appliquer tout type de méthode multivariée classique. Dans le cas de la recherche d'information, il s'agit de rechercher les l objets de la base qui sont les plus pertinents vis-à-vis d'une requête multimédia Q que l'on représente également par un vecteur composite $\mathbf{q} = (\mathbf{q}^v, \mathbf{q}^t)$. La représentation numérique de ces objets et de la requête, permettent de définir des fonctions de similarité ou de pertinence qui donnent lieu à une relation d'ordre et par conséquent à une top liste. Notons par $\text{Rel}(Q, X_i)$ la mesure de la pertinence de l'objet X_i vis-à-vis de la requête Q . Notons de plus par $\text{Sim}(\mathbf{q}, \mathbf{x}_i)$ une mesure de similarité entre les vecteurs \mathbf{q} et \mathbf{x}_i . La fusion précoce revient conceptuellement à définir une fonction de score comme suit :

$$\text{Rel}^{\text{early}}(Q, X_i) = \text{Sim}(\mathbf{q}, \mathbf{x}_i) \text{ avec } \mathbf{q} = (\mathbf{q}^v, \mathbf{q}^t) \text{ et } \mathbf{x}_i = (\mathbf{x}_i^v, \mathbf{x}_i^t).$$

Dans le cas de la **fusion tardive**, il s'agit d'appliquer en premier lieu des fonctions de similarité propres à chaque média et ensuite, de combiner ces scores de pertinence monomédia

1.1. INTRODUCTION

par le biais d'une fonction d'agrégation. Dans ce contexte, notons par Sim^v et Sim^t , des fonctions de similarité monomédia image et texte respectivement. Afin d'alléger les notations, je noterai $\text{Sim}^v(\mathbf{q}, \mathbf{x}_i)$ au lieu de $\text{Sim}^v(\mathbf{q}^v, \mathbf{x}_i^v)$ pour indiquer la similarité entre les parties images des objets. Il en va de même pour la similarité texte. Suite à ces précisions, on peut alors formaliser la fusion tardive comme suit :

$$\text{Rel}^{\text{late}}(Q, X_i) = F(\text{Sim}^v(\mathbf{q}, \mathbf{x}_i), \text{Sim}^t(\mathbf{q}, \mathbf{x}_i)),$$

où $F : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ est une fonction d'agrégation¹.

Dans la Figure 1.1 tirée de [Ah-Pine et al., 2015], sont illustrés les principes de la fusion précoce et tardive (2 première lignes) dans le cas de données image-texte.

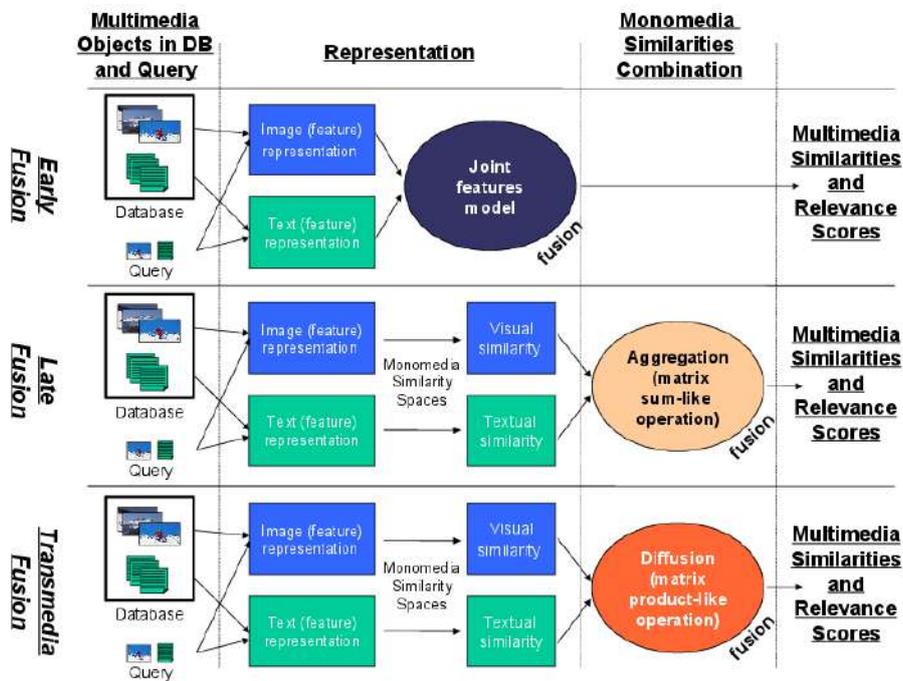


FIGURE 1.1 – Fusion précoce (*early*), tardive (*late*) et intermédiaire ou transmodale (*cross-media*)

La *early* fusion et la *late* fusion forment des stratégies simples et classiques de fusion d'informations multimédias. Dans ce qui suit, j'introduis le principe de **fusion transmédia (ou cross-média)** qui avait été défini par les collègues de XRCE dans [Clinchant et al., 2007, Clinchant et al., 2008] avant mon arrivée. L'idée principale consiste à diffuser des similarités d'un média vers l'autre afin d'avoir une matrice de similarités entre objets multimédias qui comble davantage le **fossé sémantique**². Plus précisément, ce concept formalise la déviation

1. Pour fixer les idées, on prendra F telle une moyenne pondérée mais je reviendrai plus longuement sur ces outils dans le Chapitre 3.

2. Concept discuté en introduction page 4.

entre la sémantique portée par un objet multimédia, par exemple le contenu d'une image X^v , et la représentation numérique utilisée pour indexer cette image, par exemple un vecteur \mathbf{x}^v représentant l'histogramme de couleurs de cette image.

Soit l'application Knn^k qui prend comme argument un vecteur de nombres, et donne en résultat, un vecteur de même taille rempli de zéros sauf pour les cellules contenant les k plus grandes valeurs qui sont conservées (ou alors transformées en valeur unitaire). Soient également $\mathbf{S}^v = (\text{Sim}^v(\mathbf{x}_i, \mathbf{x}_{i'}))_{i,i'=1,\dots,n}$ et $\mathbf{S}^t = (\text{Sim}^t(\mathbf{x}_i, \mathbf{x}_{i'}))_{i,i'=1,\dots,n}$ les matrices de similarité image et texte des objets de la base. Par ailleurs, désignons par un léger abus de notation, $\text{Sim}^v(\mathbf{q}, \cdot)$ et $\text{Sim}^t(\mathbf{q}, \cdot)$, les vecteurs de taille $1 \times n$ des similarités image et texte entre la requête Q et tous les objets X_i de la base. L'approche transmédia conduit dans le cas de données image-texte vers deux possibilités : soit **on propage les similarités images vers les similarités textes** ce que l'on dénote par **similarités transmodales image-texte** (Sim^{vt}), soit l'inverse et on obtient des **similarités transmodales texte-image** (Sim^{tv}). Formellement, les vecteurs des similarités image-texte et texte-image entre une requête Q et les éléments X_i sont respectivement définis par :

$$\text{Sim}^{vt}(\mathbf{q}, \cdot) = \text{Knn}^k(\text{Sim}^v(\mathbf{q}, \cdot))\mathbf{S}^t, \quad (1.1)$$

$$\text{Sim}^{tv}(\mathbf{q}, \cdot) = \text{Knn}^k(\text{Sim}^t(\mathbf{q}, \cdot))\mathbf{S}^v. \quad (1.2)$$

Le principe de propagation transmédia image-texte est illustré dans la Figure 1.2.

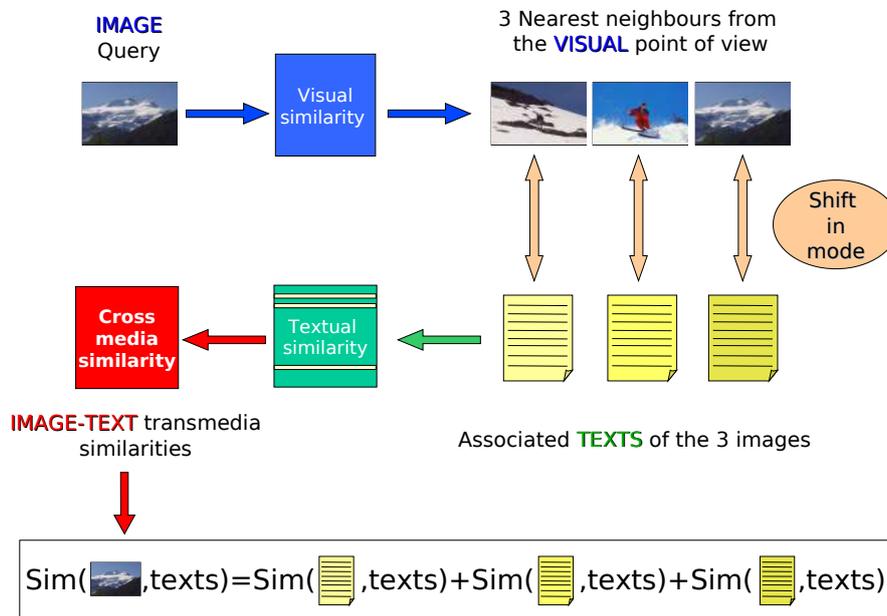


FIGURE 1.2 – Similarités image-texte en recherche d'information image-texte

Les approches transmodales ont permis à mes collègues de XRCE de remporter le challenge

Photo Retrieval de la campagne d'évaluation internationale ImageCLEF en 2007. Je présente dans la section suivante, les différents travaux auxquels j'ai contribué afin d'enrichir cette approche dans les divers contextes suivants : **promotion de la diversité dans la top liste** (sous-section 1.2.1), **information seeking et boucle de pertinence** (sous-section 1.2.2) et liens avec les **modèles en recherche d'information (RI) fondés sur les chaînes de Markov** (sous-section 1.2.3).

1.2 Contributions

1.2.1 *Reranking, clustering* et promotion de la diversité en recherche d'information

Mes contributions dans ce contexte ont été multiples. Dans un premier temps, nous avons participé à la tâche *Photo Retrieval* d'ImageCLEF 2008 et contrairement à l'année précédente, l'objectif était double : pour chaque requête il fallait retrouver le maximum d'objets pertinents de la base, mais on devait également **générer de la diversité au sein de la top liste**. Dit autrement, l'objectif était d'éviter de présenter d'affilée des objets pertinents similaires et de favoriser une présentation des différents types d'objets pertinents. Le critère d'évaluation prenait donc en compte deux éléments : la pertinence et la diversité. Notre système était composé de deux étapes. Dans un premier temps, nous avons utilisé les similarités trans-modales pour déterminer une top liste d'objets. Puis, nous avons appliqué deux approches pour favoriser la diversité. La première technique dénotée MMR pour **Maximum Margin Relevance** est connue de la littérature. Elle consiste à ré-ordonner la top liste de sorte à ce que l'objet placé en rang j soit dissimilaire aux objets classés avant lui, c'est à dire à un rang $j' < j$. Formellement, le score MMR d'un objet \mathbf{x}_i étant donné une requête \mathbf{q} est donné par :

$$\text{MMR}(\mathbf{q}, \mathbf{x}_i) = \beta(j)\text{Rel}(\mathbf{q}, \mathbf{x}_i) - (1 - \beta(j)) \max_{\mathbf{x}_{i'} \in \mathbb{L}_j} \text{Sim}(\mathbf{x}_i, \mathbf{x}_{i'}), \quad (1.3)$$

où \mathbb{L}_j est la liste des objets rangés en position 1 à j ; Rel et Sim sont des fonctions de pertinence et de similarité quelconques ; et $\beta(j) \in [0, 1]$ est un paramètre de mélange pouvant dépendre du rang j .

L'autre approche qui a été implémentée est une initiative personnelle et consiste à employer un **algorithme de clustering pour segmenter les objets de la top liste** afin de capter les différents *topics* au sein de cette dernière. On considère alors deux listes d'objets, l'une appelée prioritaire, et l'autre non prioritaire. Pour les remplir, on parcourt les éléments de la top liste en commençant par le premier. Si le *cluster id* de l'objet courant n'est pas dans la liste prioritaire alors on le rajoute dans celle-ci et autrement il est ajouté dans la liste non prioritaire. On passe au suivant et on procède ainsi jusqu'à ce que le nombre de *clusters* distincts représentés dans la liste prioritaire atteint un entier k ($k = 10$ typiquement) qui est un paramètre. Les objets de la liste prioritaire sont alors placés avant tous les autres objets (non prioritaires et complémentaires). L'ordre au sein de ces deux parties est donnée par les

1.2. CONTRIBUTIONS

scores de pertinence initiaux. On a ainsi en haut de la liste, k objets variés car appartenant à des *clusters* différents.

Une propriété importante dans le cadre de cette seconde méthode, est l'utilisation d'un **algorithme de *clustering* qui ne fixe pas le nombre de *clusters***. Pour cela, j'ai mis en avant des méthodes basées sur l'analyse relationnelle qui est l'approche sur laquelle j'ai travaillé au cours de ma thèse de doctorat. Cette propriété permet de découvrir les différents thèmes au sein de la top liste et ceci a contribué aux performances que nous avons obtenues lors de ImageCLEF 2008. XRCE a en effet également remporté le challenge *Photo Retrieval* de la session 2008 [Ah-Pine et al., 2008, Ah-Pine et al., 2009].

Je donne dans la Figure 1.3 un exemple de résultats qui compare la top liste de taille 10 sans et avec *diversity re-ranking* basé sur le *clustering*. Les images pertinentes sont encadrées et les couleurs des cadres varient selon les *clusters id*. La ligne 2 montre une plus grande variété d'images pertinentes en comparaison de la ligne 1 qui présente des images assez redondantes.



FIGURE 1.3 – Exemples de résultats sans (ligne du haut) et avec (ligne du bas) *diversity re-ranking* basé sur le *clustering*.

1.2.2 Apprentissage actif et *multimedia information seeking*

Avec Jean-Michel Renders de XRCE, nous avons ensuite collaboré dans le cadre du projet [Infom@gic](#) avec des collègues de l'INA (Institut National de l'Audiovisuel) toujours en recherche d'information multimédia mais dans un paradigme autre. Il s'agit d'*information seeking* où le besoin en recherche d'information est plus large. Dans ce contexte, on suppose deux types de besoin distincts, l'un qualifié de *browse-based search* et l'autre, plus classique, de *query-based search*. Dans le premier cas, on explore une base de données à partir d'une interface présentant les objets de façon structurée permettant à l'utilisateur.rice de positionner les éléments qui l'intéresse au sein de contextes de niveaux de granularité variables. Dans le second cas, l'utilisateur.rice peut interroger la base de données de manière itérative à l'aide de requêtes qui sont progressivement alimentées de ses *feed-back*. Le but est de trouver un sous-ensemble d'objets le plus pertinent possible vis-à-vis d'un besoin supposé plus précis que dans le premier cas.

Le système que nous avons défini met un place un *continuum* entre ces deux besoins en information. Il permet à un.e utilisateur.rice de faire des va-et-vient entre d'une

1.2. CONTRIBUTIONS

part, explorer globalement par sérendipité la base de données, et d'autre part, faire un focus sur un sous-ensemble d'objets d'intérêt qu'il s'agit d'approfondir. Ce **système** intègre les composantes suivantes :

- Une visualisation et navigation multi-échelles. Cela s'effectue par le biais de deux cartes, l'une statique et globale, et l'autre dynamique et locale, et qui dépend des résultats de la recherche par requête. Les cartes sont générées par des techniques de *graph-layout*.
- Un accès multimodal à la base de données multimédias. Dans notre cas et dans la suite de ce qui a été présenté précédemment, nous nous sommes intéressés à une base d'objets image-texte. Les vues et requêtes peuvent être basées sur l'image, sur le texte ou sur les deux (mode hybride).
- Une boucle de pertinence multimodale. Dans ce contexte, l'utilisateur.rice peut donner un *feed-back* sur les images et/ou les textes des objets qu'il.elle a trouvé pertinents.
- Un système de paramétrage de la recherche itérative. Celui-ci permet une grande flexibilité dans le choix des médias à utiliser et également dans la définition des comportements de recherche, en spécifiant des facteurs de localité et d'oubli (poids décroissant des objets sélectionnés au fur et à mesure des itérations).

Je donne dans la Figure 1.4 le diagramme précisant l'architecture de notre système.

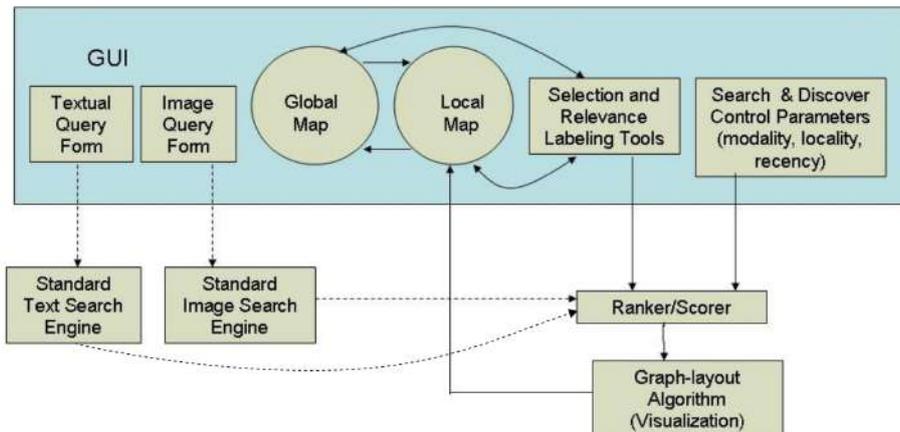


FIGURE 1.4 – Architecture de notre système d'*information seeking*.

De plus la Figure 1.5 illustre dans un cas d'usage "tourisme à Paris", le *continuum* entre le *browse-based* et le *query-based search* permettant une vue de l'information qui est soit panoramique et globale, soit précise et locale.

Au coeur de ce système se trouve un **modèle de boucle de pertinence** qui repose sur les caractéristiques suivantes :

- L'utilisateur.rice peut annoter (pertinent ou non pertinent) l'image ou le texte d'un objet multimédia de façon indépendante. Ceci lui permet une grande flexibilité dans l'expression de son besoin en information.

1.2. CONTRIBUTIONS

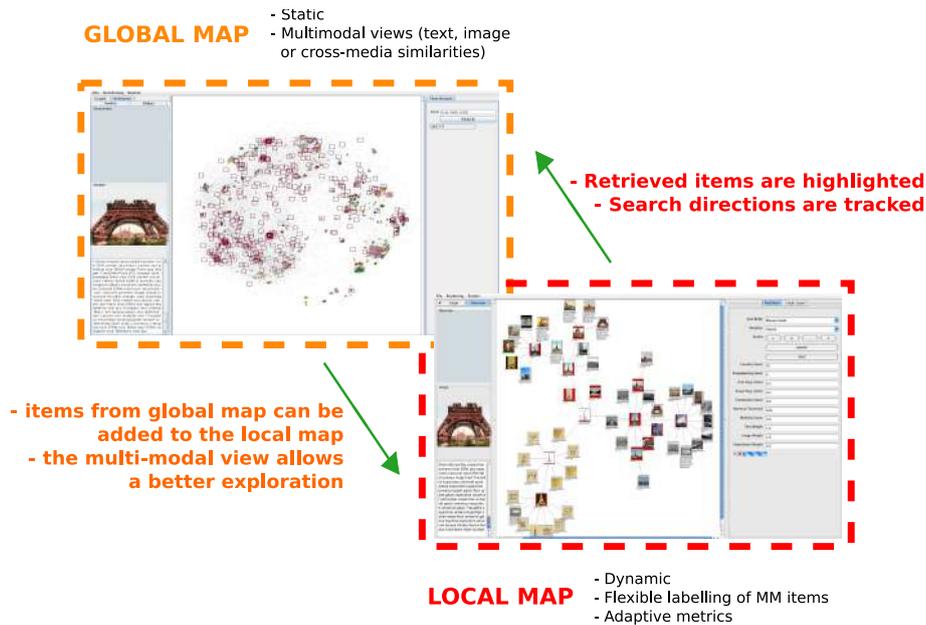


FIGURE 1.5 – Complémentarité des deux cartes et des deux modes de recherche d’information.

- De façon similaire, l’utilisateur.rice peut effectuer à tout moment de sa session, une recherche basée sur la similarité image ou texte ou transmodale.
- Le modèle intègre un *forgetting factor* qui suppose que les éléments annotés récemment ont plus d’importance dans la recherche suivante que ceux qui ont été annotés plus loin dans le passé.
- L’approche comprend également un *locality factor* qui permet à l’utilisateur.rice de sélectionner un sous-ensemble d’objets annotés à qui on peut attribuer un poids supplémentaire dans la recherche suivante. Ceci permet de réorienter la recherche à partir de ces éléments sans avoir à démarrer une nouvelle session.

L’ensemble de ces propriétés sont implicitement intégrés dans la formule que je présente ci-dessous et qui peut être vue comme une **extension de la formule de Rocchio** qui est une approche vectorielle de *relevance feed-back* introduite dans [Salton and Buckley, 1990] à partir des travaux de [Rocchio Jr, 1971].

Malgré une formulation lourde, l’approche est relativement simple dans ces constituants principaux. Il s’agit d’une part, de combiner similarités monomédias (en vert dans (1.4)) et similarités transmédias (en orange dans (1.4)) et d’autre part, d’agréger soit positivement (en rouge dans (1.4)) les textes et images annotés comme pertinents, soit négativement (en bleu dans (1.4)) les textes et images annotés comme non pertinents. Dans l’idée, cela revient à définir une nouvelle requête à partir du *feed-back* en pondérant positivement les éléments pertinents et négativement les éléments non pertinents.

Toutefois, notre approche va plus loin que les techniques initiales en conceptualisant des fonctions de pondération qui permettent de modéliser à la fois un **processus d’oubli** dans

les annotations, et un **changement de direction** dans la recherche d'information. Cette modélisation est décrite dans ce qui suit et elle correspond aux équations données en (1.5) et (1.6).

$$\begin{aligned}
 \text{Rel}_{t+1}(\mathbf{x}_i) = & \tag{1.4} \\
 & \gamma_t^t \left[\sum_{\mathbf{y} \in \mathbb{T}_t^+} \frac{\alpha_t^t(\mathbf{y})}{\sum_{\mathbf{y}' \in \mathbb{T}_t^+} \alpha_t^t(\mathbf{y}')} \left(\text{Sim}^t(\mathbf{y}, \mathbf{x}_i) + \lambda^t \frac{\sum_{\mathbf{z} \in \mathbb{B}_t^t(\mathbf{y})} \text{Sim}^t(\mathbf{y}, \mathbf{z}) \text{Sim}^v(\mathbf{z}, \mathbf{x}_i)}{\sum_{\mathbf{z}' \in \mathbb{B}_t^t(\mathbf{y})} \text{Sim}^t(\mathbf{y}, \mathbf{z}')} \right) \right. \\
 & - \sum_{\mathbf{y} \in \mathbb{T}_t^-} \frac{\beta_t^t(\mathbf{y})}{\sum_{\mathbf{y}' \in \mathbb{T}_t^-} \beta_t^t(\mathbf{y}')} \left. \left(\text{Sim}^t(\mathbf{y}, \mathbf{x}_i) + \delta^t \frac{\sum_{\mathbf{z} \in \mathbb{B}_t^t(\mathbf{y})} \text{Sim}^t(\mathbf{y}, \mathbf{z}) \text{Sim}^v(\mathbf{z}, \mathbf{x}_i)}{\sum_{\mathbf{z}' \in \mathbb{B}_t^t(\mathbf{y})} \text{Sim}^t(\mathbf{y}, \mathbf{z}')} \right) \right] \\
 & + \gamma_t^v \left[\sum_{\mathbf{y} \in \mathbb{I}_t^+} \frac{\alpha_t^v(\mathbf{y})}{\sum_{\mathbf{y}' \in \mathbb{I}_t^+} \alpha_t^v(\mathbf{y}')} \left(\text{Sim}^v(\mathbf{y}, \mathbf{x}_i) + \lambda^v \frac{\sum_{\mathbf{z} \in \mathbb{B}_t^v(\mathbf{y})} \text{Sim}^v(\mathbf{y}, \mathbf{z}) \text{Sim}^t(\mathbf{z}, \mathbf{x}_i)}{\sum_{\mathbf{z}' \in \mathbb{B}_t^v(\mathbf{y})} \text{Sim}^v(\mathbf{y}, \mathbf{z}')} \right) \right. \\
 & - \sum_{\mathbf{y} \in \mathbb{I}_t^-} \frac{\beta_t^v(\mathbf{y})}{\sum_{\mathbf{y}' \in \mathbb{I}_t^-} \beta_t^v(\mathbf{y}')} \left. \left(\text{Sim}^v(\mathbf{y}, \mathbf{x}_i) + \delta^v \frac{\sum_{\mathbf{z} \in \mathbb{B}_t^v(\mathbf{y})} \text{Sim}^v(\mathbf{y}, \mathbf{z}) \text{Sim}^t(\mathbf{z}, \mathbf{x}_i)}{\sum_{\mathbf{z}' \in \mathbb{B}_t^v(\mathbf{y})} \text{Sim}^v(\mathbf{y}, \mathbf{z}')} \right) \right].
 \end{aligned}$$

Les notations de la formule précédente sont définies comme suit :

- t et $t + 1$ sont les itérations courante et suivante.
- γ_t^t et γ_t^v sont les poids attribués au texte et à l'image à l'itération t .
- \mathbb{T}_t^+ et \mathbb{T}_t^- sont les sous-ensembles d'objets dont les textes ont été annotés pertinents et non pertinents de l'itération 1 jusqu'à l'itération t .
- \mathbb{I}_t^+ et \mathbb{I}_t^- sont les sous-ensembles d'objets dont les images ont été annotées pertinentes et non pertinentes de l'itération 1 jusqu'à l'itération t .
- \mathbf{y} et \mathbf{y}' sont des objets multimédias appartenant à \mathbb{T}_t^+ ou \mathbb{T}_t^- ou \mathbb{I}_t^+ ou \mathbb{I}_t^- .
- $\mathbb{B}_t^t(\mathbf{y})$ sont les objets non annotés dont les textes font partie des k plus proches voisins du texte de \mathbf{y} .
- $\mathbb{B}_t^v(\mathbf{y})$ sont les objets non annotés dont les images font partie des k plus proches voisins de l'image de \mathbf{y} .
- Étant donné \mathbf{y} , \mathbf{z} et \mathbf{z}' sont des objets multimédias appartenant à $\mathbb{B}_t^t(\mathbf{y})$ ou $\mathbb{B}_t^v(\mathbf{y})$.
- λ^t et δ^t sont des paramètres dans $[0, 1]$ pondérant la similarité transmodale texte-image des textes pertinents et non pertinents respectivement.

1.2. CONTRIBUTIONS

- λ^v et δ^v sont des paramètres dans $[0, 1]$ pondérant la similarité transmodale image-texte des images pertinentes et non pertinentes respectivement.
- α_t^t et β_t^t sont des fonctions attribuant des poids aux textes annotés se trouvant dans \mathbb{T}_t^+ et \mathbb{T}_t^- respectivement.
- α_t^v et β_t^v sont des fonctions attribuant des poids aux images annotées se trouvant dans \mathbb{I}_t^+ et \mathbb{I}_t^- respectivement.

Les fonctions de pondération α_t^u et β_t^u avec $u \in \{t, v\}$, sont des paramètres qui tiennent compte des facteurs de localité ($\text{Locality}_t^u \in [0, 1]$) et d'oubli ($\text{Forget}_t^u \in [0, 1]$). Je précise uniquement ci-dessous la définition de α_t^t qui est valable pour tout $\mathbf{y} \in \mathbb{T}_t^+$, c'est à dire pour tous les textes qui ont été annotés pertinents au cours de la session allant des itérations 1 à t incluse. Les autres fonctions sont définies de façon similaire. En particulier, on pourra prendre par défaut $\beta_t^u = \alpha_t^u$, dans la mesure où cette configuration donnait des résultats satisfaisant lors de nos tests.

$$\alpha_t^t(\mathbf{y}) = \begin{cases} \frac{1}{1 - \text{Locality}_t^t} & \text{si } \mathbf{y} \in \mathbb{S}_t^t, \\ (1 - \text{Forget}_t^t)^{\text{Decay}^t(\mathbf{y})}, & \text{si } \mathbf{y} \notin \mathbb{S}_t^t \text{ et } \text{Decay}^t(\mathbf{y}) \geq 0, \\ 0 & \text{si } \mathbf{y} \notin \mathbb{S}_t^t \text{ et } \text{Decay}^t(\mathbf{y}) = -1, \end{cases} \quad (1.5)$$

où, étant donné $\mathbf{y} \in \mathbb{T}_t^+ \cup \mathbb{T}_t^-$, la fonction Decay^t est définie par :

$$\text{Decay}^t(\mathbf{y}) = \begin{cases} t - \text{Date}^t(\mathbf{y}) & \text{si } \mathbb{S}_t^t = \emptyset, \\ \min_{\mathbf{z} \in \mathbb{D}_t^t(\mathbf{y})} (\text{Date}^t(\mathbf{z}) - \text{Date}^t(\mathbf{y})) & \text{si } \mathbb{S}_t^t \neq \emptyset \text{ et } \mathbb{D}_t^t(\mathbf{y}) \neq \emptyset, \\ -1 & \text{si } \mathbb{S}_t^t \neq \emptyset \text{ et } \mathbb{D}_t^t(\mathbf{y}) = \emptyset. \end{cases} \quad (1.6)$$

Je donne ci-dessous une description littérale des différents contextes considérés par ce modèle afin de formaliser les facteurs d'oubli et de "récence" des *feed-back* dans le processus itératif d'*information seeking* :

- Parmi les textes ayant été annotés pertinents et mis dans \mathbb{T}_t^+ , l'utilisateur.rice a la possibilité de sélectionner un sous-ensemble $\mathbb{S}_t^t \subset \mathbb{T}_t^+$ afin de signifier au système qu'il.elle souhaite focaliser la recherche sur les éléments de celui-ci. Lorsque ce sous-ensemble est non vide ($\mathbb{S}_t^t \neq \emptyset$), cela correspond à la 1ère condition de (1.5), et dans ce cas, le paramètre $\text{Locality}_t^t \in [0, 1]$ est activé et celui-ci permet de donner une importance spécifique aux éléments $\mathbf{y} \in \mathbb{S}_t^t$ en leur attribuant un poids $\alpha_t^t(\mathbf{y}) \geq 1$. Plus Locality_t^t est proche de 1, plus α_t^t est grand et ainsi, plus les textes $\mathbf{y} \in \mathbb{S}_t^t$ ont une forte prépondérance dans la détermination des résultats à l'itération $t + 1$.
- Pour les éléments \mathbf{y} qui ont été précédemment annotés pertinents ($\mathbf{y} \in \mathbb{T}_t^t$) mais qui n'ont pas été sélectionnés à l'itération t ($\mathbf{y} \notin \mathbb{S}_t^t$), leur poids $\alpha_t^t(\mathbf{y})$ est compris entre 0 et 1. On distingue alors trois situations qui dépendent, entre autre, de l'itération au cours de laquelle \mathbf{y} a été annoté, ce que l'on dénote par $\text{Date}^t(\mathbf{y})$. Cette dernière application renvoie un entier compris entre 1 et t :

- Si aucun élément parmi \mathbb{T}_t^+ n'est sélectionné à l'itération t ($\mathbb{S}_t^t = \emptyset$), alors l'exposant incarnant le phénomène d'oubli décroît de façon linéaire en fonction du temps : $\text{Decay}^t(\mathbf{y}) = t - \text{Date}^t(\mathbf{y})$. Ainsi, le poids donné par $\alpha_t^t(\mathbf{y}) = (1 - \text{Forget}_t^t)^{\text{Decay}^t(\mathbf{y})}$, est d'autant plus faible que $\text{Date}^t(\mathbf{y})$ est loin de l'itération courante t . Dans ce contexte et également dans le suivant, plus le paramètre Forget_t^t est proche de 1, plus le modèle aura tendance à oublier les annotations passées. Le cas extrême $\text{Forget}_t^t = 1$ correspond, par conséquent, à un processus qui n'a pas de mémoire.
- Si des éléments sont sélectionnés à l'itération t ($\mathbb{S}_t^t \neq \emptyset$), ceux-ci sont nécessairement dans \mathbb{T}_t^+ mais l'itération lors de laquelle ils ont été annotés peut varier entre 1 et t . En effet, $\forall \mathbf{y} \in \mathbb{T}_t^+ : \text{Date}^t(\mathbf{y}) \in \{1, \dots, t\}$. L'exposant $\text{Decay}^t(\mathbf{y})$ dans la 2ème condition de (1.5) est alors déterminé en tenant compte des valeurs $\text{Date}^t(\mathbf{z})$ pour tout $\mathbf{z} \in \mathbb{S}_t^t$ ayant été annoté après \mathbf{y} . Ces éléments sont ceux du sous-ensemble $\mathbb{D}_t^t(\mathbf{y}) = \{\mathbf{z} \in \mathbb{S}_t^t : \text{Date}^t(\mathbf{z}) \geq \text{Date}^t(\mathbf{y})\}$. Dans ce cas, on définit $\text{Decay}^t(\mathbf{y})$ par $\min_{\mathbf{z} \in \mathbb{D}_t^t(\mathbf{y})} (\text{Date}^t(\mathbf{z}) - \text{Date}^t(\mathbf{y}))$ afin de retranscrire un contexte local entre d'une part, les éléments annotés et d'autre part, les éléments sélectionnés.
- Enfin, si aucun élément $\mathbf{z} \in \mathbb{S}_t^t$ est tel que son annotation ait été faite après celle de \mathbf{y} ($\forall \mathbf{z} \in \mathbb{S}_t^t : \text{Date}^t(\mathbf{z}) < \text{Date}^t(\mathbf{y})$), alors $\mathbb{D}_t^t(\mathbf{y}) = \emptyset$ et on prend dans ce cas $\alpha_t^t(\mathbf{y}) = 0$. Il s'agit typiquement, d'une situation où l'utilisateur.rice décide de revenir sur des résultats annotés antérieurement, à une itération $t' < t$. Il.elle sélectionne alors un sous-ensemble d'éléments de $\mathbb{T}_{t'}^+$ ce qui indique un changement de direction dans la recherche à partir de cette itération t' . Dans ce cas nous supposons que les annotations données au cours des itérations allant de $t' + 1$ à $t - 1$ doivent être oubliées.

Ce système d'*information seeking* a fait l'objet d'un brevet [Ah-Pine et al., 2012] et d'une publication [Ah-Pine et al., 2009].

1.2.3 Marche aléatoire sur des graphes et unification de méthodes de fusion transmodale

Suite aux différents succès empiriques recueillis par l'approche des **similarités transmédi**as en **CBMIR** aux campagnes d'évaluation internationales ImageCLEF, nous avons entrepris une **analyse plus théorique** de ces techniques avec Gabriela Csurka et Stéphane Clinchant de XRCE. Nous avons rapproché ces méthodes, d'approches à base de graphe qui sont populaires en recherche d'information depuis l'avènement de l'algorithme *PageRank* de Google.

Nous avons publié dans [Ah-Pine et al., 2015] un modèle qui **unifie conceptuellement deux approches développées en recherche d'information multimédia**, l'une correspondant aux **similarités transmodales** dans le cas des bases d'objets image-texte, et l'autre fondée sur les **marches aléatoires** et appliquées dans le domaine de la vidéo. L'approche englobe également les travaux que nous avons publiés dans [Clinchant et al., 2011] en CBMIR.

Notre cadre suppose que les objets multimédias d'une base ainsi que les requêtes multimédias, sont des noeuds d'un **multigraphe** et qu'ils partagent donc entre eux plusieurs types

d'arêtes valuées correspondant à autant de médias qui les composent. Dans notre contexte, les arêtes matérialisent des relations de similarité qui peuvent être initialement de deux types : similarités images et similarités textes. Des méthodes d'analyse de graphe peuvent alors être mobilisées afin d'inférer des mesures permettant d'affiner des relations d'ordre entre noeuds. En particulier, en supposant que les valuations d'un graphe de similarités sont non négatives, on peut normer chaque ligne de sa matrice d'adjacence de sorte à obtenir des distributions de probabilités et par conséquent une matrice stochastique. Nous pouvons interpréter cette dernière comme la matrice des probabilités de transition d'un noeud à un autre et simuler des marches aléatoires sur les noeuds du graphe, ce qui revient à une chaîne de Markov. En particulier si on simule indéfiniment la **marche aléatoire sur le graphe**, ceci permet de calculer une distribution stationnaire sur l'ensemble des noeuds. La probabilité stationnaire ou d'équilibre d'un noeud indique la proportion du temps que le processus de Markov devrait passer sur ce noeud à long terme. En pratique, plus cette probabilité d'occupation est grande plus le noeud est important. Par conséquent, la distribution stationnaire indique une **mesure de centralité qui infère une relation d'ordre reflétant l'importance des noeuds du graphe**. Cette propriété est au coeur de l'algorithme *PageRank* de Google que je rappelle formellement dans ce qui suit après avoir introduit les notations suivantes :

- \mathbf{e}_n est le vecteur de taille $n \times 1$ rempli de 1.
- \mathbf{P} est une matrice de taille $n \times n$, de valeurs non négatives et stochastique : $\mathbf{P}\mathbf{e}_n = \mathbf{e}_n$.
- $\boldsymbol{\pi}_t$ est un vecteur de taille $n \times 1$, de valeurs non négatives et stochastique : $\boldsymbol{\pi}_t^\top \mathbf{e}_n = 1$. Il est calculé à l'itération t .
- \mathbf{v} est un vecteur de taille $n \times 1$, de valeurs non négatives et stochastique : $\mathbf{v}^\top \mathbf{e}_n = 1$. C'est une distribution de probabilités *a priori* qui permet de biaiser la mesure de centralité en fonction, par exemple, des préférences spécifiques d'un utilisateur.rice.

Le score *PageRank* correspond en fait à la distribution stationnaire de la matrice stochastique dite "matrice de Google", $(1 - \mu)\mathbf{P} + \mu\mathbf{e}_n\mathbf{v}^\top$ avec $\mu \in [0, 1]$:

$$\begin{aligned}\boldsymbol{\pi}_\infty^\top &= \boldsymbol{\pi}_\infty^\top((1 - \mu)\mathbf{P} + \mu\mathbf{e}_n\mathbf{v}^\top) \\ &= (1 - \mu)\boldsymbol{\pi}_\infty^\top\mathbf{P} + \mu\mathbf{v}^\top.\end{aligned}\tag{1.7}$$

Afin d'estimer $\boldsymbol{\pi}_\infty$, la *power method* est employée (voir par exemple [Langville and Meyer, 2005]). On prend pour vecteur initial $\boldsymbol{\pi}_0 = \mathbf{e}_n/n$, et on itère la formule suivante jusqu'à convergence :

$$\boldsymbol{\pi}_{t+1}^\top = \boldsymbol{\pi}_t^\top((1 - \mu)\mathbf{P} + \mu\mathbf{e}_n\mathbf{v}^\top).\tag{1.8}$$

Il est important de préciser que la matrice $((1 - \mu)\mathbf{P} + \mu\mathbf{e}_n\mathbf{v}^\top)$ et le vecteur initial $\boldsymbol{\pi}_0$ étant stochastiques, les vecteurs $\{\boldsymbol{\pi}_t\}_{t \geq 1}$ sont également stochastiques et il n'est alors pas nécessaire de normer ces derniers après chaque multiplication, contrairement au cadre général de l'algorithme des puissances itérées.

Dans [Hsu et al., 2007], ce principe est adapté dans le cas de la recherche de vidéos à

partir d'une requête texte \mathbf{q}^t . Dans ce cas, les n vidéos de la base sont vues comme des objets multimédias images-texte et on suppose des matrices (de similarités) stochastiques \mathbf{S}^v (similarités visuelles entre vidéos) et \mathbf{S}^t de taille $n \times n$. Le cas asymétrique où une requête constituée uniquement de texte est utilisée pour interroger la base, est considéré. Supposons alors un moteur de recherche texte donné. Je dénote par $\mathbf{v}^t = (\text{Sim}^t(\mathbf{q}^t, \mathbf{x}_i))_{i=1, \dots, n}$, le vecteur de taille $n \times 1$ des scores de pertinence monomédia des parties textes des objets multimédias de la base, $\{\mathbf{x}_i^t\}_{i=1, \dots, n}$ vis-à-vis de la requête texte \mathbf{q}^t . Supposons que \mathbf{v}^t contienne uniquement des valeurs non négatives et que la somme de ses éléments fasse 1 (distribution de probabilités).

Soit également $\boldsymbol{\pi}_\infty^t$ le vecteur de taille $n \times 1$ des scores de pertinence multimédia des objets de la base étant donnée la requête texte. Afin de ne pas alourdir les notations, je ne spécifie pas la requête texte \mathbf{q}^t dans les notations des deux variables $\boldsymbol{\pi}_\infty^t$ et \mathbf{v}^t . Le vecteur $\boldsymbol{\pi}_\infty^t$ dénote la méthode de CBMIR proposée dans [Hsu et al., 2007]. Celui-ci est en fait un point fixe défini par la relation suivante :

$$[\boldsymbol{\pi}_\infty^t]^\top = [\boldsymbol{\pi}_\infty^t]^\top \left((1 - \mu)[(1 - \gamma)\mathbf{S}^v + \gamma\mathbf{S}^t] + \mu\mathbf{e}_n[\mathbf{v}^t]^\top \right), \quad (1.9)$$

où $\gamma, \mu \in [0, 1]$ sont des paramètres des mélanges suivants :

- γ contrôle la combinaison convexe des matrices \mathbf{S}^v et \mathbf{S}^t représentée en **vert**,
- μ contrôle *in fine* l'arbitrage entre d'une part, la mesure de centralité multimédia basée sur $(1 - \gamma)\mathbf{S}^v + \gamma\mathbf{S}^t$, et d'autre part le *prior* donné par la similarité monomédia texte entre les objets de la base et la requête texte représenté par la couleur **orange**.

Il est important de mentionner que la matrice $(1 - \mu)[(1 - \gamma)\mathbf{S}^v + \gamma\mathbf{S}^t] + \mu\mathbf{e}_n[\mathbf{v}^t]^\top$ est stochastique, sous les hypothèses précédentes. Sa plus grande valeur propre et celle de sa transposée est donc 1 et le score de pertinence multimédia $\boldsymbol{\pi}_\infty^t$ peut alors être interprété comme le vecteur propre dominant de $((1 - \mu)[(1 - \gamma)\mathbf{S}^v + \gamma\mathbf{S}^t] + \mu\mathbf{e}_n[\mathbf{v}^t]^\top)^\top$.

Comme pour le cas classique du *PageRank*, en pratique, $\boldsymbol{\pi}_\infty^t$ est estimée par la **power method**. On initialise $\boldsymbol{\pi}_0^t$ par une distribution uniforme et on réitère la formule suivante jusqu'à convergence :

$$[\boldsymbol{\pi}_{t+1}^t]^\top = [\boldsymbol{\pi}_t^t]^\top \left((1 - \mu)[(1 - \gamma)\mathbf{S}^v + \gamma\mathbf{S}^t] + \mu\mathbf{e}_n[\mathbf{v}^t]^\top \right). \quad (1.10)$$

La méthode des puissances itérées est un ingrédient essentiel à notre cadre unificateur comme je le montrerai par la suite.

Avant cela, afin d'avoir une approche globale de CBMIR, nous généralisons le cadre applicatif de [Hsu et al., 2007] en considérant le cas dual où on interroge la base de vidéo par une requête image \mathbf{q}^v . En adaptant les définitions et notations précédentes, nous définissons le vecteur de scores de pertinence multimédia $\boldsymbol{\pi}_\infty^v$ qui est caractérisé par l'équation et la

procédure itérative qui suivent :

$$[\boldsymbol{\pi}_\infty^v]^\top = [\boldsymbol{\pi}_\infty^v]^\top \left((1 - \mu)[(1 - \gamma)\mathbf{S}^t + \gamma\mathbf{S}^v] + \mu\mathbf{e}_n[\mathbf{v}^v]^\top \right), \quad (1.11)$$

$$[\boldsymbol{\pi}_{t+1}^v]^\top = [\boldsymbol{\pi}_t^v]^\top \left((1 - \mu)[(1 - \gamma)\mathbf{S}^t + \gamma\mathbf{S}^v] + \mu\mathbf{e}_n[\mathbf{v}^v]^\top \right), \quad (1.12)$$

avec $\boldsymbol{\pi}_0^v = \mathbf{e}_n/n$ comme vecteur initial pour la formule de récurrence (1.12). Ici, $\boldsymbol{\pi}_\infty^v$ est biaisé par le score de pertinence monomédia image \mathbf{v}^v .

Suite à cet exposé, j'introduis ci-dessous les formules de récurrence au coeur de notre approche et j'explique par la suite en quoi elles généralisent à la fois les méthodes précédentes (1.10) et (1.12), et les similarités transmédias (1.1) et (1.2).

$$[\boldsymbol{\pi}_{t+1}^{tv}]^\top = \text{Knn}^k([\boldsymbol{\pi}_t^{tv}]^\top) \left((1 - \mu^t)[(1 - \gamma^t)\mathbf{S}^v + \gamma^t\mathbf{S}^t] + \mu^t\mathbf{e}_n[\mathbf{v}^t]^\top \right), \quad (1.13)$$

$$[\boldsymbol{\pi}_{t+1}^{vt}]^\top = \text{Knn}^k([\boldsymbol{\pi}_t^{vt}]^\top) \left((1 - \mu^v)[(1 - \gamma^v)\mathbf{S}^t + \gamma^v\mathbf{S}^v] + \mu^v\mathbf{e}_n[\mathbf{v}^v]^\top \right). \quad (1.14)$$

Les équations (1.10) et (1.12) sont obtenues à partir de (1.13) et (1.14) respectivement, à condition que :

- $\boldsymbol{\pi}_0^{tv} = \boldsymbol{\pi}_0^{vt} = \mathbf{e}_n/n$ (initialisation par des distributions uniformes).
- $k = n$ (pas de diffusion *via* les plus proches voisins uniquement).

Dans le cas des similarités transmodales, données par (1.1) et (1.2), celles-ci correspondent aux cas particuliers donnés par les paramétrages suivants :

- $\mu^t = \mu^v = 0$ (pas de *prior* dans la marche aléatoire sur le graphe).
- $\gamma^t = \gamma^v = 0$ (pas de mélange de matrices de similarité monomédia).
- $[\boldsymbol{\pi}_0^{tv}]^\top = [\mathbf{v}^t]^\top = \text{Sim}^t(\mathbf{q}, \cdot)$ et $[\boldsymbol{\pi}_0^{vt}]^\top = [\mathbf{v}^v]^\top = \text{Sim}^v(\mathbf{q}, \cdot)$ (initialisation donnée par les scores de pertinence monomédia -supposés non négatifs-).
- $k < n$ (propagation *via* les plus proches voisins uniquement).
- $\boldsymbol{\pi}_1^{tv}$ et $\boldsymbol{\pi}_1^{vt}$ (une seule itération de la marche aléatoire).

Il est important de mentionner que lorsque $k < n$ et/ou $\boldsymbol{\pi}_0^{tv}$ et $\boldsymbol{\pi}_0^{vt}$ ne somment pas à un, les vecteurs $\boldsymbol{\pi}_t^{tv}$ et $\boldsymbol{\pi}_t^{vt}$ n'ont pas de garantie théorique de converger. Toutefois, en pratique, nous avons **testé les similarités transmodales dans le cadre d'un processus de diffusion généralisé** et celles-ci sont stables après quelques itérations.

Par ailleurs, je n'ai pas abordé la question de la complexité de ces approches basées sur des graphes. Étant donné que la solution recherchée est le vecteur propre dominant, cette recherche a une complexité de base en $O(n^3)$. La méthode des puissances itérées permet de réduire cette complexité et obtenir un vecteur approché après $q < n$ itérations conduisant à une complexité en $O(qn^2)$. Néanmoins, le goulot d'étranglement persiste en le nombre n de noeuds dans le graphe. Avec Gabriela et Stéphane, nous avons montré du point de vue empirique, qu'en CBMIR, pour une requête donnée comprenant du texte, il était possible à la fois de réduire la complexité et d'augmenter la pertinence des résultats.

1.2. CONTRIBUTIONS

Pour cela, nous avons introduit le concept de **filtrage sémantique** qui sélectionne dans un premier temps un sous-ensemble d'objets multimédias en utilisant la similarité texte uniquement. Supposons un modèle de pertinence monomédia texte et soit $\text{Sim}^t(\mathbf{q}, \cdot)$ le vecteur des scores de pertinence de taille $1 \times n$ des éléments \mathbf{x}_i^t de la base vis-à-vis de \mathbf{q}^t , la partie texte de la requête. Le concept de *semantic filtering* que nous avons introduit dans [Clinchant et al., 2011], consiste à appliquer les similarités transmédias uniquement à la top liste de taille $l < n$ associée à $\text{Sim}^t(\mathbf{q}, \cdot)$. Dans cette perspective, il est nécessaire de normaliser les différents sous-vecteurs et sous-matrices d'ordre l afin qu'ils soient de nature stochastique. Après filtrage sémantique, la complexité des équations (1.13) et (1.14) est ainsi réduite à $O(l^2)$. La Figure 1.6 permet d'illustrer le *workflow* du système décrit précédemment.

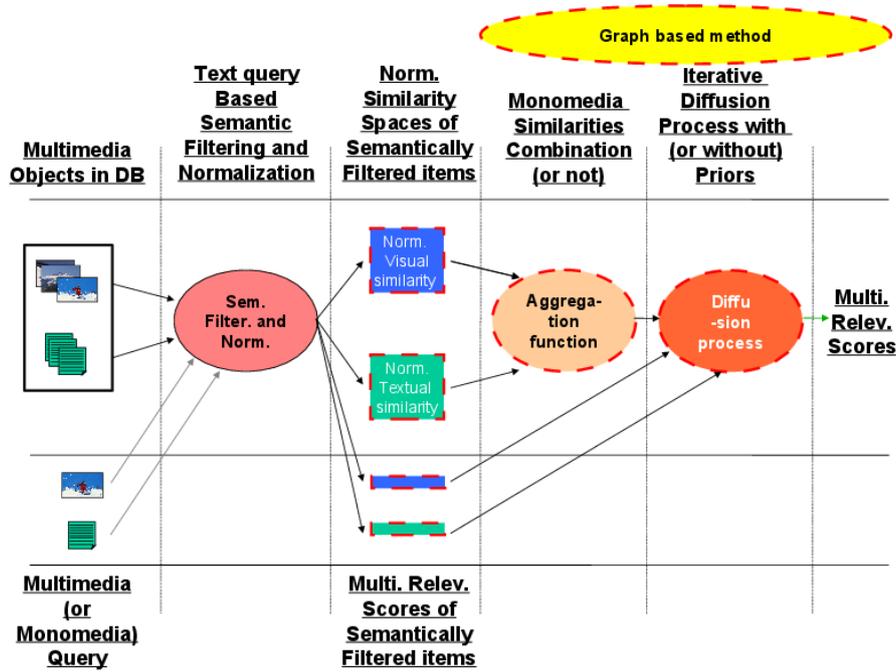


FIGURE 1.6 – *Workflow* du modèle général de fusion intermédiaire d'information multimédia basé sur le filtrage sémantique et la diffusion dans des graphes.

De plus, les liens entre notre modèle d'une part, et les **chaînes de Markov** et la **méthode des puissances itérées** d'autre part, nous permettent d'analyser différents concepts associés aux paramétrages des équations (1.13) et (1.14). En particulier, nous pouvons comparer les idées fondamentales des marches aléatoires sur des graphes définies dans [Hsu et al., 2007] et celles des similarités transmédias que nous avons développées à XRCE. Plusieurs expériences effectuées à partir de quatre jeux de données réels ont été réalisés dans [Ah-Pine et al., 2015]. Je donne ci-dessous l'essentiel des observations faites suite à ce travail empirique et les recommandations de paramétrage des équations (1.13) et (1.14) que nous en avons déduites :

- L'initialisation devrait être basée sur les similarités textes et images : $\pi_0^{tv} = \mathbf{v}^t$ et

$$\boldsymbol{\pi}_0^{\text{vt}} = \mathbf{v}^{\text{v}}.$$

- Les marches courtes donnent des résultats meilleurs que les marches longues : $\boldsymbol{\pi}_t^{\text{tv}}$ et $\boldsymbol{\pi}_t^{\text{vt}}$ sont plus pertinents pour $t = 1$ ou $t = 2$ que pour $t > 2$.
- La propagation transmodale *via* les plus proches voisins donne de meilleurs résultats que celle utilisant tous les voisins : $k = 10$ est un paramètre par défaut conduisant souvent aux meilleurs résultats.
- En CBMIR, la modalité texte est particulièrement importante du point de vue sémantique et ajouter \mathbf{v}^{t} comme vecteur *prior* dans (1.13) donne de meilleurs résultats en moyenne ($\mu^{\text{t}} = 0.3$ est une valeur de paramètre par défaut à considérer).
- Mélanger les matrices de similarités monomédias \mathbf{S}^{t} et \mathbf{S}^{v} n'apporte pas d'amélioration et nous recommandons de prendre $\gamma^{\text{t}} = \gamma^{\text{v}} = 0$.

Les techniques de fusion multimédia intermédiaire fondées sur (1.13) et (1.14), permettent de réduire le fossé sémantique en CBMIR de façon complémentaire vis-à-vis des approches classiques de fusions précoce et tardive. Les nombreuses expériences que nous avons menées, nous ont conduit à définir par défaut le modèle de pertinence général suivant :

$$\text{Rel}^{\text{inter}}(Q, \cdot) = \theta^{\text{t}} \text{Sim}^{\text{t}}(\mathbf{q}, \cdot) + \theta^{\text{v}} \text{Sim}^{\text{v}}(\mathbf{q}, \cdot) + \theta^{\text{tv}} [\boldsymbol{\pi}_t^{\text{tv}}]^{\top} + \theta^{\text{vt}} [\boldsymbol{\pi}_t^{\text{vt}}]^{\top}, \quad (1.15)$$

où les vecteurs de scores de pertinence transmédias $\boldsymbol{\pi}_t^{\text{vt}}$ et $\boldsymbol{\pi}_t^{\text{tv}}$ sont définis avec les paramètres décrits précédemment et où les coefficients de la combinaison convexe sont égaux par défaut.

Cette suggestion de poids uniforme entre les différentes composantes monomédia et transmédia laisse penser que image et texte sont d'importance symétrique en CBMIR. Or, ce n'est pas le cas, car il ne faut pas oublier le *semantic filtering* qui est appliqué en amont de la procédure ci-dessus, comme cela est indiqué dans la Figure 1.6. Dans le cas de la recherche d'images basée sur le contenu, **la modalité texte reste une vue particulièrement riche au niveau sémantique** qu'il convient donc d'exploiter en priorité et de façon appropriée. Nous avons en particulier montré dans [Clinchant et al., 2011], la prédominance du texte en CBMIR.

1.3 Discussions et perspectives

A l'ère de la *data science*, du *big data* et de l'*internet of things*, l'étude d'un phénomène qu'il soit issu des sciences physiques, naturelles, numériques ou humaines et sociales, peut être menée de façon riche au travers des données que l'on est capable aujourd'hui de récolter de façon massive et parfois quasi-continue. Or ces données sont mixtes, hétérogènes, multi-modales, multi-échelles, ... Dans ce contexte, les **méthodes de fusion d'informations** sont particulièrement pertinentes afin d'établir des approches permettant de tirer profit des **synergies entre les différentes sources d'information** disponibles pour la modélisation et la prédiction du phénomène à l'étude.

Dans le cas de données images, les **modèles de *Convolutional Neural Networks***

(CNN) appris à partir de la base ImageNet³ ont permis des avancées spectaculaires depuis les années 2010 pour résoudre des tâches de catégorisation et/ou de reconnaissance d’objets en *computer vision*...

Puis, l’utilisation d’**images massives “faiblement annotées”** issues de réseaux sociaux comme Flickr [Joulin et al., 2016] ou du web, ont permis de franchir de nouveaux caps dans le contexte plus spécifique des données image-texte. Les modèles de *deep learning* associés à ces données massives dans un paradigme d’**auto-supervision**, ont permis de résoudre de façon spectaculaire le problème de **fossé sémantique** que j’ai évoqué précédemment. Je cite en particulier le travail incontournable des chercheurs d’OpenAI publié dans [Radford et al., 2021] où l’**approche neuronale CLIP**⁴ (*Contrastive Language-Image Pre-training*) est présentée. Le modèle exploite une base intitulée *WebImageText* (WIT) qui est composée de 400 millions d’image-texte provenant du web.

L’architecture de CLIP permet d’apprendre à associer information visuelle et information textuelle de façon puissante. Ce type de modèle permet également de résoudre avec beaucoup d’efficacité des tâches de **zero-shot image classification**⁵. Cette très grande capacité de “généricité” montre la robustesse de l’apprentissage effectué par des modèles de *deep learning*.

Le modèle pré-entraîné CLIP permet d’avoir deux espaces de représentations : l’un pour le texte et l’autre pour l’image. Ces espaces sont de même dimension et sont appris de façon jointe de sorte à ce que la représentation d’une image et celle d’un texte soient proches lorsque la paire (image, texte) est dans la base de données et éloignées sinon. Cette **forme d’auto-supervision est appelée *contrastive learning***. Ces espaces de représentation de CLIP permettent ainsi de combler le fossé sémantique entre image et texte. Par ailleurs, la couverture de ce modèle est large puisqu’il provient de l’apprentissage de 400 millions de cas. On parle alors de ***Large Vision and Language models (LVLM)***.

Dans le cas de la fusion image-texte en CBMIR, on peut naturellement utiliser CLIP afin de projeter un texte et une image dans ses espaces de représentation sémantiquement riches (*embeddings*) et employer ces vecteurs obtenus pour mesurer la pertinence entre une requête mono ou multimédia et les éléments d’une collection image-texte.

Toutefois, dans le cas d’une collection ou d’une tâche spécifique, il est souvent nécessaire d’avoir une étape d’apprentissage en aval dans le but de spécifier les sorties d’un modèle pré-entraîné de façon générique à un cas plus singulier et afin de gagner en précision. On parle d’**apprentissage par transfert (*transfer learning*)**. Une première solution dans ce cas consiste en le ***fine-tuning du LVLM*** où on poursuit l’apprentissage des paramètres de ce dernier par rétro-propagation de l’erreur mais en utilisant uniquement les données de la tâche spécifique. Néanmoins, cette approche est très coûteuse en temps de traitement et en

3. Base d’images incontournable en *computer vision* qui est composée de plus de 14 millions d’images ayant été annotées manuellement : <https://www.image-net.org/>.

4. CLIP est d’ailleurs au cœur de l’application grand public DALL-E qui étend ce dernier à la génération d’images à partir de *prompts*.

5. En effet, en donnant une description textuelle d’une nouvelle catégorie, CLIP est capable de catégoriser des images appartenant à celle-ci sans n’avoir eu aucun exemple d’apprentissage au préalable.

dépense d'énergie. CLIP, par exemple, est composé de plusieurs centaines de millions de paramètres et mettre à jour ces derniers nécessitent d'importantes ressources computationnelles et financières. Une deuxième solution, plus raisonnable, consiste à utiliser le **modèle pré-entraîné comme extracteur de *features*** uniquement, et d'ajouter une deuxième phase qui consiste en un apprentissage supervisé à partir de ces *features* et à l'aide d'un *learner* annexe et approprié.

Dans ce contexte, les méthodes de fusion décrites précédemment pourraient jouer un rôle intéressant. Dans le cas de la fusion basée sur la **propagation transmodale** en particulier, il s'agirait de former des matrices de similarités textes et images sur la base de la représentation vectorielle donnée par un LVLM puis, de diffuser des similarités d'un mode vers un autre comme indiqué dans les équations (1.13) et (1.14).

Les méthodes *contrastive learning* telles que CLIP représentent une approche neuronale parmi d'autres en matière de fusion multimodale. En effet, il existe d'autres alternatives dans le champs de la **fusion d'informations multimodales à l'aide de réseaux de neurones** comme l'attestent les articles de *survey* suivants [Summaira et al., 2021, Sleeman IV et al., 2022, Manzoor et al., 2023].

Dans le contexte du *deep learning*, une autre question qui me paraît intéressante à étudier et qui serait une extension des travaux que j'ai exposés plus haut, concerne l'**apprentissage supervisé des opérateurs de diffusion transmodale**. En effet, l'ensemble des techniques exposées dans ce Chapitre sont de nature non supervisée ou active impliquant une boucle de pertinence avec l'intervention d'un humain. Supposons désormais, que nous disposions à l'avance de données annotées où, pour un ensemble de requêtes multimédias, nous ayons parmi la collection de n objets image-texte ceux qui sont pertinents. Il est alors possible de définir un **réseau de neurones dont le mécanisme est proche dans l'esprit des applications de propagation transmodale**.

Pour illustrer ce propos, je prends le cas de la propagation de similarités textes vers des similarités images comme décrit par (1.1). Dans cette approche non supervisée, la même application Knn^k de sélection de plus proches voisins est utilisée quelque soit la requête ou l'objet texte. En particulier, l'hyperparamètre k est fixé et ne dépend pas de l'*input* texte. Dans la mesure où l'approche cross-modale est fondée sur les plus proches voisins, il me semble particulièrement intéressant d'employer ici, les **Radial Basis Function Neural Networks (RBF NN)**. L'architecture que je propose est composite. Elle est composée en entrée et en sortie de n RBF NN en parallèle et entre ces deux couches spécifiques, je propose un NN de type *Multi-Layer Perceptron* (MLP). Les n RBF NN en parallèle en entrée visent à évaluer un profil de similarité texte, le MLP du milieu a pour but d'apprendre un opérateur de sélection qui soit plus flexible que l'application Knn^k et les n RBF NN en parallèle de la couche de sortie donnent un score de pertinence transmédia.

L'approche revient à une **architecture neuronale de learning to rank avec fusion transmodale**. Celle-ci est illustrée dans la Figure 1.7.

Je considère une requête texte représentée par un vecteur \mathbf{q}^t obtenu par un LVLM (ou

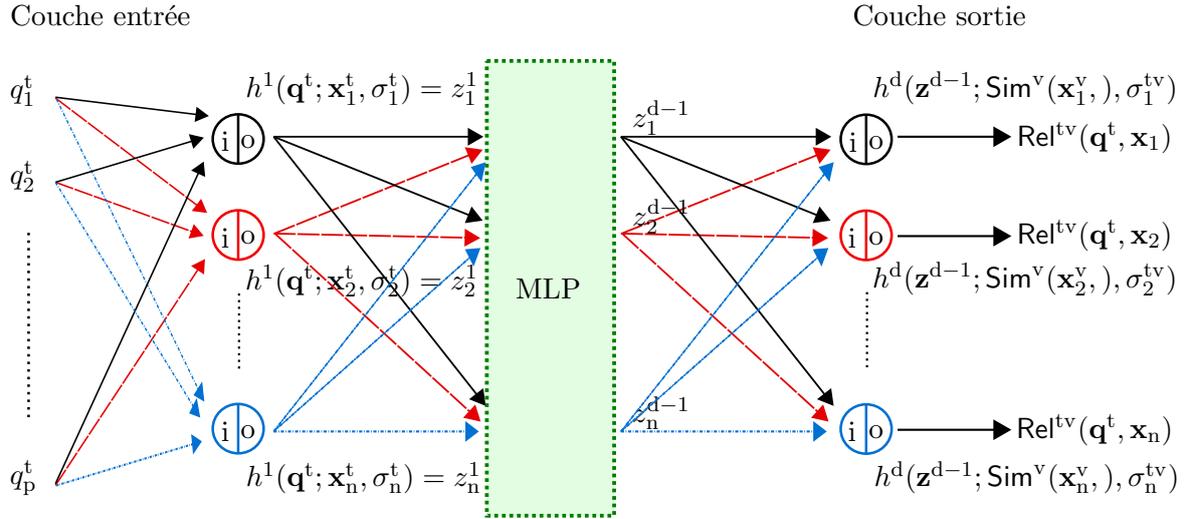


FIGURE 1.7 – Une architecture neuronale de fusion transmodale pour la tâche *learning to rank* utilisant des RBF NN en parallèle en entrée et en sortie et un MLP intermédiaire.

toute autre méthode d'extraction de *features* textes). Ce vecteur constitue la couche d'entrée du modèle proposé. Il est transmis à une 1ère couche cachée composée de n unités RBF. Chaque fonction d'activation RBF de cette couche cachée est centrée en le vecteur de la représentation texte de chaque élément de la base qui pourrait également être donnée par le LVLMM. Notons ces vecteurs de taille p par $\mathbf{x}_1^t, \dots, \mathbf{x}_n^t$.

La sortie de la 1ère couche cachée notée \mathbf{z}^1 , peut être vue comme l'estimation d'un profil de similarités textes de \mathbf{q}^t avec chaque vecteur texte de la base. Pour fixer les idées, supposons que la RBF est le noyau Gaussien. Nous aurons alors, $\forall i = 1, \dots, n$:

$$\begin{aligned} z_i^1 &= h^1(\mathbf{q}^t; \mathbf{x}_i^t, \sigma_i^t) \\ &= \exp\left(-\frac{\|\mathbf{x}_i^t - \mathbf{q}^t\|^2}{(\sigma_i^t)^2}\right) \\ &= \text{Sim}^t(\mathbf{q}^t, \mathbf{x}_i^t). \end{aligned}$$

\mathbf{z}^1 est de taille n . Il est ensuite donné en entrée d'un *Multi-Layer Perceptron* (MLP) dont le nombre d'arguments en entrée et en sortie est n , le nombre d'objets dans la base. Ce MLP vise à représenter un opérateur de sélection qui serait plus flexible que l'opérateur Knn^k utilisé par défaut dans (1.2).

La sortie de la dernière couche du MLP est notée \mathbf{z}^{d-1} . Il s'agit d'un vecteur de taille n résultant de transformations non linéaires des similarités textes \mathbf{z}^1 . Il comporte donc des informations provenant du mode texte uniquement. \mathbf{z}^{d-1} est transmis à une dernière couche cachée qui, comme la 1ère couche cachée, est composée de n unités RBF. Chaque fonction d'activation RBF est centrée en un profil de similarités images d'un élément \mathbf{x}_i^v . Notons ces

vecteurs de taille n par $\text{Sim}^v(\mathbf{x}_1^v), \dots, \text{Sim}^v(\mathbf{x}_n^v)$. Si nous supposons à nouveau que la fonction RBF h^d est le noyau Gaussien, nous aurons alors, $\forall i = 1, \dots, n$:

$$\begin{aligned} h^d(\mathbf{z}^{d-1}; \text{Sim}^v(\mathbf{x}_i^v), \sigma_i^{tv}) &= \exp\left(-\frac{\|\mathbf{z}^{d-1} - \text{Sim}^v(\mathbf{x}_i^v)\|^2}{(\sigma_i^{tv})^2}\right) \\ &= \text{Sim}^{tv}(\mathbf{q}^t, \mathbf{x}_i) \\ &= \text{Rel}^{tv}(\mathbf{q}^t, \mathbf{x}_i). \end{aligned}$$

Le vecteur de valeurs en sortie du NN, $(\text{Rel}^{tv}(\mathbf{q}^t, \mathbf{x}_i))_{i=1, \dots, n}$, infère un ordre sur les n objets de la collection ce qui permet de les *ranker*.

Ce **modèle neuronal généralise l'équation** (1.1). En effet, cette dernière formule peut être retrouvée à partir du modèle précédent en fixant de façon déterministe :

- des fonctions de similarité linéaire à la place des RBF pour les n unités de la 1ère couche cachée, $\forall i = 1, \dots, n : z_i^1 = \langle \mathbf{q}^t, \mathbf{x}_i \rangle$;
- une fonction d'activation ReLU avec seuil θ , à chaque sortie de la 1ère couche cachée et à la place du MLP⁶, qui permet de retenir uniquement les similarités avec les plus proches voisins, $\forall i = 1, \dots, n : z_i^2 = \text{ReLU}(z_i^1, \theta_k) = z_i^1$ si $z_i^1 > \theta_k$ et 0 sinon ;
- des fonctions de similarité linéaire à la place de RBF dans la couche de sortie, $\forall i = 1, \dots, n : \text{Rel}^{tv}(\mathbf{q}^t, \mathbf{x}_i) = \langle \mathbf{z}^2, \text{Sim}^v(\mathbf{x}_i^v) \rangle$.

Par conséquent, ma proposition englobe l'approche classique cross-média⁷ et l'étend suivant différentes directions dans un contexte d'apprentissage supervisé. Par ailleurs, l'approche peut bien sûr être adaptée de façon symétrique pour la propagation de l'information image vers l'information texte. Il est également possible de définir une architecture neuronale plus complète permettant de reproduire le modèle de pertinence donné par (1.15) et qui exploite plusieurs techniques de fusion.

6. Du point de vue d'un réseau de neurones, cela reviendrait à considérer n perceptrons en parallèle et mutuellement indépendants. Chaque perceptron prendrait une seule valeur en entrée, aurait un coefficient synaptique unitaire et utiliserait comme fonction d'activation une ReLU avec seuil θ .

7. Toutefois, je considère ici une sélection des plus proches voisins selon un seuil réel θ à la place d'une sélection basée sur un entier k .

Fusion d'informations linguistiques et statistiques en TALN

Sommaire du chapitre

2.1 Introduction	29
2.1.1 Contexte	29
2.1.2 Travaux antérieurs	33
2.2 Contributions	36
2.2.1 Méthodes de fusion et enrichissement d'hypergraphes linguistiques	36
2.2.2 <i>Clique Based Clustering</i> et désambiguïsation d'entités nommées	42
2.3 Discussions et perspectives	47

2.1 Introduction

2.1.1 Contexte

Les recherches que je présente dans ce deuxième Chapitre couvrent deux périodes de mon expérience professionnelle et portent sur des **contributions en Traitement Automatique du Langage Naturel (TALN -*Natural Language Processing* (NLP)-)**.

En premier lieu, au cours de mes activités à XRCE entre 2007 et 2010, j'ai, en plus de mes travaux au sein de l'équipe *Textual and Visual Pattern Analysis* (TVPA), collaboré étroitement avec Guillaume Jacquet de l'équipe *Parsing and Semantics* (ParSem) dans le cadre du TALN. Nous avons travaillé sur des **approches hybrides** et en particulier pour la **tâche de reconnaissance et de désambiguïsation d'entités nommées (EN) (*Named Entity Recognition* -NER-)**. Il s'agit de la catégorisation des mentions de personne (PERS), de lieux (LOC) ou d'organisation (ORG) (et, si besoin, d'autres types plus spécifiques) dans un texte. A cette époque, les domaines principaux de Guillaume étaient les sciences cognitives et la linguistique computationnelle. Il s'intéressait en particulier à la tâche de NER qu'il abordait de façon originale à partir de deux concepts clefs : d'une part, les similarités entre termes à partir de leur représentation distributionnelle issue des dépendances syntaxiques et d'autre part, la représentation graphique de ces relations de similarité et la

détermination de sens spécifiques de groupes de termes par recherche de cliques maximales. Dans ce contexte, Guillaume était confronté à deux problèmes : d'un côté, la surproduction de cliques maximales d'EN et d'un autre côté, l'exploitation de ces cliques d'EN pour la tâche de désambiguïsation. J'ai contribué à solutionner ces deux problèmes d'une part par mes compétences en *clustering* pour réduire le nombre de cliques et d'autre part, par mes connaissances en techniques de fusion et sélection d'informations.

Ce travail s'est effectué dans le cadre du projet européen Sync3 qui était composé de huit partenaires. L'objectif général était de créer une plateforme pour analyser et structurer autour de la notion d'évènements, les textes provenant de sites de dépêches d'une part, et de blogs sur l'actualité d'autre part. Cette plateforme avait pour but de faciliter la collaboration pour la création de contenus. J'ai contribué à plusieurs titres au projet Sync3 : j'ai participé à son montage, sa rédaction et sur l'aspect scientifique, j'ai étudié la tâche de détection d'évènements dans des flux de dépêches à l'aide de *clustering*. Je n'ai malheureusement pas pu valoriser le travail effectué dans la portée de cette tâche en terme de publication, car j'ai quitté XRCE avant la fin du projet Sync3.

L'introduction de ce Chapitre me permet par ailleurs, d'évoquer un point d'inflexion dans ma carrière. J'ai commencé dans l'industrie avec une thèse CIFRE à Thales et un postdoctorat de trois ans à *Xerox Research Centre Europe*. En sus de plusieurs publications scientifiques, j'ai également produit 5 brevets. A partir de 2009, j'ai souhaité rejoindre la recherche publique et contribuer avant tout à la connaissance ouverte. Ce changement était aussi motivé par un besoin de plus de liberté dans le choix et l'exploitation de mes travaux de recherche. C'est ainsi que j'ai passé avec succès le concours de MCF section CNU 26 que l'Université Lumières Lyon 2 (UL2) et le laboratoire ERIC avait ouvert pour la rentrée de l'année universitaire 2010-2011.

A mon arrivée au laboratoire ERIC, une thématique émergente que l'équipe de direction souhaitait développer était les **humanités numériques**. Il s'agit, en effet, d'un champ propice au contexte local puisque l'UL2 est un établissement principalement en sciences humaines et sociales. Afin de contribuer à cet axe de recherche, j'ai collaboré dans un premier temps avec Djamel Zighed alors membre du laboratoire ERIC et directeur de l'Institut des Sciences de l'Homme (ISH) de Lyon (maintenant Maison des Sciences de l'Homme de Lyon), qui souhaitait entreprendre une **activité autour du recensement, de l'analyse et de la communication des compétences en SHS développées au sein du PRES (Pôle de Recherche et d'Enseignement Supérieur) de Lyon/Saint-Etienne**. L'année même de mon recrutement, j'ai coordonné la rédaction et la soumission du projet intitulé SHS DocNet pour un financement interne BQR de l'UL2. Ce projet a été financé sur 2011 et 2012 et a regroupé les partenaires suivants : ISH, Laboratoire ERIC, ICAR (Interactions, Corpus, Apprentissages, Représentations) de l'UL2 et le LHC (Laboratoire Hubert Curien) de l'Université de Saint-Etienne. Ce projet nous a permis de produire un *proof of concept* que l'on a souhaité développer davantage en soumettant une proposition de projet au programme CONTINT 2011 de l'ANR. Djamel et moi avons approché des PME dans le but d'établir un consortium. Celui-ci a été composé des partenaires suivants : ISH, Laboratoire ERIC, LINA

2.1. INTRODUCTION

(Laboratoire d'Informatique de Nantes), Clever Age (PME de Lyon), Armadillo (PME de Paris). J'ai été le **principal contributeur de la rédaction de la proposition du projet** que j'avais intitulé RéSoCo pour **Réseau Social et de Compétences**. Je donne dans la Figure 2.1 un diagramme indiquant les principales briques technologiques du système que nous souhaitions mettre en oeuvre.

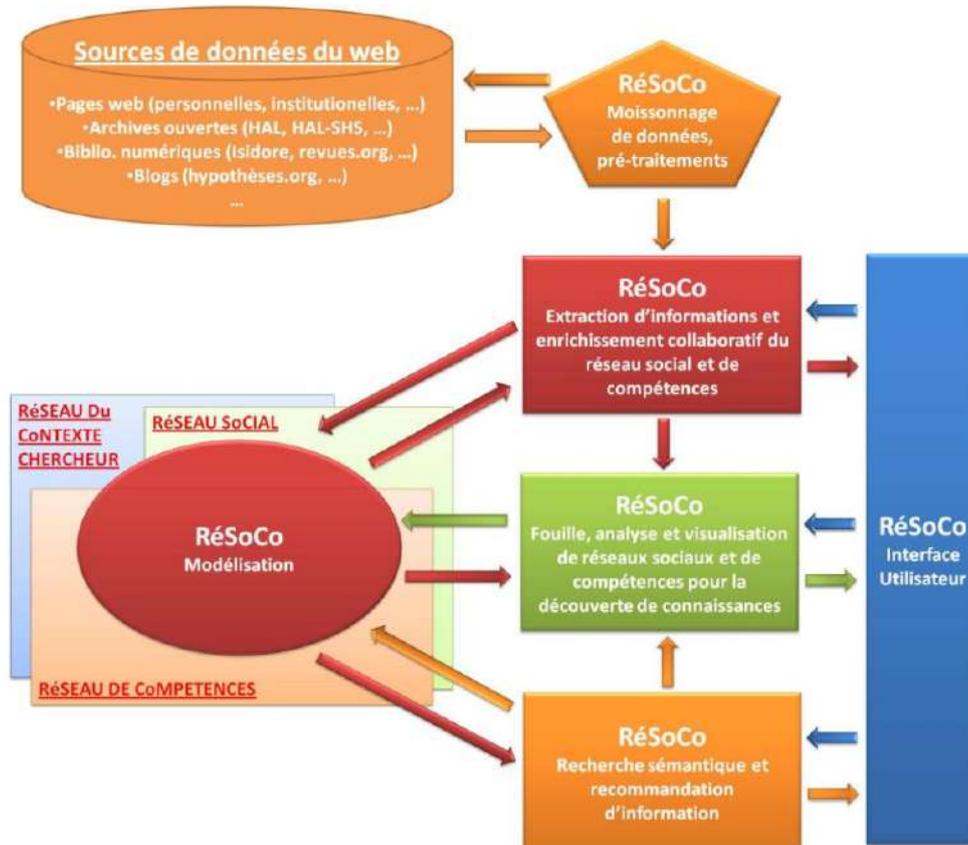


FIGURE 2.1 – Architecture et briques technologiques de la proposition RéSoCo.

Malheureusement la proposition RéSoCo n'a pas été sélectionnée par le comité de sélection de l'ANR. Des améliorations de nature scientifique nous avaient été demandées. Celles-ci étaient surmontables. C'est en fait un fort déséquilibre dans les demandes de financement des différents partenaires qui nous a lourdement pénalisé. Djamel et moi n'avons pas réussi à résoudre ce problème en amont de la soumission. Pour ma part, j'ai fini par me retirer de cette activité malgré un très important investissement.

Laissant derrière moi cette expérience, j'ai voulu rebondir et continuer à contribuer au **développement des humanités numériques au sein du laboratoire ERIC**. A la fin de l'année universitaire 2012-2013 j'ai proposé un sujet de stage de niveau Master 2 sur des **approches hybrides (linguistiques et statistiques) pour résoudre des tâches en TALN**. J'ai alors encadré Pavel Soriano Morales, qui était à ce moment étudiant au sein

du Master Erasmus Mundus DMKM (*Data Mining and Knowledge Management*) de l'UL2 dans lequel j'intervenais. Suite au stage de Pavel, j'ai proposé à Sabine Loudcher qui était à ce moment là MCF HDR au Laboratoire ERIC, si elle souhaitait qu'on co-encadre Pavel dans le cadre d'un projet de doctorat. Sabine a accepté et nous avons établi le sujet de thèse suivant, "*Weakly supervised and unsupervised information extraction from text by leveraging open sources of information*", en incluant initialement des aspects en systèmes d'information et en apprentissage automatique. Nous avons pu décrocher une bourse MENRT qui débuta à la rentrée de l'année universitaire 2013-2014.

Dans ce travail collaboratif, j'ai souhaité orienter le travail de Pavel vers des activités de recherche que j'avais menées à XRCE. J'ai proposé d'**exploiter les différentes méthodes de fusion précoce (*early*), tardive (*late*) et transmodale exposées dans le Chapitre 1 mais en TALN**. Les challenges que je voulais aborder étaient le développement de méthodes hybrides d'une part, et le **problème de parcimonie des données en linguistique quantitative** d'autre part.

Concernant les approches hybrides, l'idée centrale consistait à utiliser de façon jointe les méthodes à base de règles (connaissances linguistiques expertes) et les approches statistiques (modèles d'apprentissage automatique), afin d'avoir une représentation riche de l'information linguistique et dans le but de résoudre des tâches de TALN plus efficacement. A cette époque, les approches à base de règles avaient des taux de précision (*precision rate*) bien meilleurs que ceux obtenus par les méthodes de *machine learning*. En revanche, l'inverse était observé en ce qui concernait les taux de rappel (*recall rate*). Il y avait donc un intérêt à combiner les deux points de vue et de tenter d'exploiter les atouts de l'un pour combler les lacunes de l'autre.

En 2013, Les approches d'apprentissage auto-supervisé comme les *masked language models* n'existaient pas ou peu et les données annotées n'étaient pas suffisamment volumineuses pour bon nombre de tâches afin d'inférer des modèles du niveau des *large language models* (LLM). De ce fait, les matrices terme-document qui indiquent les associations entre les unités lexicales et les documents d'un corpus étaient très creuses. Ce problème de *data sparsity* impactait négativement l'apprentissage de modèles de *machine learning*. Dans ce contexte, il m'avait semblé opportun d'exploiter de façon flexible les différentes techniques de fusion afin de pallier le problème de manque de données. Nous avons donc travaillé sur ces questions dans le cadre de la thèse de Pavel¹ de 2013 à 2017.

Les publications avec comité de lecture qui sont concernées par ce Chapitre sont les suivantes :

- E.P. Soriano Morales, **J. Ah-Pine**, S. Loudcher. 2017. Fusion Techniques for Named Entity Recognition and Word Sense Induction and Disambiguation. *Proceedings of the 20th International Conference on Discovery Science (DS 2017)*. [Lien vers la conférence,

1. Thèse soutenue en janvier 2017, Pavel est actuellement responsable *data science* à la Haute Autorité de Santé

<http://www.iip.ist.i.kyoto-u.ac.jp/ds2017/index.html#1>].

- E.P. Soriano-Morales, **J. Ah-Pine**, S. Loudcher. 2016. Using a Heterogeneous Linguistic Network for Word Sense Induction and Disambiguation. *Computación y Sistemas*. 20(3) : 315-325. [Lien vers le journal, <http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2466>].
- E.P. Soriano Morales, **J. Ah-Pine**, S. Loudcher. 2016. Hypergraph Modelization of a Syntactically Annotated Wikipedia Dump. *Proceedings of the 10th Edition of Language Resources and Evaluation Conference (LREC 2016)*. [Lien vers la conférence, <http://lrec2016.lrec-conf.org/>].
- **J. Ah-Pine**, G. Jacquet. 2009. Clique based *clustering* for improving Named Entity Recognition systems. *Proceedings of the European chapter of the Association of Computational Linguistics (EACL 2009)* [Taux d’acceptation < 28%]. [Consultable à l’adresse suivante, www.aclweb.org/anthology/E09-1007].

2.1.2 Travaux antérieurs

Le concept sous-jacent sur lequel s’appuie le travail que je décris dans ce Chapitre et qui est central en TALN est celui de l’**hypothèse distributionnelle** que l’on attribue à Zellig S. Harris [Harris, 1954]. Ce concept est simple mais puissant : **les mots qui apparaissent dans les mêmes contextes linguistiques partagent une signification proche**. Ceci est exprimée par John R. Firth [Firth, 1957] par la formule suivante : “*You shall know a word by the company it keeps*”. Par conséquent, la signification d’un mot peut être déterminée par l’ensemble des contextes dans lesquels celui se trouve. J’illustre ce propos par un exemple tiré de [Nida, 1979] dans la Table 2.1 et qui montre le mot “tesgüino” dans différents contextes.

*A bottle of **tesgüino** is on the table.*
*Everybody likes **tesgüino**.*
***Tesgüino** makes you drunk.*
*We make **tesgüino** out of corn.*

TABLE 2.1 – Mot “tesgüino” dans quatre phrases distinctes.

On infère facilement des exemples d’occurrence donnés dans la Table 2.1, que le mot “tesgüino” est une boisson alcoolisée.

Ce qui va m’intéresser dans un premier temps c’est **la notion de contexte qui peut revêtir plusieurs formes**. On peut en effet, faire la distinction entre le **contexte lexical** et le **contexte syntaxique**. Dans le premier cas, il s’agit de considérer la cooccurrence de deux unités lexicales (deux mots typiquement) dans un voisinage prédéfini. C’est cette notion de contexte qui est la plus répandue en TALN dans le cas des techniques de *machine learning*. Le contexte syntaxique est quant à lui basé sur une analyse linguistique (*parsing*) du texte qui permet d’obtenir des relations existantes entre les mots et unités lexicales (*chunks*) d’une phrase. Dans cette catégorie on peut faire la distinction entre l’analyse syntaxique de surface

2.1. INTRODUCTION

(*shallow parsing, constituency tree*) et l'analyse syntaxique des grammaires de dépendance (*dependency tree*). Je ne rentrerai pas dans les détails mais procéderai par l'exemple afin de donner une idée de la différence entre ces trois points.

J'emploie pour cela une illustration qui a été utilisée par Pavel dans sa thèse [Soriano-Morales, 2018] et lors de sa soutenance et qui concerne la phrase, “*The report contains copies of the minutes of these meetings*”.

Dans la Figure 2.2, on considère un voisinage de taille 5. L'ensemble des mots adjacents au mot “*contains*” qui sont marqués par une flèche se trouvent donc dans le même contexte lexical que ce dernier.

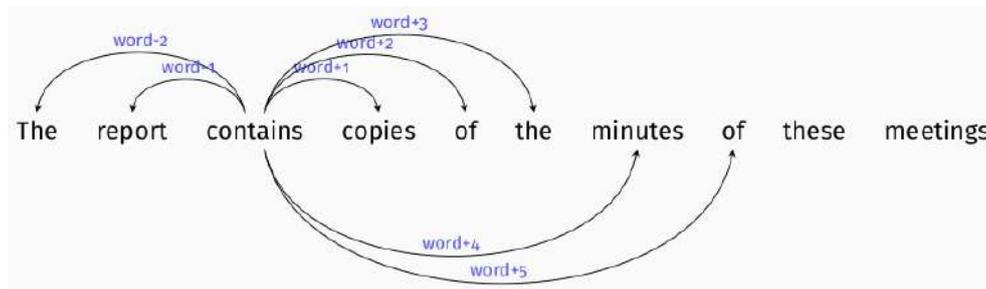


FIGURE 2.2 – Illustration de contexte lexical du mot “*contains*”.

L'analyse syntaxique de la phrase permet de produire les étiquettes grammaticales (nom -*Noun*-, verbe -*Verb*-, déterminant -*Determinant*-, ...) et les syntagmes ou groupes de mots (nominal -*Noun Phrase*-, verbal -*Verb Phrase*-, ...) constituant la phrase (*Sentence*). L'arbre des constituants de l'exemple est illustré dans la Figure 2.3.

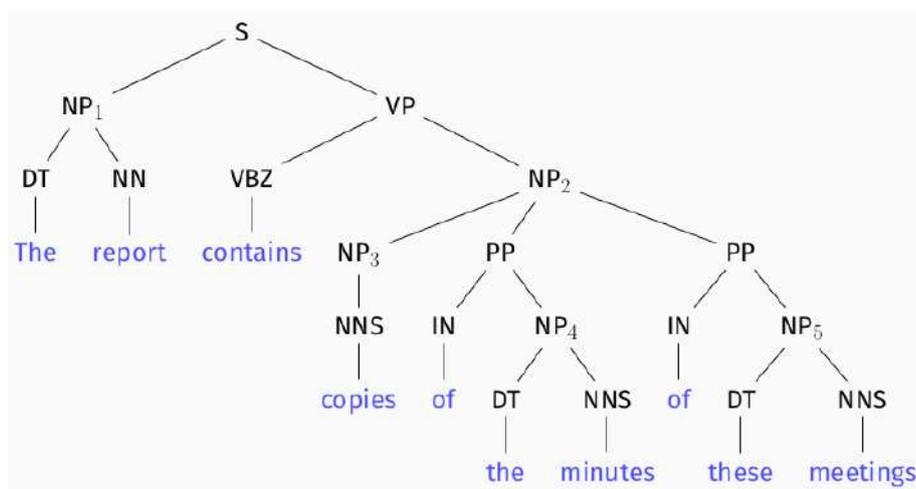


FIGURE 2.3 – Illustration d'un arbre syntaxique simple indiquant les catégories grammaticales et donnant lieu à un premier type de contexte syntaxique.

Enfin, l'arbre des dépendances permet d'établir des relations entre mots qui décrivent des

fonctions syntaxiques spécifiques entre eux. Dans ce contexte, la relation est orientée : tout mot dans un syntagme est légitimé par la présence d’un autre mot que l’on appelle gouverneur (*head* de la relation). Dans l’exemple précédent, le mot “*contains*” est le gouverneur du mot “*report*” pour la fonction syntaxique *nsubj*, c’est à dire, “nom sujet”.

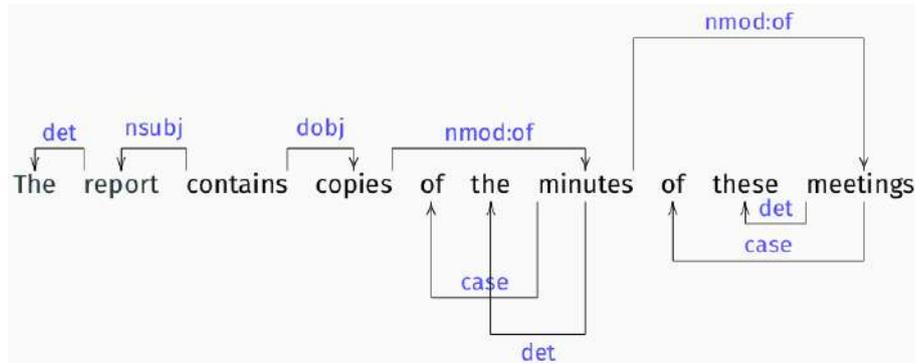


FIGURE 2.4 – Illustration d’un arbre syntaxique indiquant les grammaires de dépendance et donnant lieu à un deuxième type de contexte syntaxique.

Considérer les contextes syntaxiques en sus des contextes lexicales s’inscrit dans une **démarche hybride**. Les **cooccurrences lexicales sont une vue purement statistique de l’hypothèse distributionnelle** et ne font intervenir aucune connaissance linguistique. Or, les **dépendances syntaxiques** que l’on peut extraire à partir de systèmes à base de règles construites par des linguistes, nous **procurent une vue complémentaire de l’hypothèse distributionnelle**.

Deux aspects sont alors importants d’exposer à ce stade. Dans un premier temps, il s’agit de déterminer une **représentation formelle permettant d’encoder l’ensemble de ces informations diverses** de façon cohérente et compatible avec des méthodes d’apprentissage. Dans le cas classique des contextes lexicaux, on utilise la représentation vectorielle des mots que l’on regroupe dans une matrice dite terme-document où on observe une mesure non nulle lorsque le terme en ligne apparaît au moins une fois dans le document en colonne. Dans le paragraphe 2.2.1, j’expose le modèle que nous avons développé afin de représenter de façon unifiée les contextes lexicaux et syntaxiques. Celui-ci repose sur les **hypergraphes**.

Dans un second temps, peu importe le type de contexte que l’on exploite, **les données observées en TALN sont en générale creuses ou parcimonieuse (*sparse*)**. En effet, il est bien connu que la fréquence des mots dans un corpus suit une loi de puissance (loi de Zipf) : beaucoup de termes apparaissent très peu de fois dans des textes. Dans le cas des cooccurrence lexicales, la matrice terme-document est ainsi remplie de beaucoup de zéros. Ceci rend peu efficace la mise en pratique de l’hypothèse distributionnelle. Pour tenir compte de cet inconvénient, les méthodes classiques de lissage de modèles de langue comme celle de Jelinek-Mercer [Jelinek and Mercer, 1980] consiste à mélanger $P(W|D)$, la probabilité d’occurrence d’un terme W dans un document D , avec $P(W|C)$, la probabilité d’observer ce terme dans

tout le corpus \mathbb{C} (ou un corpus de référence) :

$$P^{\text{JM}}(W|D) = (1 - \lambda)P(W|D) + \lambda P(W|\mathbb{C}),$$

où $\lambda \in [0, 1]$ est le paramètre de mélange.

Dans ce contexte, et dans la continuité du Chapitre précédent, je montre en premier lieu dans la sous-section 2.2.1, en quoi **les méthodes de fusion que j’ai présentées dans la section précédente, peuvent être employées afin de pallier au problème de *data sparsity***.

Ensuite, dans le paragraphe 2.2.2, je reviens sur une autre contribution en TALN réalisée à XRCE en collaboration avec Guillaume Jacquet. Dans ce travail, nous nous intéressons tout particulièrement aux **entités nommées (EN) et à leur désambiguïsation (ou annotation fine)**. Par exemple, le terme “*Oxford*” peut faire référence à une ville, une université ou une équipe de sport. Afin de déterminer la bonne annotation, nous analysons les contextes syntaxiques. Je présente notre système, ***Clique Based Clustering***, qui crée de façon non supervisée une ressource permettant d’améliorer les performances de systèmes de reconnaissance d’EN existants.

2.2 Contributions

2.2.1 Méthodes de fusion et enrichissement d’hypergraphes linguistiques

Afin de représenter les différents types d’informations que nous avons rappelés précédemment, nous avons opté pour les **hypergraphes**. Dans ce cadre, l’ensemble des noeuds est celui des termes. Je précise toutefois, qu’en fouille de textes, il est pertinent de ne pas tenir compte dans l’ensemble des termes, ceux qui ne portent pas de sens, comme les déterminants ou les prépositions qui forment du bruit. Quant à l’ensemble des **hyperarêtes**, celui-ci n’est pas limité à des paires de noeuds mais est constitué de sous-ensembles de termes de cardinalités variables. Ceci nous permet donc d’**encoder des contextes linguistiques faisant intervenir un nombre arbitraire de termes**. L’ensemble des hyperarêtes que nous considérons est organisé selon plusieurs catégories. Celles-ci sont relatives au type de contexte que nous employons dans notre modèle.

A l’issue des différentes analyses linguistiques, nous extrayons **trois types de contextes** que nous dénotons par SEN, NP, et DEP pour “*SENtence*”, “*Noun Phrase*” et “*DEPendency*”. J’illustre ci-dessous ces trois catégories avec l’exemple précédent “*The report contains copies of the minutes of these meetings*” :

- SEN :
 - $S_1 = \{\text{report, contains, copies, minutes, meetings}\}$;
- NP :
 - $NP_1 = \{\text{report}\}$,
 - $NP_2 = \{\text{copies, minutes, meetings}\}$,

2.2. CONTRIBUTIONS

- $NP_3 = \{\text{minutes}\}$;
- DEP :
 - $nsubj_{contains} = \{\text{report}\}$,
 - $dobj_{contains} = \{\text{copies}\}$.

Chaque élément listé est une potentielle hyperarête de notre hypergraphe. J'illustre graphiquement ce propos avec un focus sur le terme “copies”, dans la Figure 2.5 .

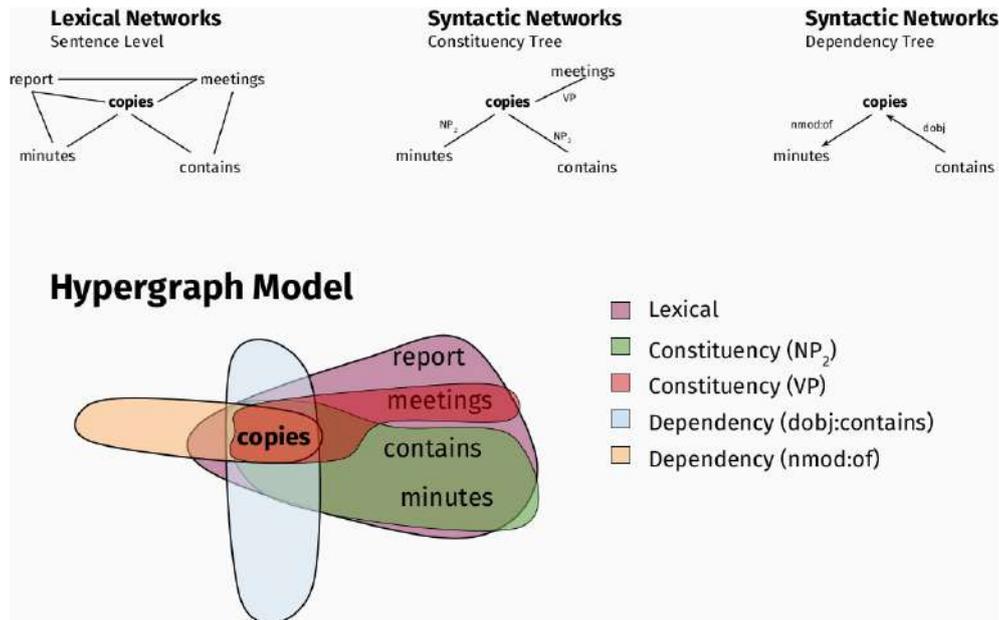


FIGURE 2.5 – Illustration de la représentation par hypergraphe des différents contextes.

L’hypergraphe montré dans la Figure 2.5 est une représentation graphique. Pour encoder le graphe en machine, nous avons recours aux **matrices d’incidence**. Dans ce cas, chaque ligne est associée à un noeud, c’est à dire à un terme, et chaque colonne correspond à une hyperarête. Les sommets appartenant à une hyperarête sont marqués par la valeur 1 dans la colonne correspondante et les autres par la valeur 0. Afin d’exploiter les informations de nature linguistique issue de l’analyse syntaxique, chaque ligne et chaque colonne de la matrice d’incidence a des métadonnées. Pour les lignes, la catégorie grammaticale (*Part-Of-Speech - POS- tagging*) du terme est utilisée. En ce qui concerne les colonnes, le type de contexte parmi les trois mentionnées ci-dessus, ainsi que la nature spécifique de la dépendance syntaxique, sont employées. J’illustre ce propos dans la Figure 2.6 tirée de [Soriano-Morales, 2018], à l’aide de l’exemple précédent.

Étant donné un corpus, les différentes analyses linguistiques nous permettent d’obtenir une matrice d’incidence que je dénote par \mathbf{X} et qui encode une information riche dans le but de **capturer les relations de proximité sémantiques entre termes sous l’hypothèse distributionnelle**. Toutefois, cette richesse de la représentation des mots vient avec un prix,

2.2. CONTRIBUTIONS

		NOUN PHRASE			DEPENDENCY		SENTENCE
		NP ₁ DT:NN	NP ₂ NP:PP:PP	NP ₃ NNS	nsubj contains	dobj contains	S ₁
NN	report	1			1		1
	copies		1	1		1	1
	minutes		1				1
	meetings		1				1
VB	contains						1

FIGURE 2.6 – Matrice d’incidence de l’hypergraphe avec divers types d’hyperarêtes selon les contextes.

celui de l’**augmentation de la dimensionnalité du problème** et de l’**aggravation du problème de *data sparsity***. En effet, les colonnes que l’on ajoute dans \mathbf{X} sont parcimonieuses.

Dans ce contexte, nous avons cherché à **densifier ces matrices sparses par des opérations matricielles similaires aux fusions tardives et transmodales** que j’ai présentées précédemment dans le contexte CBMIR. Plus précisément, notre approche vise à modéliser de façon “algébrique” des opérations sur des modèles de langues instanciés par des matrices d’incidences, à l’aide des mécanismes similaires aux fusion précoces, tardives et transmodales. Cette modélisation nous permet de “raisonner” de manière riche quant aux différentes sources d’information à notre disposition. Elle permet de combiner ces dernières de sorte à obtenir une matrice encodant judicieusement des informations sur la proximité sémantique entre termes en vue de résoudre des tâches d’apprentissage supervisé en TALN.

Dénotons par \mathbf{X}^l la matrice d’incidence de l’hypergraphe dont les hyperarêtes concernent uniquement des dépendances lexicales. Soit \mathbf{X}^s la matrice d’incidence dont les colonnes sont relatives aux dépendances syntaxiques de type NP ou DEP. Les matrices \mathbf{X}^l et \mathbf{X}^s sont rectangulaires, elles ont le même nombre de lignes n car elles représentent le même ensemble de termes mais elles possèdent deux nombres de colonnes distincts.

De plus, je suppose que nous pouvons définir des matrices de similarités \mathbf{S}^l et \mathbf{S}^s , qui nous indiquent des relations de proximité entre mots à partir des contextes lexicaux et syntaxiques respectivement. Pour fixer les idées, nous pouvons considérer $\mathbf{S}^l = \mathbf{X}^l[\mathbf{X}^l]^\top$ et $\mathbf{S}^s = \mathbf{X}^s[\mathbf{X}^s]^\top$, bien que d’autres possibilités existent. \mathbf{S}^l et \mathbf{S}^s sont ainsi carrées d’ordre n , et peuvent être vues comme des **matrices d’adjacence pondérées dont l’ensemble des sommets est celui des termes**.

A partir de ces données primaires, j’introduis ci-après les **applications matricielles E, L, P qui sont respectivement associées aux fusions *early*, *late* et par propagation transmodale ou monomodale**. Je considère des notations génériques où \mathbf{M}^u est une matrice de taille $n \times q$ qui peut être soit une matrice d’adjacence pondérée (ou de similarités) \mathbf{S}^u , soit une matrice d’incidence (ou de *features*) \mathbf{X}^u . Les indices u ci-dessus, et les indices u_1 et u_2 qui seront utilisés par la suite, sont des éléments de $\{l, s\}$, et indiquent la source de l’information qui, dans notre cas présent, peut être soit lexicale (l) soit syntaxique (s) (mais d’autres types

pourraient être employés).

La fusion précoce peut ici être formalisée comme suit :

$$\begin{aligned} E : \mathbb{R}^{n \times q_{u_1}} \times \mathbb{R}^{n \times q_{u_2}} &\rightarrow \mathbb{R}^{n \times (q_{u_1} + q_{u_2})} \\ (\mathbf{M}^{u_1}, \mathbf{M}^{u_2}) &\rightarrow E(\mathbf{M}^{u_1}, \mathbf{M}^{u_2}) = \text{Hstack}(\mathbf{M}^{u_1}, \mathbf{M}^{u_2}), \end{aligned} \quad (2.1)$$

où l'application Hstack consiste à concaténer deux matrices ayant le même nombre de lignes en les mettant côte à côte.

La fusion tardive correspond alors à l'opération suivante :

$$\begin{aligned} L : \mathbb{R}^{n \times q} \times \mathbb{R}^{n \times q} &\rightarrow \mathbb{R}^{n \times q} \\ (\mathbf{M}^{u_1}, \mathbf{M}^{u_2}) &\rightarrow L(\mathbf{M}^{u_1}, \mathbf{M}^{u_2}) = (1 - \gamma)\mathbf{M}^{u_1} + \gamma\mathbf{M}^{u_2}, \end{aligned} \quad (2.2)$$

où $\gamma \in [0, 1]$ est un paramètre de mélange.

Enfin, la fusion par propagation est introduite par l'équation ci-dessous :

$$\begin{aligned} P : \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times q} &\rightarrow \mathbb{R}^{n \times q} \\ (\mathbf{S}^{u_1}, \mathbf{M}^{u_2}) &\rightarrow P(\mathbf{S}^{u_1}, \mathbf{M}^{u_2}) = \text{Knn}^k(\mathbf{S}^{u_1})\mathbf{M}^{u_2}, \end{aligned} \quad (2.3)$$

où, par abus de notation, j'étends l'application Knn^k défini sur un vecteur page 12, à une matrice entière si bien que $\text{Knn}^k(\mathbf{S}^u)$ retourne une matrice de taille identique à \mathbf{S}^u mais pour chaque ligne, seules les k ($k < n$) plus grandes valeurs sont maintenues, alors que toutes les autres sont mises à 0.

Je n'indique pas les paramètres γ ou k afin d'alléger les notations et de présenter le plus simplement possible le propos central.

L'application E permet d'étendre l'espace de description en concaténant les colonnes de deux matrices \mathbf{M}^{u_1} et \mathbf{M}^{u_2} .

L'application L peut être vue telle une opération d'agrégation terme à terme entre deux matrices \mathbf{M}^{u_1} et \mathbf{M}^{u_2} qui doivent être de même taille. L peut contribuer à **pallier au problème de *data sparsity*** dans la mesure où la matrice résultante contient moins de cellules nulles.

En ce qui concerne P , elle permet de propager des informations d'une 1ère matrice d'adjacence pondérée \mathbf{S}^{u_1} , vers une 2ème matrice \mathbf{M}^{u_2} qui peut être soit d'adjacence soit d'incidence. Si \mathbf{M}^{u_2} est une matrice de similarités \mathbf{S}^{u_2} alors P correspond formellement à la propagation transmédia définie en sous-section 1.1.2, conduisant aux similarités transmodales. Le cas où \mathbf{M}^{u_2} est une matrice incidence \mathbf{X}^{u_2} de taille $n \times p$, est donc nouveau ici et consiste à propager des similarités entre termes pour enrichir les occurrences des termes dans des contextes. La diffusion de similarités de nature syntaxique vers des similarités de type lexical est particulièrement intéressante et renvoie clairement vers l'hypothèse distributionnelle : deux termes se retrouvant souvent dans les mêmes contextes syntaxiques, auront tendance à se retrouver dans des contextes lexicaux identiques. Il est également important d'indiquer que P contribue

2.2. CONTRIBUTIONS

à réduire le problème de données creuses puisque les matrices qu'elle donne en sortie sont plus denses que celles données en entrée en deuxième argument : dans le cas d'une matrice d'adjacence, de nouvelles arêtes entre termes apparaissent, tandis que dans le cas d'une matrice d'incidence, les hyperarêtes sont de tailles plus grandes.

Nous proposons donc d'**exploiter ces diverses applications de fusion de manière plus libre et flexible** en comparaison du Chapitre 1. D'une part, les opérations E, L et P peuvent prendre en compte des matrices de caractéristiques/d'incidence (hypergraphe et hyperarêtes de contextes linguistique), en plus des matrices de similarité/d'adjacence (graphe non orienté et arêtes pondérées non négatives). D'autre part, elles peuvent être combinées de différentes façons donnant lieu à des **procédures de fusion composites**.

Dans la Figure 2.7, j'indique la chaîne de traitements définissant notre système. Au-delà des informations lexicales et syntaxiques, il est important, pour cette tâche, d'intégrer un type d'information de base que j'indique par l'indice f pour (*stantard*) *feature* et qui comprend les informations suivantes :

- le mot lui-même,
- s'il commence par une majuscule ou pas,
- les 3 caractères précédant et succédant le mot,
- la catégorie grammaticale (*POS tagging*) du mot.

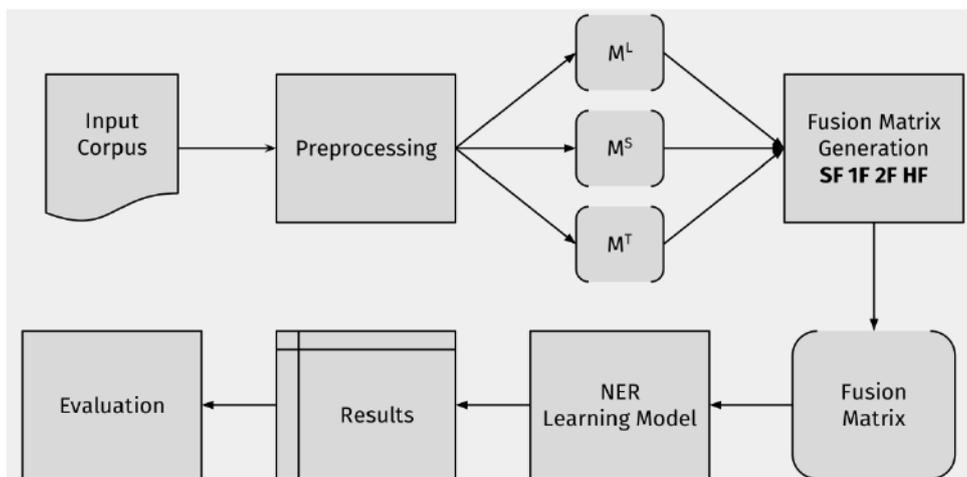


FIGURE 2.7 – Chaîne de traitements pour la tâche de reconnaissance d'entités nommées.

Afin d'illustrer l'intérêt de notre approche, je donne un exemple de fusion d'ordre supérieur à 1 qui a produit des **résultats intéressants pour la tâche de NER**. Nous avons constaté du point de vue empirique, que la matrice issue de la procédure de fusion suivante avait produit des résultats supérieurs aux *baselines* sur les trois *benchmarks* que nous avons utilisés dans

[Soriano-Morales et al., 2017].

$$E(E(\mathbf{X}^f, \underbrace{L(\mathbf{X}^f, \underbrace{P(\mathbf{S}^f, \mathbf{X}^f)})}_{1er\ ordre})}_{2nd\ ordre}), E(\mathbf{X}^l, \underbrace{L(\mathbf{X}^l, \underbrace{P(\mathbf{S}^s, \mathbf{X}^l)})}_{1er\ ordre})}_{2nd\ ordre})). \quad (2.4)$$

3ème ordre
3ème ordre

4ème ordre

Dans la procédure de fusion (2.4), les 1ères opérations $P(\mathbf{S}^f, \mathbf{X}^f)$ et $P(\mathbf{S}^s, \mathbf{X}^l)$ sont des diffusions mono et transmodale respectivement. Les matrices résultantes sont ensuite mélangées avec \mathbf{X}^f et \mathbf{X}^l par les applications de *late* fusion $L(\mathbf{X}^f, P(\mathbf{S}^f, \mathbf{X}^f))$ et $L(\mathbf{X}^l, P(\mathbf{S}^s, \mathbf{X}^l))$. Les ordres 3 et 4 mentionnés dans (2.4) consistent à employer la *early* fusion et de concaténer les matrices d’incidence non altérées \mathbf{X}^f et \mathbf{X}^l à celles obtenues précédemment.

Somme toute, la matrice que nous utilisons pour décrire les termes, est une matrice de *features* étendue, reprenant à la fois les informations initiales \mathbf{X}^f et \mathbf{X}^l et des nouvelles données provenant de la diffusion et de l’agrégation de matrices de types qui diffèrent selon la nature du graphe sous-jacent (matrice d’adjacence \mathbf{S}^u , matrice d’incidence \mathbf{X}^u) et/ou selon l’origine de l’information (*standard feature* f , lexicale l et syntaxique s).

Ci-dessous, dans la Table 2.2 j’indique les performances obtenues par cette approche sur la tâche de reconnaissance d’entités nommées à partir de trois *benchmarks* : CONLL [Sang and De Meulder, 2003], WNER [Sang and De Meulder, 2003] et WGLD [Nothman et al., 2009]. La méthode d’apprentissage supervisée utilisée est le *structured perceptron* [Collins, 2002, Daumé III, 2006], la baseline consiste en la fusion simple $E(\mathbf{X}^f, E(\mathbf{X}^l, \mathbf{X}_s))$ et la mesure de performance employée est la *F – mesure*.

<i>Input</i>	CONLL	WNER	WGLD
\mathbf{X}^f	77.41	77.50	59.66
\mathbf{X}^l	69.40	69.17	52.34
\mathbf{X}^s	32.95	28.47	25.49
$E(\mathbf{X}^f, E(\mathbf{X}^l, \mathbf{X}_s))$	78.90	80.04	63.20
$E(E(\mathbf{X}^f, L(\mathbf{X}^f, P(\mathbf{S}^f, \mathbf{X}^f))), E(\mathbf{X}^l, L(\mathbf{X}^l, P(\mathbf{S}^s, \mathbf{X}^l))))$	79.67	81.79	67.05

TABLE 2.2 – F – mesure sur trois tâches de NER en utilisant plusieurs représentations des mots.

La Table 2.2 montre les résultats de plusieurs systèmes et permet d’illustrer les bonnes performances de notre modèle et l’intérêt des fusions composites, en comparaison de représentations monomodales ou fondées sur des *early fusion* de base.

Nous avons aussi appliqué ces techniques dans le cadre de la **tâche de l’induction et de la désambiguïsation lexicale (WSI/D)**. Dans ce contexte, étant donnée une liste de mots cibles et de contextes, il s’agit de détecter automatiquement les différentes significations possibles des mots ciblés (induction). Ensuite, étant donné un mot et un contexte spécifique, la

désambiguïsation vise à préciser parmi les divers sens possibles détectées, celui qui est le plus approprié. Les méthodes de fusion composite ont produit, ici aussi, des résultats intéressants. Néanmoins, je ne commenterai pas davantage cette tâche et renvoie le lecteur à la thèse de Pavel [Soriano-Morales, 2018] et l'article suivant [Soriano-Morales et al., 2017].

Je reviens sur un point particulier illustrer dans la Table 2.2 : il est intéressant de noter que la matrice d'incidence syntaxique \mathbf{X}^s , donne des résultats décevants lorsqu'elle est utilisée seule. Toutefois, les contextes syntaxiques sont utiles lorsqu'elles sont combinées avec d'autres sources d'information. En effet, dans notre approche de fusion composite, la matrice de similarité \mathbf{S}^s est utilisée dans l'application \mathbf{P} pour enrichir par diffusion la matrice \mathbf{X}^1 . Ainsi, les bonnes performances de cette approche sont en partie dues aux contextes syntaxiques.

Dans le paragraphe suivant, je présente à nouveau un travail en TALN dans lequel les contextes syntaxiques sont également exploités en conjonction avec d'autres outils dans le but d'améliorer la résolution de la tâche NER.

2.2.2 *Clique Based Clustering* et désambiguïsation d'entités nommées

Je décris la recherche que nous avons développée avec Guillaume Jacquet en 2008-2009 lorsque nous étions à XRCE et que nous avons publié dans [Ah-Pine and Jacquet, 2009]. La tâche concerne la **reconnaissance et la désambiguïsation d'entités nommées (EN) dans des corpus spécifiques**. Par exemple, dans les dépêches d'actualité, il arrive fréquemment que de nouvelles EN apparaissent comme cela est le cas pour une affaire criminelle avec des noms de personnes ou d'organisations qui sont mentionnés dans des médias pour la toute première fois. De plus, une EN déjà connue peut posséder plusieurs annotations comme le terme "*Oxford*" qui peut faire référence à l'université, la ville, ... Même au sein d'une même catégorie d'EN, l'ambiguïté est possible. Dans le cas de "*Oxford*" à nouveau, l'annotation ORG (organisation) pourrait concerner l'université mais aussi la société française de papeterie ou également une équipe de sport. Un autre exemple classique dans ce cas, concerne l'annotation PERS (personne) où la mention d'un nom et prénom peut faire référence à des individus distincts en raison des homonymies ...

A cette époque, de nombreux systèmes de reconnaissance et de désambiguïsation d'EN reposaient sur l'utilisation de ressources externes comme Wikipedia [Bunescu and Pasca, 2006, Cucerzan, 2007]. Toutefois, cela était efficient à condition que les EN étaient déjà présentes dans ce référentiel, ce qui n'est pas toujours le cas comme pour les corpus spécifiques susmentionnés. Dans cette situation, nous avons proposé d'utiliser les contextes des dépendances syntaxiques afin d'appréhender les relations de similarité qui sont spécifiques à des EN appartenant à une même annotation, pour faciliter la désambiguïsation. Afin d'illustrer ce propos, je donne dans la Figure 2.8, un exemple de sorties de notre système concernant l'EN "*Oxford*".

Trois groupes sont mentionnés et chacun d'entre eux est constitué d'un ensemble d'EN associés à un ensemble de dépendances syntaxiques. L'EN "*Oxford*" est présente dans cha-

2.2. CONTRIBUTIONS

num clu	more significant NEs	more significant contexts
4	Oxford_NOUN 497	1.be_VERB.AT 77.17
	London_NOUN 291	1.area_NOUN.MOD 63.56
	Liverpool_NOUN 252	1.have_VERB.AT 50.66
	Manchester_NOUN 240	1.move_VERB.TO 48.23
	Newcastle_NOUN 166	1.member_NOUN.FOR 44.76
	Leeds_NOUN 135	1.magistrate_NOUN.MOD 42.19
	Edinburgh_NOUN 131	1.go_VERB.TO 41.91
	Birmingham_NOUN 125	1.live_VERB.IN 41.47
Glasgow_NOUN 123	1.be_VERB.NEAR 41.05	
58	Cambridge_NOUN 26	1.study_VERB.AT 8.76
	Oxford_NOUN 26	1.professor_NOUN.AT 8.25
	London_NOUN 7	1.student_NOUN.AT 7.27
	Edinburgh University_NOUN 6	1.graduate_NOUN.MOD 7.24
	Edinburgh_NOUN 5	1.attend_VERB.AT 6.06
	Oxford University_NOUN 5	1.be_VERB.AT 5.93
	Westminster_NOUN 4	1.degree_NOUN.MOD 5.70
	Glastonbury_NOUN 4	1.teach_VERB.AT 5.62
	Cheltenham_NOUN 4	1.educate_VERB.AT 4.88
95	Wembley_NOUN 11	1.beat_VERB.AT 4.71
	Ibrox_NOUN 10	1.play_VERB.AT 4.51
	Twickenham_NOUN 9	1.final_NOUN.AT 4.27
	Elland_NOUN road_NOUN 6	1.win_VERB.AT 4.13
	Highbury_NOUN 5	1.match_NOUN.AT 4.00
	Oxford_NOUN 5	1.game_NOUN.AT 3.52
	Wimbledon_NOUN 4	1.face_VERB.AT 3.49
	Cheltenham_NOUN 4	1.crowd_NOUN.AT 3.18
	Ascot_NOUN 3	1.the_DET game_NOUN.AT 2.84

FIGURE 2.8 – Chaîne de traitements pour la tâche de reconnaissance d’entités nommées.

cun des groupes. Cependant, il n’est pas difficile de constater, au regard des autres EN et surtout des contextes syntaxiques, que chaque groupe peut être associé à une annotation fine spécifique : le 1er à un lieu administratif (ville d’Oxford), le 2ème à une organisation universitaire (Université d’Oxford), et le 3ème à un lieu de sport (Stade à Oxford).

Les groupes que j’indique dans la Figure 2.8 sont des exemples de ressources structurées que le système que nous avons développé vise à découvrir. Ceci est réalisé par une approche hybride utilisant à la fois des connaissances expertes en linguistique, des méthodes d’apprentissage non supervisées et des techniques d’annotations semi-supervisées. En particulier, au coeur de notre approche, se trouve la **représentation par graphe de proximités entre mots** et le concept de **clique d’EN**. Nous instancions l’hypothèse distributionnelle par le biais des contextes syntaxiques uniquement. Deux EN sont alors d’autant plus proches qu’elles se retrouvent fréquemment dans les mêmes dépendances syntaxiques. Ici, l’ambiguïté d’une EN se traduit par le fait qu’elle peut être fortement liée à deux ou plusieurs autres EN appartenant à des catégories distinctes. Afin de représenter ces appartenances multiples tout en favorisant une certaine robustesse, nous utilisons les **cliques maximales d’EN à partir d’un graphe de similarités**. Une même EN peut alors appartenir à plusieurs cliques ce qui rend possible la **représentation des différents sens d’une même unité lexicale**. Ceci est illustré dans la Figure 2.9 dans le cas de l’EN “London”.

Les étapes clés de notre système appelé CBC pour *Clique Based Clustering* sont :

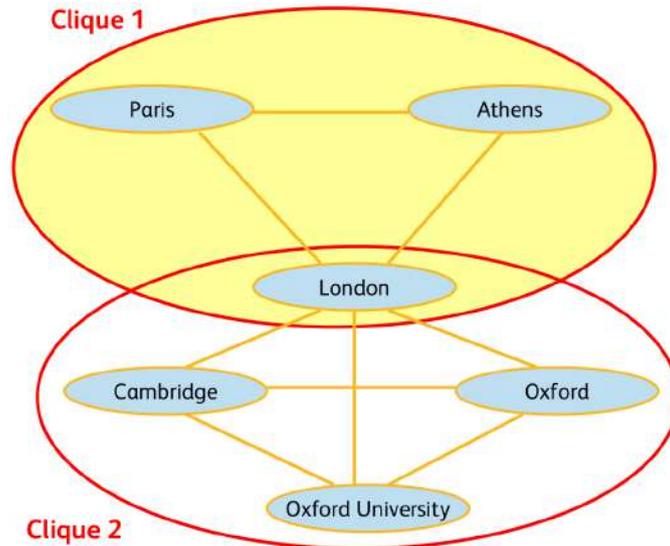


FIGURE 2.9 – Exemple de deux annotations possibles de “London” (LOC -capitale- et ORG -université-) représentée par deux cliques maximales contenant ce terme.

1. Analyse linguistique des textes et extraction des dépendances syntaxiques entre termes (par un système à base de règles ou un modèle d’apprentissage automatique).
2. Détermination automatique d’EN potentielles (sans référentiel) par application de règles lexico-syntaxiques (par exemple : termes commençant par une majuscule ou terme gouverné d’une dépendance syntaxique de type *attribute* dont le mot gouverneur est un nom comme dans *attr(president, Bush), ...*).
3. Représentation des EN potentielles dans un espace engendré par les dépendances syntaxiques (hypothèse distributionnelle) et par un graphe de similarités associé. La mesure de similarité utilisée ici est de nature entropique.
4. Détection des cliques maximales d’EN potentielles à partir du graphe de similarités en utilisant l’algorithme de Bron-Kerbosch [Bron and Kerbosch, 1973].
5. Réduction de l’ensemble des cliques d’EN potentielles par classification automatique en utilisant l’heuristique de l’analyse relationnelle [Marcotorchino and Michaud, 1981, Ah-Pine, 2009] qui permet de ne pas fixer le nombre de *clusters*. La similarité entre deux cliques correspond au nombre d’EN qu’elles ont en commun.
6. Affectation d’un *cluster* de cliques à chaque paire (EN potentielle, dépendance syntaxique) : il s’agit du *cluster* contenant l’EN potentielle pour lequel la dépendance syntaxique donnée est la plus pertinente. Précisons cette mesure de pertinence d’un contexte syntaxique pour un *cluster* de cliques. Soit \mathbb{C} un ensemble de documents (corpus), C_j une dépendance syntaxique, \mathbb{K}_l un *cluster* de cliques d’EN potentielles et E_i une EN potentielle. Soit $\mathbf{X}^s(E_i, C_j)$ le nombre d’occurrence de C_j pour E_i dans tout \mathbb{C} (par exemple le nombre de fois que “*Bush*” apparaît dans le contexte *attr_{president}* dans

le corpus). Dans ce cas, la pertinence d'un contexte syntaxique C_j pour un *cluster* de cliques \mathbb{K}_l se calcule comme suit :

$$G(C_j, \mathbb{K}_l) = \underbrace{\left(\frac{\sum_{E_i \in \mathbb{K}_l} \mathbf{X}^s(E_i, C_j)}{\sum_{E_{i'} \in \mathbb{C}} \mathbf{X}^s(E_{i'}, C_j)} \right)}_{P(C_j | \mathbb{K}_l)} \underbrace{\left(\sum_{E_i \in \mathbb{K}_l} \text{Ind}_{\{\mathbf{X}^s(E_i, C_j) > 0\}} \right)}_{\text{Nb. d'occurrences de } C_j \text{ dans } \mathbb{K}_l}.$$

où $\text{Ind}_{\{A\}}$ est la fonction indicatrice qui renvoie 1 si la proposition A est vraie et 0 sinon. Étant donné une paire (E_i, C_j) c'est à dire (EN potentielle, dépendance syntaxique), on lui affecte un *cluster* de cliques d'EN potentielles, au travers de l'application suivante :

$$\text{Cluster}(E_i, C_j) = \arg \max_{\mathbb{K}_l \in \mathbb{K} : E_i \in \mathbb{K}_l} G(C_j, \mathbb{K}_l),$$

où $\mathbb{K} = \{\mathbb{K}_1, \dots, \mathbb{K}_k\}$ est l'ensemble des *clusters* de cliques.

Ainsi, le *cluster* \mathbb{K}_l retenu est, parmi ceux qui contiennent E_i , celui qui a le score $G(C_j, \mathbb{K}_l)$, le plus grand.

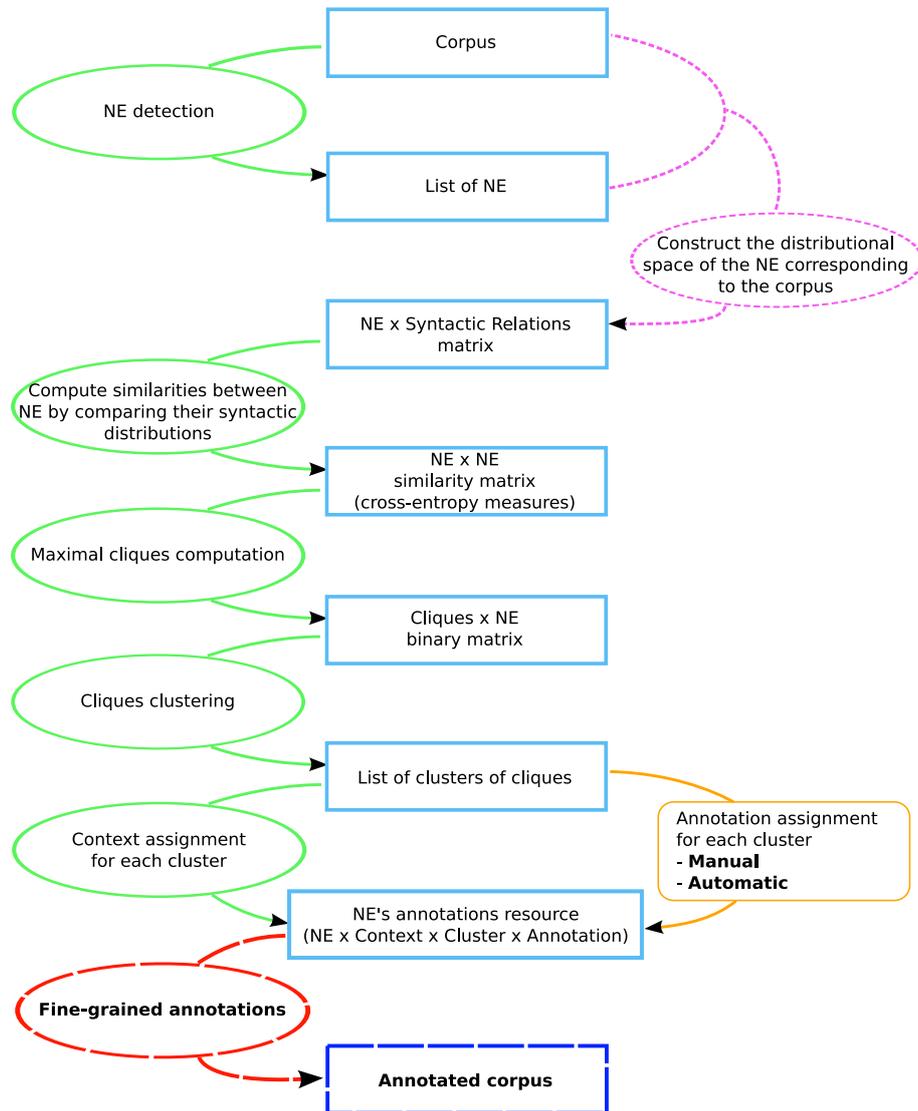
7. Attribution d'une annotation à chaque *cluster* $\mathbb{K}_l \in \mathbb{K}$. Les catégories possibles sont ORG (*organisation*), PERS (*person*) et LOC (*location*). C'est à ce stade que notre système devient semi-supervisé. Nous avons proposé deux approches :

- Annotation manuelle des *clusters* (CBC M) à partir de la liste des EN potentielles les plus fréquentes et les contextes syntaxiques rangés par ordre de pertinence. Par exemple, le système présente les *clusters* et les contextes comme dans la Figure 2.8 et l'annotateur.rice indique la catégorie qu'il.elle pense être le plus adéquat. L'annotation NONE est également prévue dans le cas où la catégorie n'est pas clairement identifiée.
- Annotation automatique (CBC A) à condition que l'on dispose en amont des catégories de plusieurs EN *via* une ressource externe comme Wikipedia par exemple ou alors *via* un système NER déjà entraîné sur un corpus annexe. Dans ce cas, on procède par un vote pondéré. Étant donné un *cluster* de cliques, on parcourt l'ensemble de ses EN potentielles et si une EN potentielle a une annotation disponible on ajoute son poids (sa fréquence au sein du cluster) au nombre de vote pour son annotation. Si une catégorie atteint plus de 50% des votes alors le *cluster* tout entier (ses EN potentielles et les dépendances syntaxiques pertinentes associées) reçoit cette annotation. Si aucune catégorie parmi ORG, PERS et LOC n'atteint la majorité absolue, alors le *cluster* reçoit l'annotation NONE.

Ces différentes étapes sont représentées schématiquement dans la Figure 2.10.

Notre système CBC construit une ressource qui consiste, *in fine*, en un **ensemble de *biclusters* recouvrants, annotés**, qui sont constitués chacun :

- d'un sous-ensemble d'EN potentielles pondérées (par leur fréquence),
- d'un sous-ensemble de contexte syntaxiques pondérés (par la mesure de pertinence),

FIGURE 2.10 – *Clique Based Clustering (CBC) System.*

- et d'une annotation parmi ORG, PERS, LOC et NONE.

La ressource CBC peut être utilisée pour annoter des EN d'un corpus en identifiant dans le texte les mentions des cas qu'il a en référence dans ses *biclusters* et en analysant les contextes syntaxiques associés. Étant donnée une mention d'une EN dans le texte, celle-ci peut être dans plusieurs dépendances syntaxiques renvoyant à plusieurs *clusters*. Si les annotations de ces *clusters* diffèrent alors la catégorie choisie est NONE sinon le label commun est utilisé. Si l'annotation de tous les *clusters* est NONE alors le label du *cluster* dans lequel l'EN concernée est le plus fréquent est choisi. Dans ce cas d'annotation par défaut, on ne tient pas compte des contextes syntaxiques et on attribue la catégorie que l'on suppose être la plus fréquente.

En comparaison des modèles supervisés utilisés à l'époque, les performances de CBC sont

	Systems	Prec.	Rec.	F-me.
1	<i>CBC-NER system M</i>	<i>71.67</i>	<i>23.47</i>	<i>35.36</i>
	<i>CBC-NER system A</i>	<i>70.66</i>	<i>32.86</i>	<i>44.86</i>
2	<i>XIP NER</i>	<i>77.77</i>	<i>56.55</i>	<i>65.48</i>
	XIP + CBC M	78.41	60.26	68.15
	XIP + CBC A	76.31	60.48	67.48
3	<i>Stanford NER</i>	<i>67.94</i>	<i>68.01</i>	<i>67.97</i>
	Stanford + CBC M	69.40	71.07	70.23
	Stanford + CBC A	70.09	72.93	71.48
4	<i>GATE NER</i>	<i>63.30</i>	<i>56.88</i>	<i>59.92</i>
	GATE + CBC M	66.43	61.79	64.03
	GATE + CBC A	66.51	63.10	64.76
5	<i>Stanford + XIP</i>	<i>72.85</i>	<i>75.87</i>	<i>74.33</i>
	Stanford + XIP + CBC M	72.94	77.70	75.24
	Stanford + XIP + CBC A	73.55	78.93	76.15
6	<i>GATE + XIP</i>	<i>69.38</i>	<i>66.04</i>	<i>67.67</i>
	GATE + XIP + CBC M	69.62	67.79	68.69
	GATE + XIP + CBC A	69.87	69.10	69.48
7	<i>GATE + Stanford</i>	<i>63.12</i>	<i>69.32</i>	<i>66.07</i>
	GATE + Stanford + CBC M	65.09	72.05	68.39
	GATE + Stanford + CBC A	65.66	73.25	69.25

TABLE 2.3 – Résultats du système CBC seul et en combinaison avec d’autres systèmes de NER.

défavorables comme cela est indiqué dans la Table 2.3 qui expose les critères de précision, rappel et de F-mesure de différentes approches et combinaisons d’approches. Notre système vise à **identifier et catégoriser des entités nommées nouvelles** d’une part, et à **améliorer la catégorisation en cas d’ambiguïté** d’autre part. **CBC est donc complémentaire à des outils existants de NER**. Dans ce contexte, les performances affichées dans la Table 2.3 indiquent que, dans un mode hybride, notre ressource permet d’améliorer les résultats de tous les systèmes existants que nous avons testés (XIP -*Xerox Incremental Parser*-, Stanford NER, GATE NER), ce qui est satisfaisant.

2.3 Discussions et perspectives

Similairement au domaine vision par ordinateur, la discipline TALN a été révolutionnée par les méthodes de *deep learning* au cours des années 2010. Dernièrement, ce sont les **architectures neuronales basées sur les *transformers*** [Vaswani et al., 2017] et les **mécanismes d’attention**, qui ont eu un impact phénoménal. Les modèles particulièrement marquants sont *BERT* (*Bidirectional Encoder Representations from Transformers*) [Devlin et al., 2018] et *GPT* (*Generative Pre-trained Transformers*) [Radford et al., 2018]. Ces approches ont été entraînées à partir de corpus provenant du web et composés de plusieurs centaines de milliards de mots. Elles permettent d’estimer des **Large Language Models (LLM)** qui donnent

avec précision des probabilités d’occurrence de mots étant donné un début de phrase ou un contexte. Ces LLM peuvent être ensuite utilisées pour résoudre des *downstream tasks* par *fine-tuning* mais aussi de solutionner avec beaucoup d’efficacité des tâches de *zero-shot learning*. Les technologies que nous utilisons dans la vie quotidienne et qui incluent ces LLM sont, par exemple, le moteur de recherche de Google qui emploie BERT, ou encore l’agent conversationnel ChatGPT qui a provoqué une onde de choc depuis son avènement en novembre 2022, en exposant au monde entier les capacités de l’**IA générative**.

La **taille de la base d’apprentissage** et l’**architecture neuronale profonde reposant sur les *Transformers*** sont les deux principaux **ingrédients expliquant les excellentes performances des LLM** susmentionnés.

Les travaux auxquels j’ai participé dans ce domaine et qui précèdent les LLM traitaient du problème de *data sparsity* dans un contexte de corpus d’apprentissage spécifique ou limité. Or cette limite est dépassée dans le cas des LLM. Ils sont principalement développés par ou en collaboration avec les GAFAM qui ont accès aux données massives du web et aux ressources computationnelles adéquates. Toutefois, toute proportion gardée, je souhaite indiquer des idées communes entre les travaux présentés en sous-section 2.2.1 et quelques principes importants sous-jacents aux LLM. Pour cela je rappelle le diagramme introduisant les *Transformers* [Vaswani et al., 2017] dans la Figure 2.11 et je reviens sur la **fusion d’ordre supérieur** suivante que j’ai introduite page 41 et qui a donné de bons résultats sur nos benchmarks en NER :

$$\begin{array}{c}
 E(E(\mathbf{X}^f, L(\mathbf{X}^f, \underbrace{P(\mathbf{S}^f, \mathbf{X}^f)}_{1\text{er ordre}})), E(\mathbf{X}^l, L(\mathbf{X}^l, \underbrace{P(\mathbf{S}^s, \mathbf{X}^l)}_{1\text{er ordre}}))) \\
 \underbrace{\hspace{10em}}_{2\text{nd ordre}} \quad \underbrace{\hspace{10em}}_{2\text{nd ordre}} \\
 \underbrace{\hspace{15em}}_{3\text{ème ordre}} \quad \underbrace{\hspace{15em}}_{3\text{ème ordre}} \\
 \underbrace{\hspace{20em}}_{4\text{ème ordre}}
 \end{array}$$

La **diffusion transmodale** donnée par $P(\mathbf{S}^s, \mathbf{X}^l)$ permet de renforcer les liens de type *terme-feature*, en tenant compte des contextes syntaxiques terme-terme. Il s’agit d’une exploitation *ad-hoc* de l’hypothèse distributionnelle qui place le contexte des mots au coeur de l’analyse des proximités sémantiques. Les applications de propagation transmodale ou même monomodale comme $P(\mathbf{S}^f, \mathbf{X}^f)$ par exemple, sont des **techniques non supervisées d’enrichissement de *features***. Dans notre cas, la propagation se fait *via* des relations de similarité diverses mais fondées principalement sur des contextes linguistiques qu’ils soient lexicaux ou syntaxiques.

Prise en compte du contexte et ajout de nouvelles variables sont des principes importants au sein des *Transformers*. Les mécanismes de *self-attention* permettent de représenter spécifiquement un mot étant donné les autres mots de la phrase ce qui précise le contexte dans lequel il est utilisé. D’ailleurs, cette architecture n’utilise pas une mais plusieurs unités d’auto-attention (*mutli-head attention*) dont les blocs (en gris clair dans la Figure 2.11) sont répliqués “Nx” fois. Ceci indique clairement qu’il est important de considérer plusieurs dimen-

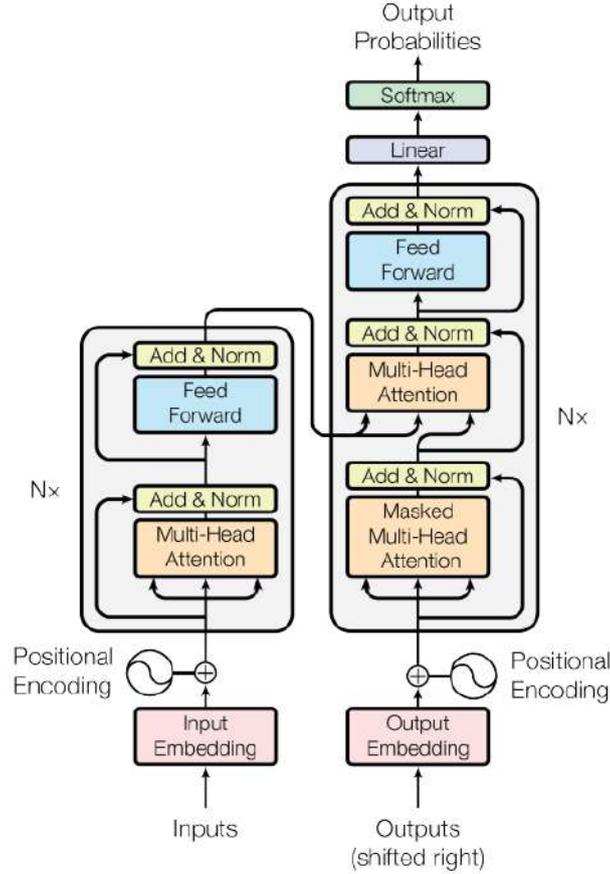


FIGURE 2.11 – Architecture des *Transformers* pour la traduction automatique [Vaswani et al., 2017].

sions possibles dans la quantification de l'importance des dépendances entre les mots d'une phrase. Dans notre cas, ce sont les différents types de contexte SEN, NP et DEP introduits en début de sous-section 2.2.1, qui donnent plusieurs vues de ces "attentions". Ensuite, les applications de fusion précoce E sont de simples concaténation de matrices conduisant à un espace de représentation de grande dimension, riche et varié. Elles sont similaires à la concaténation des sorties des différentes unités du module *Multi-Head Attention* (en orange clair dans la Figure 2.11). Enfin, les opérations de fusion tardive L sont des sommes de matrices. Dans notre approche, les fusions de 2nd ordre $L(\mathbf{X}^f, P(\mathbf{S}^f, \mathbf{X}^f))$ et $L(\mathbf{X}^l, P(\mathbf{S}^s, \mathbf{X}^l))$ sont similaires aux connexions résiduelles formalisées par les étapes *Add & Norm* des *Transformers* (en jaune clair dans la Figure 2.11).

De mon point de vue, il existe donc des **similitudes de nature conceptuelle entre les *Transformers* et les travaux que nous avons développés dans la thèse de Pavel**. Mais les *Transformers* vont bien évidemment plus loin. En particulier, contrairement à nos approches non supervisées, les LLM basés sur les *Transformers* utilisent l'**auto-supervision** comme les *masked language model*, ce qui permet d'apprendre les poids d'attention et les

représentations contextuelles des mots étant donné une phrase. Le paradigme *self-supervision* est puissant et permet d’outrepasser les limites des modèles supervisés lorsque les données annotées sont peu volumineuses ou inexistantes.

Parmi les enjeux scientifiques relatifs aux **LLM** et aux *Transformers*, je souhaite évoquer des problématiques concernant l’**interprétabilité de ces modèles neuronaux**. En particulier, le domaine du TALN est, à mon avis, particulièrement propice à l’étude de ce type de questions dans la mesure où les méthodes symboliques/*rule based* et les approches numériques/*machine learning* ont été développées par deux communautés distinctes et sont complémentaires. Les travaux que j’ai entrepris dans ce domaine visent d’ailleurs à exploiter cette complémentarité par le biais d’approches hybrides exploitant des informations linguistiques au sein d’approches statistiques. Les questions de recherche qui me semblent pertinentes dans ce contexte sont les suivantes. En premier lieu, il s’agit de confronter les unités d’attention estimées par l’architecture *Transformer* à des connaissances/règles linguistiques. Autrement dit, est-ce que les LLM apprennent en leur sein, des règles interprétables du point de vue linguistique ? De façon symétrique, comment intégrer/infuser des connaissances linguistiques dans l’apprentissage d’un LLM ? Est-ce que cette intégration permettrait de réduire la taille des modèles de LLM, de les améliorer, de mieux les interpréter ?

Ces deux axes suscitent des travaux de plus en plus nombreux dans la littérature et sont également en lien avec les sous-domaines *eXplainable Artificial Intelligence (XAI)* et *Informed Machine Learning*. Des exemples de travaux qui illustrent ces études sont respectivement [Clark et al., 2019] et [Strubell et al., 2018]. L’article [Hamilton et al., 2022] analyse plus spécifiquement en TALN, des **approches hybrides dites neuro-symboliques**. Cette fusion de paradigmes constitue un champ de recherche excitant et permettrait une meilleure modélisation, compréhension et appréhension des LLM.

Je conclus ce Chapitre en discutant de l’**approche CBC (*Clique Based Clustering*)**. Nous avons défini cette méthode de classification automatique pour contribuer à la résolution de la tâche NER. Une singularité de l’approche est qu’elle engendre un partitionnement flou des EN. En effet, une EN peut appartenir à plusieurs cliques (voir l’exemple de “*London*” dans la Figure 2.9) et ces dernières peuvent être regroupées dans des *clusters* différents. De ce fait, une même EN peut se retrouver dans plusieurs *clusters* à la fois. Je pense qu’il y a un intérêt à exploiter la méthode CBC dans d’autres contextes que celui du TALN en tant qu’**algorithme générique de partitionnement flou**. L’algorithme de Bron-Kerbosch [Bron and Kerbosch, 1973] est dans le pire des cas en $O(3^{n/3})$ ce qui n’est pas une complexité polynomiale. Toutefois, dans le cas de graphe *sparse* il est possible d’avoir un algorithme en complexité linéaire vis-à-vis du nombre de sommets n . Par ailleurs, la méthode CBC utilise un algorithme de regroupement qui scanne de façon successive les cliques. Il est alors possible d’avoir une **stratégie on-line de CBC** avec une procédure de génération de cliques maximales qui peut s’effectuer en parallèle de la procédure de *clustering* de cliques.

Enfin, j’ai expliqué que l’utilisation du *clustering* de cliques à des fins d’annotations d’EN

nécessitait des étapes intermédiaires de quantification des dépendances syntaxiques pour caractériser les *clusters*. Cela suggère une extension de l’approche au ***biclustering de bicliques maximales***. Plus précisément, l’unité informationnelle centrale serait non plus une clique maximale d’observations mais une **biclique maximale dans un graphe biparti qui est composé à la fois d’un sous-ensemble d’observations et d’un sous-ensemble de *features***. L’approche correspondrait alors à un **biclustering flou**. Dans le cas de la tâche NER, on aurait alors directement un sous-ensemble de contextes syntaxiques pertinents associés à un sous-ensemble d’EN. La complexité de l’énumération des bicliques maximales est un challenge en théorie des graphes mais dans le cas de graphes bipartis *sparse* des travaux récents permettent d’avoir des algorithmes efficaces comme par exemple [Chen et al., 2022]. L’algorithme de *clustering* de nature *on-line* peut facilement s’adapter aux bicliques. Comme indiqué précédemment, cette approche pourrait tout aussi suscité un intérêt pour des applications autres qu’en TALN.

Fonctions d'agrégation et explicabilité en apprentissage supervisé

Sommaire du chapitre

3.1	Introduction	52
3.1.1	Contexte	52
3.1.2	Travaux antérieurs	55
	Capacité, bicapacités et intégrales de Choquet associées	57
	Mesures maxitatives et finiment additives, TOWA et formules combinatoires de Poincaré et de Jordan	59
3.2	Contributions	61
3.2.1	Intégrale de Choquet bipolaire et 2-additive	61
3.2.2	Une nouvelle famille de fonctions d'agrégation	70
3.3	Discussions et perspectives	76

3.1 Introduction

3.1.1 Contexte

Dans ce Chapitre 3, je présente mes travaux se situant à l'intersection du domaine de l'**aide multicritère à la décision (AMCD)** et celui du *machine learning*. Au cours de mon travail de thèse de doctorat, je m'étais intéressé au problème d'agrégation de relations de préférence ou d'ordre que l'on rencontre en théorie des votes et du choix social. Typiquement, nous avons N votants exprimant une relation de préférence totale sur un ensemble de M candidats. L'approche sur laquelle je travaillais dans ma thèse est l'**analyse relationnelle (AR)** et celle-ci s'inspire des travaux du Marquis de Condorcet [Condorcet, 1785]. L'AR modélise les relations de préférence par des graphes orientés représentés par des matrices d'adjacence binaires (en AR on parle également de matrices de comparaison par paires ou de matrices relationnelles). La consolidation des préférences individuelles s'effectue alors par simple addition des matrices d'adjacence des votants. On obtient alors un graphe pondéré

représenté par une matrice d'adjacence valuée $\mathbf{C} = (c_{jj'})$, où pour chaque paire de candidats (j, j') , $c_{jj'}$ indique le nombre de votants ayant préféré j à j' . Comme on suppose N votants, on dira qu'une majorité stricte préfère j à j' si $c_{jj'} > \frac{N}{2}$. Dénotons par $\mathbf{X} = (x_{jj'})$ la matrice d'adjacence binaire d'ordre M encodant le graphe de la relation d'ordre sous-jacente au **classement collectif** des candidats. Si $c_{jj'} > \frac{N}{2}$, la "vraisemblance" de $x_{jj'} = 1$ est grande. Toutefois, il faut tenir compte des problèmes de transitivité conduisant au fameux paradoxe de Condorcet : une application naïve de la **règle de la majorité par paire**, $c_{jj'} > \frac{N}{2} \Rightarrow x_{jj'} = 1$, peut conduire à des circuits de type $(x_{jj'} = 1) \wedge (x_{j'j''} = 1) \wedge (x_{jj''} = 1)$ et, par conséquent, l'impossibilité d'ordonner le triplet (j, j', j'') .

La **règle de la majorité globale sous contrainte** est celle qui est promue en AR [Marcotorchino and Michaud, 1979]. Il s'agit d'une approche déjà suggérée par le Marquis de Condorcet [Michaud, 1987] mais dont le calcul de la solution est un problème NP-dur. Jean-François Marcotorchino et Pierre Michaud ont établi une méthode de calcul exacte permettant de résoudre en pratique le problème. Celle ci repose sur le programme linéaire en nombres bivalents suivant [Marcotorchino and Michaud, 1979, Michaud and Marcotorchino, 1979] :

$$\begin{aligned} \max \sum_{j,j'=1}^M \left(c_{jj'} - \frac{N}{2} \right) x_{jj'} & \quad (3.1) \\ \text{s.l.c.} \begin{cases} x_{jj} = 1, \forall j = 1, \dots, M & \text{(réflexivité),} \\ x_{jj'} + x_{j'j} \leq 1, \forall j, j' = 1, \dots, M : j \neq j' & \text{(antisymétrie),} \\ x_{jj'} + x_{j'j} \geq 1, \forall j, j' = 1, \dots, M : j \neq j' & \text{(totalité),} \\ x_{jj'} + x_{j'j''} - x_{jj''} \leq 1, \forall j, j', j'' = 1, \dots, M & \text{(transitivité),} \\ x_{jj'} \in \{0, 1\}, \forall j, j' = 1, \dots, M & \text{(binarité).} \end{cases} \end{aligned}$$

La matrice carrée binaire \mathbf{X} que l'on cherche, dite matrice relationnelle, doit vérifier les propriétés d'un ordre total. Par conséquent, on est garantie d'éviter les problèmes de non-transitivité ou effets Condorcet.

L'approche que je viens de décrire est de type "**compare puis agrège**" : pour chaque votant, on compare d'abord ses préférences pour chaque paire de candidats, puis on agrège les comparaisons par paires de tous les votants. Il existe une approche inverse en AMCD de type "**agrège puis compare**". Dans ce cas, on suppose que l'on dispose d'une table où pour chacun des M candidats, on a un score ou un rang attribué par chacun des N votants. Ici, pour chaque candidat, on agrège les scores ou rangs et la distribution de scores agrégés permet ensuite de comparer les candidats entre eux et de les ranger par ordre de préférence collective. L'agrégation des scores peut s'effectuer de façon très riche par l'utilisation de **fonctions d'agrégation** qui vont être la thématique de ce Chapitre.

Au sein du laboratoire ERIC, Antoine Rolland et moi-même avons été recrutés la même année, en 2010. Antoine a effectué sa thèse de doctorat sur les procédures d'agrégation ordi-

nale de préférences avec points de référence et il s'était intéressé, dans ce cadre, à la fonction d'agrégation donnée par l'intégrale de Choquet (dénotée CI pour *Choquet Integral*). Il était à ce moment en contact avec Brice Mayag de l'Université Paris Dauphine qui avait fait sa thèse sur ces outils mathématiques dans le cas 2-additif et pour le problème d'élicitation des préférences d'un décideur. J'ai rejoint le duo et nous avons tous les trois collaborés sur trois articles concernant l'**intégrale de Choquet bipolaire** (dénotée BCI pour *Bipolar Choquet Integral*) et **2-additive** (dénoté 2A pour *2-Additive*). J'expose plus loin une synthèse de nos collaborations.

De plus, je présente une **fonction d'agrégation que j'ai définie** et étudiée en plusieurs temps. Je l'ai d'abord appliquée lorsque j'étais à XRCE sur des tâches de *meta search* en recherche d'information où il est question d'agréger les résultats de plusieurs moteurs de recherche. Ensuite, après avoir rejoint l'UL2 et le laboratoire ERIC, j'ai entrepris un travail de nature plus fondamentale afin d'établir les propriétés mathématiques de cette fonction d'agrégation dans le cadre de l'AMCD.

Les questions de recherche abordées dans ce Chapitre sont organisées en deux blocs : l'un sur les intégrales de Choquet bipolaire et l'autre sur la famille de fonctions d'agrégation que j'ai définie. Les publications dans des revues ou conférences avec comités de lecture qui sont associées à ce Chapitre sont les suivantes :

- **J. Ah-Pine**. 2016. On aggregation functions based on linguistically quantified propositions and finitely additive set functions. *Fuzzy Sets and Systems*. 287 :1-21. [Lien vers le journal, <http://www.sciencedirect.com/science/article/pii/S0165011415002870>].
- A. Rolland, **J. Ah-Pine**, B. Mayag. 2015. Elicitation of 2-additive bicapacity parameters. *EURO Journal on Decision Processes*. 3(1). [Lien vers le journal, <http://link.springer.com/article/10.1007/s40070-015-0043-3>].
- **J. Ah-Pine**, Brice Mayag and Antoine Rolland. 2013. Identification of a 2-additive bicapacity by using mathematical programming. *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT 2013)*. [Lien vers la conférence, <http://www.adt2013.org/>].
- B. Mayag, A. Rolland, **J. Ah-Pine**. 2012. Elicitation of a 2-additive bicapacity through cardinal information on ternary actions. *Proceedings of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2012)*. [Lien vers la conférence, <http://www.ipmu2012.unict.it/>].
- **J. Ah-Pine**. 2011. On data fusion in information retrieval using different aggregation operators. *Web Intelligence and Agent Systems*. 9(1) :43-55. [Lien vers le journal, <http://www.iospress.nl/loadtop/load.php?isbn=15701263>].
- **J. Ah-Pine**. 2008. Data fusion in information retrieval using consensus aggregation operators. *Proceedings of Web Intelligence (WI 2008) held in conjunction with Intelligent Agent Technologies (IAT 2008) (IEEE/WIC/ACM)* [Taux d'acceptation < 20%].

[Lien vers la conférence, <http://datamining.it.uts.edu.au/conferences/wi08/>].

3.1.2 Travaux antérieurs

Les **fonctions d'agrégation**, que je dénoterai de façon générique par F , sont des applications multivariées à valeurs réelles dont le but est de combiner plusieurs valeurs numériques en une seule. Elles interviennent en AMCD où l'objectif est de décider une ou plusieurs alternatives dans un ensemble dénoté \mathbb{M} . On suppose que toutes les alternatives sont évaluées par rapport à un ensemble de critères dénoté \mathbb{N} . En AMCD et contrairement en statistiques, ce sont les critères qui sont les éléments d'étude au centre des modèles. J'adopte, par la suite, les notations couramment utilisées dans cette communauté où l'ensemble des critères est dénoté $\mathbb{N} = \{1, 2, \dots, N\}$ et l'indice i est utilisé comme itérateur pour parcourir cet ensemble. Toutefois, par abus de notations, i désigne également un critère et on pourra donc écrire de façon équivalente $\sum_{i=1}^N$ et $\sum_{i \in \mathbb{N}}$. De la même manière, N est utilisé à la fois pour désigner le nombre total de critères et le critère N . Ces raccourcis ne nuisent pas à la compréhension, au contraire, ils permettent d'alléger les expressions.

Pour une alternative $X \in \mathbb{M}$, une fonction d'agrégation F permet de combiner ses N différents scores en un seul afin d'avoir une mesure synthétique de la satisfaction de l'alternative au regard de l'ensemble des critères. Représentons par le vecteur $\mathbf{x} = (x_i)$ de taille N , le **profil de scores de X** , c'est à dire les N valeurs de satisfaction de X relativement à chaque critère $i = 1, \dots, N$. La fonction d'agrégation est appliquée au profil de scores de chaque alternative et la distribution des valeurs synthétiques obtenues permet de comparer, au sens de l'ensemble des critères, les alternatives les unes par rapport aux autres et de décider *in fine*, laquelle ou lesquelles choisir.

Sans perte de généralité, on considère que pour chaque alternative, son profil est un vecteur composé de valeurs comprises entre 0 et 1. Étant donné un score, plus celui-ci est proche de 1, plus l'alternative satisfait au critère. Bien que cette hypothèse de borne soit assez naturelle, je la justifie au travers du concept des **sous-ensembles flous** en suivant le cadre proposé initialement par Richard Bellman et Lotfi Zadeh dans [Bellman and Zadeh, 1970]. Il s'agit d'ailleurs d'un ingrédient central pour la définition de la famille de fonctions d'agrégation que j'ai définie et que je présente en sous-section 3.2.2. Soit alors S_i le sous-ensemble "le critère i est satisfait" et dénotons par μ_{S_i} **la fonction d'appartenance floue à S_i** . Plus précisément, étant donné une alternative X , $\mu_{S_i}(X) \in [0, 1]$ indique le degré d'appartenance de X au sous-ensemble S_i , qu'on interprète également comme le degré de satisfaction de X vis-à-vis du critère i . Nous avons donc $\mathbf{x} = (\mu_{S_1}(X), \dots, \mu_{S_N}(X))$.

Je précise qu'il existe un autre paradigme plus répandu concernant la notion de profil de scores. Il s'agit de la **théorie des utilités multi-attribut** (*Multi-Attribute Utility Theory* -MAUT-). Dans cette approche, $\mathbf{x} = (u_1(x_1), \dots, u_N(x_N))$ où $u_i : \mathbb{A}_i \mapsto [0, 1]$ est la fonction d'utilité partielle du critère i ; \mathbb{A}_i est l'ensemble des attributs possibles pour le critère i ; et $x_i \in \mathbb{A}_i$ est l'attribut du critère i observé pour l'alternative X . Dans la mesure où le cadre conceptuel de la fonction d'agrégation que j'introduis dans la sous-section 3.2.2 repose sur les

sous-ensemble flous, j'utiliserai principalement le vocabulaire relevant de ce champ. Toutefois, à de multiples reprises, j'emprunterai de façon interchangeable les termes scores, utilités et évaluations. Je reviendrai également sur l'approche MAUT dans le cadre de l'**élicitation des préférences** d'un agent à l'aide d'alternatives ternaires en page 64.

J'introduis à présent la relation de préférence stricte sur l'ensemble des alternatives dans le cadre formel rappelé précédemment. On dira que " **X est strictement préférée à X'** " ce que l'on notera par $X \succ X'$ si et seulement si leurs profils de scores $\mathbf{x} = (x_j)$ et $\mathbf{x}' = (x'_j)$ de \mathbb{R}^N vérifient :

$$X \succ X' \Leftrightarrow (\forall j \in \mathbb{N} : x_j \geq x'_j) \wedge (\exists j \in \mathbb{N} : x_j > x'_j).$$

Dans le cas de la relation " **X est indifférente à X'** " et celui de la relation " **X est préférée ou indifférente à X'** ", on a les définitions respectives suivantes :

$$\begin{aligned} X \sim X' &\Leftrightarrow \forall j \in \mathbb{N} : x_j = x'_j, \\ X \succeq X' &\Leftrightarrow \forall j \in \mathbb{N} : x_j \geq x'_j. \end{aligned}$$

Je définis avec plus de précision la notion de **fonction d'agrégation**. F est une application de $[0, 1]^N$ dans $[0, 1]$ qui doit satisfaire les conditions suivantes :

$$\begin{cases} F(0, \dots, 0) = 0, \\ F(1, \dots, 1) = 1, \\ F \text{ est monotone croissante en chaque critère.} \end{cases} \quad (3.2)$$

Ces conditions minimales sont motivées du point de vue de la **théorie de la décision**. Si une alternative ne satisfait à aucun critère alors son score global doit être au plus bas c'est à dire 0. Au contraire, si elle satisfait parfaitement à tous les critères sa mesure de synthèse doit être au maximum, c'est à dire 1. Enfin, toute chose étant égale par ailleurs, si le score d'une alternative pour un critère donné augmente alors le score globale de cette alternative ne peut pas diminuer.

Parmi les fonctions d'agrégation que nous utilisons couramment dans des contextes variés, il y a les **moyennes** (arithmétique, pondérée, généralisée, ...) et les **statistiques d'ordre** (minimum, médiane, maximum, ...). Les fonctions minimum et maximum représentent des comportements extrêmes dits conjonctifs et disjonctifs respectivement. Les fonctions moyennes sont très riches et permettent des comportements de consensus. Toutefois, elles trouvent aussi rapidement des limites lorsqu'il s'agit de modéliser des préférences complexes. En particulier, les moyennes supposent implicitement que les critères sont indépendants les uns vis-à-vis des autres. C'est un point important sur lequel je reviendrai par la suite.

Il existe des **fonctions d'agrégation paramétriques** qui généralisent les cas classiques précédents. La fonction OWA (*Ordered Weighted Averaging*) introduite par Ronald Yager [Yager, 1988], englobe les statistiques d'ordre et la moyenne arithmétique. Du point de vue axiomatique, elle satisfait au critère d'invariance par rapport aux permutations des scores

(propriété de symétrie) et à celui de l'idempotence, $F(a, \dots, a) = a, \forall a \in [0, 1]$. Plusieurs extensions de la fonction OWA ont été proposées dans la littérature comme la fonction WOWA pour *Weighted OWA* qui permet de généraliser également les moyennes pondérées. La fonction d'agrégation qui m'intéresse en particulier dans ce contexte et qui est en lien avec les travaux que j'expose dans la sous-section 3.2.2 est TOWA pour *Triangular norms OWA*. Je présente cette fonction généralisante dans le paragraphe 3.1.2. Avant cela, je discute ci-dessous des définitions de base dans le scope de l'intégrale de Choquet afin de présenter ensuite mes contributions dans ce domaine dans la sous-section 3.2.1.

Capacité, bicapacités et intégrales de Choquet associées

Une fonction d'agrégation généralisante qui englobe OWA (mais pas TOWA) est l'**intégrale de Choquet (CI)**. En AMCD, on utilise de façon sous-entendue la version discrète de l'application introduite par Gustave Choquet dans [Choquet, 1954] et qui est une intégrale par rapport à une **mesure non additive**. Dans le cas discret, cette mesure est une **fonction d'ensemble** sur $2^{\mathbb{N}} = \{S \subseteq \mathbb{N}\}$, c'est à dire une fonction à valeurs réelles définie sur l'ensemble des sous-ensembles de \mathbb{N} . On dénote par μ cette fonction d'ensemble et on a donc $\mu : 2^{\mathbb{N}} \mapsto \mathbb{R}$. Dans le cas particulier d'une CI et afin de produire une fonction d'agrégation bien fondée, la fonction d'ensemble μ doit vérifier les propriétés suivantes :

$$\left\{ \begin{array}{ll} \forall S \subseteq \mathbb{N} : \mu(S) \geq 0 & (\mu \text{ est non négative}), \\ \mu(\emptyset) = 0 & (\mu \text{ est grounded}), \\ \mu(\mathbb{N}) = 1 & (\mu \text{ est normalisée}), \\ \forall S' \subset S : \mu(S') \leq \mu(S) & (\mu \text{ est monotone croissante}). \end{array} \right. \quad (3.3)$$

Une fonction d'ensemble vérifiant ces conditions est appelée **capacité ou mesure floue**. Dans le contexte AMCD, elle peut être interprétée comme donnant une mesure de l'importance à tous les sous-ensembles de critères. La condition de monotonie implique que l'ajout d'un critère à un sous-ensemble ne peut pas faire décroître le poids du sous-ensemble augmenté. Une CI est associée à une capacité et on la dénotera ainsi par CI_{μ} . Notons par $\mathbf{x} \in [0, 1]^N$ un profil de scores, alors la CI par rapport à μ et appliquée à \mathbf{x} est définie comme suit :

$$CI_{\mu}(\mathbf{x}) = \sum_{i=1}^N x_{\tau(i)} [\mu(\{\tau(i), \dots, \tau(N)\}) - \mu(\{\tau(i+1), \dots, \tau(N)\})], \quad (3.4)$$

où τ est une permutation telle que $x_{\tau(1)} \leq x_{\tau(2)} \leq x_{\tau(3)} \leq \dots \leq x_{\tau(N)}$.

Des instances spéciales de capacité permettent de retrouver comme cas particulier la fonction OWA et la moyenne pondérée. Dans ces deux derniers cas, la **capacité est additive (ou 1-additive)**, c'est à dire que l'ajout d'un critère à un sous-ensemble augmente le poids du sous-ensemble résultant, exactement du poids du critère ajouté. La CI associée à des **capacités non additives** permet donc de définir des fonctions d'agrégation plus complexes

que OWA et les moyennes pondérées. Dans ce contexte, on peut formaliser des **notions de redondance et de complémentarité entre critères**.

La CI brièvement introduite ci-dessus est dite **unipolaire** car elle suppose que les scores appartiennent numériquement à une échelle positive, ce qui permet uniquement d'exprimer un avis favorable. En théorie de la décision, plusieurs travaux, comme la *Cumulative Prospect Theory* (CPT) d'Amos Tversky et du prix Nobel en économie Daniel Kahneman [Tversky and Kahneman, 1992], montrent que les agents expriment davantage leurs évaluations vis-à-vis d'une **échelle bipolaire**. Plus précisément, un agent émet un avis dans un référentiel au sein duquel un point de référence définit une utilité neutre, et de part et d'autre de ce point, les utilités sont positives (gain/satisfaction) et négatives (perte/insatisfaction).

Dans le contexte des échelles bipolaires, on va supposer, sans perte de généralité, que les scores sont dans l'intervalle $[-1, 1]$. Afin de modéliser et d'agréger des préférences représentées par des valeurs dans $[-1, 1]$, l'intégrale de Choquet a été généralisée. Dans cette perspective, le concept clef est celui de **bicapacité** qui généralise celui de capacité. Ces extensions ont été introduites par Michel Grabisch et Christophe Labreuche dans [Grabisch and Labreuche, 2002a]. Soit $3^{\mathbb{N}} = \{(A, B) \in 2^{\mathbb{N}} \times 2^{\mathbb{N}} : A \cap B = \emptyset\}$, l'ensemble des couples disjoints de sous-ensembles de \mathbb{N} . On définit sur $3^{\mathbb{N}}$ la relation \sqsubseteq suivante, $\forall (A_1, A_2), (B_1, B_2) \in 3^{\mathbb{N}}$:

$$(A_1, A_2) \sqsubseteq (B_1, B_2) \Leftrightarrow [A_1 \subseteq B_1 \text{ et } B_2 \subseteq A_2]. \quad (3.5)$$

Une application $\nu : 3^{\mathbb{N}} \rightarrow [-1, 1]$ est une **bicapacité normalisée** sur $3^{\mathbb{N}}$ si elle vérifie les conditions suivantes [Grabisch and Labreuche, 2002a] :

$$\left\{ \begin{array}{ll} \nu(\emptyset, \emptyset) = 0 & (\nu \text{ est grounded}), \\ \nu(\mathbb{N}, \emptyset) = 1 \text{ et } \nu(\emptyset, \mathbb{N}) = -1 & (\nu \text{ est normalisée}), \\ \forall (A_1, A_2), (B_1, B_2) \in 3^{\mathbb{N}} : [(A_1, A_2) \sqsubseteq (B_1, B_2) \Rightarrow \nu(A_1, A_2) \leq \nu(B_1, B_2)] & (\nu \text{ est monotone croissante}). \end{array} \right. \quad (3.6)$$

L'intégrale de Choquet bipolaire (BCI) associée à ν et dénotée par BCI_ν est alors définie par l'expression suivante :

$$\text{BCI}_\nu(\mathbf{x}) = \sum_{i=1}^{\mathbb{N}} |x_{\tau(i)}| \left[\nu(\mathbb{N}_{\tau(i)} \cap \mathbb{N}^+, \mathbb{N}_{\tau(i)} \cap \mathbb{N}^-) - \nu(\mathbb{N}_{\tau(i+1)} \cap \mathbb{N}^+, \mathbb{N}_{\tau(i+1)} \cap \mathbb{N}^-) \right], \quad (3.7)$$

où $\mathbb{N}^+ = \{i \in \mathbb{N} : x_i \geq 0\}$, $\mathbb{N}^- = \mathbb{N} \setminus \mathbb{N}^+$, $\mathbb{N}_{\tau(i)} = \{\tau(i), \dots, \tau(N)\}$ et τ est une permutation sur \mathbb{N} telle que $|x_{\tau(i)}| \leq |x_{\tau(i+1)}| \leq \dots \leq |x_{\tau(n)}|$.

Une capacité et une bicapacité normalisées contiennent respectivement $2^N - 2$ et $3^N - 3$ paramètres. Il s'agit d'**objets mathématiques hautement combinatoires**. Des concepts supplémentaires permettent de définir des sous-ensembles de mesures floues qui comportent implicitement moins de paramètres à déterminer. Je reviendrai sur ces notions dans la section suivante.

En AMCD, on utilise ces fonctions d'agrégation afin de modéliser les préférences d'un

agent/décideur. Pour cela il faut estimer les valeurs d’une capacité ou d’une bicapacité. En pratique, dans ce domaine, une interaction avec l’agent permet de construire de façon itérative une mesure floue appropriée et on parle d’**élicitation des préférences** du décideur. J’ai alors suggéré à Antoine et Brice d’étudier ce problème sous l’angle de l’**apprentissage supervisé** en supposant que nous disposions des profils de scores et aussi des scores agrégés donnés par le décideur. Ce faisant, il devient opportun d’explorer des concepts définis en apprentissage supervisé pour l’estimation des paramètres d’une bicapacité. Je développe ces aspects dans la sous-section 3.2.1.

Mesures maxitives et finiment additives, TOWA et formules combinatoires de Poincaré et de Jordan

Dans ce paragraphe, j’introduis des notions, identités et une fonction d’agrégation qui sont en lien avec la fonction d’agrégation que j’expose par la suite dans la sous-section 3.2.2. Dans cette perspective, je reprends les concepts de sous-ensembles flous $S_i =$ “le critère i est satisfait”, ainsi que la fonction d’appartenance floue μ_{S_i} à valeur dans $[0, 1]$. Soit alors la collection de sous-ensembles flous $\mathbb{S} = \{S_1, \dots, S_N\}$ et dénotons par \mathcal{S} l’algèbre des ensembles qui en découle. Par la suite, par abus de notation, j’utilise $\mu(S)$, la mesure floue de S , et μ_S , la fonction d’appartenance à S , de façon interchangeable.

Comme indiqué en introduction de ce Chapitre, au cours de ma thèse de doctorat, je me suis intéressé à la théorie des votes et du choix social. Dans ce contexte, j’ai étudié le calcul de probabilité des événements de type “au moins k votants sur N préfèrent j à j' ” pour une paire de candidats (j, j') donnée. Ce type d’évènement est également employé dans le contexte des fonctions d’agrégation. En effet, l’approche **Triangular norm OWA** définie par Ronald Yager dans [Yager, 2005], est fondée sur les **sous-ensembles flous “au moins k critères sur N sont satisfaits”** que je dénote par E_k^N . Du point de vue ensembliste, on a par définition, $\forall k = 1, \dots, N$:

$$E_k^N = \bigcup_{1 \leq i_1 < \dots < i_k \leq N} (S_{i_1} \cap \dots \cap S_{i_k}). \quad (3.8)$$

Soit alors $\mu_{E_k^N}$ la fonction d’appartenance au sous-ensemble flou E_k^N et soit \mathbf{x} le profil de scores d’une alternative X . La valeur $\mu_{E_k^N}(\mathbf{x}) \in [0, 1]$ indique le degré d’appartenance de X au sous-ensemble E_k^N , ce que l’on pourra également interpréter comme étant le niveau de satisfaction de X à au moins k critères sur N (tout sous-ensemble de k critères confondu).

On peut définir, de façon générale, une application qui **combine linéairement les fonctions d’appartenance** $\mu_{E_k^N}$ et on introduit ainsi le fonction générique $F_{\mathbf{w}} = [0, 1]^N \mapsto \mathbb{R}$ suivante :

$$F_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^N w_k \mu_{E_k^N}(\mathbf{x}), \quad (3.9)$$

3.1. INTRODUCTION

où $\mathbf{w} = (w_k)_{k=1,\dots,N}$ est un vecteur de poids tel que $w_k \geq 0, \forall k = 1, \dots, N$ et $\sum_k w_k = 1$.

L'approche *Triangular norm OWA* est un cas particulier de (3.9). En effet, plusieurs types de fonctions d'ensemble μ sur \mathcal{S} peuvent être considérées pour déterminer $\{\mu_{E_k^N}\}_{k=1,\dots,N}$. La méthode *Triangular norm OWA* utilise dans ce cas une **fonction d'ensemble maxitive** pour laquelle $\mu_{S \cup S'} = \max(\mu_S, \mu_{S'}), \forall S, S' \in \mathcal{S}$. Dans ce cas précis, (3.8) se réduit à :

$$\mu_{E_k^N} = \max_{1 \leq i_1 < \dots < i_k \leq N} \mu_{S_{i_1} \cap \dots \cap S_{i_k}}. \quad (3.10)$$

Ensuite, pour définir $\mu_{S \cap S'}$, la mesure floue de l'intersection entre deux sous-ensembles flous S et S' , on peut utiliser une **triangular norm (t-norm)**. Il s'agit d'applications binaires introduites par Berthold Schweizer et Abe Sklar dans [Schweizer and Sklar, 1958, Schweizer and Sklar, 1983], suite à des travaux pionniers de Karl Menger [Menger, 1942]. Les *t-norms* ont ensuite été utilisées pour définir des fonctions mesurant les valeurs d'appartenance d'un ensemble flou formé par l'intersection de deux sous-ensembles flous. Une *t-norm* est une application $\mathbb{T} : [0, 1]^2 \mapsto [0, 1]$ qui vérifie les conditions suivantes :

$$\left\{ \begin{array}{ll} \mathbb{T}(x, y) = \mathbb{T}(y, x) & (\mathbb{T} \text{ est commutative}), \\ \mathbb{T}(\mathbb{T}(x, y), z) = \mathbb{T}(x, \mathbb{T}(y, z)) & (\mathbb{T} \text{ est associative}), \\ \forall y \leq z : \mathbb{T}(x, y) \leq \mathbb{T}(x, z) & (\mathbb{T} \text{ est monotone croissante}), \\ \mathbb{T}(x, 1) = x & (1 \text{ est l'élément neutre}). \end{array} \right. \quad (3.11)$$

Les quatre *t-norms* fondamentales sont les suivantes :

- La *t-norm* minimum¹ : $\forall x, y \in [0, 1] : \mathbb{T}_M(x, y) = \min(x, y)$.
- La *t-norm* produit : $\forall x, y \in [0, 1] : \mathbb{T}_P(x, y) = xy$.
- La *t-norm* de Lukasiewicz : $\forall x, y \in [0, 1] : \mathbb{T}_L(x, y) = \max(x + y - 1, 0)$.
- La *t-norm* drastique : $\forall x, y \in [0, 1] : \mathbb{T}_D(x, y) = \min(x, y)$ si $\max(x, y) = 1$, et $\mathbb{T}_D(x, y) = 0$ sinon.

Il existe également des familles paramétriques de *t-norms* comme celle de Maurice J. Frank² qui dépend du paramètre $\lambda \in [0, \infty]$ et dont la définition est donnée par, $\forall a, b \in [0, 1]$:

$$\mathbb{T}_\lambda^F(a, b) = \begin{cases} \mathbb{T}_M(a, b) & \text{if } \lambda = 0, \\ \mathbb{T}_P(a, b) & \text{if } \lambda = 1, \\ \mathbb{T}_L(a, b) & \text{if } \lambda = \infty, \\ \log_\lambda \left(1 + \frac{(\lambda^a - 1)(\lambda^b - 1)}{\lambda - 1} \right) & \text{otherwise.} \end{cases} \quad (3.12)$$

Suite à cette digression nécessaire sur les *t-norms*, je reviens sur l'équation (3.10). Pour un profil de scores \mathbf{x} , supposons que τ soit une permutation de $\{1, \dots, N\}$ de sorte que $\mu_{S_{\tau(1)}}(\mathbf{x}) \geq \mu_{S_{\tau(2)}}(\mathbf{x}) \geq \dots \geq \mu_{S_{\tau(N)}}(\mathbf{x})$. Étant donné qu'une *t-norm* \mathbb{T} est monotone crois-

1. Aussi appelé produit de Gödel ou de Zadeh.

2. Les *t-norms* sont des applications proches des copules. Les *t-norm* de Frank sont également des copules.

3.2. CONTRIBUTIONS

sante, on a la propriété suivante qui simplifie considérablement le calcul des $\mu_{E_k^N}$:

$$\max_{1 \leq i_1 < \dots < i_k \leq N} \mathbb{T}(\mu_{S_{i_1}}(\mathbf{x}), \dots, \mu_{S_{i_k}}(\mathbf{x})) = \mathbb{T}(\mu_{S_{\tau(1)}}(\mathbf{x}), \dots, \mu_{S_{\tau(k)}}(\mathbf{x})). \quad (3.13)$$

Sous l'hypothèse d'une mesure maxitive et de l'utilisation d'une t -norm \mathbb{T} pour la mesure des intersections, l'application $F_{\mathbf{w}}$ définie par (3.9) correspond à la fonction d'agrégation $\text{TOWA}_{\mathbf{w}, \mathbb{T}}$ (*Triangular norms OWA*) et on a :

$$\text{TOWA}_{\mathbf{w}, \mathbb{T}}(\mathbf{x}) = \sum_{1 \leq k \leq N} w_k \mathbb{T}(\mu_{S_{\tau(1)}}(\mathbf{x}), \dots, \mu_{S_{\tau(k)}}(\mathbf{x})). \quad (3.14)$$

Contrairement à $\text{TOWA}_{\mathbf{w}, \mathbb{T}}$, je propose d'utiliser une **fonction d'ensemble μ sur \mathcal{S} qui soit finiment additive**. Dans ce cas, les calculs des $\mu_{E_k^N}$ deviennent hautement combinatoires. A première vue, il est donc peu intéressant de raisonner avec un tel type de mesure dont la pratique risque d'être complexe. Je montre en sous-section 3.2.2, qu'en fixant des vecteurs de poids particuliers \mathbf{w} , il est possible d'obtenir des fonctions d'agrégation $F_{\mathbf{w}}$ dont la complexité de calcul peut être réduite drastiquement. De plus, cette simplification ne nuit pas à la capacité de cette approche à représenter des préférences complexes.

3.2 Contributions

3.2.1 Intégrale de Choquet bipolaire et 2-additive

Je commence par présenter des contributions concernant la tâche d'élicitation d'une intégrale de Choquet bipolaire (BCI). Une bicapacité normalisée ν sur $3^{\mathbb{N}}$ comporte $3^{\mathbb{N}} - 3$ paramètres. Estimer l'ensemble des coefficients de ν requiert des ressources computationnelles importantes voire insurmontables si N est trop grand. Une stratégie permettant d'outrepasser cette difficulté est de nature théorique et consiste à travailler avec un sous-ensemble de bicapacités dont la complexité est moindre. Il s'agit des **bicapacités k -additives** définies dans [Grabisch and Labreuche, 2005a, Grabisch and Labreuche, 2005b].

Avant de pouvoir présenter cette propriété, il est nécessaire d'introduire au préalable la **transformée de Möbius**. Initialement, il s'agit d'une application bijective définie sur les fonctions d'ensemble sur $2^{\mathbb{N}}$ et qui intervient dans l'étude des propriétés d'une capacité et de l'intégrale de Choquet unipolaire. Cette transformée a été généralisée dans le cas des bicapacités et de la BCI par [Grabisch and Labreuche, 2003] et [Fujimoto, 2004]. Il s'agit de deux approches distinctes mais leur équivalence a été établie par la suite dans [Fujimoto and Murofushi, 2005]. Nous avons préféré utiliser l'approche dite **transformée bipolaire de Möbius** (dénotée BMT pour *Bipolar Möbius Transform*), introduite par Katsushige Fujimoto [Fujimoto, 2004] et que je rappelle ci-dessous. Soit ν une bicapacité sur $3^{\mathbb{N}}$. La BMT de ν , dénotée par \mathbf{b}^{ν} , est une fonction d'ensemble sur $3^{\mathbb{N}}$ à valeurs dans \mathbb{R} et qui pour

3.2. CONTRIBUTIONS

tout $(A_1, A_2) \in 3^{\mathbb{N}}$ est donnée par :

$$\begin{aligned} \mathbf{b}^\nu(A_1, A_2) &= \sum_{\substack{B_1 \subseteq A_1 \\ B_2 \subseteq A_2}} (-1)^{|A_1 \setminus B_1| + |A_2 \setminus B_2|} \nu(B_1, B_2) \\ &= \sum_{(\emptyset, A_2) \sqsubseteq (B_1, B_2) \sqsubseteq (A_1, \emptyset)} (-1)^{|A_1 \setminus B_1| + |A_2 \setminus B_2|} \nu(B_1, B_2), \end{aligned} \quad (3.15)$$

où $A \setminus B$ est le sous-ensemble des éléments de A qui ne sont pas dans B et $|A|$ est le cardinal de A .

Réciproquement, pour tout $(A_1, A_2) \in 3^{\mathbb{N}}$, nous avons :

$$\nu(A_1, A_2) = \sum_{\substack{B_1 \subseteq A_1 \\ B_2 \subseteq A_2}} \mathbf{b}^\nu(B_1, B_2). \quad (3.16)$$

Remarquons que si $\nu(\emptyset, \emptyset) = 0$ alors nous avons nécessairement :

$$\mathbf{b}^\nu(\emptyset, \emptyset) = 0. \quad (3.17)$$

On peut alors exprimer de façon équivalente la BCI donnée par 3.7 en fonction de \mathbf{b}^ν :

$$\text{BCI}_{\mathbf{b}^\nu}(\mathbf{x}) = \sum_{(A_1, A_2) \in 3^{\mathbb{N}}} \mathbf{b}^\nu(A_1, A_2) \left(\bigwedge_{i \in A_1} x_i^+ \wedge \bigwedge_{j \in A_2} x_j^- \right) \quad (3.18)$$

où $\begin{cases} x_i^+ = x_i & \text{si } x_i > 0 \\ x_i^+ = 0 & \text{si } x_i \leq 0 \end{cases}$, et $\begin{cases} x_i^- = -x_i & \text{si } x_i < 0 \\ x_i^- = 0 & \text{si } x_i \geq 0 \end{cases}$, et $\bigwedge_{i \in A} x_i = \min((x_i)_{i \in A})$.

Ensuite, une bicapacité est **k -additive** pour $k \in \{1, \dots, N\}$ si et seulement si les deux conditions suivantes sont vérifiées :

- $\forall (A_1, A_2) \in 3^{\mathbb{N}} : |A_1 \cup A_2| > k \Rightarrow \mathbf{b}^\nu(A_1, A_2) = 0$.
- $\exists (A_1, A_2) \in 3^{\mathbb{N}} : |A_1 \cup A_2| = k \wedge \mathbf{b}^\nu(A_1, A_2) \neq 0$.

Autrement dit, ν est k -additive si sa BMT \mathbf{b}^ν a des valeurs non nulles uniquement pour les paires de sous-ensembles dont l'union est de cardinal plus petit ou égale à k . Pour k petit, les valeurs de \mathbf{b}^ν sont nulles en grande majorité. Toutefois, il est important de souligner qu'une bicapacité ν k -additive n'est en aucun cas parcimonieuse.

Dans le cas particulier $k = 1$ on parle de bicapacité additive (ou 1-additive) et on a alors :

$$\forall (A_1, A_2) \in 3^{\mathbb{N}} : \nu(A_1, A_2) = \sum_{i \in A_1} \nu(\{i\}, \emptyset) + \sum_{j \in A_2} \nu(\emptyset, \{j\}).$$

Les bicapacités additives se réduisent à des règles de décision linéaires où les critères sont supposés indépendants. Avec Brice et Antoine, nous nous sommes donc focalisés sur le cas $k = 2$, et dans ce cas une **bicapacité 2-additive (2ABC pour 2-Additive Bicapacity)**

3.2. CONTRIBUTIONS

vérifie :

$$\forall (A_1, A_2) \in 3^{\mathbb{N}} : |A_1| + |A_2| > 2 \Rightarrow \mathbf{b}^\nu(A, B) = 0, \quad (3.19)$$

$$\exists (A_1, A_2) \in 3^{\mathbb{N}} : |A_1| + |A_2| = 2 \wedge \mathbf{b}^\nu(A, B) \neq 0. \quad (3.20)$$

Je donne dans la Figure 3.1 une matrice indiquant schématiquement, pour $\mathbb{N} = \{1, 2, 3\}$, les paires d'éléments de $3^{\mathbb{N}}$ et la différence entre la transformée bipolaire de Möbius d'une bicapacité quelconque et celle d'une 2ABC. Les paires associées à un symbole \cdot ne sont pas des éléments de $3^{\mathbb{N}}$. Celles qui sont marquées du symbole \checkmark concernent les éléments non nuls de \mathbf{b}^ν dans le cas 2-additif. Par conséquent, les couples indiqués par $*$ sont les paires de $3^{\mathbb{N}}$ pour lesquelles \mathbf{b}^ν est nulle si ν est 2-additif.

(A, B)	\emptyset	1	2	3	12	13	23	123
\emptyset	\checkmark	$*$						
1	\checkmark	\cdot	\checkmark	\checkmark	\cdot	\cdot	$*$	\cdot
2	\checkmark	\checkmark	\cdot	\checkmark	\cdot	$*$	\cdot	\cdot
3	\checkmark	\checkmark	\checkmark	\cdot	$*$	\cdot	\cdot	\cdot
12	\checkmark	\cdot	\cdot	$*$	\cdot	\cdot	\cdot	\cdot
13	\checkmark	\cdot	$*$	\cdot	\cdot	\cdot	\cdot	\cdot
23	\checkmark	$*$	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
123	$*$	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot

FIGURE 3.1 – Représentation sous forme matricielle de $3^{\mathbb{N}}$ (paires marquées par \checkmark et $*$), des valeurs de \mathbf{b}^ν lorsque ν est quelconque (\checkmark et $*$) et lorsqu'elle est 2-additive (\checkmark uniquement).

Dans cet exemple avec $N = 3$, en cas de 2-additivité, \mathbf{b}^ν comporterait 19 valeurs non nulles contre 27 si la bicapacité ν était quelconque. La diminution du nombre de paramètres semble limitée à première vue, mais dans le cas général, une bicapacité ν 2-additive comporte $2N^2 - 1$ valeurs non nulles pour sa transformée bipolaire de Möbius, contre $3^N - 1$ dans le cas quelconque. La 2-additivité fait ainsi passer le problème de l'estimation d'une bicapacité d'un ordre exponentiel à un ordre quadratique.

Par la suite, je vais présenter différentes expressions et propriétés faisant intervenir la BMT. Ceci permet de rendre explicite la complexité réduite de la 2ABC et de faciliter la modélisation du problème d'élicitation. Afin d'alléger les expressions, j'introduis les notations suivantes, $\forall i, j \in \mathbb{N}$:

$$\left\{ \begin{array}{l} \mathbf{b}_{i|}^\nu = \mathbf{b}^\nu(\{i\}, \emptyset), \\ \mathbf{b}_{i|j}^\nu = \mathbf{b}^\nu(\{i, j\}, \emptyset), \\ \mathbf{b}_{i|j}^\nu = \mathbf{b}^\nu(\{i\}, \{j\}). \end{array} \right. \quad \text{et} \quad \left\{ \begin{array}{l} \nu_{i|} = \nu(\{i\}, \emptyset), \\ \nu_{i|j} = \nu(\{i, j\}, \emptyset), \\ \nu_{i|j} = \nu(\{i\}, \{j\}). \end{array} \right.$$

L'intégrale de Choquet bipolaire par rapport à une 2ABC ν , mais exprimée de manière

équivalente en fonction de \mathbf{b}^ν , est donnée par la formule suivante :

$$\begin{aligned} \text{BCI}_{\mathbf{b}^\nu}(\mathbf{x}) &= \sum_{i=1}^N \mathbf{b}_{|i}^\nu x_i^+ + \sum_{i=1}^N \mathbf{b}_{|i}^\nu x_i^- + \sum_{i,j=1}^N \mathbf{b}_{|ij}^\nu (x_i^+ \wedge x_j^-) \\ &+ \sum_{\{i,j\} \subseteq \mathbb{N}} \mathbf{b}_{|ij}^\nu (x_i^+ \wedge x_j^+) + \sum_{\{i,j\} \subseteq \mathbb{N}} \mathbf{b}_{|ij}^\nu (x_i^- \wedge x_j^-). \end{aligned} \quad (3.21)$$

A des fins de modélisation, il est également important de déterminer les conséquences sur \mathbf{b}^ν des différentes propriétés de ν dans le cas 2-additif. Nous avons alors les résultats suivants :

- Si ν est 2-additive alors ν est normalisée si et seulement si :

$$\begin{cases} \sum_{i \in \mathbb{N}} \mathbf{b}_{|i}^\nu + \sum_{\{i,j\} \subseteq \mathbb{N}} \mathbf{b}_{|ij}^\nu = 1, \\ \sum_{i \in \mathbb{N}} \mathbf{b}_{|i}^\nu + \sum_{\{i,j\} \subseteq \mathbb{N}} \mathbf{b}_{|ij}^\nu = -1. \end{cases} \quad (3.22)$$

- Si ν est 2-additive alors ν est monotone croissante si et seulement si :

$$\begin{cases} \forall (A, B) \in 3^{\mathbb{N}} \text{ tel que } |A| + |B| > 2, \forall k \in A : \mathbf{b}_{|k}^\nu + \sum_{j \in B} \mathbf{b}_{|kj}^\nu + \sum_{i \in A \setminus k} \mathbf{b}_{|ik}^\nu \geq 0, \\ \forall (A, B) \in 3^{\mathbb{N}} \text{ tel que } |A| + |B| > 2, \forall k \in A : \mathbf{b}_{|k}^\nu + \sum_{j \in B} \mathbf{b}_{|j|k}^\nu + \sum_{i \in A \setminus k} \mathbf{b}_{|ik}^\nu \leq 0. \end{cases} \quad (3.23)$$

Je discute maintenant des différentes procédures d'élicitation d'une 2ABC que nous avons étudiées avec Brice et Antoine. Nous avons défini trois procédures. La première se situe dans la continuité du travail de thèse de Brice. Nous avons proposé un cadre exclusivement ancré dans le champ de l'AMCD où l'on suppose que l'on peut collecter auprès du décideur des préférences de nature ordinale dans le cadre d'une procédure itérative [Mayag et al., 2012].

Dans ce cas, je dénote par \mathbb{A}_i l'ensemble des valeurs que peut prendre le critère i . Les valeurs de \mathbb{A}_i peuvent appartenir à plusieurs types d'échelle : nominale, ordinale, numérique. Prenons pour illustration, le cas où le décideur doit choisir une nouvelle voiture et parmi l'ensemble des critères, supposons un cas discret avec $i = \text{Carburant}$. Dans ce cas, l'ensemble des attributs de i pourrait être $\mathbb{A}_i = \{\text{hybride, électrique, diesel, essence, GPL}\}$. Dans l'approche MAUT, les scores x_i sont obtenus par des fonctions d'utilités bipolaires partielles $u_i : \mathbb{A}_i \mapsto [-1, 1]$. Dans notre illustration précédente, un décideur soucieux de la pollution de l'air indiquerait par exemple : $u_i(\text{électrique}) = 1$, $u_i(\text{hybride}) = 0.5$, $u_i(\text{GPL}) = 0$, $u_i(\text{essence}) = -0.7$, $u_i(\text{diesel}) = -1$.

La procédure d'élicitation impliquant l'agent se décline alors en 4 étapes :

1. Déterminer avec le décideur, **3 niveaux de référence** pour chaque critère $i \in \mathbb{N}$ qui sont des attributs de \mathbb{A}_i définis respectivement comme suit :

- Un niveau de référence 1_i que le décideur considère comme être complètement satisfaisant.

3.2. CONTRIBUTIONS

- Un niveau de référence 0_i que le décideur considère comme neutre.
- Un niveau de référence -1_i que le décideur considère comme totalement insatisfaisant.

On pose de plus par simplicité, $u_i(1_i) = 1$, $u_i(0_i) = 0$ et $u_i(-1_i) = -1$. En reprenant l'exemple $i = \text{Carburant}$, on aurait donc dans ce cas $1_i = \text{électrique}$, $0_i = \text{GPL}$, $-1_i = \text{diesel}$.

2. Construire un ensemble d'**alternatives (ou actions) ternaires** qui sont des alternatives fictives dont les attributs pour chaque critère i sont fixés à des niveaux de référence 1_i ou 0_i ou -1_i . Plus précisément, on appelle une alternative ternaire un élément parmi l'ensemble $\{\mathbf{x}_|, \mathbf{x}_{i|}, \mathbf{x}_{|i}, \mathbf{x}_{i|j}, \mathbf{x}_{j|i}, \mathbf{x}_{ij|}\}$ où :

- $\mathbf{x}_|$ est une alternative d'attributs 0_k sur tous les critères $k = 1, \dots, N$.
- $\mathbf{x}_{i|}$ est une alternative d'attribut 1_i sur le critère i et d'attributs 0_k sur tous les autres critères $k \neq i$.
- $\mathbf{x}_{|i}$ est une alternative d'attribut -1_i sur le critère i et d'attributs 0_k sur tous les autres critères $k \neq i$.
- $\mathbf{x}_{i|j}$ est une alternative d'attributs 1_i et -1_j sur les critères i et j respectivement et d'attributs 0_k sur tous les autres critères $k \neq i, j$.
- $\mathbf{x}_{j|i}$ est une alternative d'attributs 1_i et 1_j sur les critères i et j et d'attributs 0_k sur tous les autres critères $k \neq i, j$.
- $\mathbf{x}_{ij|}$ est une alternative d'attributs -1_i et -1_j sur les critères i et j et d'attributs 0_k sur tous les autres critères $k \neq i, j$.

La terminologie "alternative ternaire" vient du fait que l'on construit des cas dont les profils de scores (ou d'utilités partielles) ne comportent que 3 types de valeurs $\{-1, 0, 1\}$. Par ailleurs, en raison de (3.21) et (3.16), la valeur de $\text{BCI}_{\mathbf{b}^\nu}$ pour une alternative ternaire est égale à la valeur unitaire de ν pour le cas considéré. Par exemple, $\text{BCI}_{\mathbf{b}^\nu}(\mathbf{x}_{i|j}) = \mathbf{b}_{i|}^\nu + \mathbf{b}_{j|}^\nu + \mathbf{b}_{i|j}^\nu = \nu_{i|j}$.

3. Demander au décideur ses préférences pour chaque paire d'alternatives ternaires. Dans ce cas, étant donné deux alternatives ternaires, il peut, soit préférer strictement l'un vis-à-vis de l'autre, soit être indifférent entre les deux. C'est la **collection d'information de nature ordinale**. Ces relations de préférences entre alternatives ternaires permettent d'inférer des relations d'ordre sur la bicapacité. Par exemple si le décideur indique $\mathbf{x}_{i|j} \succ \mathbf{x}_{k|l}$ ($\mathbf{x}_{i|j}$ est strictement préféré à $\mathbf{x}_{k|l}$) cela implique $\nu_{i|j} > \nu_{k|l}$.
4. Déterminer les paramètres du modèle de préférence à l'aide d'un **modèle d'optimisation qui intègre dans ses contraintes les différentes propriétés requises et les informations ordinales fournies par l'agent**. La programmation par contraintes peut être une approche possible tout comme le modèle Split ci-dessous.

3.2. CONTRIBUTIONS

Dans ce qui suit, j'introduis les deux autres approches d'élicitation des paramètres d'une 2ABC. Il s'agit de deux **modèles d'optimisation dénotés Split et Rss**. Dans ces deux cas, et contrairement à la première procédure, on suppose que le décideur fournit pour un ensemble d'alternatives \mathbb{M} , les utilités partielles (ou scores de satisfaction) pour chaque critère $i \in \mathbb{N}$. De plus, pour un sous-ensemble $\mathbb{M}' \subseteq \mathbb{M}$, on fait l'hypothèse que le décideur procure un score global. La notation $\mathbf{x} \in \mathbb{R}^N$, représente le vecteur des utilités partielles (ou profil de scores) (x_1, \dots, x_N) d'une alternative $X \in \mathbb{M}$. Si de plus, $X \in \mathbb{M}'$, alors son score global est dénoté par y . Notons que ce contexte est proche de celui rencontré en apprentissage supervisé où nous disposons d'exemples numériques permettant de faire de l'inférence de paramètres de modèles.

Dans le cadre du **modèle Split**, on suppose de plus que le décideur procure, pour plusieurs paires d'alternatives (X, X') dans $(\mathbb{M} \setminus \mathbb{M}') \times (\mathbb{M} \setminus \mathbb{M}')$, des relations de préférences strictes ou d'indifférence qui sont dénotées respectivement par $X \succ X'$ et $X \sim X'$. Notons que pour toute paire $(X, X') \in \mathbb{M}' \times \mathbb{M}'$, nous pouvons inférer ces relations en comparant les scores globaux y et y' .

Dans ce contexte, la fonction objectif du modèle **Split**, introduit initialement par Jean-Luc Marichal et Marc Roubens dans [Marichal and Roubens, 2000], consiste à maximiser l'écart $\text{BCI}_{\mathbf{b}'}(\mathbf{x}) - \text{BCI}_{\mathbf{b}'}(\mathbf{x}')$ pour toute paire $(X, X') \in \mathbb{M} \times \mathbb{M}$ tel que $X \succ X'$.

Plus formellement, il s'agit de résoudre le programme mathématique suivant où les variables sont les composantes non nulles de \mathbf{b}' :

$$\begin{aligned} & \max \epsilon & (3.24) \\ \text{s.l.c.} & \left\{ \begin{array}{l} \epsilon > 0, \\ \forall (X, X') \in \mathbb{M} \times \mathbb{M} \text{ tel que } X \succ X' : \text{BCI}_{\mathbf{b}'}(\mathbf{x}) - \text{BCI}_{\mathbf{b}'}(\mathbf{x}') \geq \epsilon, \\ (3.17), \\ (3.21), \\ (3.22), \\ (3.23). \end{array} \right. \end{aligned}$$

Dans le calcul de l'intégrale de Choquet bipolaire avec une 2ABC donnée par l'équation (3.21), on peut pré-calculer les valeurs positives x_i^+ et x_i^- ainsi que les minima. Par conséquent, la fonction objectif et l'ensemble des contraintes de (3.24) sont linéaires et le modèle **Split** est un programme linéaire.

Dans la troisième **approche Rss**, on traite la tâche comme un **problème de régression**. La fonction objectif dépend des éléments de \mathbb{M}' et est égale à la somme des carrés des écarts entre la prédiction donnée par la BCI et les scores globaux indiqués par l'agent. Des informations ordinales peuvent être incorporées dans la modélisation mais en termes de contraintes comme pour le modèle **Split** ci-dessus. L'approche conduit à un problème quadratique convexe

3.2. CONTRIBUTIONS

et il se formalise comme suit :

$$\begin{aligned} & \min_{\mathbf{b}^{\nu}} \sum_{X \in \mathbb{M}'} (\text{BCI}_{\mathbf{b}^{\nu}}(\mathbf{x}) - y)^2 & (3.25) \\ \text{s.l.c.} & \left\{ \begin{array}{l} (3.17), \\ (3.21), \\ (3.22), \\ (3.23), \\ \forall (X, X') \in \mathbb{M} \times \mathbb{M} \text{ tel que } X \succ X' : \text{BCI}_{\mathbf{b}^{\nu}}(\mathbf{x}) - \text{BCI}_{\mathbf{b}^{\nu}}(\mathbf{x}') \geq \epsilon. \end{array} \right. \end{aligned}$$

Contrairement au modèle **Split**, ϵ est ici un paramètre non négatif fixé par le décideur et non pas une variable. Par défaut, on pourra prendre $\epsilon = 0$.

Il est important d'indiquer que pour les modèles **Split** et **Rss**, les contraintes d'inégalités issues des informations ordinales procurées par le décideur, peuvent conduire à des **incohérences**. En effet, prenons le cas où le décideur préfère strictement X à X' , $X \succ X'$ mais que les profils de scores \mathbf{x} et \mathbf{x}' sont tels que $x_i \leq x'_i, \forall i = 1, \dots, N$. Dans la mesure où la BCI est monotone croissante, il est donc impossible dans ce cas de satisfaire $\text{BCI}_{\mathbf{b}^{\nu}}(\mathbf{x}) - \text{BCI}_{\mathbf{b}^{\nu}}(\mathbf{x}') \geq \epsilon > 0$, et le problème est alors impossible à résoudre.

La violation de ce type de contrainte renvoie aux **concepts de soft margin et soft error** pour les problèmes non linéairement séparables dans le cas du modèle SVM en apprentissage supervisé. J'ai donc proposé à Antoine et Brice d'intégrer dans les modèles précédents des *slack variables* $\xi_{XX'} \geq 0$ pour toute paire (X, X') , permettant de **relaxer les contraintes sur les informations ordinales en cas de préférences incohérentes (inconsistencies)**. De plus, similairement au principe de *soft error*, **un terme de pénalisation est ajouté à la fonction objectif** afin de limiter autant que possible l'impact de ces préférences incohérentes. Je donne ci-dessous l'extension du modèle **Rss** qui tient compte des *inconsistencies*. Celle-ci est dénotée **Rss – flex**. La même démarche peut être appliquée pour définir le modèle **Split – flex**.

$$\begin{aligned} & \min_{\mathbf{b}^{\nu}} \sum_{\mathbf{x} \in \mathbb{M}'} (\text{BCI}_{\mathbf{b}^{\nu}}(\mathbf{x}) - y)^2 + \sum_{(X, X') : X \succ X'} \xi_{XX'} & (3.26) \\ \text{s.l.c.} & \left\{ \begin{array}{l} (3.17), \\ (3.21), \\ (3.22), \\ (3.23), \\ \forall (X, X') \in \mathbb{M} \times \mathbb{M} \text{ tel que } X \succ X' : \text{BCI}_{\mathbf{b}^{\nu}}(\mathbf{x}) - \text{BCI}_{\mathbf{b}^{\nu}}(\mathbf{x}') \geq \epsilon - \xi_{XX'}, \\ \forall (X, X') \in \mathbb{M} \times \mathbb{M} \text{ tel que } X \succ X' : \xi_{XX'} \geq 0. \end{array} \right. \end{aligned}$$

Pour la mise en pratique, l'ensemble des modèles d'optimisation **Split**, **Rss**, **Split – flex** et **Rss – flex** ont été implémenté par mes soins en **AMPL (A Mathematical Programming Language)**. Il s'agit d'un langage de nature algébrique permettant de décrire un problème

3.2. CONTRIBUTIONS

d'optimisation dans une syntaxe proche du langage mathématique et qui peut être alors traité par de nombreux solveurs libres (GLPK par exemple) ou propriétaires (CPLEX, MINOS par exemple).

J'illustre le comportement de ces modèles sur un exemple simple de [Grabisch et al., 2008]. Il s'agit de notes sur $N = 5$ matières obtenues par un ensemble de $M = 7$ étudiants. Les matières qui jouent le rôle de critère, sont les suivantes : statistiques (S), probabilité (P), économie (E), management (M), et anglais (En). Les notes sont exprimées dans l'échelle classique $[0, 20]$ mais dans cet exemple, elles ne varient que dans l'intervalle $[11, 18]$. On transforme ces notes pour qu'elles s'expriment dans une échelle bipolaire. On suppose le contexte suivant : un élève obtient son diplôme avec mention si sa note globale est supérieure ou égale à 14. On retranche alors 14 à toutes les valeurs. Suite à cette translation, les notes appartiennent à $[-3, 4]$ et on attribue la mention à condition que la note agrégée soit non négative. Remarquons qu'ici, on utilise, sans perte de généralité, des degrés d'insatisfaction et de satisfaction qui ne sont pas dans l'intervalle $[-1, 1]$.

La Table 3.1 indique les données initiales présentées dans [Grabisch et al., 2008]. Les notes translattées sont montrées dans la sous-table (a). Dans la sous-table (b), la première colonne y indique les notes agrégées fournis par le cas d'étude. Les résultats de **Split** et **Split – flex** sont exposées ensuite. On constate que les notes ne sont pas reproduites. Ce phénomène est attendu puisque la fonction objectif consiste à respecter les relations ordinales, et à maximiser l'écart entre les scores prédits. On remarquera par exemple, que l'écart initial entre X_a et X_g est de $1 - (-2) = 3$ tandis qu'il est de $1.68 - (-2.14) = 3.82$ et $1.02 - (-2.8) = 3.82$ pour **Split** et **Split – flex** respectivement. De façon plus générale, les écarts entre les scores prédits pour chaque couple sont les mêmes pour **Split** et **Split – flex**. En fait, ces modèles ne sont pas en général strictement convexes et il existe donc plusieurs solutions optimales possibles mais équivalentes entre elles. Par ailleurs, dans cet exemple, il n'y a pas de préférence incohérente. Ainsi, la somme des variables d'écarts $\sum_{(X, X') : X > X'} \xi_{XX'}$ est nulle et n'a donc aucun impact sur la fonction objectif.

Student	S	P	E	M	En	y	Split	Split flex	Rss	Rss flex
X_a	4	-3	-3	-3	4	1	1.68	1.02	1	1
X_b	4	-3	4	-3	-3	0.5	1.04	0.38	0.5	0.5
X_c	-3	-3	4	-3	4	0	0.41	-0.25	0	0
X_d	4	4	-3	-3	-3	-0.5	-0.23	-0.89	-0.5	-0.5
X_e	-3	-3	4	4	-3	-1	-0.86	-1.53	-1	-1
X_f	-3	-3	4	-3	-3	-1.5	-1.5	-2.16	-1.5	-1.5
X_g	-3	-3	-3	-3	4	-2	-2.14	-2.8	-2	-2

TABLE 3.1 – (a) Table des notes partielles translattées; (b) Note agrégée translattée et prédictions des différents modèles.

3.2. CONTRIBUTIONS

Les méthodes Rss et Rss – flex visent, au contraire, à approximer les scores donnés par le décideur. Dans cet exemple, l’erreur des moindres carrés est nul pour les deux modèles qui, comme précédemment, ne sont pas strictement convexes et produisent des 2ABC distinctes. De façon similaire, le fait qu’il n’y ait pas d’*inconsistencies* implique que les deux modèles Rss et Rss – flex sont équivalents.

Afin d’illustrer le comportement des approches Split – flex et Rss – flex en cas de préférences incohérentes, nous avons altéré la valeur du score agrégé de l’individu X_g ce qui est indiqué en rouge dans la Table 3.2. Si on compare les profils de score de X_c et X_g , on voit que le premier domine le second : $X_c \succ X_g$. Or dans la Table 3.2, le score agrégé de X_g , 0.5, est plus grand que celui de X_c , 0. Ceci montre que les préférences d’un tel décideur ne sont pas monotones croissantes. Clairement, les modèles Split et Rss sont alors insolubles. Les versions flex peuvent en revanche gérer ces incohérences et produire malgré tout une BCI qui approxime le modèle de préférence du décideur.

Student	S	P	E	M	En	y'	Split	Split flex	Rss	Rss flex
X_a	4	-3	-3	-3	4	1	.	0.22	.	1.12
X_b	4	-3	4	-3	-3	0.5	.	-0.28	.	0.62
X_c	-3	-3	4	-3	4	0	.	-0.78	.	0.12
X_d	4	4	-3	-3	-3	-0.5	.	-1.28	.	-0.5
X_e	-3	-3	4	4	-3	-1	.	-1.78	.	-1
X_f	-3	-3	4	-3	-3	-1.5	.	-2.28	.	-1.5
X_g	-3	-3	-3	-3	4	0.5	.	-0.78	.	0.12

(a) (b')

TABLE 3.2 – (a) Table des notes partielles translattées; (b) Note agrégée translattée avec modification pour l’élève g pour un cas d’*inconsistency* et prédictions des différents modèles.

Je donne dans la Table 3.3, les valeurs estimées des BMT des 2ABC des modèles Split – flex et Rss – flex dans le cas des données présentant des préférences incohérentes. Remarquons que dans nos modèles, les termes $b^\nu(A, B)$ pour tout couple $(A_1, A_2) \in 3^{\mathbb{N}} : |A_1| + |A_2| > 2$ ne sont pas représentés. Par conséquent, la 2-additivité est implicitement imposée et la BMT obtenue ne peut donc pas être k -additive pour $k \geq 3$. En revanche, la condition (3.20) qui impliquerait que $b^\nu(A, B)$ soit exactement 2-additive n’est pas inscrite dans nos modèles. Ainsi, la solution pourrait également être 1-additive si l’inférence aboutissait à $b^\nu(A, B) = 0$ pour tout $(A_1, A_2) \in 3^{\mathbb{N}} : |A_1| + |A_2| = 2$.

Intuitivement, la valeur k de la **k -additivité de ν indique un niveau de complexité de la classe d’hypothèses associée à une BCI**. Par principe de rasoir d’Occam et également afin d’éviter des problèmes de sur-apprentissage, il serait intéressant d’encourager l’inférence d’un modèle moins complexe au sein d’une classe d’hypothèses flexible. J’ai introduit cette idée à l’intersection entre l’AMCD et le *machine learning* dans [Ah-Pine et al., 2013].

3.2. CONTRIBUTIONS

		Split – flex		Rss – flex	
A_1	A_2	$b^\nu(A_1, A_2)$	$b^\nu(A_2, A_1)$	$b^\nu(A_1, A_2)$	$b^\nu(A_2, A_1)$
\emptyset	S			-0.077	0.47
\emptyset	P	-0.19		-0.53	0.077
\emptyset	E			-0.077	0.023
\emptyset	M		0.031	-0.14	0.077
\emptyset	En	-0.031			0.19
\emptyset	SP	-0.26		0.077	0.09
\emptyset	SE		0.19		0.263
\emptyset	SM		0.73		
\emptyset	SEn	-0.16	0.055		-0.018
\emptyset	PE				
\emptyset	PM			0.14	
\emptyset	PEn	0.031			-0.077
\emptyset	EM				
\emptyset	EEn	-0.24		-0.39	-0.023
\emptyset	MEEn	-0.16			-0.077
S	P			0.06	
S	E				
S	M		-0.031	-0.33	
S	En				
P	E				
P	M				0.14
P	En		0.19	-0.077	0.31
E	M				
E	En				0.077
M	En			-0.077	

TABLE 3.3 – Valeurs estimées de b^ν des le cas des données de la Table 3.2.

Je reviendrai sur celle-ci en section 3.3.

3.2.2 Une nouvelle famille de fonctions d’agrégation

Par commodité, je commence par rappeler l’application générique $F_{\mathbf{w}}$ introduite dans (3.9) :

$$F_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^N w_k \mu_{E_k^N}(\mathbf{x}),$$

où, $\mu_{E_k^N}(\mathbf{x})$ est la mesure d’appartenance de l’alternative X , représentée par le profil de scores \mathbf{x} , au sous-ensemble flou $E_k^N =$ “au moins k critères sur N sont satisfaits” et \mathbf{w} est un vecteur de poids, c’est-à-dire une collection de N valeurs non négatives et sommant à 1.

Il est important d’indiquer que le vecteur $\mathbf{w} = (w_1, \dots, w_N)$ n’attribue pas des poids à chaque critère mais à des mesures d’“événements emboîtés” E_k^N où k indique le nombre

3.2. CONTRIBUTIONS

minimal de critères qui doit être satisfait.

J'ai rappelé précédemment la fonction d'agrégation $\text{TOWA}_{\mathbf{w},\mathbb{T}}$ définie dans [Yager, 2005], qui est un cas particulier de $F_{\mathbf{w}}$ où on utilise une mesure maxitive μ sur \mathcal{S} . Contrairement à cette approche, je propose d'employer une **fonction d'ensemble finiment additive** qui vérifie $\mu_{S \cup S'} = \mu_S + \mu_{S'} - \mu_{S \cap S'}$ pour deux sous-ensembles flous $S, S' \in \mathcal{S}$. En particulier, si S et S' sont disjoints on a, $\mu_{S \cup S'} = \mu_S + \mu_{S'}$. Toutefois, il est judicieux de mentionner que selon la ou les *t-norm* utilisées, on n'est pas garantie que $\forall S \in \mathcal{S} : \mu_S \in [0, 1]$ (voir par exemple [Perovic et al., 2011]). Par conséquent, on parlera de fonction d'ensemble plutôt que de mesure μ sur \mathcal{S} .

Le cas $k = 1$ correspond à la **formule d'Henri Poincaré** également connu comme le **principe d'inclusion-exclusion** ou la **formule du crible**. Afin de rappeler cette expression (et sa généralisation), j'introduis les notions de sommes symétriques dénotées et définies comme suit, $\forall k = 1, \dots, N$:

$$\begin{aligned} \Sigma_k^N &= \sum_{1 \leq i_1 < \dots < i_k \leq N} \mu_{S_{i_1} \cap \dots \cap S_{i_k}} \\ &= \sum_{1 \leq i_1 < \dots < i_k \leq N} \mathbb{T}(\mu_{S_{i_1}}, \dots, \mu_{S_{i_k}}), \end{aligned} \quad (3.27)$$

où \mathbb{T} est une *t-norm*.

En considérant μ comme étant une mesure finiment additive et en utilisant les notations exposées précédemment, la formule de Henri Poincaré s'exprime alors par :

$$\mu_{E_1^N} = \sum_{1 \leq l \leq N} (-1)^{l-1} \Sigma_l^N \quad (3.28)$$

Cette formule a été généralisée par Charles Jordan [Jordan, 1926, Jordan, 1939] et Lajos Takács [Takács, 1967] dans les cas $k > 1$ et nous avons, $\forall k = 1, \dots, N$:

$$\mu_{E_k^N} = \sum_{k \leq l \leq N} (-1)^{l-k} \binom{l-1}{k-1} \Sigma_l^N. \quad (3.29)$$

L'approche que je propose sera dénotée $A_{\mathbf{w},\mathbb{T}}$. Elle correspond au cas particulier de la fonction générique (3.9) qui utilise une fonction d'ensemble μ sur \mathcal{S} qui est finiment additive. En injectant les formules de Charles Jordan (3.29), l'approche s'exprime comme suit :

$$A_{\mathbf{w},\mathbb{T}} = \sum_{1 \leq k \leq N} w_k \sum_{k \leq l \leq N} (-1)^{l-k} \binom{l-1}{k-1} \Sigma_l^N. \quad (3.30)$$

De façon générale, la complexité du calcul de $A_{\mathbf{w},\mathbb{T}}$ est d'ordre exponentiel puisqu'il est globalement nécessaire d'énumérer les 2^N sous-ensembles de \mathbb{N} .

J'étudie alors des **cas particuliers pour \mathbb{T} d'abord puis pour \mathbf{w} ensuite**, afin d'analyser des propriétés intéressantes du modèle. Dans le cas où je fais varier \mathbb{T} , je suis resté prudent

3.2. CONTRIBUTIONS

en raison du fait que si μ est une fonction d'ensemble finiment additive, alors selon la t -norm employée, on n'a pas de garantie que $\mu_S \in [0, 1]$ pour tout $S \in \mathcal{S}$. Je me suis limité aux cas où $\mathsf{T} = \mathsf{T}_M$ et $\mathsf{T} = \mathsf{T}_P$. J'ai alors établi les résultats suivants dont les preuves détaillées sont données dans [Ah-Pine, 2016].

Théorème. 1. *Si $\mathsf{T} = \mathsf{T}_M$ pour toutes les intersections de sous-ensembles flous $S_{i_1} \cap \dots \cap S_{i_k}$ avec $1 \leq i_1 < \dots < i_k \leq N$ et $k = 1, \dots, N$ alors $A_{\mathbf{w}, \mathsf{T}_M} = \text{TOWA}_{\mathbf{w}, \mathsf{T}_M} = \text{OWA}_{\mathbf{w}}$.*

Ainsi tout comme $\text{TOWA}_{\mathbf{w}, \mathsf{T}_M}$, mon approche $A_{\mathbf{w}, \mathsf{T}_M}$ avec comme paramètre $\mathsf{T} = \mathsf{T}_M = \min$, englobe la fonction d'agrégation $\text{OWA}_{\mathbf{w}}$ qui elle-même généralise les statistiques d'ordre et la moyenne arithmétique. Ce résultat a pour conséquence la propriété suivante.

Corollaire. 1. *Pour tout vecteur de poids \mathbf{w} tel que $w_k \geq 0, \forall k$ et $\sum_k w_k = 1$, l'application $A_{\mathbf{w}, \mathsf{T}_M}$ est une fonction d'agrégation et vérifie donc les conditions (3.2).*

Un résultat similaire est établi pour le cas $\mathsf{T} = \mathsf{T}_P$.

Théorème. 2. *Pour tout vecteur de poids \mathbf{w} tel que $w_k \geq 0, \forall k$ et $\sum_k w_k = 1$, l'application $A_{\mathbf{w}, \mathsf{T}_P}$ est une fonction d'agrégation et vérifie donc les conditions (3.2).*

Dans ce qui suit, c'est le vecteur \mathbf{w} que je fixe et je laisse libre le paramètre T . Je m'intéresse en particulier aux deux vecteurs de poids suivants :

$$\mathbf{w}^\uparrow = \frac{2}{N(N+1)}(1, 2, \dots, N), \quad (3.31)$$

$$\mathbf{w}^\downarrow = \frac{2}{N(N+1)}(N, N-1, \dots, 1). \quad (3.32)$$

Le vecteur \mathbf{w}^\uparrow affecte un poids qui croît linéairement en fonction de k et donne ainsi de plus en plus d'importance aux événements E_k^N lorsque k augmente. L'évènement qui traduit l'unanimité, E_N^N , a le poids le plus important et le comportement de $A_{\mathbf{w}^\uparrow, \mathsf{T}}$ est plutôt de type conjonctif. Le cas \mathbf{w}^\downarrow a la sémantique opposée et dans ce cas, c'est l'évènement E_1^N = "au moins 1 critère parmi N " qui a le plus d'importance. Dans cette situation, on est plus proche d'un comportement disjonctif.

Je montre que l'application de ces deux vecteurs de poids aboutit à des **cas particuliers de $A_{\mathbf{w}, \mathsf{T}}$ dont le calcul est d'ordre quadratique.**

Proposition. 1. *Soient \mathbf{w}^\uparrow and \mathbf{w}^\downarrow les deux vecteurs de poids définis par (3.31) et (3.32)*

respectivement alors on a :

$$\begin{aligned} A_{\mathbf{w}^\uparrow, \mathbb{T}} &= \frac{1}{N(N+1)/2} \left(\sum_{1 \leq i \leq N} \mu_{S_i} + \sum_{1 \leq i < j \leq N} \mu_{S_i \cap S_j} \right) \\ &= \frac{1}{N(N+1)/2} \left(\sum_{1 \leq i \leq N} \mu_{S_i} + \sum_{1 \leq i < j \leq N} \mathbb{T}(\mu_{S_i}, \mu_{S_j}) \right), \end{aligned} \quad (3.33)$$

$$\begin{aligned} A_{\mathbf{w}^\downarrow, \mathbb{T}} &= \frac{1}{N(N+1)/2} \left(N \sum_{1 \leq i \leq N} \mu_{S_i} - \sum_{1 \leq i < j \leq N} \mu_{S_i \cap S_j} \right) \\ &= \frac{1}{N(N+1)/2} \left(N \sum_{1 \leq i \leq N} \mu_{S_i} - \sum_{1 \leq i < j \leq N} \mathbb{T}(\mu_{S_i}, \mu_{S_j}) \right). \end{aligned} \quad (3.34)$$

Cette expression réduite invite à définir une fonction d'agrégation encore plus flexible en permettant aux intersections entre deux sous-ensembles flous d'être modélisées par des t -norm distinctes. Ceci peut se faire par l'utilisation d'une famille paramétrique de t -norm comme celle de Maurice J. Frank que j'ai rappelée dans 3.12. Dans cette perspective, soit λ_{ij} la valeur du paramètre de la t -norm paramétrique \mathbb{T}_λ spécifiant la fonction d'appartenance floue de $S_i \cap S_j$ pour tout $i, j = 1, \dots, N$ tel que $i < j$:

$$\mu_{S_i \cap S_j} = \mathbb{T}_{\lambda_{ij}}(\mu_{S_i}, \mu_{S_j}). \quad (3.35)$$

Soit alors la matrice triangulaire supérieure $\mathbf{\Lambda} = \{\lambda_{ij}\}_{1 \leq i < j \leq N}$ composée de $N(N-1)/2$ valeurs réelles appartenant au domaine des paramètres de la famille paramétrique \mathbb{T}_λ . On peut alors définir les fonctions suivantes :

$$A_{\mathbf{w}^\uparrow, \mathbb{T}_\lambda, \mathbf{\Lambda}} = \frac{1}{N(N+1)/2} \left(\sum_{1 \leq i \leq N} \mu_{S_i} + \sum_{1 \leq i < j \leq N} \mathbb{T}_{\lambda_{ij}}(\mu_{S_i}, \mu_{S_j}) \right), \quad (3.36)$$

$$A_{\mathbf{w}^\downarrow, \mathbb{T}_\lambda, \mathbf{\Lambda}} = \frac{1}{N(N+1)/2} \left(N \sum_{1 \leq i \leq N} \mu_{S_i} - \sum_{1 \leq i < j \leq N} \mathbb{T}_{\lambda_{ij}}(\mu_{S_i}, \mu_{S_j}) \right). \quad (3.37)$$

J'ai alors montré les propriétés suivantes.

Théorème. 3. Soit \mathbf{w}^\uparrow le vecteur de poids défini par (3.31) et soit \mathbb{T}_λ une famille paramétrique de t -norms avec $\mathbf{\Lambda} = \{\lambda_{ij}\}_{1 \leq i < j \leq N}$ l'ensemble des $N(N-1)/2$ valeurs réelles appartenant au domaine des paramètres de \mathbb{T}_λ , spécifiant la fonction d'appartenance floue de toutes les intersections $S_i \cap S_j$, $\forall i < j$. Alors $A_{\mathbf{w}^\uparrow, \mathbb{T}_\lambda, \mathbf{\Lambda}}$ défini par (3.36) est une fonction d'agrégation et vérifie donc les conditions (3.2).

Théorème. 4. Soit \mathbf{w}^\downarrow le vecteur de poids défini par (3.32) et soit \mathbb{T}_λ une famille paramétrique de t -norms qui satisfait à la condition de Lipschitz suivante : $\mathbb{T}_\lambda(b, c) - \mathbb{T}_\lambda(a, c) \leq b - a$

3.2. CONTRIBUTIONS

pour tout $a \leq b$. Soit $\Lambda = \{\lambda_{ij}\}_{1 \leq i < j \leq N}$ l'ensemble des $N(N-1)/2$ valeurs réelles appartenant au domaine des paramètres de \mathbb{T}_λ . Alors $A_{\mathbf{w}^\downarrow, \mathbb{T}_\lambda, \Lambda}$ défini par (3.37) est une fonction d'agrégation et vérifie donc les conditions (3.2).

D'autres propriétés de ces deux fonctions d'agrégation sont étudiées dans [Ah-Pine, 2016].

Je conclus cette sous-section en montrant les apports pratiques des méthodes introduites ci-dessus sur un exemple connu de la littérature. Il s'agit du *dean problem* présenté dans [Grabisch and Labreuche, 2002b]. Quatre étudiants X_1, X_2, X_3, X_4 sont évalués sur trois matières : mathématiques, physique et littérature. Le doyen souhaite classer les étudiants selon les deux règles suivantes :

- Pour un étudiant bon en mathématiques, la littérature est plus importante que la physique.
- Pour un étudiant mauvais en mathématiques, la physique est plus importante que la littérature.

Les notes sur une échelle unipolaire continue $[0, 1]$ sont données ci-dessous. Ici, on interprète ces notes comme la valeur d'appartenance de chaque étudiant au sous-ensemble flou le "critère i est satisfait" avec $i = 1, 2, 3$ et 1 =mathématiques, 2 =physique et 3 =littérature.

<i>Etu.</i>	μ_{S_1}	μ_{S_2}	μ_{S_3}
X_1	0.75	0.9	0.3
X_2	0.75	0.8	0.4
X_3	0.3	0.65	0.1
X_4	0.3	0.55	0.2

Les préférences du doyen impliquent la relation d'ordre totale suivante : $X_2 \succ X_1 \succ X_3 \succ X_4$. Michel Grabisch et Christophe Labreuche montrent dans [Grabisch and Labreuche, 2002b], qu'il n'est pas possible de représenter les préférences du doyen par une intégrale de Choquet unipolaire. A titre illustratif, je donne ci-dessous les notes générales pour la moyenne arithmétique et la moyenne pondérée dont le vecteur des poids est $(5, 3, 2)/10$. On peut constater que l'ordre induit par les notes globales ne correspond pas aux préférences du doyen.

<i>Etu.</i>	$\frac{1}{3}\mu_{S_1} + \frac{1}{3}\mu_{S_2} + \frac{1}{3}\mu_{S_3}$	$\frac{5}{10}\mu_{S_1} + \frac{3}{10}\mu_{S_2} + \frac{2}{10}\mu_{S_3}$
X_1	0.65	0.705
X_2	0.65	0.695
X_3	0.35	0.365
X_4	0.35	0.355

Ici, la fonction $A_{\mathbf{w}, \mathbb{T}_M}$ ne permet pas non plus de représenter les préférences du doyen car elle est équivalente à la fonction $OWA_{\mathbf{w}}$ (Théorème 1) qui est un cas particulier de l'intégrale de Choquet unipolaire.

3.2. CONTRIBUTIONS

En revanche, les autres propositions que j'ai introduites permettent de résoudre le problème (tout comme l'intégrale de Choquet bipolaire). Je présente uniquement le cas de $A_{\mathbf{w}, \mathcal{T}_P}$. Je précise ci-dessous les valeurs des calculs intermédiaires des sommes symétriques Σ_k^N et des mesures $\mu_{E_k^N}$, afin d'illustrer les différentes définitions introduites plus haut à l'aide d'un exemple numérique.

<i>Etu.</i>	Σ_1^3	Σ_2^3	Σ_3^3
X_1	1.95	1.17	0.2025
X_2	1.95	1.22	0.24
X_3	1.05	0.29	0.0195
X_4	1.05	0.335	0.033

<i>Etu.</i>	$\mu_{E_1^3}$	$\mu_{E_2^3}$	$\mu_{E_3^3}$
X_1	0.9825	0.765	0.2025
X_2	0.97	0.74	0.24
X_3	0.7795	0.251	0.0195
X_4	0.748	0.269	0.033

Dans le cas de la fonction $A_{\mathbf{w}, \mathcal{T}_P}$, l'élicitation consiste à déterminer un vecteur de poids \mathbf{w} permettant de produire des scores agrégés qui respectent les préférences du doyen. Pour cela, le modèle Split [Marichal and Roubens, 2000] introduit précédemment, permet d'employer la programmation linéaire pour trouver une telle solution.

$$\begin{aligned} & \max \epsilon && (3.38) \\ \text{s.l.c.} & \begin{cases} \epsilon > 0, \\ \forall (X, X') \in \mathbb{M} \times \mathbb{M} \text{ tel que } X \succ X' : A_{\mathbf{w}, \mathcal{T}_P}(\mathbf{x}) - A_{\mathbf{w}, \mathcal{T}_P}(\mathbf{x}') \geq \epsilon, \\ w_k \geq 0, \forall k = 1, \dots, N, \\ \sum_{k=1}^N w_k = 1. \end{cases} \end{aligned}$$

La solution optimale obtenue est donnée par $\epsilon^* = 0.01066$, $\mathbf{w}^* = (0.536842, 0, 0.463158)$ et les scores agrégés des étudiants obtenus sont :

<i>Etu.</i>	$A_{\mathbf{w}^*, \mathcal{T}_P}$
X_1	0.621237
X_2	0.631895
X_3	0.4275
X_4	0.416842

Des solutions distinctes mais respectant les préférences des doyens peuvent être obtenues avec $A_{\mathbf{w}^\uparrow, \mathcal{T}_{\lambda, \Lambda}}$ et $A_{\mathbf{w}^\downarrow, \mathcal{T}_{\lambda, \Lambda}}$. Dans ces deux cas, le modèle d'élicitation consiste à déterminer

les valeurs de Λ et pour cela, on peut utiliser à nouveau l’approche Split et la programmation linéaire.

3.3 Discussions et perspectives

Au cours de ces travaux, j’ai été particulièrement intéressé par les apports croisés pouvant exister entre l’aide multicritère à la décision (AMCD) d’une part, et l’apprentissage supervisé d’autre part. En effet, la généralisation des moyennes pondérées par des fonctions d’agrégation comme celle de l’intégrale de Choquet, fait écho au concept de *structural risk minimization* de Vladimir Vapnik et Alexey Chervonenkis. Autrement dit, l’augmentation de la complexité de la classe d’hypothèses permet ici de surmonter le problème de sous-apprentissage relatif à l’impossibilité pour une fonction d’agrégation additive de représenter le modèle de préférence d’un agent et la nécessité d’employer des fonctions d’agrégation plus flexibles.

Par ailleurs, le développement des fonctions d’agrégation est motivé par la définition de nouveaux outils permettant de représenter de façon riche les modèles de décision et de préférence complexes d’un décideur. Les fonctions d’agrégation étudiées en AMCD sont donc *de facto* interprétables du point de vue de la théorie de la décision. Dans ce contexte, il existe de mon point de vue, une **réelle opportunité de fertilisation croisée entre AMCD et machine learning**, notamment dans le cadre du domaine XAI (*eXplainable Artificial Intelligence*). Ce sujet attire d’ailleurs de plus en plus d’intérêt comme l’atteste les références suivantes : [Murray et al., 2020, Murray et al., 2021].

Dans cette perspective, un premier axe que je suggère est l’utilisation de fonctions d’agrégation comme classe d’hypothèses dans le cadre de tâches supervisées. Cette question de recherche a été saisie par [Fallah Tehrani et al., 2012] dans le cas de l’utilisation de l’intégrale de Choquet pour des problèmes de catégorisation. Dans cet article, des bornes de la VC-dimension de la classe d’hypothèses associée à l’intégrale de Choquet (CI) sont établies et démontrent l’intérêt de ces fonctions vis-à-vis des combinaisons linéaires. Les auteurs introduisent également la *Choquistic regression* qui est un type de régression logistique multinomiale où le modèle linéaire est remplacé par la CI. Des résultats d’expérience montrent l’apport à la fois en précision et en interprétabilité de la méthode. De la même manière que les méthodes classiques ont été “kernelisées” par l’usage de fonctions noyaux non linéaires à la place du produit scalaire usuel, il me semble particulièrement attrayant d’**utiliser la CI à la place de la combinaison linéaire usuelle dans de nombreuses techniques classiques**.

Que ce soit pour le problème de régression ou celui de catégorisation, une difficulté sous-jacente à la CI tient en sa nature combinatoire et sa structure hiérarchique/de treillis, imposée par les contraintes de monotonie. Tout comme pour le cas de nos travaux sur l’élucation d’une BCI que j’ai présentés plus haut, l’approche introduite dans [Fallah Tehrani et al., 2012] exploite également la k -additivité afin de diminuer le nombre de paramètres à estimer. Il existe dans la littérature plusieurs travaux définissant des procédures d’optimisation pour l’estimation d’une capacité dans le cas général. Un article de recherche incontournable dans ce contexte

est celui de Michel Grabisch, Ivan Kojadinovic et Patrick Meyer dans [Grabisch et al., 2008]. Dans cette contribution, la **procédure HLMS** (*Heuristic Least Mean Squares*) y est définie. Il s’agit d’une approche basée sur les moindres carrés. Mais contrairement au modèle R_{ss} exposé dans la sous-section 3.2.1, il s’agit d’une heuristique dont la stratégie est similaire à une **descente de gradient stochastique**. Lors d’une itération, une alternative est employée afin de corriger les valeurs des paramètres de la capacité de sorte à faire diminuer la somme des carrés des résidus. L’utilisation d’une unique alternative donne lieu à la modification d’un chemin dans le treillis sous-jacent à une capacité. La procédure HLMS a été mise au point de sorte à effectuer les mises à jour des paramètres de ce chemin et de ceux qui lui sont adjacents dans le but de maintenir la propriété de monotonie de la capacité. D’autres stratégies comme les algorithmes génétiques [Islam et al., 2019a], ont également été étudiées pour l’inférence des paramètres d’une capacité.

J’ai également entamé des travaux sur cette question de recherche en utilisant une approche de type **descente de gradient par blocs**. Il existe en effet des propriétés entre les différentes strates du treillis associée à une capacité qui peuvent être exploitées pour estimer et mettre à jour efficacement les paramètres d’une capacité. J’ai formalisé ces propriétés, implémenté l’approche en R, testé sur des données simulées et obtenus des résultats expérimentaux intéressants et complémentaires à HLMS. Toutefois, je n’ai pas su trouver le temps pour préparer un article sur ce sujet. Je compte néanmoins le faire dans les mois à venir.

Supposons que l’on représente la CI par rapport à la capacité μ , en fonction de la transformée de Möbius de cette dernière m^μ . La notion de **k -additivité** d’une capacité est similaire à celle d’une bicapacité : μ est k -additive si les valeurs de m^μ pour les sous-ensembles de taille strictement supérieure à k sont nulles et qu’il existe au moins un sous-ensemble de taille k pour lequel m^μ est différent de 0. Dans ce contexte, il serait intéressant de **fixer k petit mais plus grand que 2 ($k = 3$ ou 4 par exemple) et de pénaliser le problème par la fonction de régularisation $\|m^\mu\|_{\ell_1}$** . De ce fait, on obtiendrait une capacité qui soit au plus k -additive mais, par principe de rasoir d’Occam, le modèle cherchera à favoriser les valeurs petites de k . Autrement dit, les concepts de régularisation par norme ℓ_1 pour favoriser la parcimonie, est un concept intéressant dans le cadre de l’estimation d’une capacité dans le but de favoriser l’interprétabilité car il est plus aisé d’appréhender un modèle possédant moins de paramètres.

Ensuite, un axe que je souhaiterais également poursuivre concerne la famille de **fonctions d’agrégation** que j’ai détaillée dans la sous-section 3.3. Dans ce contexte, j’ai établi de **nouvelles extensions avec l’utilisation de nouveaux vecteurs de poids particuliers** qui conduisent, à l’instar de $A_{w^\uparrow, T_\lambda, \Lambda}$ et $A_{w^\downarrow, T_\lambda, \Lambda}$, à une réduction de la complexité de calcul. En effet, les identités combinatoires qui m’ont permis de démontrer les résultats énoncés dans les Théorèmes 3 et 4 peuvent être généralisées. On aboutit alors à des fonctions d’agrégation avec une granularité encore plus fine. Je souhaite là aussi préparer un article dans les mois à venir.

3.3. DISCUSSIONS ET PERSPECTIVES

Dans le cas de ces fonctions d'agrégation, une autre extension qui serait pertinente de développer concerne l'ajout de poids non uniformes aux différents constituants de l'expression (3.36). En effet, il y a $\frac{N(N+1)}{2}$ termes qui sont additionnés et la constante multiplicative $\frac{1}{N(N+1)/2}$ suggère qu'à chacun d'entre eux est attribué un poids identique. Supposons un vecteur de poids noté \mathbf{p} constitué de $\frac{N(N+1)}{2}$ valeurs non négatives sommant à 1 et qui attribue à chaque terme un poids particulier. Quelle serait alors la sémantique d'une telle approche ?

Enfin, un axe de recherche qui me paraît tout aussi intéressant à examiner est la possibilité de modéliser des fonctions d'agrégation d'un certain type par des réseaux de neurones. Je prends pour exemple l'ensemble des fonctions $A_{\mathbf{w}^\dagger, \tau_\lambda, \Lambda}$ et son expression réduite donnée par (3.36). Dans ce cas, toute paire de valeurs d'appartenance aux sous-ensembles flous (S_i, S_j) pourrait être fusionnée par une *t-norm* au sein d'un “**fuzzy perceptron**”. Il existe, au sein de la communauté, un intérêt grandissant pour les **architectures associant les réseaux de neurones et les systèmes à base de logique flou, afin d'améliorer l'interprétabilité**. Le cas de l'intégrale de Choquet a déjà été étudié dans l'article [Islam et al., 2019b]. Plus généralement, les articles suivants présentent des *survey* dans ce domaine encore émergent [Das et al., 2020, Zheng et al., 2021].

Mesures de proximité et critères de partitionnement en clustering

Sommaire du chapitre

4.1 Introduction	79
4.1.1 Contexte	79
4.1.2 Travaux antérieurs	81
Analyse relationnelle, <i>clustering</i> , mesures d'association et critères de partitionnement	82
Indices de similarité entre vecteurs binaires et interprétations géométriques	86
4.2 Contributions	89
4.2.1 Maximisation de mesures d'association en <i>clustering</i>	89
4.2.2 Normalisation de mesures de similarité	96
4.3 Discussions et perspectives	104

4.1 Introduction

4.1.1 Contexte

Je présente plusieurs travaux de recherche sur **des mesures de similarité et des critères de partitionnement en *clustering***. Les idées sources de ces recherches proviennent initialement de mon travail de thèse. Le domaine de la classification automatique a en effet, été un axe de recherche permanent tout au long de ma carrière. Le présent Chapitre mais également le suivant y sont consacrés.

Dans le cadre de mon travail de thèse de doctorat, j'ai contribué sur différents aspects de l'analyse relationnelle (AR). J'ai eu l'occasion de présenter quelques éléments de l'AR en aide multicritère à la décision dans la section 3.1 du Chapitre précédent. En bref, l'agrégation de relations de préférence s'effectue par le biais de matrices d'adjacence binaires (matrices relationnelles) que l'on additionne et qui donne lieu à une matrice d'adjacence pondérée collective qui indique pour chaque paire de candidats le nombre de votants qui préfèrent

le premier au second. Ensuite, l'AR permet de déterminer un consensus au travers d'un programme linéaire en nombres bivalents. Pour cela, elle modélise par des contraintes linéaires, chaque propriété d'une relation d'ordre totale : réflexivité, antisymétrie, totalité et transitivité.

Quel est le lien avec le *clustering*? Dans le cadre général de l'AR, on s'intéresse aux relations binaires structurées ce qui inclut non seulement les relations d'ordre, mais également les **relations d'équivalence**. Prenons le cas d'un ensemble de n individus $\{X_i\}_{i=1,\dots,n}$ décrits par un ensemble de p variables qualitatives nominales $\{X^j\}_{j=1,\dots,p}$. Supposons que l'on souhaite partitionner l'ensemble des individus. Chaque variable nominale infère une relation d'équivalence sur $\{X_i\}_i$. Par ailleurs, il existe une relation biunivoque entre l'ensemble des partitions d'un ensemble d'objets d'une part, et l'ensemble des relations d'équivalence sur ce même ensemble d'objets d'autre part. Le problème du *clustering* dans ce cas précis, peut alors être vu comme une tâche d'agrégation et de recherche de consensus dans un ensemble de relations d'équivalence. La même démarche de l'AR rappelée dans le Chapitre précédent dans le cas des préférences, peut alors être appliquée dans le cas des partitions. Je donne plus d'éléments formels dans la section suivante.

Au-delà de cette modélisation du problème de classification automatique, l'AR s'intéresse également aux mesures d'association entre variables qualitatives nominales. La représentation de ces dernières par le biais de leur matrice d'adjacence binaire permet un certain nombre de propriétés. Celles-ci rendent possible la définition de **critères de partitionnement associés à des mesures d'association** de la littérature comme celles de William Rand [Rand, 1971] ou celle de Svan Janson et Jan Vegelius [Janson and Vegelius, 1982] appelée *J-index*. Ces critères de partitionnement sont valables pour les individus décrits par des variables qualitatives nominales. Dans ma thèse de doctorat, j'ai suggéré l'extension de ces critères au **cas des individus décrits par des variables quantitatives** (ou continues ou réelles). J'ai, par la suite, développé cette idée et introduit le concept d'**écart à une tendance centrale** qui permet de mieux interpréter les différences entre les divers critères de partitionnement. J'ai également appliqué ces approches dans le cadre de la **détection de communautés dans des graphes** et comparer celles-ci au **critère de modularité** de Michelle Girvan et de Mark Newman [Newman and Girvan, 2004, Newman, 2006a].

Par ailleurs, dans le champ du *clustering*, les concepts de **mesure de proximité entre vecteurs** sont primordiaux. Dans le cas de l'AR, des travaux avaient également été entrepris sur ces sujets notamment au travers du concept de similarité régularisée introduite par Hamid Benhadda et Jean-François Marcotorchino [Benhadda and Marcotorchino, 1998]. Pour ma part, j'ai travaillé dans ma thèse sur une interprétation géométrique des indices de similarité entre vecteurs binaires tels que les mesures de Dice, Jaccard, Ochiaï, Sokal-Sneath-Anderberg, ... Ceci m'a permis de mettre en évidence deux paramètres centraux dans la comparaison de ces mesures que sont le **cosinus de l'angle formé par les vecteurs** d'une part, et le **rapport des normes des vecteurs** d'autre part. Cette interprétation géométrique à l'avantage de s'étendre à des vecteurs quelconques.

Dans la continuité de mes travaux de thèse, j'ai appliqué ces concepts dans différents

contextes : pour la définition d'**indices de similarité pour des relations binaires hétérogènes**, la **normalisation de matrice de noyaux** et une **généralisation de la normalisation de la matrice Laplacienne** en *spectral clustering*.

Je présente de façon synthétique ces différentes contributions qui couvrent les publications suivantes :

- **J. Ah-Pine**. 2017. Sur la normalisation de la matrice Laplacienne en partitionnement spectral. *Actes des XXIVèmes Rencontres de la Société Francophone de Classification (SFC 2017)*. [Lien vers la conférence, <http://polytech-sfc2017.univ-lyon1.fr/>].
- **J. Ah-Pine**. 2013. Graph Clustering by Maximizing Statistical Association Measures. *Proceedings of the 12th International Symposium on Intelligent Data Analysis (IDA 2013)* [Taux d'acceptation < 23%]. [Lien vers la conférence, <http://www.ida2013.org/>].
- **J. Ah-Pine**. 2013. A general framework for comparing heterogeneous binary relations. *Proceedings of the 1st International Conference on Geometric Sciences of Information (GSI 2013)*. [Lien vers la conférence, <http://www.gsi2013.org/>]. **J. Ah-Pine**. 2010. Normalized kernels as similarity indices. *Proceedings of the 14th Pacific Asia conference on Knowledge Discovery and Data Mining (PAKDD 2010)* [Taux d'acceptation < 11%]. [Lien vers la conférence, <http://www.iiit.ac.in/conferences/pakdd2010/>].
- **J. Ah-Pine**, F. Marcotorchino. 2010. Unifying some association criteria between partitions using relational matrices. *Communications in Statistics - Theory and Methods*. 39(3) :531-542. [Lien vers le journal, <http://www.iospress.nl/loadtop/load.php?isbn=15701263>].
- **J. Ah-Pine**. 2009. Cluster analysis based on the central tendency deviation principle. *Proceedings of Advanced Data Mining and Applications (ADMA 2009), Lecture Notes in Artificial Intelligence (LNAI 5678)* [Taux d'acceptation < 12%]. [Lien vers le journal, <http://www.springerlink.com/content/08357122g0831064/>].

4.1.2 Travaux antérieurs

Dans cette sous-section, je rappelle des résultats en AR et des contributions de ma thèse de doctorat. Le contenu est organisé en deux paragraphes. Dans la première partie page 82, je discute des programmes linéaires en nombres binaires pour la détermination des relations de consensus ; des relations entre codage contingenciel et codage relationnel pour la formalisation de mesures d'association ; et de la propriété de linéarité permettant de définir des critères de partitionnement dans le cas multivarié. Ensuite dans la seconde partie page 86, je rappelle la genèse des similarités d'ordre t qui est une famille de mesures de proximité que j'ai mise en place dans ma thèse et qui sont valables à la fois pour des vecteurs binaires et des vecteurs réels.

Analyse relationnelle, *clustering*, mesures d'association et critères de partitionnement

En AR, étant donnée une table de données individu-attribut, on infère des relations binaires (RB) à partir des variables en colonnes. Ces RB sont représentées par des matrices d'adjacence binaires que l'on appelle matrices relationnelles. Dans le contexte de l'aide multicritère à la décision, les variables ordinales sont interprétées tels des scores de satisfaction à des critères et on les représente donc par des relations d'ordre.

Dans le cas de la classification automatique, je considère dans un premier temps, des variables qualitatives nominales. Considérons n individus $\{X_1, \dots, X_n\}$, décrits par p variables qualitatives nominales $\{X^1, \dots, X^p\}$. Ces dernières sont représentées par des matrices relationnelles d'ordre n que l'on dénote par $\mathbf{X}^1, \dots, \mathbf{X}^p$. Pour tout $j = 1, \dots, p$, le terme général de $\mathbf{X}^j = (x_{ii'}^j)$ est défini par $x_{ii'}^j = 1$ si X_i et $X_{i'}$ ont la même modalité pour X^j et 0 sinon.

Peu importe la nature et l'hétérogénéité des ensembles de modalités des variables $\{X^j\}_{j=1, \dots, p}$, le codage relationnel transforme ces variables dans une représentation uniforme $\{\mathbf{X}^j\}_{j=1, \dots, p}$ consistant en des matrices carrées binaires de tailles identiques. Ce codage rend alors possible l'agrégation des informations contenues dans les variables qualitatives hétérogènes en sommant leurs matrices relationnelles et évite ainsi le problème d'incommensurabilité.

La sommation donne lieu à une matrice relationnelle non binaire dite collective qui sera dénotée $\mathbf{C} = (c_{ii'})$: $\mathbf{C} = \sum_{j=1}^p \mathbf{X}^j$. Pour chaque paire $(X_i, X_{i'})$, $c_{ii'}$ donne le nombre de variables ayant mis dans la même catégorie ces deux individus. \mathbf{C} peut être vue comme la matrice d'adjacence d'un graphe non dirigé valué. Le consensus revient à déterminer une relation d'équivalence (ou partition) que l'on va représenter par une matrice d'adjacence binaire \mathbf{X} . Intuitivement, plus $c_{ii'}$ est grand, plus les individus X_i et $X_{i'}$ sont similaires et on s'attend à ce que plus la vraisemblance que $x_{ii'} = 1$ (X_i et $X_{i'}$ dans le même *cluster* de consensus) soit forte. Ce principe de "maximisation de vraisemblance" est similaire au critère de Condorcet dans le cas des votes.

Par conséquent, on peut naturellement étendre le principe de comparaison par paires et le critère de consensus introduits par le Marquis de Condorcet au cas des relations d'équivalence. Dans cette perspective, l'AR propose une méthode de calcul exacte fondée sur la programmation linéaire en nombres binaires. Comme précédemment, elle modélise par des contraintes linéaires les propriétés d'une relation d'équivalence. Ce qui change, en comparaison de l'agrégation de relations d'ordre, concerne la propriété d'antisymétrie qui devient une propriété de symétrie ; et l'abandon de la propriété de totalité qui n'est pas requise ici. Plus formellement, le modèle de l'AR pour le problème de *clustering* d'individus décrits par des variables qualitatives nominales (sans données manquantes) peut être exprimé par le programme linéaire en nombres

bivalents suivant :

$$\begin{aligned} & \max \sum_{i,i'=1}^n \left(c_{ii'} - \frac{p}{2} \right) x_{ii'} & (4.1) \\ \text{s.l.c.} \quad & \begin{cases} x_{ii} = 1, \forall i = 1, \dots, n & (\text{réflexivité}), \\ x_{ii'} - x_{i'i} = 0, \forall i, i' = 1, \dots, n & (\text{symétrie}), \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1, \forall i, i', i'' = 1, \dots, n & (\text{transitivité}), \\ x_{ii'} \in \{0, 1\}, \forall i, i' = 1, \dots, n & (\text{binarité}). \end{cases} \end{aligned}$$

La constante $\frac{p}{2}$ dans la fonction objectif indique la majorité absolue ce qui renvoie au critère de Condorcet dans le cas de la théorie des votes.

Par ailleurs, la matrice relationnelle \mathbf{X} qui représente une relation d'équivalence et qui vérifie les propriétés relationnelles mentionnées ci-dessus, peut être réorganisée par permutation symétrique de ses lignes et de ses colonnes de sorte à faire apparaître des blocs le long de sa diagonale. Je reviendrai par ailleurs sur cette structure "bloc diagonale" dans le Chapitre suivant à la sous-section 5.2.1.

Au-delà de cette modélisation originale du problème de *clustering*, l'AR a également engendré de nombreux résultats sur les mesures d'association entre variables qualitatives nominales. En effet, dans le cadre de l'analyse de dépendance entre ce type de données, elle met en lumière une certaine dualité entre d'un côté, l'analyse de la table de contingence et de l'autre, l'analyse du croisement des matrices d'adjacence binaires. Grâce au codage relationnel, on montre que de nombreuses mesures d'association sont en fait des coefficients de covariance ou de corrélation linéaire.

Pour illustrer ces propos je m'intéresse aux critères \mathbf{B} de Belson [Belson, 1959], \mathbf{E} d'écart carré à l'indépendance [Marcotorchino, 1984a], \mathbf{J} de Charles Jordan¹ [Jordan, 1927, Marcotorchino, 1984a], et \mathbf{LM} de Light-Margolin [Light and Margolin, 1971] que j'utiliserai en particulier par la suite dans la section 4.2. Soit deux variables qualitatives nominales X^k et X^l ayant respectivement p_k et p_l modalités. Dénotons par \mathbf{N} la table de contingence de taille $p_k \times p_l$ croisant ces deux variables et n_{uv} , le nombre d'individus ayant la modalité u de X^k et la modalité v de X^l . On

1. Il s'agit en fait d'une interprétation de la mesure de Charles Jordan décrite dans [Jordan, 1927] donnée par Jean-François Marcotorchino dans [Marcotorchino, 1984a].

a alors les définitions suivantes :

$$B(X^k, X^l) = \sum_{u=1}^{p_k} \sum_{v=1}^{p_l} \left(n_{uv} - \frac{n_u \cdot n_v}{n} \right)^2. \quad (4.2)$$

$$E(X^k, X^l) = \sum_{u=1}^{p_k} \sum_{v=1}^{p_l} \left(n_{uv}^2 - \frac{n_u^2 \cdot n_v^2}{n^2} \right). \quad (4.3)$$

$$J(X^k, X^l) = \frac{1}{n} \sum_{u=1}^{p_k} \sum_{v=1}^{p_l} \left(n_{uv} \left(n_{uv} - \frac{n_u \cdot n_v}{n} \right) \right). \quad (4.4)$$

$$LM(X^k, X^l) = \sum_{u=1}^{p_k} \sum_{v=1}^{p_l} \frac{n_{uv}^2}{n_u} - \frac{1}{n} \sum_{v=1}^{p_l} n_v^2. \quad (4.5)$$

Il existe des formules de passage permettant d'exprimer certaines quantités impliquant la matrice \mathbf{N} en fonction des matrices relationnelles \mathbf{X}^k et \mathbf{X}^l [Marcotorchino, 1984b]. Je les rappelle ci-dessous :

$$\sum_{u=1}^{p_k} \sum_{v=1}^{p_l} n_{uv}^2 = \sum_{i=1}^n \sum_{i'=1}^n x_{ii'}^k x_{ii'}^l, \quad (4.6)$$

$$\sum_{u=1}^{p_k} n_u^2 = \sum_{i=1}^n \sum_{i'=1}^n x_{ii'}^k, \quad (4.7)$$

$$\sum_{u=1}^{p_k} \sum_{v=1}^{p_l} n_{uv} n_u \cdot n_v = \sum_{i=1}^n \sum_{i'=1}^n \left(\frac{x_{ii'}^k + x_{ii'}^l}{2} \right) x_{ii'}^l, \quad (4.8)$$

$$\sum_{u=1}^{p_k} \sum_{v=1}^{p_l} \frac{n_{uv}^2}{n_u} = \sum_{i=1}^n \sum_{i'=1}^n \left(\frac{2x_{ii'}^k}{x_{ii'}^k + x_{ii'}^l} \right) x_{ii'}^l, \quad (4.9)$$

$$= \sum_{i=1}^n \sum_{i'=1}^n \left(\frac{x_{ii'}^k}{x_{ii'}^k} \right) x_{ii'}^l, \quad (4.10)$$

où $x_{ii'}^k = \sum_{i''=1}^n x_{ii''}^k x_{i''i'}^k$ et $x_{ii}^k = x_{ii}^k$ étant donné que \mathbf{X}^k est symétrique.

L'équivalence entre les expressions (4.9) et (4.10) provient du fait que $x_{ii'}^k = 1$ si et seulement si X_i et $X_{i'}$ ont la même modalité pour X^k , et que dans ce cas $x_{ii}^k = x_{i'i}^k$, ce qui établit l'égalité $\frac{x_{ii}^k + x_{i'i}^k}{2} = x_{ii}^k$ et engendre ces deux versions.

En appliquant les formules de passage contingenciel-relationnel, on peut démontrer sans

difficulté les expressions suivantes des critères susmentionnés :

$$\mathbb{B}(\mathbf{X}^k, \mathbf{X}^l) = \sum_{i=1}^n \sum_{i'=1}^n \left(x_{ii'}^k - \frac{x_{i.}^k + x_{.i'}^k}{n} + \frac{x_{..}^k}{n^2} \right) \left(x_{ii'}^l - \frac{x_{i.}^l + x_{.i'}^l}{n} + \frac{x_{..}^l}{n^2} \right) \quad (4.11)$$

$$= \sum_{i=1}^n \sum_{i'=1}^n \left(x_{ii'}^k - \frac{x_{i.}^k + x_{.i'}^k}{n} + \frac{x_{..}^k}{n^2} \right) x_{ii'}^l, \quad (4.12)$$

$$\mathbb{E}(\mathbf{X}^k, \mathbf{X}^l) = \sum_{i,i'} \left(x_{ii'}^k - \sum_{i,i'} \frac{x_{ii'}^k}{n^2} \right) \left(x_{ii'}^l - \sum_{i,i'} \frac{x_{ii'}^l}{n^2} \right) \quad (4.13)$$

$$= \sum_{i,i'} \left(x_{ii'}^k - \sum_{i,i'} \frac{x_{ii'}^k}{n^2} \right) x_{ii'}^l, \quad (4.14)$$

$$\mathbb{J}(\mathbf{X}^k, \mathbf{X}^l) = \frac{1}{n} \sum_{i,i'} \left(x_{ii'}^k - \frac{x_{i.}^k}{n} \right) \left(x_{ii'}^l - \frac{x_{.i'}^l}{n} \right) \quad (4.15)$$

$$= \frac{1}{n} \sum_{i,i'} \left(x_{ii'}^k - \frac{x_{i.}^k}{n} \right) x_{ii'}^l, \quad (4.16)$$

$$\mathbb{LM}(\mathbf{X}^k, \mathbf{X}^l) = \sum_{i=1}^n \sum_{i'=1}^n \left(\frac{2x_{ii'}^k}{x_{i.}^k + x_{.i'}^k} - \frac{1}{n} \right) x_{ii'}^l \quad (4.17)$$

$$= \sum_{i=1}^n \sum_{i'=1}^n \left(\frac{x_{ii'}^k}{x_{i.}^k} - \frac{1}{n} \right) x_{ii'}^l. \quad (4.18)$$

Concernant le critère LM, deux expressions équivalentes (4.17) et (4.18) peuvent être utilisées. Ces versions sont dues à la correspondance entre les formules de passage (4.9) et (4.10) que j'ai explicitée plus haut. Il est toutefois important de garder en tête que cette équivalence vient du fait que \mathbf{X}^k et \mathbf{X}^l sont des matrices d'adjacence représentant des relations d'équivalence. En particulier, si le premier argument \mathbf{X}^k n'encode pas une partition, les deux formulations sont alors distinctes.

Pour les trois premières mesures \mathbb{B} , \mathbb{E} et \mathbb{J} , je donne deux formulations qui sont équivalentes également. La première est de nature symétrique et permet de montrer que ces mesures sont similaires à des covariances. La deuxième expression indique que l'on peut aussi faire jouer un rôle asymétrique aux deux matrices relationnelles. Ce type de formulation permet de mettre en perspective le principe d'**association maximale** lorsqu'il y a plusieurs variables qualitatives $\{\mathbf{X}^j\}_{k=1,\dots,p}$.

Soit la notation générique Δ indiquant une mesure d'association parmi $\{\mathbb{B}, \mathbb{E}, \mathbb{J}\}$. Le principe d'association maximale consiste à déterminer une **matrice relationnelle centrale (ou de consensus)** \mathbf{X} , résumant l'information apportée par les p matrices relationnelles individuelles $\{\mathbf{X}^j\}_{k=1,\dots,p}$, en maximisant la somme (ou la moyenne) $\sum_{j=1}^p \Delta(\mathbf{X}^j, \mathbf{X})$ (respectivement, $\frac{1}{p} \sum_{j=1}^p \Delta(\mathbf{X}^j, \mathbf{X})$).

Les expressions relationnelles des critères \mathbb{B} , \mathbb{E} , \mathbb{J} données par les équations (4.12), (4.14),

(4.16) respectivement, permettent de mettre en lumière la propriété de linéarité suivante :

$$\sum_{j=1}^p \Delta(\mathbf{X}^j, \mathbf{X}) = \Delta\left(\sum_{j=1}^p \mathbf{X}^j, \mathbf{X}\right) \quad (4.19)$$

$$= \Delta(\mathbf{C}, \mathbf{X}), \quad (4.20)$$

où j'ai repris la notation $\mathbf{C} = \sum_{j=1}^p \mathbf{X}^j$ introduite plus haut.

Le problème d'optimisation (4.1) peut donc être mis en oeuvre avec des fonctions objectif $\Delta \in \{\mathbf{B}, \mathbf{E}, \mathbf{J}\}$ qui sont distinctes du critère de Condorcet. On a alors le modèle générique suivant qui illustre la panoplie de méthodes de *clustering* pouvant être pratiquée en AR :

$$\begin{aligned} & \max \Delta(\mathbf{C}, \mathbf{X}) && (4.21) \\ \text{s.l.c.} & \begin{cases} x_{ii} = 1, \forall i = 1, \dots, n & (\text{réflexivité}), \\ x_{ii'} - x_{i'i} = 0, \forall i, i' = 1, \dots, n & (\text{symétrie}), \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1, \forall i, i', i'' = 1, \dots, n & (\text{transitivité}), \\ x_{ii'} \in \{0, 1\}, \forall i, i' = 1, \dots, n & (\text{binarité}). \end{cases} \end{aligned}$$

En ce qui concerne mes contributions post thèse de doctorat sur ces sujets, je montre en sous-section 4.2.1, comment on peut étendre les critères B, E et J, du cas d'individus décrits par des variables qualitatives nominales, au cas d'individus décrits par des variables quantitatives, au travers du **principe d'écart à une tendance centrale**.

Je m'intéresse ensuite à l'extension au cas d'individus dont les relations de similarité sont représentées par un graphe non orienté et non pondéré. Dans cette perspective, j'étudie, en plus des trois fonctions objectifs précédentes, le critère engendré par la mesure LM et compare ces derniers au critère de modularité de Newman incontournable pour le problème de partitionnement de graphe et la détection de communautés.

Indices de similarité entre vecteurs binaires et interprétations géométriques

Le caractère logique sous-jacent à l'AR a naturellement amené ses contributeurs à s'intéresser aux indices de similarité entre vecteurs binaires. Ceux-ci sont employés en AR soit directement sur des vecteurs binaires pour mesurer des similarités, soit sur des matrices relationnelles pour évaluer des associations. Ces indices sont basés sur la combinaison de différentes quantités qui comptent le nombre de motifs d'accords 1-1 ou 0-0 et le nombre de motifs de désaccord 0-1 ou 1-0. Les matrices relationnelles étant binaires les indices évoqués précédemment peuvent donc être adaptées à ce type de structures.

Il existe de nombreux indices de similarité pour vecteurs binaires. J'étudie en particulier les indices de Dice, d'Ochiaï et de Kulczynski (voir par exemple [Lesot et al., 2009] pour un article de *survey*). Ces méthodes jouent un rôle particulier dans le contexte de mes contributions de thèse. Je les rappelle brièvement ci-après. Je présenterai plus loin en sous-section 4.2.2, les extensions de ces indices que j'ai proposées à la suite de ma thèse de doctorat.

4.1. INTRODUCTION

Je considère deux vecteurs binaires de taille p , $\mathbf{x}_i, \mathbf{x}_{i'} \in \{0, 1\}^p$. En parcourant les dimensions de ces deux vecteurs, on peut compter le nombre de cas pour chacun des 4 motifs suivants : 1-1, 0-0, 1-0 et 0-1. On introduit alors les notations suivantes :

- $11_{ii'}$ = Nombre de 1 en commun entre \mathbf{x}_i et $\mathbf{x}_{i'}$ (accords positifs 1-1),
- $00_{ii'}$ = Nombre de 0 en commun entre \mathbf{x}_i et $\mathbf{x}_{i'}$ (accords négatifs 0-0),
- $10_{ii'}$ = Nombre de fois où on rencontre 1 pour \mathbf{x}_i et 0 pour $\mathbf{x}_{i'}$ (désaccords relatifs 1-0),
- $01_{ii'}$ = Nombre de fois où on rencontre 0 pour \mathbf{x}_i et 1 pour $\mathbf{x}_{i'}$ (désaccords relatifs 0-1).

Les indices de Dice, Ochiaï et de Kulczynski sont alors définis comme suit :

$$\text{Dice}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{11_{ii'}}{11_{ii'} + \frac{1}{2}(10_{ii'} + 01_{ii'})}, \quad (4.22)$$

$$\text{Ochiaï}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{11_{ii'}}{\sqrt{(11_{ii'} + 10_{ii'})(11_{ii'} + 01_{ii'})}}, \quad (4.23)$$

$$\text{Kulczynski}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{1}{2} \left(\frac{11_{ii'}}{11_{ii'} + 10_{ii'}} + \frac{11_{ii'}}{11_{ii'} + 01_{ii'}} \right). \quad (4.24)$$

En utilisant les vecteurs binaires \mathbf{x}_i et $\mathbf{x}_{i'}$ que l'on interprète comme étant des éléments de l'espace ambiant \mathbb{R}^p muni du produit scalaire canonique $\langle \cdot, \cdot \rangle$, on constate sans difficulté les expressions et interprétations géométriques suivantes :

$$\begin{aligned} 11_{ii'} &= \langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle, \\ 11_{ii} &= \|\mathbf{x}_i\|^2, \\ 10_{ii'} &= \langle \mathbf{x}_i, \mathbf{x}_i - \mathbf{x}_{i'} \rangle, \\ 01_{ii'} &= \langle \mathbf{x}_{i'} - \mathbf{x}_i, \mathbf{x}_{i'} \rangle, \\ 10_{ii'} + 01_{ii'} &= \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2, \\ 10_{ii'} + 11_{ii'} &= \|\mathbf{x}_i\|^2, \\ 01_{ii'} + 11_{ii'} &= \|\mathbf{x}_{i'}\|^2. \end{aligned}$$

En utilisant les relations précédentes, il vient :

$$\text{Dice}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle}{\frac{1}{2} (\|\mathbf{x}_i\|^2 + \|\mathbf{x}_{i'}\|^2)}, \quad (4.25)$$

$$\text{Ochiaï}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle}{\sqrt{\|\mathbf{x}_i\|^2 \|\mathbf{x}_{i'}\|^2}}, \quad (4.26)$$

$$\text{Kulczynski}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle}{\left(\frac{2\|\mathbf{x}_i\|^2 \|\mathbf{x}_{i'}\|^2}{\|\mathbf{x}_i\|^2 + \|\mathbf{x}_{i'}\|^2} \right)}. \quad (4.27)$$

On reconnaît aux dénominateurs les moyennes arithmétiques, géométriques et harmoniques des normes au carré, $\|\mathbf{x}_i\|^2$ et $\|\mathbf{x}_{i'}\|^2$. J'ai alors introduit dans ma thèse de doctorat,

la **famille de similarité d'ordre t** définie comme suit et qui généralise les cas précédents :

$$S^t(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle}{M^t(\langle \mathbf{x}_i, \mathbf{x}_i \rangle, \langle \mathbf{x}_{i'}, \mathbf{x}_{i'} \rangle)}, \quad (4.28)$$

où $M^t(a, b) = (\frac{1}{2}(a^t + b^t))^{1/t}$ est la **moyenne généralisée d'ordre t**.

J'ai examiné plusieurs propriétés de cette famille de similarité. Une première expression et interprétation géométrique fait intervenir les coefficients de projections orthogonales d'un vecteur sur l'autre et *vice-versa* :

$$S^t(\mathbf{x}_i, \mathbf{x}_{i'}) = M^{-t} \left(\frac{\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle}, \frac{\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle}{\langle \mathbf{x}_{i'}, \mathbf{x}_{i'} \rangle} \right). \quad (4.29)$$

Ensuite, introduisons les notations suivantes :

$$\theta_{ii'} = \text{l'angle formé par les vecteurs } \mathbf{x}_i \text{ et } \mathbf{x}_{i'}, \quad (4.30)$$

$$\gamma_{i'}^i = \text{le rapport des normes } \frac{\|\mathbf{x}_i\|}{\|\mathbf{x}_{i'}\|}, \quad (4.31)$$

$$\gamma_i^{i'} = \text{le rapport des normes } \frac{\|\mathbf{x}_{i'}\|}{\|\mathbf{x}_i\|}, \quad (4.32)$$

$$\gamma_{ii'} = \text{le rapport des normes } \frac{\max(\|\mathbf{x}_i\|, \|\mathbf{x}_{i'}\|)}{\min(\|\mathbf{x}_i\|, \|\mathbf{x}_{i'}\|)}. \quad (4.33)$$

Remarquons en particulier que $\gamma_{ii'} \geq 1$ et que plus les normes $\|\mathbf{x}_i\|$ et $\|\mathbf{x}_{i'}\|$ diffèrent, plus $\gamma_{ii'}$ est grand.

Il vient alors une deuxième expression et interprétation de $S^t(\mathbf{x}_i, \mathbf{x}_{i'})$:

$$S^t(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{\cos(\theta_{ii'})}{M^t(\gamma_{i'}^i, \gamma_i^{i'})} \quad (4.34)$$

$$= \cos(\theta_{ii'}) \left(\frac{2^{1/t} \gamma_{ii'}}{(1 + \gamma_{ii'}^{2t})^{1/t}} \right). \quad (4.35)$$

Considérons l'expression (4.35) qui dépend de $\cos(\theta_{ii'})$ et de $\gamma_{ii'}$. Voici une façon d'interpréter les similarités d'ordre t :

- Pour $t < 0$: $S^t(\mathbf{x}_i, \mathbf{x}_{i'})$ est d'autant plus grand que le rapport des normes $\gamma_{ii'}$ est grand.
- Pour $t \rightarrow 0$: $S^t(\mathbf{x}_i, \mathbf{x}_{i'})$ ne fait jouer aucun rôle au rapport des normes $\gamma_{ii'}$: l'indice d'Ochiaï représente le cosinus de l'angle et est indépendant du rapport des normes.
- Pour $t > 0$: $S^t(\mathbf{x}_i, \mathbf{x}_{i'})$ est d'autant plus grand que le rapport des normes $\gamma_{ii'}$ est proche de 1.

Par conséquent, la valeur de t indique le "rôle" et l'"intensité du rôle" que l'indice $S^t(\mathbf{x}_i, \mathbf{x}_{i'})$ fait jouer au rapport des normes $\gamma_{ii'}$. Typiquement, les cas intéressants pour le *clustering* correspondent à $t \rightarrow 0$ et $t > 0$. La famille S^t généralise la mesure cosinus puisque la moyenne géométrique correspond au cas limite $t \rightarrow 0$. Lorsque $t > 0$, on a alors la propriété

que deux vecteurs positivement colinéaires atteignent la valeur de similarité maximale de 1, que s'ils ont exactement la même norme et auquel cas ils se confondent totalement. Ainsi, quand $t > 0$, l'indice S^t permet de discriminer des vecteurs positivement colinéaires en comparant leurs normes ce qui en fait une mesure de similarité plus subtile que la mesure classique de cosinus. De plus, toute chose étant égale par ailleurs, plus les normes sont éloignées, plus la valeur S^t converge vers 0 et cette convergence est d'autant plus rapide que t est grand.

Dans mes contributions post thèse que je décris dans la sous-section 4.2.2, j'emploie plusieurs principes sous-jacents aux idées exposées ci-dessus pour **normaliser des mesures de similarité dans différents contextes** : dans le cas des **relations binaires hétérogènes**, dans celui des **matrices de noyaux** et dans le cadre de la **matrice Laplacienne**.

4.2 Contributions

4.2.1 Maximisation de mesures d'association en *clustering*

Suite à mes travaux de thèse en AR, il y avait un fort intérêt à développer l'extension des résultats valables de façon inhérente pour des variables qualitatives nominales à des variables continues. Je me suis intéressé à cette question de recherche dans le cas des critères de partitionnement issus des mesures d'associations $\{\mathbf{B}, \mathbf{E}, \mathbf{J}\}$ que j'ai rappelées précédemment dans les équations (4.12), (4.14) et (4.16).

Dans cette perspective, mon approche est similaire à celle exposée dans la sous-section précédente 4.1.2. Je vais chercher d'abord à interpréter géométriquement les matrices relationnelles individuelles $\mathbf{X}^1, \dots, \mathbf{X}^p$ ainsi que la matrice collective $\mathbf{C} = \sum_{j=1}^p \mathbf{X}^j$. Cela va me permettre ensuite, de mettre en évidence une interprétation du **principe d'association maximale** décrit dans l'équation (4.19) qui va donner lieu au **principe d'écart à une tendance centrale**.

En ce qui concerne la première étape, on s'aperçoit aisément qu'une matrice relationnelle \mathbf{X}^j peut être formulée à partir de la représentation disjonctive de la variable qualitative X^j . Ainsi, soit $\mathbf{D}^j = \begin{pmatrix} \mathbf{d}^{j,1} & \dots & \mathbf{d}^{j,p_j} \end{pmatrix}$ une matrice binaire de taille $n \times p_j$, où les $\mathbf{d}^{j,u}, \forall u = 1, \dots, p_j$, sont des vecteurs binaires de taille $n \times 1$ indiquant les individus ayant la modalité u de X^j . Nous avons alors :

$$\mathbf{X}^j = \sum_{u=1}^{p_j} \mathbf{d}^{j,u} [\mathbf{d}^{j,u}]^\top. \quad (4.36)$$

Puis, si on dénote par $\mathbf{d}_i^j = (d_{i,u}^j)_{u=1, \dots, p_j}$ le vecteur binaire de taille $p_j \times 1$ tel que $d_{i,u}^j = 1$ si l'individu X_i a la modalité u de X^j et 0 sinon, alors on a également :

$$\mathbf{X}^j = (\langle \mathbf{d}_i^j, \mathbf{d}_{i'}^j \rangle)_{i,i'=1, \dots, n}, \quad (4.37)$$

où $\langle \cdot, \cdot \rangle$ est le produit scalaire canonique de \mathbb{R}^{p_j} vu comme l'espace ambiant de $\{0, 1\}^{p_j}$.

4.2. CONTRIBUTIONS

Cette dernière expression nous permet d'interpréter \mathbf{X}^j comme une matrice de produits scalaires utilisant l'espace défini par les modalités de la variable X^j .

La matrice agrégée $\mathbf{C} = \sum_{j=1}^p \mathbf{X}^j$ peut donc être interprétée telle une matrice de produits scalaires dans un espace étendu, qui est engendré par l'ensemble des modalités de toutes les variables qualitatives nominales X^1, \dots, X^p , et dont la dimension est $q = \sum_{j=1}^p p_j$. Prenons pour l'individu X_i , le vecteur binaire \mathbf{d}_i de taille $q \times 1$, qui concatène l'ensemble des vecteurs \mathbf{d}_i^j pour $j = 1, \dots, p$. Alors clairement nous avons :

$$\mathbf{C} = (\langle \mathbf{d}_i, \mathbf{d}_{i'} \rangle)_{i, i' = 1, \dots, n}. \quad (4.38)$$

Soit alors $\{\mathbf{x}_i\}_{i=1, \dots, n}$ des vecteurs quelconques de \mathbb{R}^p . Ceux-ci sont décrits par p variables X^1, \dots, X^p continues et on note par \mathbf{x}^j le vecteur de taille $n \times 1$ comportant les valeurs des n individus pour la variable X^j pour $j = 1, \dots, p$. Dénotons désormais par \mathbf{X}^j , les matrices carrées d'ordre n définies par, $\forall j = 1, \dots, p$:

$$\mathbf{X}^j = \mathbf{x}^j [\mathbf{x}^j]^\top. \quad (4.39)$$

Continuant le parallélisme avec le cas des variables qualitatives nominales, je dénote par \mathbf{S} la matrice agrégée suivante similaire, dans l'esprit, à la matrice \mathbf{C} :

$$\mathbf{S} = \sum_{j=1}^p \mathbf{X}^j. \quad (4.40)$$

Il est clair que $\mathbf{S} = (s_{ii'})$ est également une matrice de produits scalaires, $\forall i, i' = 1, \dots, n$:

$$\mathbf{S} = (\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle)_{i, i' = 1, \dots, n}. \quad (4.41)$$

Je m'intéresse alors à l'interprétation de la maximisation des critères de partitionnement $\Delta(\mathbf{S}, \mathbf{X})$ avec $\Delta \in \{\mathbf{B}, \mathbf{E}, \mathbf{J}\}$. Il s'agit donc de résoudre (4.21) dans le cas de données continues, où \mathbf{S} est définie par (4.41) et \mathbf{X} est la matrice relationnelle de la partition inconnue. Les critères de partitionnement s'écrivent alors comme suit :

$$\mathbf{B}(\mathbf{S}, \mathbf{X}) = \sum_{i=1}^n \sum_{i'=1}^n \left(s_{ii'} - \frac{s_{i.} + s_{.i'}}{n} + \frac{s_{..}}{n^2} \right) x_{ii'}, \quad (4.42)$$

$$\mathbf{E}(\mathbf{S}, \mathbf{X}) = \sum_{i, i'} \left(s_{ii'} - \sum_{i, i'} \frac{s_{ii'}}{n^2} \right) x_{ii'}, \quad (4.43)$$

$$\mathbf{J}(\mathbf{S}, \mathbf{X}) = \frac{1}{n} \sum_{i, i'} \left(s_{ii'} - \frac{1}{2} \left(\frac{s_{i.}}{n} + \frac{s_{.i'}}{n} \right) \right) x_{ii'}. \quad (4.44)$$

J'introduis alors le **principe d'écart à une tendance centrale**. En effet, pour l'ensemble

4.2. CONTRIBUTIONS

des trois critères précédents, on peut les mettre sous la forme générique suivante :

$$\Delta(\mathbf{S}, \mathbf{X}) = \sum_{i=1}^n \sum_{i'=1}^n \left(\tilde{s}_{ii'} - \mu(\tilde{\mathbf{S}}, i, i') \right) x_{ii'}, \quad (4.45)$$

où $\tilde{\mathbf{S}}$ est une transformation de \mathbf{S} et $\mu(\tilde{\mathbf{S}}, i, i')$ est une application qui prend en entrée la matrice $\tilde{\mathbf{S}}$ et les indices de deux individus, et donne en sortie un réel qui représente une tendance centrale de similarités. Cette notion de tendance centrale peut être de nature globale ou locale.

L'équation (4.45) s'interprète comme suit : plus la similarité (transformée) $\tilde{s}_{ii'}$ entre deux individus X_i et $X_{i'}$ est grande relativement à une tendance centrale (globale ou locale) $\mu(\tilde{\mathbf{S}}, i, i')$, plus la vraisemblance que $x_{ii'} = 1$ (X_i et $X_{i'}$ dans le même *cluster* de consensus) est forte. Chaque critère repose sur la définition d'une transformation de la similarité initiale et sur l'utilisation d'un modèle de tendance centrale. Je précise ces éléments pour chacun des critères dans la Table 4.1 où j'utilise la notation $s_i = \sum_{i'=1}^n s_{ii'}$.

	Transformation	Tendance centrale
B	$\tilde{s}_{ii'}^B = s_{ii'} - \frac{s_i + s_{i'}}{n} + \frac{s}{n^2}$	$\mu^B(\tilde{\mathbf{S}}^B, i, i') = \frac{1}{n^2} \sum_{i, i'=1}^n \tilde{s}_{ii'}^B = 0$
E	$\tilde{s}_{ii'}^E = s_{ii'}$	$\mu^E(\tilde{\mathbf{S}}^E, i, i') = \frac{1}{n^2} \sum_{i, i'=1}^n \tilde{s}_{ii'}^E$
J	$\tilde{s}_{ii'}^J = s_{ii'}$	$\mu^J(\tilde{\mathbf{S}}^J, i, i') = \frac{1}{2} \left(\frac{1}{n} \sum_{i'=1}^n \tilde{s}_{ii'}^J + \frac{1}{n} \sum_{i=1}^n \tilde{s}_{ii'}^J \right)$

TABLE 4.1 – Modèles de tendance centrale des 3 critères de partitionnement B, E et J.

Pour B, le critère de Belson, la matrice $\tilde{\mathbf{S}}^B$ est connue en analyse de donnée et notamment en MDS (*Multidimensional Scaling*). Elle correspond à la transformation de Warren Torgerson [Torgerson, 1952] également employée sous le nom de *double centering operator*. Étant donnée une matrice de produits scalaires quelconque, cette transformation renvoie la matrice de produits scalaires entre vecteurs centrés vis-à-vis du barycentre du nuage de points et on a donc, $\forall i, i' = 1, \dots, n$:

$$\tilde{s}_{ii'}^B = \langle \mathbf{x}_i - \mathbf{m}, \mathbf{x}_{i'} - \mathbf{m} \rangle,$$

avec $\mathbf{m} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$.

Dans ce cas, nous avons la moyenne arithmétique des valeurs de $\tilde{\mathbf{S}}^B$ qui vaut 0. Ainsi, la notion de tendance centrale dans le cas de Belson est de nature géométrique et est relative au barycentre.

Ensuite dans le cas de la mesure E, initialement fondée sur l'écart carré à l'indépendance,

son interprétation dans le cas des variables quantitatives, est associée à une transformation identité et à une tendance centrale globale qui est simplement la moyenne arithmétique des valeurs de \tilde{S}^E . Notons que cette valeur moyenne est généralement non nulle contrairement à ce que l'on observe pour le critère précédent B.

En ce qui concerne J, le critère de Jordan, il est également associé à la transformation identité. Toutefois, contrairement à E, la tendance centrale est locale. La quantité $\frac{s_i}{n}$ est la moyenne arithmétique du profil de similarité de X_i . Ainsi, la tendance centrale $\mu^J(\tilde{S}, i, i')$ est la moyenne arithmétique des moyennes arithmétiques des profils de similarité de X_i et $X_{i'}$. Par conséquent, les valeurs des tendances centrales sont distinctes pour chaque paire $(X_i, X_{i'})$.

Dans [Ah-Pine, 2009], j'étudie ces critères de partitionnement B, E et J dans le cas d'individus décrits par des variables quantitatives. Je considère des approches mixtes en supposant par défaut des données centrées (ce qui est motivé par B), et en calculant les tendances centrales sur les mesures de similarité strictement positives. Dans ce cas, deux individus auront tendance à être dans le même *cluster*, si leur mesure de similarité est très discriminante.

Le programme linéaire en nombres bivalents (4.21) est NP-dur. Par conséquent, en pratique, on utilise une heuristique. Cette dernière permet de ne pas fixer le nombre de *clusters* mais il est requis de donner une borne supérieure. Elle repose sur des **opérations de transfert des individus vers un *cluster*** en vue de maximiser la fonction objectif. L'**heuristique de l'AR**² procède comme suit :

- Elle prend le premier individu de la liste comme le premier élément du premier *cluster*.
- Elle parcourt la liste des individus suivants et pour chaque individu :
 - Elle calcule sa contribution à chaque *cluster* existant et retient l'identité du *cluster* auquel il contribue le plus au sens de la maximisation du critère de partitionnement utilisé.
 - Ensuite, elle compare la contribution de ce potentiel transfert vis-à-vis de la contribution obtenue par la création d'un nouveau *cluster* dont l'individu serait le premier élément. Si le gain de cette dernière situation est plus grand et si la borne du nombre de *clusters* n'est pas atteinte alors l'heuristique crée un nouveau *cluster*, sinon elle transfère l'individu vers le *cluster* optimal trouvé auparavant.
- L'algorithme effectue plusieurs itérations³ jusqu'à un critère d'arrêt.

Cette heuristique permet d'optimiser n'importe lequel des critères de partitionnement que j'ai introduit précédemment. En revanche, elle est très dépendante de l'ordre de traitement des individus qui influe fortement sur le résultat obtenu. Dans [Ah-Pine, 2009], je propose une stratégie déterministe utilisant des statistiques simples sur les profils de similarités des individus pour fixer un ordre.

Après avoir proposé l'extension des critères de partitionnement développés en AR au cas des individus représentés par des variables quantitatives, j'ai approfondi, dans l'article

2. Notons que cette procédure est proche de la technique nommée *first leader* dans [Hartigan, 1975].

3. Notons que l'approche *first leader* ne fait qu'une seule passe.

[Ah-Pine, 2013b], le **point de vue graphe** qui est sous-jacent à l'ensemble de ces propositions. La matrice de produits scalaires \mathbf{S} précédente peut être vue telle une matrice d'affinités possédant une structure métrique qui se traduit par le fait qu'elle est semi-définie positive. Je me place désormais dans un cadre où je suppose que les individus X_1, \dots, X_n , représentent l'ensemble des sommets d'un **graphe non orienté et non valué**. Une arête existant entre deux sommets indique que ces individus partagent une affinité avérée sans que cette information provienne d'un espace géométrique. Dans ce contexte purement relationnel, le *clustering* revient à **partitionner les sommets d'un graphe en sous-graphes complets (cliques) qui sont mutuellement non connectés**. Ce problème a connu un essor important dans les années 2000 suite au développement des réseaux sociaux. En effet, le partitionnement de graphe permet de **détecter des communautés** et de structurer un réseau en composantes homogènes et ceci peut grandement faciliter son étude. L'analyse des réseaux sociaux et la détection de communautés intéressent de nombreuses disciplines y compris en sciences humaines et sociales.

Dans ce contexte, j'interprète la matrice d'adjacence du graphe observé comme étant une **partition bruitée** qui ne respecte pas la propriété de transitivité. L'objectif est alors de **retrouver la partition bien formée en maximisant un des critères de partitionnement introduits précédemment**. Rappelons que ceux-ci sont initialement fondés à partir de la notion d'**indépendance statistique** entre deux variables qualitatives. Par conséquent, il s'agit implicitement de déterminer la partition la plus associée au graphe en maximisant une quantité qui dépend de la mesure de l'écart à la situation d'indépendance statistique.

Soit alors $\mathbf{A} = (a_{ii'})$ une matrice binaire carrée d'ordre n représentant la matrice d'adjacence du graphe non orienté et non pondéré des affinités au sein de l'ensemble des individus. Je m'intéresse aux critères de partitionnement de graphe ci-dessous :

$$\mathbf{B}(\mathbf{A}, \mathbf{X}) = \sum_{i=1}^n \sum_{i'=1}^n \left(a_{ii'} - \frac{a_{i.} + a_{.i'}}{n} + \frac{a_{..}}{n^2} \right) x_{ii'}, \quad (4.46)$$

$$\mathbf{E}(\mathbf{A}, \mathbf{X}) = \sum_{i,i'} \left(a_{ii'} - \sum_{i,i'} \frac{a_{ii'}}{n^2} \right) x_{ii'}, \quad (4.47)$$

$$\mathbf{J}(\mathbf{A}, \mathbf{X}) = \frac{1}{n} \sum_{i,i'} \left(a_{ii'} - \frac{1}{2} \left(\frac{a_{i.}}{n} + \frac{a_{.i'}}{n} \right) \right) x_{ii'}, \quad (4.48)$$

$$\mathbf{LM}(\mathbf{A}, \mathbf{X}) = \sum_{i=1}^n \sum_{i'=1}^n \left(\frac{2a_{ii'}}{a_{i.} + a_{.i'}} - \frac{1}{n} \right) x_{ii'}. \quad (4.49)$$

Il s'agit des mêmes méthodes qu'auparavant à l'exception du critère de Light-Margolin que j'ajoute à l'étude. Je donne quelques précisions sur cette mesure. J'ai indiqué deux expressions équivalentes de $\mathbf{LM}(\mathbf{X}^k, \mathbf{X}^l)$: (4.17) et (4.18). Cependant, j'ai précisé que l'équivalence venait du fait que les matrices relationnelles \mathbf{X}^k et \mathbf{X}^l encodent des relations d'équivalence. Or, dans le cas présent, la matrice d'adjacence binaire observée \mathbf{A} n'est pas transitive et n'encode donc pas une partition. Les expressions n'étant pas ici équivalentes, j'opte finalement pour

celle de l'équation (4.17) qui est plus appropriée. En effet, cette formulation est symétrique en X_i et $X_{i'}$ ce qui est plus adéquat. Du point de vue empirique elle procure de très bonnes performances comme je l'évoquerai ultérieurement. Enfin, elle présente des liens avec les méthodes de normalisation de la matrice Laplacienne d'un graphe que j'introduirai dans le paragraphe suivant.

A ce stade, j'évoque des concepts importants dans la champ de la détection de communautés dans des graphes. Il s'agit des travaux de Marc Newman [Newman, 2018] et de l'introduction du critère de **modularité** par Michelle Girvan et ce dernier dans [Newman and Girvan, 2004], qui ont généré une très vaste littérature. Il existe en fait des liens intéressants entre le modèle probabiliste sous-jacent au critère classique de modularité et le principe d'écart à une tendance centrale inhérent aux mesures d'association statistique.

Le concept de modularité est une mesure qui vise à modéliser la notion de communauté. Je dénote cette mesure par Q dans la suite et le principe qui le fonde peut être grossièrement exprimé comme suit :

Q = "Nb. d'arêtes dans une communauté – Espérance du nb. d'arêtes dans cette communauté".

Par conséquent, une communauté est d'autant plus avérée que sa mesure de modularité est grande. Dénotons par $\mathbf{P} = (p_{ii'})$ une matrice carrée d'ordre n avec $p_{ii'}$ indiquant l'espérance du nombre d'arêtes entre X_i et $X_{i'}$. Étant donné que le graphe est non pondéré, $p_{ii'}$ peut être interprétée telle la probabilité qu'il y ait une arête entre ces deux sommets. Supposons qu'il y ait m arêtes au sein du graphe. Le concept de modularité s'exprime formellement comme suit [Newman, 2006b] :

$$Q(\mathbf{A}, \delta) = \frac{1}{2m} \sum_{i=1}^n \sum_{i'=1}^n (a_{ii'} - p_{ii'}) \delta(g_i, g_{i'}), \quad (4.50)$$

où g_i est l'application qui renvoie le *cluster* de X_i et $\delta(g_i, g_{i'}) = 1$ si $g_i = g_{i'}$ et 0 sinon.

De cette formulation générale, Mark Newman fait différentes hypothèses afin d'instancier le modèle probabiliste représenté par \mathbf{P} :

- Comme le graphe est non orienté, on doit avoir : $p_{ii'} = p_{i'i}, \forall i, i' = 1, \dots, n$.
- Q doit être nul lorsque tous les individus sont dans un seul *cluster* ce qui implique : $\sum_{i,i'} a_{ii'} = \sum_{i,i'} p_{ii'} = 2m$.
- La distribution des degrés du modèle aléatoire représenté par \mathbf{P} doit être similaire à la distribution empirique des degrés et on pose : $\sum_{i'} p_{ii'} = \sum_{i'} a_{ii'} = a_i, \forall i = 1, \dots, n$.
- Les arêtes sont placés aléatoirement ce qui implique que la probabilité d'observer une arête entre X_i et $X_{i'}$ doit être indépendante de la probabilité d'observer une arête incidente à X_i et la probabilité d'observer une arête incidente à $X_{i'}$.

4.2. CONTRIBUTIONS

Sous ces hypothèses, le modèle le plus simple est alors donné par, $\forall i, i' = 1, \dots, n$:

$$p_{ii'} = \frac{a_i a_{i'}}{2m}. \quad (4.51)$$

Ceci conduit à la définition usuelle de la mesure de modularité :

$$Q(\mathbf{A}, \mathbf{X}) = \frac{1}{a_{..}} \sum_{i=1}^n \sum_{j=1}^n \left(a_{ij} - \frac{a_i a_j}{a_{..}} \right) x_{ij}. \quad (4.52)$$

où j'ai utilisé la relation $a_{..} = 2m$ et la notation \mathbf{X} qui est la matrice relationnelle de la partition définissant les communautés recherchées après avoir facilement constaté que $x_{ii'} = \delta(g_i, g_{i'})$, $\forall i, i' = 1, \dots, n$.

L'expression relationnelle de Q dans (4.52) est à rapprocher des expressions des mesures B , E , J et LM données par (4.46), (4.47), (4.48), (4.49) respectivement. Tous ces critères de partitionnement sont définis à partir de principes très proches. D'ailleurs, pour aller plus loin dans ces relations, j'ai établi l'expression de la modularité en terme de table de contingence. Je suppose dans ce cas, que \mathbf{A} et \mathbf{X} sont les matrices relationnelles de deux variables qualitatives nominales X^k et X^l respectivement. J'emploie les formules de passage contingenciel-relationnel données en page 84 et celle qui suit :

$$\sum_{v=1}^{p_l} \left(\sum_{u=1}^{p_k} n_{uv} n_u \right)^2 = \sum_{i=1}^n \sum_{i'=1}^n x_i^k x_{i'}^k x_{ii'}^l. \quad (4.53)$$

Je montre alors l'identité ci-dessous [Ah-Pine, 2013b] :

$$Q(X^k, X^l) = \frac{1}{\sum_{u=1}^{p_k} n_u^2} \left(\sum_{u=1}^{p_k} \sum_{v=1}^{p_l} n_{uv}^2 - \frac{1}{\sum_{u=1}^{p_k} n_u^2} \left(\sum_{v=1}^{p_l} \left(\sum_{u=1}^{p_k} n_u n_{uv} \right)^2 \right) \right). \quad (4.54)$$

Il est intéressant d'indiquer que dans ce contexte, la modularité est nulle si les variables sont statistiquement indépendantes : $X^k \perp X^l \Rightarrow Q(X^k, X^l) = 0$.

Afin de tester l'intérêt des critères introduits et de les comparer à la modularité, j'ai mené des expériences à partir de données simulées. Il s'agit des *LFR benchmarks* proposés par Andrea Lancichinetti, Santo Fortunato et Filippo Radicchi dans [Lancichinetti et al., 2008]. Ces auteurs mettent également à disposition un outil libre qui génère des graphes possédant les caractéristiques des réseaux de terrain comme par exemple l'hétérogénéité des distributions des degrés des noeuds (loi de puissance) et des tailles de *clusters*. En particulier, le *mixing parameter* $\mu \in [0, 1]$ de leur modèle permet de contrôler graduellement la présence ou l'absence de communautés au sein du graphe à mesure que sa valeur augmente. Ainsi, le cas $\mu = 1$ est la situation où les arêtes sont générées aléatoirement sans aucune structure.

Dans la Figure 4.1 je montre les résultats obtenus avec chacun des critères, pour plusieurs nombres de noeuds (500, 1000 et 2000 de gauche à droite) et en faisant varier le paramètre μ en abscisse. Les graphiques de la ligne du haut montrent les valeurs de *Normalized Mutual*

4.2. CONTRIBUTIONS

Information (NMI) qui est une mesure d'évaluation externe classique en *clustering*. Celle-ci varie entre 0 et 1 et plus le score est grand, plus la partition trouvée et la vérité terrain sont similaires. A l'aide de ce critère, nous pouvons observer la plus ou moins grande capacité des méthodes de *graph clustering* à détecter les communautés de la vérité terrain.

Les graphiques de la ligne du bas, indiquent le nombre de *clusters* déterminés par chaque méthode et pour chaque *setting* utilisé. En effet, l'algorithme de *clustering* employé ici, qui est le même pour chaque fonction objectif, est une simple classification ascendante hiérarchique. On part de n singletons, on fusionne la paire de *clusters* permettant d'augmenter au plus le critère de partitionnement et on s'arrête lorsque plus aucune fusion n'aboutit à une amélioration de ce dernier. De ce fait, l'heuristique permet de ne pas fixer le nombre de *clusters*.

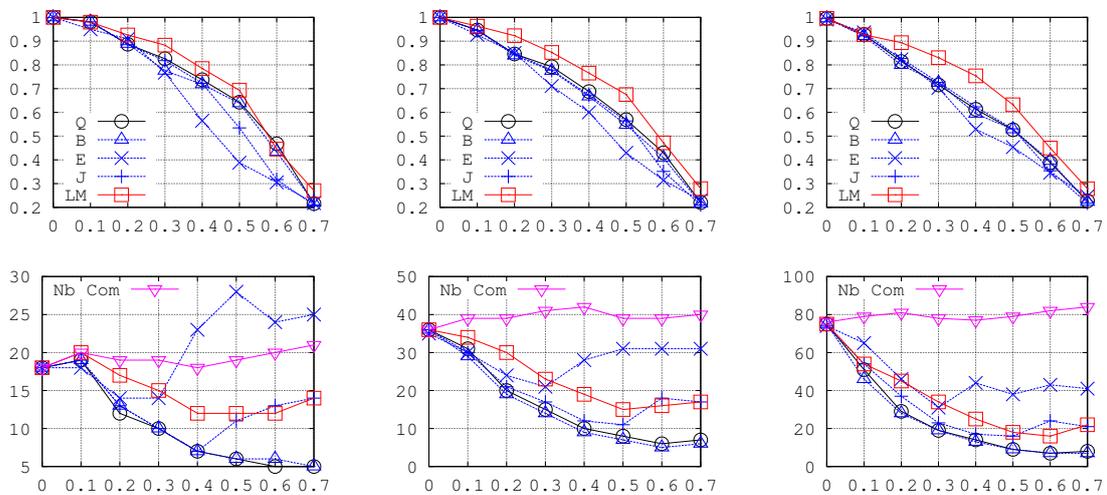


FIGURE 4.1 – Ligne du haut : valeurs NMI (axe vertical) *versus* paramètre de mélange μ (axe horizontal). Ligne du bas : Nombre de *clusters* trouvés (axe vertical) *versus* paramètre de mélange μ (axe horizontal). Les courbes pour chaque critère de partitionnement évoqué sont exposées. De gauche à droite, les graphiques correspondent à des graphes comportant 500, 1000 et 2000 sommets.

Au vu de ces résultats, le critère issu de la mesure de Light-Margolin est particulièrement performant notamment en comparaison de la modularité. Cette supériorité semble se confirmer au fur et à mesure que le nombre de sommets grandit. La transformation de la matrice d'adjacence associé au critère LM paraît être une stratégie intéressante pour le partitionnement de graphe. Je reviens sur ce type d'approche dans le cadre de la normalisation de la matrice Laplacienne d'un graphe dans le contexte du *spectral clustering* dans la section suivante.

4.2.2 Normalisation de mesures de similarité

J'ai rappelé dans le deuxième paragraphe de la sous-section 4.1.2 mes travaux antérieurs sur la similarité d'ordre t qui étend la mesure cosinus entre deux vecteurs en faisant jouer un rôle au rapport des normes de ces derniers.

Suite à ma thèse de doctorat, j'ai exploité cette approche dans différentes situations. J'ai tout d'abord appliqué l'idée centrale des similarités d'ordre t dans le cadre des méthodes à noyaux. Dans [Ah-Pine, 2010], j'introduis ainsi une famille de normalisation de matrices à noyaux. Soit $\mathbf{K} = (k_{ii'})$ une matrice de noyaux d'un ensemble de n vecteurs $\{\mathbf{x}_i\}_{i=1,\dots,n}$ de \mathbb{R}^p . Alors il existe une application dite *feature map* $\phi : \mathbb{R}^p \rightarrow \mathbb{F}$ qui projette les vecteurs de l'*input space* \mathbb{R}^p dans un *feature space* \mathbb{F} tel que $\forall i, i' = 1, \dots, n$, $k_{ii'} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'}) \rangle_{\mathbb{F}}$. Le *feature space* \mathbb{F} peut être associée à un *Reproducing Kernel Hilbert Space* (RKHS).

Je définis alors la normalisation d'ordre $t > 0$ de \mathbf{K} dénotée $\mathbf{K}^t = (k_{ii'}^t)$, par :

$$k_{ii'}^t = \frac{k_{ii'}}{M^t(k_{ii}, k_{i'i'})}. \quad (4.55)$$

Les interprétations géométriques mentionnées dans la section 4.1.2 s'étendent naturellement dans le cas des *feature vectors* dans l'espace \mathbb{F} . Par ailleurs, comme pour toute méthode à noyaux, l'approche bénéficie de l'avantage computationnel permis par le *kernel trick*. Nous avons en premier lieu, l'interprétation de $k_{ii'}^t$ en terme de symétrisation des coefficients de projection orthogonale :

$$k_{ii'}^t = M^{-t} \left(\frac{k_{ii'}}{k_{ii}}, \frac{k_{ii'}}{k_{i'i'}} \right). \quad (4.56)$$

Soient les notations suivantes :

$$\theta_{ii'} = \text{l'angle formé par } \phi(\mathbf{x}_i) \text{ et } \phi(\mathbf{x}_{i'}) \text{ dans } \mathbb{F}, \quad (4.57)$$

$$\gamma_{i'}^i = \text{le rapport des normes } \frac{\|\phi(\mathbf{x})_i\|_{\mathbb{F}}}{\|\phi(\mathbf{x}_{i'})\|_{\mathbb{F}}}, \quad (4.58)$$

$$\gamma_i^{i'} = \text{le rapport des normes } \frac{\|\phi(\mathbf{x}_{i'})\|_{\mathbb{F}}}{\|\phi(\mathbf{x}_i)\|_{\mathbb{F}}}, \quad (4.59)$$

$$\gamma_{ii'} = \text{le rapport des normes } \frac{\max(\|\phi(\mathbf{x}_i)\|_{\mathbb{F}}, \|\phi(\mathbf{x}_{i'})\|_{\mathbb{F}})}{\min(\|\phi(\mathbf{x}_i)\|_{\mathbb{F}}, \|\phi(\mathbf{x}_{i'})\|_{\mathbb{F}})}. \quad (4.60)$$

Nous avons alors en second lieu, l'expression équivalente suivante :

$$k_{ii'}^t = \frac{\cos(\theta_{ii'})}{M^t(\gamma_{i'}^i, \gamma_i^{i'})} \quad (4.61)$$

$$= \cos(\theta_{ii'}) \left(\frac{2^{1/t} \gamma_{ii'}}{(1 + \gamma_{ii'}^{2t})^{1/t}} \right). \quad (4.62)$$

Comme pour les similarités d'ordre $t > 0$, la normalisation d'ordre $t > 0$ permet de discriminer la mesure cosinus en fonction de l'écart entre les normes des vecteurs. Plus les normes $\|\phi(\mathbf{x}_i)\|_{\mathbb{F}}$ et $\|\phi(\mathbf{x}_{i'})\|_{\mathbb{F}}$ sont éloignées, plus la valeur $k_{ii'}^t$ tend vers 0. Cette convergence est d'autant plus rapide que t est grand. Ce comportement est illustré dans la Figure 4.2 dans le cas de deux vecteurs colinéaires négativement $\cos(\theta_{ii'}) = -1$ (graphique de gauche) et positivement $\cos(\theta_{ii'}) = 1$ (graphique de droite).

4.2. CONTRIBUTIONS

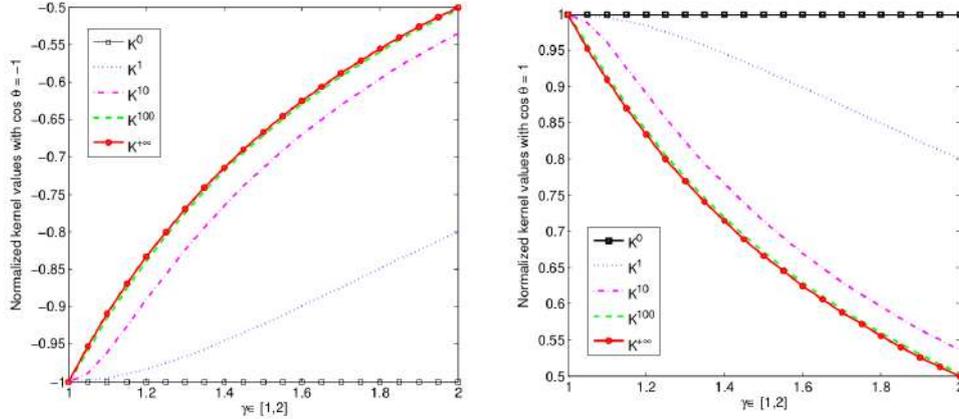


FIGURE 4.2 – Courbes représentant les valeurs de $k_{ii'}^t$ pour $t \rightarrow 0, t = 1, t = 10, t = 100, t \rightarrow \infty$; $\gamma_{ii'} \in [1, 2]$ (en abscisse); $\cos(\theta_{ii'}) = -1$ (à gauche) et $\cos(\theta_{ii'}) = 1$ (à droite).

Un aspect particulièrement important dans ce contexte concerne les propriétés de métricité de \mathbf{K}^t . En effet, dans le cas des méthodes à noyaux, il est fondamental que la matrice soit semi-définie positive. J'ai alors montré dans [Ah-Pine, 2010], le résultat suivant.

Théorème. 5. *Soit \mathbf{K} une matrice de noyaux. La matrice de noyaux normalisée \mathbf{K}^t dont le terme général est donné par (4.56), est symétrique et semi-définie positive pour tout $t > 0$.*

Ainsi, à condition que t soit positif, la normalisation d'ordre t d'une matrice de noyaux est une matrice de noyaux. La preuve de ce résultat peut être établie à l'aide du Théorème des disques de Gershgorin [Gershgorin, 1931], comme je l'ai indiqué dans [Ah-Pine, 2010]. Allen Russell et Christopher Upton donnent également une preuve directe de ce résultat dans [Russell and Upton, 1987].

J'illustre par ailleurs l'intérêt de la normalisation d'ordre t , au-delà du cas classique $t \rightarrow 0$, dans le cadre de tâches de *clustering*. Les résultats d'expériences exposés dans [Ah-Pine, 2010] indiquent que discriminer les mesures de similarité en tenant compte de la différence entre les normes en sus de la mesure angulaire, peut donner de meilleurs résultats. Toutefois, estimer l'hyperparamètre t est une question que je n'ai pas traitée. Je reviendrai sur ce point ultérieurement.

Ensuite, j'ai exploité les concepts au coeur des similarités d'ordre t dans le contexte de la mesure de similarité entre deux relations binaires structurées. J'entends par ce dernier terme, les relations d'ordre d'une part et les relations d'équivalence d'autre part. Le cadre conceptuel est à nouveau fondé sur l'analyse relationnelle (AR) et je manipulerai dans ce qui suit des matrices d'adjacence de graphes associés à ces relations binaires (RB).

Je fais alors le lien entre les développements menés en AR sur les mesures d'associations

4.2. CONTRIBUTIONS

utilisant les matrices relationnelles⁴, et le coefficient de corrélation général dénoté Γ dans l'ouvrage *Rank Correlation Methods* de Maurice Kendall [Kendall, 1948] et attribué à Henry Daniels [Daniels, 1944]. Je rappelle l'expression de Γ ci-dessous :

$$\Gamma(\mathbf{x}^k, \mathbf{x}^l) = \frac{\sum_{i,i'} x_{ii'}^k x_{ii'}^l}{\sqrt{\sum_{i,i'} [x_{ii'}^k]^2} \sqrt{\sum_{i,i'} [x_{ii'}^l]^2}}, \quad (4.63)$$

où $\mathbf{x}^k = (x_i^k)$ et $\mathbf{x}^l = (x_i^l)$ sont les vecteurs de deux variables distinctes recensant les valeurs prises par n individus, et $X^k = (x_{ii'}^k)$ et $X^l = (x_{ii'}^l)$ sont deux matrices carrées d'ordre n dérivées de \mathbf{x}^k et \mathbf{x}^l .

$\Gamma(\mathbf{x}^k, \mathbf{x}^l)$ a la forme d'un coefficient de corrélation et vise à mesurer une notion de dépendance entre les deux variables \mathbf{x}^k et \mathbf{x}^l . Selon la façon dont X^k est définie à partir de \mathbf{x}^k , on retrouve plusieurs mesures d'association connues. Prenons d'abord le cas de deux variables quantitatives ou ordinales. Dans le contexte non paramétrique des corrélations de rangs, on a les cas particuliers suivants :

- Si $\forall i, i' = 1, \dots, n : x_{ii'}^k = 1$ si $x_i^k < x_{i'}^k$, et $x_{ii'}^k = -1$ si $x_i^k > x_{i'}^k$, alors $\Gamma(\mathbf{x}^k, \mathbf{x}^l)$ est équivalent au τ de Kendall.
- Supposons que \mathbf{x}^k encode les rangs des individus lorsqu'on ordonne ces derniers selon leurs valeurs initiales. Dans ce cas, si $\forall i, i' = 1, \dots, n : x_{ii'}^k = x_{i'}^k - x_i^k$, alors $\Gamma(\mathbf{x}^k, \mathbf{x}^l)$ est équivalent au ρ de Spearman.

Si \mathbf{x}^k encode les valeurs des individus et non leurs rangs on a le cas classique suivant :

- Si $\forall i, i' = 1, \dots, n : x_{ii'}^k = x_{i'}^k - x_i^k$, alors $\Gamma(\mathbf{x}^k, \mathbf{x}^l)$ est équivalent à $r(\mathbf{x}^k, \mathbf{x}^l)$, le coefficient de corrélation linéaire de Bravais-Pearson.

Regardons à présent le cas de deux variables qualitatives nominales. Dans [Janson and Vegelius, 1982], les auteurs montrent que le coefficient T de Tchuprow (ϕ^2 normalisé) est un cas particulier de Γ et proposent un indice qui n'est autre que le *J-index* déjà rappelé précédemment :

- Notons par n_u^k le nombre d'individus ayant la modalité u de X^k avec $u = 1, \dots, p_k$ où p_k est le nombre de modalités. Si $\forall i, i' = 1, \dots, n : x_{ii'}^k = (n/n_u^k) - 1$ si $x_i^k = x_{i'}^k$, et $x_{ii'}^k = -1$ si $x_i^k \neq x_{i'}^k$, alors $\Gamma(\mathbf{x}^k, \mathbf{x}^l)$ est équivalent au coefficient T de Tchuprow.
- Si $\forall i, i' = 1, \dots, n : x_{ii'}^k = p_k - 1$ si $x_i^k = x_{i'}^k$, et $x_{ii'}^k = -1$ si $x_i^k \neq x_{i'}^k$, alors $\Gamma(\mathbf{x}^k, \mathbf{x}^l)$ est par définition le *J-index* de Janson et Vegelius.

Dans [Ah-Pine, 2013a], afin d'établir une **formule qui unifie relations d'ordre et relations d'équivalence** tout en permettant une interprétation relationnelle motivée, j'ai

4. Dans le premier paragraphe de la sous-section 4.1.2, j'ai rappelé des travaux en AR sur les mesures d'associations entre variables qualitatives nominales. Il existe également en AR des travaux qui traitent des mesures d'association entre variables ordinales en utilisant les matrices relationnelles. Le lecteur intéressé pourrait consulter mon manuscrit de thèse de doctorat [Ah-Pine, 2007], qui présente également quelques apports dans cet axe, ainsi que les références qui y sont mentionnées.

4.2. CONTRIBUTIONS

proposé de définir la variable \mathbf{X}^k de façon générique comme suit, $\forall i, i' = 1, \dots, n$:

$$\mathbf{x}_{ii'}^k = m_{ii'}^k x_{ii'}^k - \check{m}_{ii'}^k \check{x}_{ii'}^k, \quad (4.64)$$

où :

- $\mathbf{X}^k = (x_{ii'}^k)$ est la matrice relationnelle de la **RB sous-jacente à \mathbf{x}^k** . Si \mathbf{x}^k est quantitative ou ordinale alors pour toute paire (i, i') , $x_{ii'}^k = 1$ si $x_i^k < x_{i'}^k$, et $x_{ii'}^k = 0$ sinon. Si \mathbf{x}^k est qualitative nominale alors $x_{ii'}^k = 1$ si $x_i^k = x_{i'}^k$, et $x_{ii'}^k = 0$ sinon ;
- $\check{\mathbf{X}}^k = (\check{x}_{ii'}^k)$ est la matrice relationnelle de la **RB inverse de la RB sous-jacente à \mathbf{x}^k** . Si \mathbf{x}^k est quantitative ou ordinale alors pour toute paire (i, i') , $\check{x}_{ii'}^k = 1$ si $x_i^k > x_{i'}^k$ et $\check{x}_{ii'}^k = 0$ sinon. Si \mathbf{x}^k est qualitative nominale alors $\check{x}_{ii'}^k = 1$ si $x_i^k \neq x_{i'}^k$ et $\check{x}_{ii'}^k = 0$ sinon ;
- $m_{ii'}^k$ est le poids que l'on attribue à la paire (i, i') lorsque celle-ci vérifie la RB sous-jacente à \mathbf{x}^k , c'est à dire lorsque $x_{ii'}^k = 1$;
- $\check{m}_{ii'}^k$ est le poids que l'on attribue à la paire (i, i') lorsque celle-ci vérifie la RB inverse à la RB sous-jacente à \mathbf{x}^k , c'est à dire lorsque $\check{x}_{ii'}^k = 1$.

J'introduis alors la forme spécifique suivante du coefficient de corrélation général, que je dénote par Λ qui dépend alors des matrices relationnelles \mathbf{X}^k et \mathbf{X}^l mais également des matrices de poids $\mathbf{M}^k, \mathbf{M}^l, \check{\mathbf{M}}^k$ et $\check{\mathbf{M}}^l$ permettant de paramétrer le coefficient de corrélation :

$$\Lambda(\mathbf{X}^k, \mathbf{X}^l, \mathbf{M}^k, \check{\mathbf{M}}^k, \mathbf{M}^l, \check{\mathbf{M}}^l) = \frac{\sum_{i,i'} (m_{ii'}^k x_{ii'}^k - \check{m}_{ii'}^k \check{x}_{ii'}^k)(m_{ii'}^l x_{ii'}^l - \check{m}_{ii'}^l \check{x}_{ii'}^l)}{\sqrt{\sum_{i,i'} (m_{ii'}^k x_{ii'}^k - \check{m}_{ii'}^k \check{x}_{ii'}^k)^2} \sqrt{\sum_{i,i'} (m_{ii'}^l x_{ii'}^l - \check{m}_{ii'}^l \check{x}_{ii'}^l)^2}}. \quad (4.65)$$

Distinguer deux systèmes de poids $(\mathbf{M}^k, \check{\mathbf{M}}^k)$ et $(\mathbf{M}^l, \check{\mathbf{M}}^l)$ pour X^k et X^l , permet une grande flexibilité. Cependant, les mesures de corrélation classiques obtenues à partir de (4.65) utilisent la même stratégie de pondération pour les deux variables. J'indique ces stratégies dans la Table 4.2 et les cas particuliers qu'elles permettent d'obtenir.

Les résultats de la Table 4.2 permettent de mieux comprendre les différences entre plusieurs mesures de corrélation distinctes au travers de l'analyse des pondérations inhérentes à chaque méthode. De plus, le formalisme commun exprimé par l'équation (4.65) met en lumière un autre concept développé en AR qui est celui de l'**indétermination**. En statistiques, il est classique de quantifier la relation entre deux variables en se référant à la situation d'indépendance. Si l'indépendance est de nature probabiliste et multiplicative, l'indétermination est de nature "logique" et additive. Considérons le cas de variables binaires ce qui est notre contexte ici puisque nous employons des matrices relationnelles, c'est à dire des matrices d'adjacence binaires. J'utilise les notations introduites plus haut, page 86, et qui correspondent aux comptages des 4 configurations 1-1, 0-0, 1-0 et 0-1. J'applique toutefois celles-ci à des matrices d'adjacence binaires et non pas à des vecteurs binaires.

On dira que deux matrices relationnelles \mathbf{X}^k et \mathbf{X}^l sont statistiquement indépendantes si

4.2. CONTRIBUTIONS

	$m_{ii'}^k$	$\check{m}_{ii'}^k$
τ de Kendall	1	1
ρ de Spearman	$\sum_{i''} x_{ii''}^k - \sum_{i''} x_{i''i'}^k$	$\sum_{i''} x_{ii''}^k - \sum_{i''} x_{i''i'}^k$
T de Tchuprow	$1/\sum_{i'} x_{ii'}^k - 1/n$	$1/n$
J -index de Janson et Vegelius	$1 - 1/p_k$	$1/p_k$
Indice de Rand	$1 - \sum_{i,i'} x_{ii'}^k/n^2$	$\sum_{i,i'} x_{ii'}^k/n^2$

TABLE 4.2 – Matrices de poids \mathbf{M}^k (même formules pour \mathbf{M}^l) et mesures correspondantes : corrélations de rangs (variables quantitatives ou ordinales) en haut et mesures d'association (variables qualitatives) en bas.

leur *odd-ratio* $\text{OR}(\mathbf{X}^k, \mathbf{X}^l)$ vaut 1 ce qui implique la relation suivante :

$$\begin{aligned}
 \mathbf{X}^k \text{ et } \mathbf{X}^l \text{ sont indépendantes} &\Leftrightarrow \text{OR}(\mathbf{X}^k, \mathbf{X}^l) = 1 \\
 &\Leftrightarrow \frac{11_{kl}00_{kl}}{10_{kl}01_{kl}} = 1 \\
 &\Leftrightarrow 11_{kl}00_{kl} - 10_{kl}01_{kl} = 0. \tag{4.66}
 \end{aligned}$$

De façon différente, on dira que deux matrices relationnelles sont en **situation d'indétermination**, s'il y a autant d'accords que de désaccords, ce qui se traduit par la relation suivante :

$$\mathbf{X}^k \text{ et } \mathbf{X}^l \text{ sont en situation d'indétermination} \Leftrightarrow (11_{kl} + 00_{kl}) - (10_{kl} + 01_{kl}) = 0. \tag{4.67}$$

Des liens étroits entre indépendance et indétermination existent. Dans [Ah-Pine and Marcotorchino, 2010], j'ai contribué à mettre en évidence une certaine forme de dualité entre ces deux concepts en exploitant les formules de passage contingenciel-relationnel. Si on s'intéresse au numérateur de (4.65), on établit facilement la relation suivante :

$$\begin{aligned}
 &\sum (m_{ii'}^k x_{ii'}^k - \check{m}_{ii'}^k \check{x}_{ii'}^k)(m_{ii'}^l x_{ii'}^l - \check{m}_{ii'}^l \check{x}_{ii'}^l) = 0 \\
 &\Leftrightarrow \\
 &\underbrace{\sum m_{ii'}^k m_{ii'}^l x_{ii'}^k x_{ii'}^l}_{\text{Poids total de } 11_{kl}} + \underbrace{\sum \check{m}_{ii'}^k \check{m}_{ii'}^l \check{x}_{ii'}^k \check{x}_{ii'}^l}_{\text{Poids total de } 00_{kl}} = \underbrace{\sum m_{ii'}^k \check{m}_{ii'}^l x_{ii'}^k \check{x}_{ii'}^l}_{\text{Poids total de } 10_{kl}} + \underbrace{\sum \check{m}_{ii'}^k m_{ii'}^l \check{x}_{ii'}^k x_{ii'}^l}_{\text{Poids total de } 01_{kl}}. \tag{4.68}
 \end{aligned}$$

Ainsi, Λ est nul en cas de situation d'indétermination pondérée dans le cadre de

laquelle le poids total des accords positifs et négatifs égale le poids total des deux types de désaccords.

Au-delà de cette unification, j'ai proposé une extension supplémentaire du coefficient de corrélation général de Daniels et Kendall, en intégrant les principes sous-jacents aux indices de similarité d'ordre t . *In fine*, la forme générique que je suggère est un **coefficient de similarité général d'ordre t** que je dénote ici Λ^t . Celui-ci est fondé à la fois sur l'indétermination pondérée et une normalisation basée sur les moyennes généralisées d'ordre t :

$$\Lambda^t(\dots) = \frac{\sum_{i,i'}(m_{ii'}^k x_{ii'}^k - \check{m}_{ii'}^k \check{x}_{ii'}^k)(m_{ii'}^l x_{ii'}^l - \check{m}_{ii'}^l \check{x}_{ii'}^l)}{\left[\frac{1}{2} \left([\sum_{i,i'}(m_{ii'}^k x_{ii'}^k - \check{m}_{ii'}^k \check{x}_{ii'}^k)^2]^t + [\sum_{i,i'}(m_{ii'}^l x_{ii'}^l - \check{m}_{ii'}^l \check{x}_{ii'}^l)^2]^t \right) \right]^{1/t}}. \quad (4.69)$$

On reconnaît au dénominateur de (4.69), la moyenne généralisée d'ordre t des termes $\sum_{i,i'}(m_{ii'}^k x_{ii'}^k - \check{m}_{ii'}^k \check{x}_{ii'}^k)^2$ et $\sum_{i,i'}(m_{ii'}^l x_{ii'}^l - \check{m}_{ii'}^l \check{x}_{ii'}^l)^2$. Dans le contexte des RB, l'avantage de Λ^t sur Λ est qu'il permet de tenir compte de l'hétérogénéité pouvant exister entre deux RB de même type. Prenons l'exemple de deux variables qualitatives X^k et X^l engendrant deux matrices relationnelles d'équivalence \mathbf{X}^k et \mathbf{X}^l . Si leurs nombres de modalités p_k et p_l sont très différents, \mathbf{X}^k et \mathbf{X}^l représentent des graphes avec un fort déséquilibre entre leurs nombres d'arêtes. La densité du graphe associé à \mathbf{X}^k est captée par la quantité $\sum_{i,i'}(m_{ii'}^k x_{ii'}^k - \check{m}_{ii'}^k \check{x}_{ii'}^k)^2$. Λ^t permet alors de tenir compte de la différence de densités par une sorte de pénalisation dont l'intensité est contrôlée par t .

Enfin, j'ai appliqué les idées sous-jacentes aux similarités d'ordre t pour la **normalisation de la matrice Laplacienne en *spectral clustering***. En bref, cette approche moderne de classification automatique, repose sur des concepts et résultats provenant de la théorie spectrale de graphe que l'on applique pour le partitionnement de graphe. On cherche à déterminer une collection de sous-ensembles denses de noeuds qui soient mutuellement disjoints. L'objectif est de respecter autant que possible le graphe initial et par conséquent, on cherche à minimiser le nombre d'arêtes que l'on doit supprimer du graphe afin d'obtenir une décomposition en composantes connexes. Ce problème dit de *min cut* est NP-dur et pour le résoudre de façon approximative, on peut exploiter le spectre de la matrice Laplacienne associé au graphe. En effet, on montre que celle-ci est toujours semi-définie positive et que l'ordre de multiplicité de la valeur propre nulle correspond au nombre de composantes connexes du graphe. De plus, les vecteurs propres associés à la valeur propre nulle définissent les différentes composantes connexes. Ce résultat théorique a été exploité en *data mining* et *machine learning* et a donné lieu à l'approche dite *spectral clustering*. C'est notamment dans le contexte de la segmentation d'images que cette méthode a été initialement suggérée par Jianbo Shi et Jitendra Malik [Shi and Malik, 2000]. Étant donné un graphe d'affinités (arêtes valuées) entre individus (sommets du graphe), la procédure classique se déroule comme suit :

- Former la matrice Laplacienne notée \mathbf{L} à partir de la matrice d'adjacence notée \mathbf{X} .
- Normaliser la matrice Laplacienne (optionnel).

- Effectuer la décomposition spectrale de la matrice Laplacienne (normalisée, le cas échéant).
- Représenter les individus dans le sous-espace engendré par les vecteurs propres associées aux k plus petites valeurs propres (*spectral embedding*).
- Appliquer l'algorithme des *k-means* pour partitionner les individus représentés dans cet espace vectoriel.

On suppose ici que le graphe est non orienté et pondéré par des valeurs non négatives si bien que la matrice d'adjacence est telle que $a_{ii'} \geq 0$ pour tout $i, i' = 1, \dots, n$. On obtient la **matrice Laplacienne \mathbf{L}** à partir de \mathbf{A} comme suit :

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \quad (4.70)$$

où $\mathbf{D} = (d_{ii'})$ est la matrice des degrés défini par $d_{ii} = \sum_{i'=1}^n a_{ii'}$ pour tout $i = 1, \dots, n$ et $d_{ii'} = 0$ pour tout couple (i, i') tel que $i \neq i'$.

La matrice \mathbf{L} possède les propriétés exposées précédemment. Cependant, on lui préfère une version normalisée qui présente ces mêmes caractéristiques mais qui a de meilleures propriétés de convergence asymptotique (voir [Von Luxburg et al., 2008] par exemple). Il existe deux types classiques de normalisation : *random walk* [Meila and Shi, 2000] et *symmetric* [Ng et al., 2001]. Ceux-ci sont étroitement liés. Je m'intéresserai exclusivement à la **normalisation symétrique**. La matrice Laplacienne normalisée sera notée dans ce cas par \mathbf{L}_{sym} et elle est définie par :

$$\mathbf{L}_{sym} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \quad (4.71)$$

J'introduis alors \mathbf{N} , la matrice carrée d'ordre n de terme général, $\forall i, i' = 1, \dots, n$:

$$n_{ii'} = \frac{1}{\sqrt{d_{ii}d_{i'i'}}}. \quad (4.72)$$

On peut alors définir \mathbf{L}_{sym} à partir de \mathbf{L} et \mathbf{N} par le biais du **produit terme à terme de Hadamard** dénoté \odot :

$$\mathbf{L}_{sym} = \mathbf{L} \odot \mathbf{N}. \quad (4.73)$$

Dans [Ah-Pine, 2017], je propose d'employer cette dernière expression afin d'introduire une **famille de normalisation de la matrice Laplacienne**. Elle est définie à partir de $\mathbf{N}^t = (n_{ii'}^t)$ une matrice carrée d'ordre n qui dépend d'un paramètre $t > 0$. Je dénomme celle-ci par **matrice de normalisation d'ordre $t > 0$** et son terme général est donné par $\forall i, i' = 1, \dots, n$:

$$n_{ii'}^t = \frac{1}{M^t(d_{ii}, d_{i'i'})}, \quad (4.74)$$

où M^t est la moyenne généralisée d'ordre t introduite en page 88.

4.3. DISCUSSIONS ET PERSPECTIVES

La **matrice Laplacienne normalisée d'ordre** $t > 0$, \mathbf{L}_{sym}^t , est alors donnée par :

$$\mathbf{L}_{sym}^t = \mathbf{L} \odot \mathbf{N}^t. \quad (4.75)$$

Nous avons un résultat similaire au Théorème 5 concernant la normalisation d'ordre $t > 0$ d'une matrice de noyaux.

Théorème. 6. *Soit \mathbf{A} la matrice d'adjacence d'un graphe non orienté pondéré de valuations non négatives et soit \mathbf{L} la matrice Laplacienne associée. La matrice Laplacienne normalisée d'ordre t , \mathbf{L}_{sym}^t , définie par (4.75) est semi-définie positive pour tout $t > 0$.*

Afin d'appréhender l'impact du paramètre $t > 0$, j'étudie la forme quadratique associée à \mathbf{L}_{sym}^t . Soit \mathbf{f} un vecteur quelconque de \mathbb{R}^n , on a alors :

$$\mathbf{f}^\top \mathbf{L}_{sym}^t \mathbf{f} = \sum_{i=1}^n f_i^2 - \sum_{i,i'=1}^n \frac{x_{ii'}}{\left(\frac{1}{2}([d_{ii}]^t + [d_{i'i'}]^t)\right)^{\frac{1}{t}}} f_i f_{i'}.$$

En *spectral clustering*, on vise à minimiser $\mathbf{f}^\top \mathbf{L}_{sym}^t \mathbf{f}$. Les sommets X_i et $X_{i'}$ auront une vraisemblance d'être dans la même classe d'autant plus forte que $x_{ii'}/\left(\frac{1}{2}([d_{ii}]^t + [d_{i'i'}]^t)\right)^{\frac{1}{t}}$ est grand. Comme pour tout $t > 0$, $\sqrt{d_{ii}d_{i'i'}} \leq \left(\frac{1}{2}([d_{ii}]^t + [d_{i'i'}]^t)\right)^{\frac{1}{t}}$, et que l'égalité est atteinte si $d_{ii} = d_{i'i'}$, nous voyons que la moyenne généralisée d'ordre $t > 0$ pénalise la différence entre d_{ii} et $d_{i'i'}$ (toute chose étant égale par ailleurs). Autrement dit, pour que X_i et $X_{i'}$ soient dans la même classe, non seulement $x_{ii'}$ doit être grand mais d_{ii} et $d_{i'i'}$ doivent être proches. Ceci est totalement cohérent avec l'objectif de regrouper des individus similaires : s'ils devaient appartenir à une même classe alors leurs degrés respectifs devraient être identiques.

J'ai effectué un premier ensemble d'expériences afin de valider l'intérêt de l'approche. Elles concernent cinq jeux de données classiques disponibles en ligne. Une fois la partition obtenue je la compare avec la vérité terrain à l'aide du critère de Rand ajusté (*Adjusted Rand Index* - ARI). Les résultats de la Figure 4.3 indiquent que certaines valeurs de $t > 0$ permettent d'obtenir de meilleures performances que la *baseline* donnée par \mathbf{L}_{sym} qui est en fait le cas particulier \mathbf{L}_{sym}^t quand $t \rightarrow 0$. Ces résultats sont encourageants et mettent en perspective le problème de l'estimation de la valeur de $t > 0$ qui reste ouvert.

4.3 Discussions et perspectives

Dans ce Chapitre, j'ai exposé plusieurs critères de partitionnement à partir de travaux antérieurs en AR. Dans le cas d'individus décrits par des variables quantitatives, j'ai mis en avant le principe d'écart à une tendance centrale. Ces mesures de centralité sont diverses et peuvent être tantôt globale, tantôt locale. J'ai également appliqué ces approches dans le cas d'individus dont les relations d'affinité sont données directement par un graphe afin de détecter des communautés.

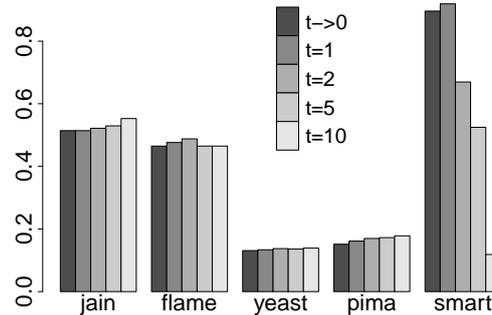


FIGURE 4.3 – Valeurs ARI sur 5 jeux de données classiques et pour 5 normalisation distinctes : $t \rightarrow 0$ (*baseline*), $t = 1, 2, 5, 10$.

Le principe d'écart à une tendance centrale formalisé dans (4.45) implique des “matrices de profits” de terme général $\tilde{s}_{ii'} - \mu(\tilde{\mathbf{S}}, i, i')$, qui peuvent prendre des valeurs positives ou négatives. En particulier, on voit que les paires de valeurs positives indiquent les individus qui ont le plus de vraisemblance d'être regroupés ensemble.

Il me semble alors pertinent d'**exploiter les critères B, E et J de façon collective au travers d'une approche ensembliste**. En effet, chaque méthode donne un point de vue différent sur la notion de *cluster* ou celle de communauté. Chacune d'elle peut alors cerner un aspect singulier et il serait ainsi intéressant de tenir compte de l'ensemble de ces points de vue lors du partitionnement. Dans cette perspective, on pourrait d'ailleurs ajouter la liste complémentaire des critères étudiés dans la thèse de Patricia Condé-Cespedes [Cespedes, 2013].

Afin de traiter cette tâche, l'**optimisation multiobjectif** pourrait être une piste intéressante. Toutefois, dans l'esprit de l'approche de l'AR, je propose de poser le problème en terme d'**agrégation de graphes d'affinités** avec la procédure suivante :

- Binarisation de chaque “matrice de profits” associé à chaque critère : les valeurs positives sont mises à 1 et les valeurs négatives à 0.
- Somme des matrices binarisées : on obtient une matrice d'adjacence collective pondérée qui indique pour chaque paire d'individus le nombre de critères qui considère qu'ils devraient être dans le même *cluster*.
- Partitionnement du graphe associé à l'aide d'un algorithme de *clustering* à base de graphe tel que le *spectral clustering* par exemple.

Ensuite, concernant plus spécifiquement la détection de communautés, **le critère LM de Light-Margolin est particulièrement performant** notamment en comparaison du critère de modularité Q. Contrairement aux critères B, E, J et Q, la mesure LM donnée par (4.49), utilise une transformation de la matrice d'adjacence \mathbf{A} qui est en lien direct avec la matrice Laplacienne normalisée d'ordre $t = 1$ définie par (4.75). Cette coïncidence invite à **étudier**

plus profondément, en théorie et en pratique, la famille de normalisation d'ordre $t > 0$ des matrices Laplaciennes pour le problème de partitionnement de graphes.

Sur l'aspect théorique, divers travaux montrent que dans le cadre de plusieurs modèles probabilistes, la matrice Laplacienne normalisée \mathbf{L}_{sym} a de meilleures **propriétés de convergence** que la matrice Laplacienne non normalisée \mathbf{L} [Von Luxburg et al., 2008],[Sarkar and Bickel, 2015]. Il me semble alors attrayant d'étudier le comportement asymptotique de \mathbf{L}_{sym}^t en comparaison de celui de \mathbf{L}_{sym} qui correspond au cas limite $t \rightarrow 0$. Dans cette perspective, la question de **l'estimation du paramètre t** serait centrale.

Les nombreux critères de partitionnement discutés ci-dessus et la famille de normalisation d'ordre t de matrices de noyaux et de matrices Laplaciennes conduisent à une multitude de modèles de *clustering*. Comment choisir l'approche la plus appropriée étant données les caractéristiques des données? Les expériences que j'ai menées jusqu'à présent, avec des données simulées ou avec des données réelles, restent trop peu limitées et ne permettent pas de répondre à la question. Ainsi, **des benchmarks plus conséquents accompagnés de développements conceptuels plus profonds** devraient être menés afin de mieux **cerner les propriétés des différentes méthodes**.

Enfin, je pense qu'il serait intéressant de développer une **approche générale des indices de similarité entre variables, qui serait fondée à la fois sur les matrices carrées de comparaison par paires, le concept d'indétermination et la normalisation d'ordre t** . Un premier pas a été fait en regroupant le coefficient de corrélation (des rangs) général Γ de Daniels et Kendall et les mesures d'association exprimées dans le codage relationnel comme cela est indiqué dans la Table 4.2. Le Γ de Daniels et Kendall englobe également *r* **le coefficient de corrélation de Bravais-Pearson** comme je l'ai soulevé en page 99. Ce dernier peut également se mettre sous la forme (4.65) comme je le montre ci-dessous.

Soient \mathbf{x}^k et \mathbf{x}^l deux vecteurs donnant pour deux variables quantitatives X^k et X^l les valeurs prises par n individus. Supposons, sans perte de généralité, qu'il n'y pas de doublons. Dans ce cas X^k et X^l engendrent deux relations d'ordre totale que l'on encode par les matrices relationnelles $\mathbf{X}^k = (x_{ii'}^k)$ et $\mathbf{X}^l = (x_{ii'}^l)$. Dans le cas de \mathbf{X}^k , on a pour tout $i, i' = 1, \dots, n$: $x_{ii'}^k = 1$ si $x_i^k < x_{i'}^k$, et $x_{ii'}^k = 0$ sinon. De plus \mathbf{X}^k vérifient⁵ les contraintes suivantes : $\forall i, i' = 1, \dots, n, x_{ii'}^k + x_{i'i}^k = 1$. \mathbf{X}^l est définie de la même façon et satisfait aux mêmes équations linéaires.

5. Voir le Problème 3.1 et la combinaison des contraintes d'asymétrie et de totalité.

4.3. DISCUSSIONS ET PERSPECTIVES

Le numérateur de $\Gamma(\mathbf{x}^k, \mathbf{x}^l)$ donné par (4.63) peut alors être reformulé comme suit :

$$\begin{aligned} \sum_{i,i'} x_{ii'}^k x_{ii'}^l &= \sum_{i,i'} (x_{i'}^k - x_i^k)(x_{i'}^l - x_i^l) \\ &= \sum_{i,i'} (x_{i'}^k - x_i^k)(x_{ii'}^k + x_{i'i}^k)(x_{i'}^l - x_i^l)(x_{ii'}^l + x_{i'i}^l) \\ &= \sum_{i,i'} \underbrace{(x_{i'}^k - x_i^k)}_{m_{ii'}^k} x_{ii'}^k - \underbrace{(x_i^k - x_{i'}^k)}_{\check{m}_{ii'}^k} \underbrace{x_{i'i}^k}_{\check{x}_{ii'}^k} \underbrace{(x_{i'}^l - x_i^l)}_{m_{ii'}^l} x_{ii'}^l - \underbrace{(x_i^l - x_{i'}^l)}_{\check{m}_{ii'}^l} \underbrace{x_{i'i}^l}_{\check{x}_{ii'}^l}. \end{aligned}$$

En posant $\check{\mathbf{X}}^k = (\check{x}_{ii'}^k)$ avec $\check{x}_{ii'}^k = x_{i'i}^k, \forall i, i' = 1, \dots, n$ (la transposée de \mathbf{X}^k), on voit que l'on peut interpréter le coefficient de corrélation linéaire de Bravais-Pearson en terme d'indétermination pondérée et amendé la Table 4.2 par les éléments de la Table 4.3.

	$m_{ii'}^k$	$\check{m}_{ii'}^k$
r de Bravais-Pearson	$x_{i'}^k - x_i^k$	$x_i^k - x_{i'}^k$

TABLE 4.3 – Matrices de poids \mathbf{M}^k (même formules pour \mathbf{M}^l) pour r le coefficient de corrélation linéaire de Bravais-Pearson.

Ce coefficient de corrélation peut aussi être étendu dans le cadre du coefficient de similarité général Λ^t défini par (4.69).

Les cas précédents concernent la similarité entre des couples de vecteurs $(\mathbf{x}^k, \mathbf{x}^l)$. Un autre concept, qui me paraît pertinent dans le cadre d'une théorie générale des coefficients de similarité, est celui du **coefficient RV d'Yves Escoufier** [Escoufier, 1970, Escoufier, 1973]. Il s'agit d'un coefficient de corrélation mesurant la similarité entre deux descriptions vectorielles d'un même ensemble d'individus. Pour cela, il mesure la similarité cosinus entre les matrices de Gram engendrées par les deux matrices de données à l'étude. Le coefficient RV est central en analyse de données multivariées [Robert and Escoufier, 1976]. Il me semble alors pertinent de proposer une **extension du coefficient RV au travers d'une normalisation d'ordre t** qui conduirait à une mesure plus fine de la similarité entre les deux descriptions des données.

Par ailleurs, le coefficient Λ^t défini par 4.69, est fondé sur la notion d'indétermination. Il s'agit d'un concept simple mais intrigant. Plusieurs articles en AR ont déjà montré les liens entre l'indétermination et l'indépendance statistique. Des travaux plus récents établissent, en outre, des **relations entre l'indétermination et différentes variantes du problème du transport optimal**. C'est notamment le cas du travail de thèse de Pierre Bertrand [Bertrand, 2021] dont j'ai été examinateur de soutenance. Dans [Bertrand et al., 2022], Pierre

4.3. DISCUSSIONS ET PERSPECTIVES

Bertrand ainsi que ses coauteurs, Michel Broniatowski et Jean-François Marcotorchino, abordent ces questions dans le contexte du *clustering*. Ces travaux encouragent une poursuite de l'étude des liens entre indépendance et indétermination dans un contexte unifié et le coefficient de similarité général Λ^t pourrait être un cadre formel approprié à cet égard.

Je termine en abordant un point sous-jacent aux méthodes de *clustering* étudiées dans ce Chapitre. Les méthodes à base de graphe comme le *spectral clustering*, prennent en entrée une matrice d'affinités qui représente un graphe non orienté pondéré entre individus. Aucune propriété spécifique n'est requise. Notamment, il n'est pas nécessaire que la matrice soit semi-définie positive ce qui impliquerait une structure d'espace métrique. Au contraire, lorsque les individus appartiennent à des sous-espaces non linéaires, les méthodes classiques telles que les *k-means* ne sont pas adaptées. En restreignant le graphe aux relations entre plus proches voisins uniquement, le *spectral embedding* effectué par le *spectral clustering*, via la décomposition spectrale de la matrice Laplacienne (normalisée ou pas), permet de projeter les individus dans un sous-espace linéaire dont les distances Euclidiennes tendent à respecter les géodésiques dans les sous-espaces non linéaires de départ. Dans le Chapitre qui suit, j'étudie plus explicitement le problème de variétés non linéaires en apprentissage non supervisé.

Variétés non linéaires en apprentissage non supervisé

Sommaire du chapitre

5.1	Introduction	109
5.1.1	Contexte	109
5.1.2	Travaux antérieurs	112
	Problème NP-dur de minimisation du SSE, modèles de relaxation et théorème de Sinkhorn	112
	Classification ascendante hiérarchique et formule de Lance et Williams	115
5.2	Contributions	119
5.2.1	Plongement et partitionnement spectral de graphes	119
5.2.2	Un nouveau modèle générique de classification ascendante hiérarchique	125
5.3	Discussions et perspectives	133

5.1 Introduction

5.1.1 Contexte

Le Chapitre précédent concernait des contributions en analyse de données inspirées par l'analyse relationnelle et présentait des extensions de mes travaux de thèse de doctorat. Dans ce nouveau Chapitre, j'expose également des contributions en analyse de données mais avec des cadres conceptuels différents. De plus, je traite plus explicitement le **challenge des variétés non linéaires** qui a pris de l'ampleur avec le développement des méthodes de *pattern recognition* en intelligence artificielle et qui est sous-jacent à l'**analyse de données complexes**.

Ceci est notamment stipulé au sein du concept de *manifold hypothesis* : les données complexes sont décrites dans des espaces linéaires de très grande dimension mais appartiennent en fait à des sous-espaces non linéaires de dimension plus petite. Cette hypothèse a souvent été confirmée en pratique dans le cas de données images, textes, audio, graphes, ... Ceci explique le succès de nombreuses méthodes non linéaires en apprentissage non supervisé,

comme *Isometry feature mapping* (ISOMAP) [Tenenbaum et al., 2000], *Locally Linear Embedding* (LLE) [Roweis and Saul, 2000], *Diffusion Maps* [Coifman and Lafon, 2006], *spectral clustering* [Shi and Malik, 2000, Ng et al., 2001, Von Luxburg, 2007], ... en comparaison des méthodes classiques tels que l'analyse en composantes principales et le *k-means* qui sont très limitées dans le cas de la détection de *clusters* de formes non ellipsoïdales.

Dans ce contexte, les méthodes à base de graphe sont particulièrement efficaces. En effet, le **graphe des plus proches voisins** peut être vu comme un *proxy* de la topologie des variétés non linéaires sous-jacentes aux *clusters*. Comme j'ai pu le mentionner dans le Chapitre précédent, la **décomposition spectrale de la matrice Laplacienne du graphe** permet alors un plongement des individus dans un espace linéaire au sein duquel les distances Euclidiennes approximent les géodésiques initiales entre observations.

L'analyse théorique du *spectral clustering* montre qu'il existe des liens étroits avec le problème de *min cut* (et ses variantes) dans des graphes (voir [Von Luxburg, 2007] et les références qui y sont mentionnées). Ce **problème est NP-dur**. De nombreuses **relaxations** ont été proposées dans la littérature afin de le solutionner de façon approchée. Je présente dans la sous-section 5.2.1 ma contribution dans ce contexte [Ah-Pine, 2022]. Celle-ci promeut le rôle central des **matrices bistochastiques et idempotentes en clustering** et met en avant un résultat de Richard Sinkhorn de 1968 [Sinkhorn, 1968] qui est passé inaperçu au sein de la communauté. Je discuterai en premier lieu de l'article [Ah-Pine, 2022] dans la mesure où il est en continuité avec les travaux que j'ai présentés dans le Chapitre précédent.

J'exposerai, ensuite, une autre approche basée sur les graphes qui est la **classification ascendante hiérarchique (CAH)**. J'introduis dans la sous-section 5.2.2 un nouveau cadre général pour ce type de méthodes. Contrairement à ce qui est pratiqué communément, la méthode que j'expose emploie une matrices de noyaux à la place d'une matrice de dissimilarités. Je montre en quoi cette approche permet non seulement de traiter les données appartenant à des **variétés non linéaires**, mais aussi d'**améliorer la scalabilité de la CAH**.

Le passage à l'échelle est dans ce contexte une motivation importante. En effet, malgré les avantages avérés de la CAH, une contrainte forte de cette technique est sa complexité, trop grande pour son application à des données massives. Je précise le contexte dans lequel mes travaux sur ces sujets se sont déroulés. J'ai examiné cette question de scalabilité dans le cadre du projet de recherche collaboratif PIA¹/FSN² intitulé Request (*REcursive QUERy and Scalable Technologies*). Ce projet a rassemblé 13 partenaires et a porté sur la recherche et le développement de technologies *Big Data* et leurs applications dans les domaines suivants : *cloud computing*, *big analytics*, *visual analytics*, *smart transport* et *cybersecurity*. J'ai participé à la rédaction de la proposition Request et, par la suite, j'ai été responsable de la tâche "requêtage intelligent".

Le projet Request m'a permis de financer le recrutement d'une doctorante, Xinyu Wang,

1. Projet d'Investissement d'Avenir.
2. Fonds national pour la Société Numérique.

que j'ai co-encadrée de 2013 à 2017, avec Jérôme Darmont Professeur et Directeur du laboratoire ERIC à ce moment. Xinyu était étudiante au sein du Master Erasmus Mundus DMKM (*Data Mining and Knowledge Management*) qui était administré principalement par l'Université Lyon 2 (UL2) et les membres du laboratoire ERIC. L'objectif de son travail de thèse était de poursuivre le développement de l'approche que j'avais initialement proposée pour traiter la CAH à partir de matrices de similarités, et aussi d'appliquer cette approche en fouille de textes et recherche d'information (RI). L'implémentation devait se faire en Spark qui était une technologie de calcul distribué émergente en 2013 et qui présentait de forts avantages pour le *machine learning* en comparaison du paradigme classique de *Map-Reduce*. Le travail de thèse de Xinyu nous a permis plusieurs contributions : un premier modèle opérationnel de CAH utilisant les similarités, une extension de cette approche au problème de *co-clustering* en *text-mining* et une étude du *cluster hypothesis* en RI. Ce dernier concept stipule que des documents similaires et appartenant à un même *cluster*, tendent à être pertinentes pour les mêmes requêtes. Par conséquent, sous cette hypothèse, il est judicieux de *clusteriser* la collection de documents en amont de la RI, afin d'améliorer les performances tant du point de vue du temps de traitement que de celui de la qualité. Malgré l'intérêt de ces différents travaux, je ne les présenterai pas dans ce qui suit. Je me focaliserai sur un deuxième modèle opérationnel de CAH utilisant les similarités. Il s'agit d'une approche que j'ai établie en parallèle du travail de thèse de Xinyu lorsque cette dernière était en fin de doctorat. J'ai alors défini un nouveau cadre formel plus riche que le premier et qui m'a permis d'introduire de nouveaux concepts et de nouvelles propriétés. Ces résultats ont été publiés dans [Ah-Pine, 2018].

Les publications dans des journaux ou conférences avec comités de lecture en lien avec les thématiques de ce Chapitre sont les suivantes :

- **J. Ah-Pine.** 2022. Learning Doubly Stochastic and Nearly Idempotent Affinity Matrix for Graph-Based Clustering. *European Journal of Operational Research*, 299(3), 1069-1078. [Lien vers le journal, <https://www.sciencedirect.com/science/article/abs/pii/S0377221721010900>].
- **J. Ah-Pine.** Sur l'apprentissage d'une matrice d'affinité bistochastique en clustering. In *52ème Journées de Statistiques, (JDS 2021)*. [Lien vers les proceedings, https://jds2021.sciencesconf.org/data/pages/book_jds2021_fr_compressed.pdf].
- **J. Ah-Pine.** 2018. An Efficient and Effective Generic Agglomerative Hierarchical Clustering Approach. *Journal of Machine Learning Research*. 19. [Lien vers le journal, <http://www.jmlr.org/papers/v19/18-117.html>].
- X. Wang, **J. Ah-Pine**, Jérôme Darmont. 2017. SHCoClust, a scalable similarity-based hierarchical co-clustering method and its application to textual collections. *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2017)*. [Lien vers la conférence, <http://www.fuzzieee2017.org/>].
- X. Wang, **J. Ah-Pine**, Jérôme Darmont. 2017. A New Test of Cluster Hypothesis Using a Scalable Similarity-Based Agglomerative Hierarchical Clustering Framework.

Proceedings of the 14th “Conférence en Recherche d’Informations et Applications - Rencontres des Jeunes Chercheurs” (CORIA-RJCRI 2017). [Lien vers la conférence, <http://www.lsis.org/coria2017/>].

- **J. Ah-Pine**, X. Wang. 2016. Similarity Based Hierarchical Clustering with an Application to Text Collections. *Proceedings of the 15th International Symposium on Intelligent Data Analysis (IDA 2016)*. [Lien vers la conférence, <https://ida2016.blogs.dsv.su.se/>].
- **J. Ah-Pine**, X. Wang. 2015. Classification ascendante hiérarchique à noyaux et pistes pour un meilleur passage à l’échelle. *Communication aux Journées de Statistiques (JDS 2015)*. [Lien vers la conférence, http://papersjds15.sfds.asso.fr/submission_116.pdf].

5.1.2 Travaux antérieurs

Les recherches que je présente dans ce Chapitre en apprentissage non supervisé, reposent sur deux paradigmes différents de la classification automatique : les méthodes de partitionnement et les méthodes hiérarchiques. Ainsi, la suite du texte est organisée en deux blocs qui font écho à ces deux cadres distincts.

Problème NP-dur de minimisation du SSE, modèles de relaxation et théorème de Sinkhorn

Je m’intéresse à nouveau à la classification automatique mais plus spécifiquement à des **méthodes à base de graphe**, c’est à dire lorsque l’on a comme information en entrée un graphe binaire ou pondéré indiquant la similarité pour toute paire d’un ensemble de n individus. Dans ce contexte, la complexité du problème dépend du critère de partitionnement. Contrairement aux fonctions objectifs étudiées dans le Chapitre précédent, je vais me focaliser sur le **critère classique de *Sum of Squared Errors* dénoté SSE** employé par la technique classique du *k-means*. Dans ce cas, **le problème de partitionnement de graphe (ou clique partitioning) est NP-dur** (voir [Aloise et al., 2009] et les références qui y sont mentionnées). Je m’intéresse spécifiquement ici aux **relaxations possibles de ce problème combinatoire**.

En premier lieu, je rappelle le critère SSE lorsque les individus $\{X_i\}_i$ sont décrits par des vecteurs $\{\mathbf{x}_i\}_i$ de \mathbb{R}^P . Afin d’intégrer plus généralement le problème de la détection de variétés non linéaires, je suppose implicitement une méthode à noyaux et je considère, comme dans la sous-section 4.2.2, qu’il existe une application $\phi : \mathbb{R}^P \rightarrow \mathbb{F}$ qui projette les vecteurs de l’*input space* \mathbb{R}^P dans un *feature space* \mathbb{F} qui est de grande dimension et au sein duquel les *clusters* sont plus facilement séparables par des méthodes linéaires. Une partition des n éléments en k *clusters* sera dénotée ici $\mathbb{C} = \{\mathbb{C}_1, \dots, \mathbb{C}_k\}$. Compte tenue des distances dans \mathbb{F} , la SSE de \mathbb{C}

est définie par :

$$\text{SSE}(\mathbb{C}) = \sum_{j=1}^k \sum_{\mathbf{x}_i \in \mathbb{C}_j} \|\phi(\mathbf{x}_i) - \mathbf{c}_j\|_{\mathbb{F}}^2, \quad (5.1)$$

où $\|\cdot\|_{\mathbb{F}}$ est la distance dans \mathbb{F} , \mathbf{c}_j est le barycentre dans \mathbb{F} des individus de \mathbb{C}_j de cardinal n_j : $\mathbf{c}_j = \frac{1}{n_j} \sum_{\mathbf{x}_i \in \mathbb{C}_j} \phi(\mathbf{x}_i)$.

Afin de représenter numériquement la partition, on introduit la **matrice d'affectation** $\mathbf{Y} = (y_{ij})$ de taille $n \times k$ et de terme général défini par, $\forall i = 1, \dots, n, \forall j = 1, \dots, k$:

$$y_{ij} = \begin{cases} 1 & \text{si } X_i \text{ est dans } \mathbb{C}_j, \\ 0 & \text{sinon.} \end{cases} \quad (5.2)$$

Notons également par \mathbf{e}_n le vecteur de taille $n \times 1$ rempli de 1. Avec ces notations, le problème de *clustering* qui minimise la SSE peut être formulé comme suit :

$$\begin{aligned} \min_{\mathbf{Y} \in \{0,1\}^{n \times k}} \text{SSE}(\mathbf{Y}) &= \sum_{j=1}^k \sum_{i=1}^n \left\| \phi(\mathbf{x}_i) - \frac{1}{\sum_{i'=1}^n y_{i'j}} \sum_{i'=1}^n \phi(\mathbf{x}_{i'}) y_{i'j} \right\|^2 \\ \text{s.l.c. } \mathbf{Y}\mathbf{e}_k &= \mathbf{e}_n, \mathbf{Y}^\top \mathbf{e}_n \geq \mathbf{e}_k. \end{aligned} \quad (5.3)$$

La contrainte $\mathbf{Y}\mathbf{e}_k = \mathbf{e}_n$ force chaque vecteur à être affecté à un et un seul *cluster* tandis que la contrainte $\mathbf{Y}^\top \mathbf{e}_n \geq \mathbf{e}_k$ indique que chaque *cluster* doit contenir au moins un individu.

Il est alors possible de reformuler de façon équivalente le problème précédent en faisant apparaître un graphe dont la matrice d'adjacence pondérée par des réels est donnée par la matrice de noyaux $\mathbf{K} = (k_{ii'})$ avec $k_{ii'} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'}) \rangle_{\mathbb{F}}$ pour tout couple d'individus $(X_i, X_{i'})$. On a dans ce cas [Dhillon et al., 2004, Zha et al., 2002] :

$$\text{SSE}(\mathbf{Y}) = \sum_{i=1}^n k_{ii} - \sum_{i=1}^n \sum_{i'=1}^n k_{ii'} \sum_{j=1}^k \frac{1}{n_j} y_{ij} y_{i'j}. \quad (5.4)$$

Ensuite, nous introduisons la matrice carrée d'ordre n , $\mathbf{X} = (x_{ii'})$, dont le terme général est donné par, $\forall i, i' = 1, \dots, n$:

$$x_{ii'} = \sum_{j=1}^k \frac{1}{n_j} y_{ij} y_{i'j}. \quad (5.5)$$

Les matrices \mathbf{X} et \mathbf{Y} sont liées par la relation fondamentale suivante :

$$\mathbf{X} = \mathbf{Y}(\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top. \quad (5.6)$$

\mathbf{X} est ainsi la **matrice de projection orthogonale** du sous-espace engendré par les

vecteurs caractéristiques de chaque *cluster*. Nous avons par ailleurs la propriété suivante :

$$\text{Tr}(\mathbf{X}) = \text{Rk}(\mathbf{X}) = k, \quad (5.7)$$

où Tr et Rk sont les applications trace et rang, respectivement.

L'ensemble de ces résultats conduisent à la formulation équivalente du problème (5.3) en fonction de \mathbf{K} et \mathbf{X} proposée par Jiming Peng et Yu Xia dans [Peng and Xia, 2005] :

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \text{SSE}(\mathbf{X}) &= \text{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \\ \text{s.l.c. } \mathbf{X} &\geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X} = \mathbf{X}^2, \text{Tr}(\mathbf{X}) = k. \end{aligned} \quad (5.8)$$

La matrice \mathbf{X} doit être remplie de valeurs non négatives ($\mathbf{X} \geq \mathbf{0}_n \Leftrightarrow \forall i, i' : x_{ii'} \geq 0$), elle doit être symétrique et la somme de chaque ligne doit être égale à 1. Ceci implique que \mathbf{X} est **bistochastique**. De plus, \mathbf{X} doit être **idempotente** c'est à dire, $\mathbf{X} = \mathbf{X}^2$. Dans ce cadre, la contrainte $\text{Tr}(\mathbf{X}) = k$ permet de fixer le nombre de *clusters* à k . La fonction objectif et les contraintes du problème 5.8 sont linéaires en \mathbf{X} à l'exception de l'idempotence. C'est en effet la condition $\mathbf{X} = \mathbf{X}^2$ qui rend particulièrement difficile le problème.

Afin de relaxer ce dernier, il semble ainsi propice d'abandonner la contrainte d'idempotence. Revenons alors sur la relation (5.6) qui indique que \mathbf{X} est une matrice de projection orthogonale. Supposons, de façon générale, que l'image de \mathbf{X} soit \mathbb{E} de dimension k . Alors \mathbf{X} possède les propriétés suivantes :

$$\mathbf{X} \text{ est symétrique : } \mathbf{X}^\top = \mathbf{X}. \quad (5.9)$$

$$\mathbf{X} \text{ est idempotente : } \mathbf{X}^2 = \mathbf{X}. \quad (5.10)$$

$$\text{Rk}(\mathbf{X}) = \text{Tr}(\mathbf{X}) = k. \quad (5.11)$$

$$\text{Il existe } \mathbf{G} \in \mathbb{R}^{n \times k} \text{ et } k = \text{Rk}(\mathbf{X}), \text{ tels que : } \mathbf{X} = \mathbf{G}\mathbf{G}^\top, \mathbf{G}^\top\mathbf{G} = \mathbf{I}_k. \quad (5.12)$$

$$\mathbf{X} \text{ a } \text{Rk}(\mathbf{X}) \text{ valeurs propres égales à } 1 \text{ et } n - \text{Rk}(\mathbf{X}) \text{ valeurs propres égales à } 0. \quad (5.13)$$

$$\mathbf{X} \text{ est semi-définie positive : } \mathbf{X} \succeq \mathbf{0}_n. \quad (5.14)$$

$$\mathbf{I}_n - \mathbf{X} \text{ est la matrice de projection orthogonale sur } \mathbb{E}^\perp. \quad (5.15)$$

$$\mathbf{X}(\mathbf{I}_n - \mathbf{X}) = (\mathbf{I}_n - \mathbf{X})\mathbf{X} = \mathbf{0}_n. \quad (5.16)$$

où \mathbf{I}_n est la matrice identité d'ordre n et $\mathbf{0}_n$ est la matrice nulle d'ordre n .

De nombreuses heuristiques en *clustering* reviennent en fait à remplacer les contraintes du problème 5.8 par un ensemble de conditions nécessaires comprenant certaines citées ci-dessus. Par exemple la relation 5.12 engendre la méthode par **factorisation de matrices de faibles rangs avec ou sans contrainte d'orthogonalité** [Zha et al., 2002, Kuang et al., 2012]. Un autre exemple concerne la **relaxation SDP** qui s'appuie donc sur la condition nécessaire 5.14 [Peng and Wei, 2007, Kulis et al., 2007, Peng and Wei, 2007]. Plus proches du travail que je présenterai par la suite, les articles [Zass and Shashua, 2005, Zass and Shashua, 2007,

Nie et al., 2016, Wang et al., 2016, Park and Kim, 2017] se concentrent sur la propriété de bistochasticité. Dans ces travaux, la démarche consiste à approximer une matrice d'affinités par une **matrice bistochastique de faible rang** et d'appliquer ensuite, le *spectral clustering* à cette dernière. A l'inverse de ces méthodes, je décris dans la sous-section 5.2.1 une relaxation qui omet la contrainte sur le rang mais qui intègre de façon approximative l'idempotence.

Il est important de noter que même si Jiming Peng et Yu Xia démontrent en 2005 dans [Peng and Xia, 2005], l'équivalence entre les problèmes 5.3 et 5.8, la correspondance entre l'ensemble des matrices d'affectation \mathbf{Y} et l'ensemble des matrices \mathbf{X} bistochastiques et idempotentes avaient été déjà établie par Richard Sinkhorn en 1968 dans [Sinkhorn, 1968]. Soit \mathbf{J}_n la matrice carrée d'ordre n remplie de la valeur $1/n$. Le théorème de Richard Sinkhorn est le suivant.

Théorème. 7 ([Sinkhorn, 1968]). $\mathbf{X} \in \mathbb{R}^{n \times n}$ est bistochastique et idempotente si et seulement si il existe des entiers positifs n_1, n_2, \dots, n_k qui somment à n et une matrice de permutation \mathbf{P} tels que :

$$\mathbf{X} = \mathbf{P} \begin{pmatrix} \mathbf{J}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{J}_{n_k} \end{pmatrix} \mathbf{P}^\top. \quad (5.17)$$

Dans le théorème précédent, la matrice de permutation \mathbf{P} permet de représenter \mathbf{X} sous forme de **matrice "bloc diagonale"**. Pour tout $j = 1, \dots, k$, chaque bloc \mathbf{J}_{n_j} est associé à un sous-ensemble d'individus. Les blocs étant sans recouvrement les uns vis-à-vis des autres, ils définissent collectivement une partition de l'ensemble des individus. Du point de vue graphe, il s'agit bien d'un **partitionnement des sommets en k cliques mutuellement indépendantes** ou encore une **décomposition du graphe en k composantes connexes**.

Classification ascendante hiérarchique et formule de Lance et Williams

Supposons que nous disposions d'une matrice carrée d'ordre n , $\mathbf{D} = (d_{ii'})$, qui indique pour chaque paire d'individus $(X_i, X_{i'})$, une **mesure de dissimilarité** avec les propriétés suivantes : $\mathbf{D} = \mathbf{D}^\top$, $\mathbf{D} \geq 0$ et $\text{Tr}(\mathbf{D}) = 0$ (diagonale nulle).

La **classification ascendante hiérarchique (CAH)** vise à extraire de \mathbf{D} une structure arborescente appelée **dendrogramme** qui est précisément un **arbre binaire** au sens de la théorie des graphes. Il est composé de $2n - 1$ sommets au total dont n sans successeur (les feuilles de l'arbre) et un sans prédécesseur (la racine de l'arbre). Chaque sommet représente un sous-ensemble d'individus, et par abus de langage on parlera par la suite de sommet ou *cluster* de façon interchangeable. Dans un dendrogramme, deux sommets distincts k et l sont tels que $k \subset l$ ou $l \subset k$ ou $k \cap l = \emptyset$. Par ailleurs, à chaque sommet on lui attribue une valeur non négative appelée la **hauteur (height)**. Cette caractéristique permet l'étude de propriétés de méthodes de CAH et je reviendrai sur celle-ci dans la sous-section 5.2.2.

A partir de la partition triviale constituée de n singletons (les feuilles de l'arbre), la procédure classique de la CAH est la suivante :

- Recherche de la paire de *clusters* ayant la plus petite dissimilarité.
- Union de la paire de *clusters* sélectionnés et calcul du profil de dissimilarités entre le nouveau *cluster* et l'ensemble des *clusters* existants.
- Répétition des étapes précédentes jusqu'à obtenir la partition avec un seul *cluster* (la racine de l'arbre). L'arbre étant binaire, il y a nécessairement $n - 1$ itérations.

La CAH permet de ne pas fixer *a priori* le nombre de *clusters*. De plus, le dendrogramme permet d'appréhender les relations entre *clusters* et une certaine "dynamique" dans la constitution de ces derniers. L'approche est également déterministe et ne souffre pas de la dépendance vis-à-vis de l'ordre de traitement des individus ou du choix aléatoire d'une partition initiale. En outre, il est possible d'utiliser plusieurs mesures de dissimilarités avec des propriétés diverses. Les méthodes de CAH les plus courantes sont des cas particuliers de la formule paramétrique de Godfrey Lance et Williams Williams [Lance and Williams, 1967].

Explicitons plus formellement la procédure de la CAH. L'algorithme procède en $n - 1$ itérations. Soit $\mathbb{T} = \{1, \dots, n - 1\}$ et dénotons par $t \in \mathbb{T}$ l'itération courante. Dénotons également par $\mathbf{D}^t = (d_{ij}^t)$ et \mathbb{C}^t , la matrice carrée d'ordre $n - t$ et la partition de $n - t$ *clusters*, que l'on observe à l'itération t . Supposons qu'en t , ce sont les sommets k et l qui sont réunis au sein d'un nouveau *cluster* que l'on dénote par (kl) . La paire (k, l) vérifie donc :

$$(k, l) = \underset{(i,j) \in \mathbb{C}^t \times \mathbb{C}^t, i \neq j}{\operatorname{arg\,min}} d_{ij}^t. \quad (5.18)$$

Au sein du dendrogramme, un nouveau noeud composé des individus de l'union $k \cup l$ est alors ajouté et la hauteur de ce dernier est fixée à d_{kl}^t .

La **formule de Lance-Williams (LW)** permettant de calculer les nouvelles dissimilarités de (kl) avec les *clusters* $m \in \mathbb{C}^t \setminus \{k, l\}$ est alors donnée par :

$$d_{(kl)m}^{t+1} = \alpha'(k, l, m)d_{km}^t + \alpha'(l, k, m)d_{lm}^t + \beta'(k, l, m)d_{kl}^t + \gamma'|d_{km}^t - d_{lm}^t|, \quad (5.19)$$

où γ' est un réel et α', β' sont des fonctions d'ensemble définies sur l'ensemble des triplets d'éléments de \mathbb{C}^t et qui sont à valeurs réelles positives.

Les différentes dissimilarités classiques que la formule de LW permet de prendre en compte sont décrites dans la Table 5.1.

Supposons le **cas particulier où $\mathbf{D} = (d_{ii'})$ est une matrice de distances Euclidiennes au carré**, c'est à dire telle que $d_{ii'} = \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2$ pour tout couple d'individus $(X_i, X_{i'})$ représenté par le couple de vecteurs $(\mathbf{x}_i, \mathbf{x}_{i'})$ dans $\mathbb{R}^p \times \mathbb{R}^p$. Dans ce cas, les méthodes *Median*, *Centroid* et *Ward* ont des interprétations géométriques où les dissimilarités entre *clusters* s'expriment en fonction de distances Euclidiennes au carré entre des vecteurs "moyens" qui représentent ces derniers. Dans le cas de *Median*, le vecteur "moyen" est le vecteur central (*midpoint*) tandis que pour *Centroid* et *Ward*, il s'agit du barycentre. Ce cadre géométrique

Method	$\alpha'(k, l, m)$	$\beta'(k, l, m)$	γ'
Single link.	1/2	0	-1/2
Complete link.	1/2	0	1/2
Group aver.	$\frac{ k }{ k + l }$	0	0
Mcquitty	1/2	0	0
Centroid	$\frac{ k }{ k + l }$	$-\frac{ k l }{(k + l)^2}$	0
Median	1/2	-1/4	0
Ward	$\frac{ k + m }{ k + l + m }$	$-\frac{ m }{ k + l + m }$	0

TABLE 5.1 – Cas particuliers de la formule initiale de LW (5.19).

est important pour l'approche que j'introduirai en sous-section 5.1.2.

Les autres méthodes sont dites de type graphe. Remarquons en particulier que les approches *Single* et *Complete linkage* sont les seules pour lesquelles le paramètre γ' de la formule initiale de LW est non nul. Ces deux cas sont singuliers et peuvent en fait être traités de façon efficace par des algorithmes de graphes classiques utilisés pour la détermination d'arbres recouvrant [Gower and Ross, 1969, Defays, 1977]. J'exclus alors ces deux techniques de mon étude et considère par la suite que $\gamma' = 0$.

Il existe alors une deuxième façon d'exprimer la procédure CAH. C'est notamment sur cette dernière que je m'appuierai par la suite. Elle met en jeu un critère légèrement modifié pour la sélection des *clusters* à fusionner et également une formule différente pour la mise à jour des dissimilarités. J'appellerai cette équation, **formule partielle de LW** dans la suite. En premier lieu, la détermination du couple de *clusters* à regrouper est fondée sur une dissimilarité qui est désormais pondérée par une nouvelle fonction d'ensemble p définie sur l'ensemble des couples de *clusters* distincts :

$$(k, l) = \arg \min_{(i, j) \in \mathbb{C}^t \times \mathbb{C}^t, i \neq j} p(i, j) d_{ij}^t. \quad (5.20)$$

Ensuite, la formule partielle de LW implique des fonctions d'ensemble α et β qui diffèrent des fonctions α' et β' précédentes. Elles sont plus simples et ne dépendent que de deux arguments au lieu de trois. La formule partielle de LW s'exprime comme suit, $\forall m \in \mathbb{C}^t \setminus \{k, l\}$:

$$d_{(kl)m}^{t+1} = \alpha(k, l) d_{km}^t + \alpha(l, k) d_{lm}^t + \beta(k, l) d_{kl}^t. \quad (5.21)$$

Pour être plus précis, je donne dans la Table 5.2, les différentes valeurs des fonctions d'ensemble p , α et β définissant les techniques classiques de la CAH dans le cadre formel reposant sur les équations (5.20) et (5.21). Ces définitions peuvent être déduites de l'approche dite *stored data* présentée dans [Murtagh and Contreras, 2012a]. Par la suite, je dénoterai cette approche de base par **D-AHC** (*Dissimilarity matrix based Agglomerative Hierarchical Clustering*).

La formulation D-AHC permet de mieux comprendre les relations entre la méthode *Cen-*

Method	$\alpha(k, l)$	$\beta(k, l)$	$p(i, j)$
Group aver.	$\frac{ k }{ k + l }$	0	1
Mcquitty	1/2	0	1
Centroid	$\frac{ k }{ k + l }$	$-\frac{ k l }{(k + l)^2}$	1
Median	1/2	-1/4	1
Ward	$\frac{ k }{ k + l }$	$-\frac{ k l }{(k + l)^2}$	$\frac{ i j }{ i + j }$
W-Median	1/2	-1/4	$\frac{ i j }{ i + j }$

TABLE 5.2 – Cas particuliers de la formule partielle de LW (5.21) (D-AHC).

troid et la méthode de *Ward*. Les distances sous-jacentes à ces deux approches sont les mêmes. C'est la pondération appliquée lors de la recherche du minimum qui change. L'approche de *Ward* peut être vue comme une version pondérée de l'approche *Centroid*. Dans ce même esprit, j'ai proposé une **version pondérée de la méthode *Median*** que j'ai dénotée par ***W-Median*** dans la Table 5.2.

Afin de comparer les différentes dissimilarités permises par la formule de LW, plusieurs notions ont été mises en place. Dans ce contexte, le concept de **monotonie d'un dendrogramme** est particulièrement intéressant et permet de cerner des comportements inadéquats de certaines techniques. Un dendrogramme est non monotone lorsque la hauteur d'un sommet formé à une itération t est plus petite que celle d'un sommet constitué à une itération $s < t$. En pratique, les dendrogrammes non monotones sont à éviter car ils sont difficiles à interpréter. On montre que les méthodes *Centroid* et *Median* peuvent produire des dendrogrammes non monotones [Milligan, 1979].

Malgré le champ riche développé autour de la CAH, l'approche D-AHC possède plusieurs limites. La procédure de construction du dendrogramme est une stratégie gloutonne. Par conséquent, du point de vue optimisation, elle est sous-optimale. Par ailleurs, **D-AHC a une complexité mémoire en $O(n^2)$ et une complexité temps de traitement en $O(n^3)$** . Dans ce dernier cas, le goulot d'étranglement est la recherche, à chaque itération, d'un minimum dans un ensemble dont la taille est au pire des cas $\frac{n(n-1)}{2}$. La complexité désavantageuse de l'approche classique de la CAH limite sévèrement son utilisation dans un contexte *big data*. Plusieurs approches ont été proposées afin d'**améliorer la scalabilité**. Certains travaux utilisent des structures de données avancées telles que les *priority queues* afin d'accéder de façon efficace aux plus proches voisins [Day and Edelsbrunner, 1984, Müllner et al., 2013]. D'autres approches exploitent plus spécifiquement la propriété de *reducibility* de certaines méthodes et donnent lieu à une procédure *ad-hoc* basée sur les *nearest neighbor chains* [Bruynooghe, 1978, Murtagh, 1984, Müllner et al., 2013]. Certains auteurs imposent également des relations de contiguïté entre objets qu'il est nécessaire de satisfaire pour que deux *clusters* puissent être regroupés [Lebart, 1978, Grimm, 1987, Randriamihamison et al., 2021].

L'approche nouvelle de la CAH que j'ai définie dans [Ah-Pine, 2018] et que je résume en sous-section 5.2.2, poursuit une piste qui est substantiellement différente de ces travaux.

5.2 Contributions

5.2.1 Plongement et partitionnement spectral de graphes

L'approche que je développe dans [Ah-Pine, 2022] s'apparente à une méthode d'**apprentissage non supervisé d'une matrice d'affinités bistochastique et quasi-idempotente**. Deux points distinguent ma démarche vis-à-vis des travaux antérieurs sur ce sujet. D'une part, je vais tenir compte de la contrainte d'idempotence de façon approximative et d'autre part, je ne fais pas d'hypothèse sur le nombre de *clusters* et je n'émetts donc aucune contrainte sur la trace ou sur le rang de la matrice \mathbf{X} recherchée.

Concernant le nombre de *clusters*, remarquons que si on supprime la contrainte $\text{Tr}(\mathbf{X}) = k$ du problème (5.8), alors la solution optimale est immédiate et correspond à la partition triviale composée de n singletons. En effet, $\text{SSE}(\mathbf{X}) = \text{Tr}(\mathbf{K}) - \text{Tr}(\mathbf{K}\mathbf{X})$ et la fonction est nulle pour $\mathbf{X} = \mathbf{I}_n$. Le critère de partitionnement $\text{SSE}(\mathbf{X})$ n'est donc pas adapté dans ce cas. Considérons alors la **distance de Frobenius au carré** entre la matrice d'affinités initiale \mathbf{K} et la matrice recherchée \mathbf{X} . Il vient :

$$\begin{aligned} \|\mathbf{K} - \mathbf{X}\|_F^2 &= \langle \mathbf{K}, \mathbf{K} \rangle_F + \langle \mathbf{X}, \mathbf{X} \rangle_F - 2\langle \mathbf{K}, \mathbf{X} \rangle_F \\ &= \text{Tr}(\mathbf{K}^\top \mathbf{K}) - 2\text{Tr}(\mathbf{K}^\top \mathbf{X}) + \text{Tr}(\mathbf{X}^\top \mathbf{X}) \\ &= \text{Tr}(\mathbf{K}^2) - 2\text{Tr}(\mathbf{K}^\top \mathbf{X}) + \text{Tr}(\mathbf{X}). \end{aligned}$$

Minimiser $\|\mathbf{K} - \mathbf{X}\|_F^2$ revient alors à minimiser la $\text{SSE}(\mathbf{X})$ (rappelons que $\mathbf{K} = \mathbf{K}^\top$) pénalisée par la fonction $\frac{1}{2}\text{Tr}(\mathbf{X})$. La distance de Frobenius intègre donc naturellement un arbitrage entre la SSE et un terme favorisant un nombre faible de *clusters* ce qui permet d'éviter le cas trivial susmentionné. J'utilise donc dans mon modèle la distance de Frobenius à la place de la SSE :

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \|\mathbf{K} - \mathbf{X}\|_F^2 & \tag{5.22} \\ \text{s.l.c. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X}^2 = \mathbf{X}. \end{aligned}$$

J'interprète \mathbf{X} comme étant la matrice d'adjacence d'un graphe non orienté pondéré et je considère également sa matrice Laplacienne. \mathbf{X} étant bistochastique, la matrice des degrés que j'ai définie par l'équation (4.70) page 103, est la matrice identité \mathbf{I}_n et on a donc :

$$\mathbf{L} = \mathbf{I}_n - \mathbf{X}. \tag{5.23}$$

Ensuite, on peut faire un changement de variable dans (5.22) en prenant $\mathbf{X} = \mathbf{I}_n - \mathbf{L}$ et

exprimer, de façon équivalente, le problème en fonction de \mathbf{L} :

$$\begin{aligned} & \min_{\mathbf{L}_X \in \mathbb{R}^{n \times n}} \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_X\|_F^2 & (5.24) \\ \text{s.l.c. } & \mathbf{L}_X \leq \mathbf{I}_n, \mathbf{L}_X = \mathbf{L}_X^\top, \mathbf{L}_X \mathbf{e}_n = \mathbf{n}_n, \mathbf{L}_X^2 = \mathbf{L}_X. \end{aligned}$$

où \mathbf{n}_n est le vecteur nul de taille $n \times 1$.

On peut mélanger (5.22) et (5.24) et considérer le problème étendu suivant :

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{L} \in \mathbb{R}^{n \times n}} \|\mathbf{K} - \mathbf{X}\|_F^2 + \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}\|_F^2 & (5.25) \\ \text{s.l.c. } & \begin{cases} \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X} \mathbf{e}_n = \mathbf{e}_n, \mathbf{X}^2 = \mathbf{X}, \\ \mathbf{L} \leq \mathbf{I}_n, \mathbf{L} = \mathbf{L}^\top, \mathbf{L} \mathbf{e}_n = \mathbf{n}_n, \mathbf{L}^2 = \mathbf{L}. \end{cases} \end{aligned}$$

Le problème (5.25) n'est pas intéressant à première vue puisque \mathbf{X} et \mathbf{L} étant indépendants, cela revient à traiter deux fois la même tâche. A ce stade, l'astuce consiste à lier \mathbf{X} et \mathbf{L} en posant explicitement $\mathbf{L} = \mathbf{I}_n - \mathbf{X}$. En raison de cette dépendance linéaire, j'adopte par la suite la notation \mathbf{L}_X à la place de \mathbf{L} . Je peux alors exprimer de plusieurs façons les contraintes d'idempotence de \mathbf{X} et \mathbf{L}_X . J'emploie la propriété suivante qui fait écho à l'équation (5.16) :

$$\begin{cases} \mathbf{X} + \mathbf{L}_X = \mathbf{I}_n, \\ \mathbf{X}^2 = \mathbf{X}. \end{cases} \Leftrightarrow \begin{cases} \mathbf{X} + \mathbf{L}_X = \mathbf{I}_n, \\ \mathbf{L}_X^2 = \mathbf{L}_X. \end{cases} \Leftrightarrow \begin{cases} \mathbf{X} + \mathbf{L}_X = \mathbf{I}_n, \\ \mathbf{X} \mathbf{L}_X = \mathbf{0}_n. \end{cases} \quad (5.26)$$

Sous la condition que $\mathbf{X} + \mathbf{L}_X = \mathbf{I}_n$, l'idempotence de \mathbf{X} et celle de \mathbf{L}_X peut donc être remplacée par $\mathbf{X} \mathbf{L}_X = \mathbf{0}_n$.

J'introduis alors le problème suivant, équivalent à (5.22), qui apprend de façon jointe \mathbf{X} et sa matrice Laplacienne associée avec $\mathbf{X} + \mathbf{L}_X = \mathbf{I}_n$ et qui utilise la contrainte $\mathbf{X} \mathbf{L}_X = \mathbf{0}_n$:

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{L}_X \in \mathbb{R}^{n \times n}} \|\mathbf{K} - \mathbf{X}\|_F^2 + \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_X\|_F^2 & (5.27) \\ \text{s.l.c. } & \begin{cases} \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X} \mathbf{e}_n = \mathbf{e}_n, \\ \mathbf{L}_X \leq \mathbf{I}_n, \mathbf{L}_X = \mathbf{L}_X^\top, \mathbf{L}_X \mathbf{e}_n = \mathbf{n}_n, \\ \mathbf{X} + \mathbf{L}_X = \mathbf{I}_n, \mathbf{X} \mathbf{L}_X = \mathbf{0}_n. \end{cases} \end{aligned}$$

Le problème (5.27) reste NP-dur mais ouvre la porte à de nouvelles stratégies de relaxation. Je propose ainsi d'enlever la condition $\mathbf{X} \mathbf{L}_X = \mathbf{0}_n$ de l'ensemble des contraintes et d'ajouter un **terme de pénalisation dans la fonction objectif**. Le modèle suivant vise alors à approximer \mathbf{K} par une matrice \mathbf{X} qui est bistochastique et quasi-idempotente. Cette nouvelle

approche est intitulée **DSNI pour *Doubly Stochastic and Nearly Idempotent*** :

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{L}_\mathbf{X} \in \mathbb{R}^{n \times n}} \frac{1}{2} \|\mathbf{K} - \mathbf{X}\|_F^2 + \frac{1}{2} \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_\mathbf{X}\|_F^2 + \frac{\mu}{2} \|\mathbf{X}\mathbf{L}_\mathbf{X}\|_F^2 & (5.28) \\ \text{s.l.c. } & \begin{cases} \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \\ \mathbf{L}_\mathbf{X} \leq \mathbf{I}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n = \mathbf{n}_n, \\ \mathbf{X} + \mathbf{L}_\mathbf{X} = \mathbf{I}_n, \end{cases} \end{aligned}$$

où $\mu \geq 0$ est un coefficient de pénalité.

Notons que si $\mu = 0$, alors il n'y a aucun intérêt à garder deux variables, le modèle peut être exprimé en fonction de \mathbf{X} uniquement et il est alors équivalent à l'approche exposée par Ron Zass et Amnon Shashua dans [Zass and Shashua, 2005] qui est intitulée **DSN pour *Doubly Stochastic Normalisation***.

Le problème **DSNI** (5.28) possède les propriétés suivantes : **la fonction objectif est bi-convexe**, c'est à dire qu'elle est convexe en \mathbf{X} lorsque $\mathbf{L}_\mathbf{X}$ est fixée et elle est convexe en $\mathbf{L}_\mathbf{X}$ lorsque \mathbf{X} est fixée. De plus, **les contraintes du problème DSNI sont toutes linéaires** en ces deux variables. Je propose alors d'employer l'algorithme **ADMM (*Alternating Direction Method of Multipliers*)** comme procédure d'optimisation. Les algorithmes précurseurs de cette approche datent des années 1950 et ont été développés dans le cadre de plusieurs travaux en optimisation. Ces idées ont trouvé un nouvel écho au cours des années 2000 au sein de la communauté *machine learning*, en permettant de mettre en oeuvre des stratégies d'optimisation distribuée dans le cadre de tâches à grande échelle [Boyd et al., 2011].

Dans notre contexte, l'application de la procédure ADMM (*scaled version*) conduit à la procédure suivante [Ah-Pine, 2022] :

0. Initialiser $\mathbf{X}^0 \leftarrow \mathbf{K}$ et $\mathbf{U}^0 \leftarrow \mathbf{0}_n$.

1. Résoudre pour $\mathbf{L}_\mathbf{X}^{t+1}$ le problème partiel suivant avec \mathbf{X}^t fixé :

$$\begin{aligned} \mathbf{L}_\mathbf{X}^{t+1} & \leftarrow \arg \min_{\mathbf{L}_\mathbf{X} \in \mathbb{R}^{n \times n}} \frac{1}{2} \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_\mathbf{X}\|_F^2 + \frac{\mu}{2} \|\mathbf{X}^t \mathbf{L}_\mathbf{X}\|_F^2 + \frac{\rho}{2} \|\mathbf{X}^t + \mathbf{L}_\mathbf{X} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 & (5.29) \\ \text{s.l.c. } & \mathbf{L}_\mathbf{X} \leq \mathbf{I}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n = \mathbf{n}_n. \end{aligned}$$

2. Résoudre pour \mathbf{X}^{t+1} le sous-problème suivant avec $\mathbf{L}_\mathbf{X}^{t+1}$ fixé :

$$\begin{aligned} \mathbf{X}^{t+1} & \leftarrow \arg \min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \frac{1}{2} \|\mathbf{K} - \mathbf{X}\|_F^2 + \frac{\mu}{2} \|\mathbf{X}\mathbf{L}_\mathbf{X}^{t+1}\|_F^2 + \frac{\rho}{2} \|\mathbf{X} + \mathbf{L}_\mathbf{X}^{t+1} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 & (5.30) \\ \text{s.l.c. } & \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n. \end{aligned}$$

3. Mettre à jour \mathbf{U}^{t+1} :

$$\mathbf{U}^{t+1} \leftarrow \mathbf{U}^t + \mathbf{X}^{t+1} + \mathbf{L}_\mathbf{X}^{t+1} - \mathbf{I}_n. \quad (5.31)$$

4. Répéter 1., 2., 3., jusqu'à ce qu'une condition d'arrêt soit satisfaite.

Les problèmes partiels (5.29) and (5.30) sont convexes et peuvent être abordés par une stratégie de type **POCS** (*Projection on Convex Sets*). En bref, il s'agit de déterminer un point appartenant à l'intersection d'un ensemble de sous-ensembles convexes. Dans ce cas, à partir d'un point initial, POCS consiste à projeter séquentiellement et de façon cyclique jusqu'à obtenir un point fixe (voir [Bauschke and Borwein, 1996] pour un état de l'art et [Combettes and Pesquet, 2011]).

Afin d'alléger les développements qui vont suivre, je définis les sous-ensembles convexes de $\mathbb{R}^{n \times n}$ suivants :

- $\mathcal{U}_{\mathbf{I}} = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X} \leq \mathbf{I}_n\}$,
- $\mathcal{L}_0 = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X} \geq \mathbf{0}_n\}$,
- $\mathcal{S} = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X} = \mathbf{X}^\top\}$,
- $\mathcal{D}_n = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X}\mathbf{e}_n = \mathbf{X}^\top \mathbf{e}_n = \mathbf{n}_n\}$,
- $\mathcal{D}_e = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X}\mathbf{e}_n = \mathbf{X}^\top \mathbf{e}_n = \mathbf{e}_n\}$.

Je note par $\Pi_{\mathcal{A}}$ l'opérateur de projection sur \mathcal{A} où \mathcal{A} est un sous-ensemble convexe de $\mathbb{R}^{n \times n}$ parmi les éléments introduits ci-dessus.

On peut alors traiter la problème partiel (5.29) comme suit :

1. Résoudre le problème non contraint suivant :

$$\widehat{\mathbf{L}}_{\mathbf{X}} \leftarrow \arg \min_{\mathbf{L}_{\mathbf{X}} \in \mathbb{R}^{n \times n}} \frac{1}{2} \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_{\mathbf{X}}\|_F^2 + \frac{\mu}{2} \|\mathbf{X}^t \mathbf{L}_{\mathbf{X}}\|_F^2 + \frac{\rho}{2} \|\mathbf{X}^t + \mathbf{L}_{\mathbf{X}} - \mathbf{I}_n + \mathbf{U}^t\|_F^2. \quad (5.32)$$

2. Projeter $\widehat{\mathbf{L}}_{\mathbf{X}}$ sur \mathcal{S} :

$$\widehat{\mathbf{L}}_{\mathbf{X}} \leftarrow \Pi_{\mathcal{S}} \widehat{\mathbf{L}}_{\mathbf{X}}. \quad (5.33)$$

3. Projeter $\widehat{\mathbf{L}}_{\mathbf{X}}$ sur \mathcal{D}_n :

$$\widehat{\mathbf{L}}_{\mathbf{X}} \leftarrow \Pi_{\mathcal{D}_n} \widehat{\mathbf{L}}_{\mathbf{X}}. \quad (5.34)$$

4. Projeter $\widehat{\mathbf{L}}_{\mathbf{X}}$ sur $\mathcal{U}_{\mathbf{I}}$:

$$\widehat{\mathbf{L}}_{\mathbf{X}} \leftarrow \Pi_{\mathcal{U}_{\mathbf{I}}} \widehat{\mathbf{L}}_{\mathbf{X}}. \quad (5.35)$$

5. Répéter 3. et 4. jusqu'à ce qu'une condition d'arrêt soit satisfaite.

$\Pi_{\mathcal{D}_n}$ et $\Pi_{\mathcal{U}_{\mathbf{I}}}$ préservent la symétrie. Par conséquent, il n'est pas utile d'appliquer $\Pi_{\mathcal{S}}$ à nouveau suite à la première itération. Ensuite, il est intéressant de préciser que les problèmes partiels (5.32)-(5.35) ont des **solutions analytiques**.

Proposition. 2. *Les solutions optimales de (5.32), (5.33), (5.34) et (5.35) sont respectivement données par :*

$$\widehat{\mathbf{L}}_{\mathbf{X}} \leftarrow ((1 + \rho)\mathbf{I}_n + \mu[\mathbf{X}^t]^2)^{-1} (\mathbf{I}_n - \mathbf{K} + \rho(\mathbf{I}_n - \mathbf{X}^t - \mathbf{U}^t)). \quad (5.36)$$

$$\widehat{\mathbf{L}}_{\mathbf{X}} \leftarrow \frac{\widehat{\mathbf{L}}_{\mathbf{X}} + \widehat{\mathbf{L}}_{\mathbf{X}}^\top}{2}. \quad (5.37)$$

$$\widehat{\mathbf{L}}_{\mathbf{X}} \leftarrow (\mathbf{I}_n - \mathbf{J}_n)\widehat{\mathbf{L}}_{\mathbf{X}}(\mathbf{I}_n - \mathbf{J}_n). \quad (5.38)$$

$$\widehat{\mathbf{L}}_{\mathbf{X}} \leftarrow \min(\widehat{\mathbf{L}}_{\mathbf{X}}, \mathbf{I}_n) \text{ (opérateur min terme à terme)}. \quad (5.39)$$

Notons que (5.38) est le *double centering operator*.

Je fournis les preuves dans les annexes de [Ah-Pine, 2022].

Le problème partiel (5.30) peut également être traité de façon similaire avec la procédure et les résultats en forme close énoncés dans la Proposition 3 qui suivent.

1. Résoudre le problème non contraint suivant :

$$\widehat{\mathbf{X}} \leftarrow \arg \min_{\mathbf{L}_{\mathbf{X}} \in \mathbb{R}^{n \times n}} \frac{1}{2} \|\mathbf{K} - \mathbf{X}\|_F^2 + \frac{\mu}{2} \|\mathbf{X}\mathbf{L}_{\mathbf{X}}^{t+1}\|_F^2 + \frac{\rho}{2} \|\mathbf{X} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n + \mathbf{U}^t\|_F^2. \quad (5.40)$$

2. Projeter $\widehat{\mathbf{X}}$ sur \mathcal{S} :

$$\widehat{\mathbf{X}} \leftarrow \Pi_{\mathcal{S}} \widehat{\mathbf{X}}. \quad (5.41)$$

3. Projeter $\widehat{\mathbf{X}}$ sur \mathcal{D}_e :

$$\widehat{\mathbf{X}} \leftarrow \Pi_{\mathcal{D}_n} \widehat{\mathbf{X}}. \quad (5.42)$$

4. Projeter $\widehat{\mathbf{X}}$ sur \mathcal{L}_0 :

$$\widehat{\mathbf{X}} \leftarrow \Pi_{\mathcal{L}_0} \widehat{\mathbf{X}}. \quad (5.43)$$

5. Répéter 3. et 4. jusqu'à ce qu'une condition d'arrêt soit satisfaite.

Proposition. 3. *Les solutions optimales de (5.40), (5.41), (5.42) and (5.43) sont respectivement données par :*

$$\widehat{\mathbf{X}} \leftarrow (\mathbf{K} + \rho(\mathbf{I}_n - \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{U}^t)) ((1 + \rho)\mathbf{I}_n + \mu(\mathbf{L}_{\mathbf{X}}^{t+1})^2)^{-1}. \quad (5.44)$$

$$\widehat{\mathbf{X}} \leftarrow \frac{\widehat{\mathbf{X}} + \widehat{\mathbf{X}}^\top}{2}. \quad (5.45)$$

$$\widehat{\mathbf{X}} \leftarrow (\mathbf{I}_n - \mathbf{J}_n)\widehat{\mathbf{X}}(\mathbf{I}_n - \mathbf{J}_n) + \mathbf{J}_n. \quad (5.46)$$

$$\widehat{\mathbf{X}} \leftarrow \max(\widehat{\mathbf{X}}, \mathbf{0}_n) \text{ (opérateur max terme à terme)}. \quad (5.47)$$

Enfin, une propriété intéressante de la matrice obtenue à la suite de la procédure DSNI est exposée ci-dessous.

Proposition. 4. *Supposons que $(\mathbf{X}^*, \mathbf{L}_{\mathbf{X}}^*)$ soient des solutions du problème (5.28). Alors la matrice \mathbf{K}^* définie par :*

$$\begin{aligned}\mathbf{K}^* &= 2\mathbf{I}_n - \mathbf{L}_{\mathbf{X}}^* \\ &= \mathbf{I}_n + \mathbf{X}^*\end{aligned}\tag{5.48}$$

vérifie $\mathbf{K}^ \geq \mathbf{0}_n$, $\mathbf{K}^* = [\mathbf{K}^*]^\top$ et $\mathbf{K}^* \succeq \mathbf{0}_n$.*

Ainsi, en ajoutant la matrice identité \mathbf{I}_n à la matrice bistochastique et quasi-idempotente \mathbf{X}^* , on obtient une matrice de noyaux et la procédure DSNI peut être vue telle une **méthode d'apprentissage non supervisé de métrique**.

Afin de valider l'intérêt de la procédure DSNI, je compare celle-ci à deux méthodes existantes d'apprentissage non supervisé de matrices bistochastique à partir d'une matrice d'affinités \mathbf{K} , dans le contexte du *spectral clustering*. Les approches concurrentes sont la procédure DSN de Zass-Shashua déjà introduite précédemment et la version symétrique de l'algorithme Sinkhorn-Knopp [Sinkhorn and Knopp, 1967] discutée également dans [Zass and Shashua, 2005]. Tout comme DSNI, les deux méthodes sont des procédures itératives. DSN vise également à minimiser la distance de Frobenius alors que SSK s'appuie sur la divergence de Kullback-Leibler. Les deux modèles estiment des matrices bistochastiques mais elles ne tiennent pas compte ni l'une, ni l'autre, de l'idempotence contrairement à DSNI. Rappelons d'ailleurs, que DSN est un cas particulier de DSNI correspondant à la situation où $\mu = 0$ dans (5.28).

Étant donné une matrice d'affinités \mathbf{K} , on transforme celle-ci en une matrice bistochastique avec DSN ou SSK ou DSNI. La matrice bistochastique est alors employée comme entrée d'un *spectral clustering* classique. J'ai également testé avec des approches classiques où (i) la matrice \mathbf{K} est employée sans aucune transformation, et (ii) lorsque qu'on utilise le graphe des k plus proches voisins en sparsifiant \mathbf{K} . Pour comparer les partitions obtenues avec la vérité terrain, j'utilise, comme dans le Chapitre précédent, la mesure *Normalized Mutual Information* (NMI). Les performances obtenues sur plusieurs jeux de données classiques de la littérature sont exposées dans la Table 5.3. Par défaut, le noyau Gaussien est utilisé (voir [Ah-Pine, 2022] pour plus de détails sur les paramètres utilisés).

Afin de faciliter la comparaison entre les différentes méthodes, je transforme les valeurs NMI pour chaque tâche en une distribution de rangs sur les méthodes et je détermine un rang moyen.

La *baseline* \mathbf{K} qui est le *spectral clustering* utilisant la matrice de noyaux Gaussiens sans écrêtage, est classée dernière. Les trois techniques produisant des matrices bistochastiques sont meilleures que la *baseline* et ceci valide empiriquement l'intérêt de la condition de bistochasticité en *clustering*.

DSNI est la méthode qui est classée première avec un rang moyen de 1.33. Elle donne ainsi de meilleurs résultats que ses concurrentes DSN et SSK. De plus, ses performances sont robustes produisant soit la meilleure performance soit la deuxième meilleure performance.

Dataset	K	k -NN	SSK	DSN	DSNI
Glass	0.253	0.281	0.276	0.243	0.297
Ionosphere	0.038	0.082	0.066	0.076	0.131
Olivetti Faces	0.782	0.755	0.786	0.803	0.855
Vowel	0.382	0.412	0.321	0.206	0.423
Breast cancer	0.010	0.677	0.010	0.010	0.670
Vehicle	0.013	0.132	0.135	0.203	0.171
Yeast	0.070	0.329	0.258	0.256	0.263
Digits	0.015	0.552	0.044	0.743	0.767
Segment.scale	0.012	0.010	0.341	0.449	0.522

TABLE 5.3 – NMI performances.

Ceci montre l'intérêt de tenir compte de l'idempotence même de façon approximative.

Enfin, la méthode utilisant le graphe des k plus proches voisins issu de \mathbf{K} est seconde au classement général avec un rang moyen de 2.77. La restriction des relations d'affinités aux k plus proches voisins, vise à favoriser la reconnaissance des variétés non linéaires sous-jacentes aux différents *clusters*. Le fait que DSNI produise en moyenne de meilleurs résultats que cette méthode, indique empiriquement que **la méthode serait adaptée pour détecter des *non linear manifolds***.

5.2.2 Un nouveau modèle générique de classification ascendante hiérarchique

L'objectif de cette recherche est de définir un **cadre formel de la CAH aussi riche que celui rappelé en sous-section 5.1.2**, mais favorisant un **meilleur passage à l'échelle** et une plus grande capacité à **reconnaître des variétés non linéaires**.

Dans cette perspective, l'idée de départ que je promeus est simple. Il s'agit de **raisonner à partir d'une matrice de similarités ou d'affinités à la place d'une matrice de dissimilarités**. Dans ce cas, la sparsification de la matrice de similarités revient à remplacer par zéro des valeurs de similarité déjà très petites afin de focaliser l'analyse sur le graphe de plus proches voisins. En effet, ceci permet d'une part, d'alléger la complexité mémoire et d'autre part, de favoriser la détection des *non linear manifolds* similairement au *spectral clustering*.

Pourquoi raisonner avec des matrices de similarités et non pas avec des matrices de dissimilarités comme pour D-AHC, l'approche classique de la CAH rappelée en sous-section 5.1.2 ? Sparsifier un graphe de dissimilarités reviendrait à supprimer les arêtes dont les poids sont très grands, enlevant ainsi les connexions entre paires d'objets très distants qui ont un intérêt limité pour le *clustering*. Mais dans ce cas, la formule partielle de LW (5.21) n'est alors pas bien définie. Supposons que nous calculions la dissimilarité entre un nouveau cluster (kl) et un cluster existant m alors qu'il n'existe aucune arête entre ces deux groupes. Nous avons dans ce cas $d_{km}^t = d_{lm}^t = 0$, et en appliquant (5.21) avec $\beta(k, l) < 0$ (cf Table 5.2), on obtiendrait une dissimilarité $d_{(kl)m}^{t+1}$ négative ce qui n'est pas souhaité. Comme je le montrerai par la suite,

cette incohérence peut être évitée si on raisonne avec un graphe de similarités.

Avant cela, le premier challenge à relever est de déterminer une expression équivalente de la formule partielle de LW (5.21) mais en fonction des termes d'une matrice d'affinités. Je me positionne alors dans le cas spécifique mais classique où $\mathbf{D} = (d_{ii'})$ est une matrice de distances au carré calculées dans un espace de Hilbert. Je suppose alors que $\mathbf{S} = (s_{ii'})$ est la matrice de produits scalaires associée à \mathbf{D} .

Comme précédemment, je m'intéresse à des **méthodes à noyaux** qui permettent d'analyser de façon riche les données en les représentant dans des espaces de grande dimension. Ainsi, je suppose implicitement qu'il existe une application ϕ projetant les données initiales dans un *feature space* \mathbb{F} qui peut être mis en correspondance avec un RKHS associé à \mathbf{S} . Plus formellement, soit $\langle \cdot, \cdot \rangle_{\mathbb{F}}$ le produit scalaire dans \mathbb{F} . Je suppose que, $\forall i, i' = 1, \dots, n$:

$$\begin{cases} s_{ii'} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'}) \rangle_{\mathbb{F}}, \\ d_{ii'} = s_{ii} + s_{i'i'} - 2s_{ii'}. \end{cases} \quad (5.49)$$

Par conséquent, \mathbf{S} est une matrice de Gram et on a $\mathbf{S} = \mathbf{S}^{\top}$ et $\mathbf{S} \succeq \mathbf{0}_n$.

L'expression de la formule partielle de LW (5.21) en terme de noyaux, peut se faire de différentes façons. Une première approche avait été établie dans [Ah-Pine and Wang, 2016]. Mais celle que je vais présenter ci-après est différente et m'a permis d'établir plusieurs résultats théoriques pertinents. Ce travail a été publié dans [Ah-Pine, 2018]. La méthode est en fait constituée de deux équations. La première sert à **mettre à jour les produits scalaires *inter-clusters***, c'est à dire l'affinité entre le nouveau *cluster* et tous les *clusters* existants. La deuxième vise à **modifier l'affinité *intra-cluster*** autrement dit, la mesure de l'homogénéité au sein du nouveau *cluster*. Du point de vue matriciel, la première équation concerne des termes hors de la diagonale de la nouvelle matrice de similarités tandis que la deuxième s'applique à une cellule de la diagonale de cette dernière.

Soit une matrice de noyaux initiale $\mathbf{S} = (s_{ii'})$. On initialise la procédure en prenant $\mathbf{S}^t = \mathbf{S}$ pour $t = 1$. Supposons qu'à l'itération $t > 1$, les *clusters* k et l sont regroupés et forment le nouveau *cluster* (kl) . J'introduis alors les équations suivantes pour la mise à jour de la matrice \mathbf{S}^{t+1} :

$$\mathbf{S}_{(kl)m}^{t+1} = \mathbf{a}(k, l)\mathbf{S}_{km}^t + \mathbf{a}(l, k)\mathbf{S}_{lm}^t, \quad \forall m \in \mathbb{C}^{t+1}, m \neq (kl); \quad (5.50)$$

$$\mathbf{S}_{(kl)(kl)}^{t+1} = \mathbf{b}(k, l)\mathbf{S}_{kl}^t + \mathbf{c}(k, l)\mathbf{S}_{kk}^t + \mathbf{c}(l, k)\mathbf{S}_{ll}^t. \quad (5.51)$$

où \mathbf{a} , \mathbf{b} et \mathbf{c} sont des fonctions d'ensemble à valeurs réelles positives définies sur l'ensemble des paires d'éléments de \mathbb{C}^t .

Similairement à l'approche classique D-AHC, je suppose que les matrices \mathbf{S}^t sont symétriques et par conséquent, si on dénote par (kl) les nouveaux *clusters* formés à chaque itération, on a $\mathbf{S}_{m(kl)}^t = \mathbf{S}_{(kl)m}^t, \forall t \in \mathbb{T}, \forall m \in \mathbb{C}^{t+1}$.

En outre, pour chaque $t \in \mathbb{T}$, je définis la matrice carrée $\mathbf{\Lambda}^t = (\lambda_{ij}^t)$ qui est d'ordre $n - t + 1$

tout comme \mathbf{S}^t , et dont le terme général est donné par, $\forall i, j \in \mathbb{C}^t$:

$$\lambda_{ij}^t = s_{ij}^t - \frac{1}{2}(s_{ii}^t + s_{jj}^t). \quad (5.52)$$

Le Lemme qui suit établit une correspondance entre les quantités utilisées dans l’approche D-AHC lorsque \mathbf{D} est une matrice de distances Euclidiennes au carré d’une part, et l’approche fondée sur les trois équations précédentes que je dénote **K-AHC pour Kernel matrix based Agglomerative Hierarchical Clustering** d’autre part .

Lemme. 1. *Soient $\{\mathbf{D}^t\}_{t \in \mathbb{T}}$ et $\{\mathbf{\Lambda}^t\}_{t \in \mathbb{T}}$ deux suites de matrices initialisées par \mathbf{D} et \mathbf{S} respectivement et dont les éléments successifs sont définis d’une part par (5.21) et d’autre part par (5.52), (5.50), (5.51). Supposons que les fonctions d’ensemble α, β d’un côté, et $\mathbf{a}, \mathbf{b}, \mathbf{c}$ de l’autre, remplissent les conditions suivantes :*

$$\begin{aligned} \mathbf{a} &= \alpha, \\ \mathbf{b} &= -2\beta, \\ \mathbf{c} &= \alpha + \beta. \end{aligned}$$

Si \mathbf{D} et \mathbf{S} vérifient (5.49) et si pour tout couple (k, l) de sous-ensembles disjoints d’éléments on a $\mathbf{a}(k, l) + \mathbf{a}(l, k) = 1$, alors $\forall t \in \mathbb{T}, \forall i, j \in \mathbb{C}^t, i \neq j$:

$$\lambda_{ij}^t = -\frac{1}{2}d_{ij}^t. \quad (5.53)$$

Sous les conditions définies dans le Lemme 1, les séquences de matrices $\{\mathbf{D}^t\}_{t \in \mathbb{T}}$ et $\{\mathbf{\Lambda}^t\}_{t \in \mathbb{T}}$ sont en correspondance biunivoque. Il reste à préciser l’étape de sélection des *clusters* à regrouper dans K-AHC. A chaque itération $t \in \mathbb{T}$, il s’agit de **rechercher la paire qui correspond à la valeur la plus grande dans $\mathbf{\Lambda}^t$** . Plus formellement, la sélection est donnée par le critère suivant :

$$(k, l) = \arg \max_{(i, j) \in \mathbb{C}^t \times \mathbb{C}^t, i \neq j} \mathbf{p}(i, j) \lambda_{ij}^t, \quad (5.54)$$

où \mathbf{p} est une fonction d’ensemble.

Lorsque le nouveau *cluster* (kl) est créé à l’itération t , la valeur de la “hauteur” qui lui est attribuée est $H((kl)) = \mathbf{p}(k, l) \lambda_{kl}^t$. Ici, la “hauteur” dans K-AHC varie dans le sens opposé de la hauteur telle que définie dans D-AHC. On devrait plutôt parler de “profondeur”. Dans le cas de K-AHC, si le dendrogramme est monotone alors cette quantité diminue au fur et à mesure que l’arbre binaire grandit.

Suite aux développements et Lemme précédents, le Théorème qui suit synthétise les conditions dans lesquelles **K-AHC et D-AHC sont équivalents**.

Théorème. 8. *Si les conditions du Lemme 1 sont satisfaites et si $\mathbf{p} = p$, alors la CAH basée sur les affinités et définie par (5.54), (5.50) et (5.51) (K-AHC), produit la même succession*

5.2. CONTRIBUTIONS

de regroupements de clusters que la CAH basée sur les dissimilarités et définie par (5.20) et (5.21) (*D-AHC*).

Les paramètres des techniques classiques de CAH dans le cadre de K-AHC sont précisés dans la Table 5.4.

Method	$\mathbf{a}(k, l)$	$\mathbf{b}(k, l)$	$\mathbf{c}(k, l)$	$\mathbf{p}(i, j)$
Group average	$\frac{ k }{ k + l }$	0	$\frac{ k }{ k + l }$	1
Mcquitty	1/2	0	1/2	1
Centroid	$\frac{ k }{ k + l }$	$\frac{2 k l }{(k + l)^2}$	$\frac{ k ^2}{(k + l)^2}$	1
Median	1/2	1/2	1/4	1
Ward	$\frac{ k }{ k + l }$	$\frac{2 k l }{(k + l)^2}$	$\frac{ k ^2}{(k + l)^2}$	$\frac{ i j }{ i + j }$
W-Median	1/2	1/2	1/4	$\frac{ i j }{ i + j }$

TABLE 5.4 – Cas particuliers connus dans le cadre de K-AHC défini par (5.54), (5.50) et (5.51).

Ensuite, j'ai étudié la **propriété de monotonie dans le scope de ce nouveau modèle générique** de CAH. Glenn Milligan donne dans le cas de la formule usuelle de LW, des conditions suffisantes sur les fonctions d'ensemble de (5.19) afin que le dendrogramme soit monotone [Milligan, 1979]. Dans la même veine, je donne ci-dessous des conditions suffisantes sur \mathbf{a} , \mathbf{b} , \mathbf{c} et \mathbf{p} intervenant dans K-AHC, permettant de satisfaire à la condition de monotonie du dendrogramme [Ah-Pine, 2018].

Proposition. 5. *Soit $\{\Lambda^t\}_{t \in \mathbb{T}}$ la suite de matrices carrées initialisée par \mathbf{S} et dont les éléments successifs sont donnés par (5.52), (5.50), (5.51). Supposons que \mathbf{S} soit une matrice de produits scalaires conformément à l'hypothèse (5.49). Supposons, de plus, que les fonctions d'ensemble \mathbf{a} , \mathbf{b} , \mathbf{c} et \mathbf{p} sont telles que pour tout couple (k, l) de sous-ensembles disjoints d'éléments, les relations suivantes soient vérifiées :*

$$\left\{ \begin{array}{l} \mathbf{a}(k, l), \mathbf{b}(k, l), \mathbf{c}(k, l), \mathbf{p}(k, l) \geq 0, \\ \mathbf{a}(k, l) + \mathbf{a}(l, k) = 1, \\ \mathbf{b}(k, l) - \mathbf{b}(l, k) = 0, \\ \mathbf{c}(k, l) - \mathbf{a}(k, l) + \frac{1}{2}\mathbf{b}(k, l) = 0, \\ \frac{\mathbf{a}(k, l)}{\mathbf{p}(k, m)} + \frac{\mathbf{a}(l, k)}{\mathbf{p}(l, m)} - \frac{\mathbf{b}(k, l)}{2\mathbf{p}(k, l)} \geq 0, \\ \mathbf{p}((kl), m) \left(\frac{\mathbf{a}(k, l)}{\mathbf{p}(k, m)} + \frac{\mathbf{a}(l, k)}{\mathbf{p}(l, m)} - \frac{\mathbf{b}(k, l)}{2\mathbf{p}(k, l)} \right) \geq 1. \end{array} \right. \quad (5.55)$$

Alors, nous avons la propriété de monotonie suivante, $\forall t \in \mathbb{T}, \forall k, l, m \in \mathbb{C}^t$:

$$\mathbf{p}((kl), m) \Lambda_{(kl)m}^{t+1} \leq \mathbf{p}(k, l) \Lambda_{kl}^t. \quad (5.56)$$

La Proposition 5 permet de montrer que la méthode *W-Median* que j'ai introduite est garantie de produire un dendrogramme monotone contrairement à *Median*.

J’analyse également dans [Ah-Pine, 2018] la possibilité d’exprimer la procédure en fonction de vecteurs décrits explicitement dans un *input* ou *feature space* et non pas en fonction de la matrice de noyaux. Je ne détaillerai pas ces développements ici et renvoie le lecteur à l’article. Je mentionne néanmoins le fait que ce formalisme, dit **stored data**, permet de raisonner avec des vecteurs “moyens” représentant les *clusters*. Cette approche peut être pertinente pour diminuer la complexité mémoire dans un contexte *big data*.

Ensuite, afin de mieux interpréter le mécanisme sous-jacent aux formules (5.50) et (5.51), j’introduis dans [Ah-Pine, 2018] le **concept de similarité pénalisée**. Au préalable, notons que les fonctions d’ensemble de toutes les méthodes définies dans la Table 5.4, remplissent les conditions suivantes pour tout couple (k, l) de sous-ensembles disjoints d’éléments :

$$\begin{cases} \mathbf{a}(k, l), \mathbf{b}(k, l), \mathbf{c}(k, l) \geq 0, \\ \mathbf{a}(k, l) + \mathbf{a}(l, k) = 1, \\ \mathbf{b}(k, l) + \mathbf{c}(k, l) + \mathbf{c}(l, k) = 1. \end{cases} \quad (5.57)$$

Par conséquent, **les fonctions d’ensemble définissent des systèmes de pondération** (nombres positifs sommant à 1) à la fois pour la mise à jour des affinités *inter-clusters* (5.50) et pour celle des affinités *intra-cluster* (5.51). Les différences entre les méthodes correspondent ainsi à des stratégies diverses de moyennisation de ces deux quantités. Lors de la recherche de la paire de *clusters* la plus appropriée à fusionner à une itération t , on cherche le maximum parmi les valeurs au sein de la matrice $\mathbf{\Lambda}^t$ définie par (5.52). $\mathbf{\Lambda}^t$ peut être interprétée comme une matrice dont les valeurs sont des **similarités *inter-clusters* pénalisées par la moyenne arithmétique des similarités *intra-clusters***. En d’autres termes, pour que deux *clusters* k et l soient regroupés, il faut que leurs individus respectifs soient très connectés entre eux mais si les *clusters* forment séparément des ensembles fortement homogènes cela amoindrit le gain associé à leur fusion.

Je m’intéresse désormais aux apports de ce nouveau cadre K-AHC, dans la poursuite de l’objectif initial d’**améliorer la scalabilité de la CAH**. Sans perte de généralité, je suppose par la suite, que la matrice de produits scalaires initiale \mathbf{S} , vérifie les conditions suivantes qui sont communément admises dans le cas d’une similarité, $\forall i, i' = 1, \dots, n$:

$$\begin{cases} \mathbf{S}_{ii'} \geq 0 \text{ (non-négativité),} \\ \mathbf{S}_{ii'} = \mathbf{S}_{i'i} \text{ (symétrie),} \\ \mathbf{S}_{ii} \geq \mathbf{S}_{ii'} \text{ (auto-similarité maximale),} \\ \mathbf{S}_{ii} = \mathbf{S}_{ii'} \text{ (auto-similarité constante).} \end{cases}$$

Le noyau Gaussien, que j’utiliserai par défaut ici, satisfait à ces conditions. Toutefois, si \mathbf{S} est issue d’un noyau différent, alors on peut appliquer la normalisation d’ordre t que j’ai énoncée au Chapitre précédent dans la sous-section 4.2.2 pour obtenir l’auto-similarité maximale et constante. En outre, afin de respecter la non-négativité, on peut transformer la matrice \mathbf{S} par l’application $u\mathbf{S} + v\mathbf{1}_n$ avec $u > 0$ et $v \in \mathbb{R}$ et obtenir le résultat escompté.

Cette opération n'a aucune incidence sur le résultat de la procédure de K-AHC. Je montre en effet dans [Ah-Pine, 2018, Proposition 6], que cette dernière est invariante vis-à-vis de ce type d'opérateur linéaire.

Après avoir normalisé \mathbf{S} (si nécessaire), je sparsifie cette dernière en remplaçant par zéro les valeurs d'affinités les plus petites. Plusieurs méthodes d'écrêtage peuvent être utilisées. Une approche classique est le graphe des k plus proches voisins que j'ai déjà évoqué dans la sous-section précédente. Spécifiquement, pour chaque objet X_i , on conserve la similarité $s_{ii'}$ si $X_{i'}$ fait partie des k plus proches voisins de X_i ou si ce dernier fait partie des k plus proches voisins de $X_{i'}$. Une alternative est la sparsification en fonction d'un seuil θ et auquel cas, on conserve la similarité $s_{ii'}$ si $s_{ii'} \geq \theta$.

Je dénote l'approche de CAH basée sur les similarités mais appliquée à une **matrice de noyaux normalisée et sparsifiée**, par **SNK-AHC pour *Sparse Normalized Kernel matrix based Agglomerative Hierarchical Clustering***. Tout comme K-AHC, SNK-AHC utilise les formules de mises à jour des similarités *inter-clusters* et *intra-clusters* données par (5.50) et (5.51) respectivement. Cependant, la règle de SNK-AHC pour sélectionner les *clusters* à réunir est légèrement modifiée en comparaison de celle de K-AHC. Il s'agit d'un changement minime mais central pour l'amélioration de la scalabilité. Soit à l'itération $t \in \mathbb{T}$, $\mathbf{S}^t = (s_{ij}^t)$ la matrice de similarités entre les paires de *clusters* de \mathbb{C}^t . Pour tout $t \in \mathbb{T}$, notons alors par \mathbb{S}^t , **l'ensemble des paires de *clusters* dont la similarité est strictement positive** :

$$\mathbb{S}^t = \{(i, j) \in \mathbb{C}^t \times \mathbb{C}^t : s_{ij}^t > 0\}. \quad (5.58)$$

SNK-AHC sélectionne la paire de *clusters* à regrouper à l'itération t en cherchant uniquement parmi les éléments dans \mathbb{S}^t :

$$(k, l) = \arg \max_{(i, j) \in \mathbb{S}^t, i \neq j} \mathbf{p}(i, j) \lambda_{ij}^t. \quad (5.59)$$

Remarquons que si la matrice de noyaux n'est pas sparsifiée alors SNK-AHC est similaire à K-AHC.

Clairement, la règle (5.59) peut **améliorer drastiquement la scalabilité de la CAH** si la matrice \mathbf{S} donnée en entrée est très sparse. Au-delà d'un meilleur passage à l'échelle, la règle (5.59) est aussi motivée par les arguments suivants. Premièrement, les paires $(i, j) \notin \mathbb{S}^t$ ont des valeurs λ_{ij}^t qui sont sous-optimales vis-à-vis du problème (5.54). Deuxièmement, regrouper des *clusters* qui ne partagent aucune arête n'est pas justifié du point de vue du *clustering* et contribuerait, au contraire, à créer du bruit.

A ce stade, il est important de préciser que SNK-AHC s'arrête à l'itération t si $\mathbb{S}^t = \emptyset$. Dans ce cas, si $t < n - 1$ alors SNK-AHC donne en sortie non pas un arbre avec une racine mais une forêt d'arbres comportant $n - t$ racines. Je montre dans [Ah-Pine, 2018] que ces arbres sont les **composantes connexes du graphe écrêté**.

Il est aussi important de mentionner que la mise à jour des similarités *inter-clusters* des matrices \mathbf{S}^t donnée par (5.50), conserve la sparsité et n'implique **aucune distorsion de l'es-**

pace contrairement au cas des matrices de dissimilarités comme je l'avais évoqué au début de cette sous-section.

Dans ce qui suit, j'expose des résultats d'expérience où je m'intéresse aux propriétés de scalabilité de la procédure SNK-AHC ainsi qu'à sa capacité à détecter des variétés non linéaires. Je commence par illustrer ce deuxième point en montrant les résultats obtenus par SNK-AHC sur le jeu de données synthétique donné dans la Figure 5.1 dénommé *Compound* et composé de 399 points de \mathbb{R}^2 . Les symboles et couleurs indiquent les six *clusters* qu'il faut arriver à détecter automatiquement. La tâche est compliquée en raison des formes (non convexes) et échelles variées des différents groupes.

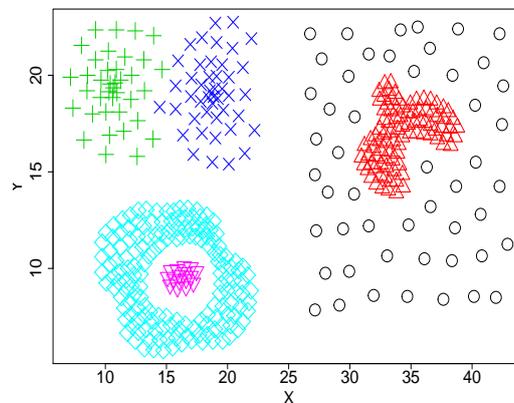


FIGURE 5.1 – Jeu de données *Compound* comportant 399 points de \mathbb{R}^2 répartis en 6 *clusters*.

Dans la Figure 5.2, j'expose les résultats obtenus par SNK-AHC avec un noyau Gaussien et une matrice de similarité à 99% creuse. Autrement dit, je ne garde que les 1% valeurs d'affinités les plus grandes dans la matrice de noyaux. Toutes les techniques décrites par les paramètres donnés dans la Table 5.4 donnent alors le même résultat qui est composé de 99 composantes connexes. La grande majorité de ces *clusters* ont un cardinal très petit. Si un objet appartient à un *cluster* de taille inférieure ou égale à 3, alors je le représente par une étoile grise dans la Figure 5.2.

SNK-AHC est capable de détecter les régions denses de la majorité des *clusters* malgré leurs formes non linéaires et non convexes. Toutefois, si la densité des points varie comme pour le groupe des + en vert et celui des × en bleu, SNK-AHC ne retrouve pas tous les membres lorsqu'ils se situent dans des régions moins peuplées. Le *cluster* des o en noir de la Figure 5.1 est celui qui pose le plus de difficulté. Il est caractérisé par des éléments diffus qui se concentrent en aucune région précise de l'espace. Ce dernier cas est particulièrement ardu et SNK-AHC échoue à l'identifier. Malgré ces inconvénients, il est important d'indiquer que SNK-AHC produit des résultats bien meilleurs que ce que l'on obtiendrait avec une méthode classique utilisant un noyau linéaire comme le *k-means*.

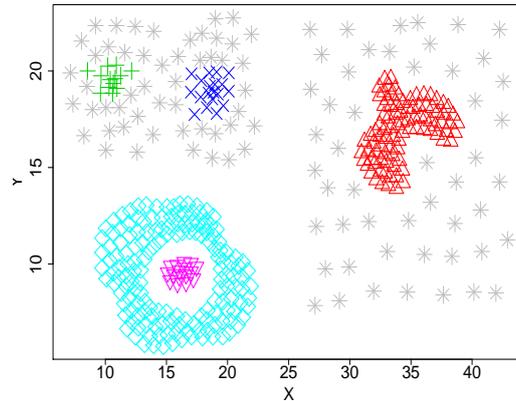


FIGURE 5.2 – Résultat de SNK-AHC sur le jeu de données *Compound* (les *clusters* de taille ≤ 3 sont indiqués de façon indifférenciée par des * en gris).

SNK-AHC donne ainsi des résultats très satisfaisants sur le cas *Compound*. Il est intéressant de rappeler que ceux-ci ont été obtenus à partir d’une matrice extrêmement sparse. Comme en *spectral clustering*, se restreindre au graphe des plus proches voisins permet de focaliser l’analyse sur les relations topologiques robustes et d’encoder la géométrie intrinsèque des données. De plus, dans le cas de SNK-AHC, cela a également pour effet de réduire significativement les complexités mémoire et temps de traitement.

J’illustre davantage par la suite, les propriétés de scalabilité de SNK-AHC à partir d’un jeu de données réelles. Il s’agit des données *Landsat* dont les 6435 observations sont des patches de taille 3×3 d’images satellitaires de terrain. Chaque pixel d’un patch est représenté par 4 valeurs de bandes spectrales. L’objectif est de segmenter automatiquement l’ensemble des patches en 6 groupes distincts correspondant à 6 différents types de sols. J’utilise à nouveau ici un noyau Gaussien. L’hyperparamètre σ^2 est fixé à 36 qui est la dimension de l’espace de description ($3 \times 3 \times 4$). Je compare la vérité terrain avec la partition en 6 *clusters* extrait d’un dendrogramme. La mesure ARI pour *Adjusted Rand Index* est utilisée à cet effet. Plus la mesure ARI est grande plus la partition obtenue est proche de la vérité terrain. La Figure 5.3 indique des statistiques sur les dendrogrammes obtenus avec chacune des six techniques *Group average*, *Mcquitty*, *Centroid*, *Median*, *Ward* et *W-Median*.

On constate que plus la matrice \mathbf{S} en entrée est sparse, plus les complexités mémoire et temps de traitement sont améliorées. SNK-AHC met donc en oeuvre une approche de la CAH qui présente des propriétés de scalabilité intéressantes. De surcroît, dans le cas du jeu de données *Landsat*, ce passage à l’échelle ne provoque pas nécessairement une dégradation des performances en terme de qualité. Au contraire, dans le cas de *Group average*, le meilleur score ARI est obtenu avec le graphe le plus parcimonieux. Autrement dit, **SNK-AHC permet à la fois de diminuer globalement la complexité de la CAH, et d’améliorer la reconnaissance des *clusters***. Dans ce contexte, l’écritage du graphe permet de renforcer

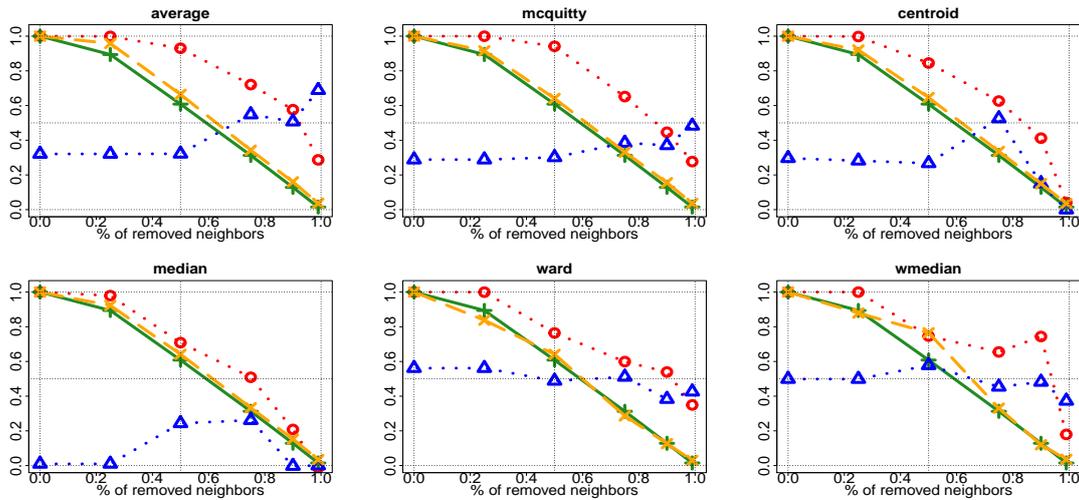


FIGURE 5.3 – Résultats de SNK-AHC avec un noyau Gaussien pour le jeu de données réelles *Landsat*. L’axe des abscisses correspond au % de voisins supprimés. L’axe des ordonnées correspond à des valeurs d’indicateurs variant dans $[0, 1]$. Les traits continus verts avec + représentent la mémoire relative employée, les traits discontinus jaunes avec × indiquent le temps de traitement relatif, les traits pointillés bleu avec Δ montrent la mesure ARI *Adjusted Rand Index*. Je ne commente pas les traits rouges avec ○.

la prise en compte de la géométrie intrinsèque des données conduisant à une **meilleure reconnaissance des cas correspondants à des variétés non linéaires**.

5.3 Discussions et perspectives

J’ai introduit des nouveaux modèles pour le problème de *clustering* avec une attention particulière pour la détection de *non linear manifolds*. A cet effet, je me suis reposé sur deux concepts centraux : les fonctions noyaux d’une part et le point de vue graphe d’autre part.

L’approche DSNI se situe dans le cadre des méthodes de partitionnement. Les résultats empiriques que j’ai obtenus sont encourageant et invitent à développer ce modèle selon plusieurs axes. Je reviens en premier lieu sur la Proposition 4 qui établit la **propriété de métricité de la matrice $\mathbf{K}^* = \mathbf{I}_n + \mathbf{X}^*$** où \mathbf{X}^* est la matrice bistochastique quasi-idempotente qui est solution du problème DSNI. Il me semble intéressant d’étudier ce modèle pour la **tâche de metric learning en apprentissage supervisé**. Par exemple, DSNI pourrait être utilisée en amont d’un SVM afin de produire une matrice de noyaux spécifiquement adaptée aux données à l’étude. Ce contexte me permet d’indiquer un fait que j’ai observé concernant les expériences menées avec DSNI : les matrices \mathbf{X}^* obtenues par l’algorithme sont sparses. Cette propriété pourrait avoir des conséquences intéressantes sur les performances du SVM (ou toute autre méthode supervisée à noyaux) en terme de qualité et en terme de scalabilité.

Ensuite, la normalisation bistochastique de matrices de noyaux fait l’objet de plusieurs études théoriques récentes. L’article [Landa et al., 2021] établit des propriétés avantageuses de la normalisation bistochastique en comparaison de la normalisation stochastique (des lignes), lorsque les données sont corrompues par un **bruit hétéroscédastique**. Dans la même veine, les travaux exposés dans [Landa and Cheng, 2023], s’intéressent à la **robustesse de la normalisation bistochastique** pour l’estimation de densité sur des variétés non linéaires en présence de bruit et d’*outliers*. Dans le cas plus général d’un mélange de plusieurs variétés non linéaires, il me semble propice d’étudier ces questions de **robustesse dans le cadre du modèle DSNI**. En sus de la normalisation bistochastique, DSNI tient compte de la propriété d’idempotence qui pourrait se révéler profitable ici.

L’approche **DSNI** définie par (5.22), est un **modèle relaxé** du problème discret NP-dur du partitionnement de graphe en cliques. La procédure d’optimisation continue basée sur l’approche ADMM que j’ai définie, offre une stratégie intéressante pour solutionner DSNI. Il serait intéressant d’analyser la capacité de cet algorithme à **résoudre le problème initial** (5.22). Une piste, classique en **optimisation combinatoire**, consisterait à résoudre une séquence d’instances de DSNI avec des valeurs de plus en plus grandes de l’hyperparamètre de pénalité μ dans (5.28). A chaque itération t , on utiliserait alors la solution \mathbf{X}^* trouvée à l’étape $t - 1$ comme solution initiale (*warm start*).

En ce qui concerne la CAH, la méthode SNK-AHC présente des propriétés intéressantes et des perspectives de recherche tant sur le plan empirique que théorique. La scalabilité, l’utilisation de fonctions noyaux et la capacité à détecter des *clusters* de formes diverses permettraient à SNK-AHC de trouver de très nombreuses applications dans l’analyse de données complexes et massives et d’outrepasser les limites de l’approche classique de la CAH. Par exemple, l’**analyse de données fonctionnelles**, que j’aborderai dans le prochain et dernier Chapitre, fait partie des applications au sein desquelles j’envisage d’étudier les apports de SNK-AHC.

Le modèle mathématique sous-jacent à **SNK-AHC** est riche. J’ai donné en Proposition 5 des conditions suffisantes permettant de qualifier si une similarité définie par les équations (5.50) et (5.51) aboutissaient à un dendrogramme monotone. Ce résultat est inspiré de l’article [Milligan, 1979] qui s’appuie sur la formule initiale de Lance-Williams (5.19). Il existe une littérature intéressante sur les **propriétés topologiques** associées à des sous-ensembles de valeurs pour les fonctions d’ensemble impliquées dans la formule de LW. Les concepts de **space-contracting** ou **space-dilating** également introduits dans [Lance and Williams, 1967] et étudiés par exemple dans [Chen and Van Ness, 1994], en sont des illustrations significatives. Il serait alors pertinent d’étudier ces notions dans le cadre de SNK-AHC ce qui pourrait aboutir à une meilleure compréhension des propriétés de ce modèle générique.

J’évoque à présent une propriété spécifique à SNK-AHC que je n’ai pas discutée précédemment. Lorsque la matrice de noyaux \mathbf{S} est sparsifiée, elle reste symétrique mais, en général, elle n’est plus semi-définie positive. Or, l’équivalence entre D-AHC et SNK-AHC repose notamment sur le fait que \mathbf{D} est la matrice de distances Euclidiennes au carré associée à la matrice

de produits scalaires \mathbf{S} . Si cette dernière n'est plus semi-définie positive, alors la matrice \mathbf{D} n'encode plus une représentation Euclidienne des objets. Pour pallier ce problème il est toujours possible d'augmenter la diagonale de \mathbf{S} d'une constante suffisamment grande avant de la rendre semi-définie positive. Je montre dans [Ah-Pine, 2018, Théorème 2] que pour les méthodes *Group average, Mcquitty et Ward*, la procédure **SNK-AHC est invariante aux translations de la diagonale de \mathbf{S}** . Autrement dit, pour ces techniques, SNK-AHC produit le même résultat pour \mathbf{S} et pour $\mathbf{S} + w\mathbf{I}_n$ avec $w \in \mathbb{R}$. Il n'est donc pas nécessaire de se soucier de la semi-définie positivité de \mathbf{S} dans ces cas. Par ailleurs, il est intéressant de noter que les résultats les plus pertinents obtenus pour les données *Landsat* sont justement pour ces trois méthodes. J'ai pu faire cette même observation avec un autre cas d'étude présenté dans [Ah-Pine, 2018]. Je pense qu'il y a un fort intérêt à étudier ce phénomène plus dans le détail. Une première question de recherche dans cette perspective que je suggère consiste, comme pour la monotonie, à déterminer des conditions sur les fonctions d'ensemble impliquées dans (5.50) et (5.51) qui permettent de **garantir une méthode invariante aux translations de la diagonale de \mathbf{S}** .

Enfin, le dendrogramme obtenu à l'issue d'une CAH a des propriétés qui sont utiles pour améliorer la représentation efficace des données. En particulier, le dendrogramme permet d'inférer une **ultramétrie** sur l'ensemble des objets. Pour deux individus X_i et $X_{i'}$, soit $H(X_i, X_{i'})$ la valeur absolue de la hauteur du noeud du dendrogramme correspondant au plus petit *cluster* contenant X_i et $X_{i'}$. Alors H est une ultramétrie, c'est à dire que pour tout triplet d'objets $X_i, X_{i'}$ et $X_{i''}$, nous avons : $H(X_i, X_{i'}) \leq \max(H(X_i, X_{i''}), H(X_{i'}, X_{i''}))$. Dans un tel espace, les relations de distance entre objets correspondent soit à des triangles équilatéraux, soit à des triangles isocèles avec une base de plus petite taille. Plusieurs travaux de Fionn Murtagh s'intéressent à la **structure topologique de nature ultramétrique, des données décrites dans des espaces de très grande dimension** [Murtagh, 2004, Murtagh and Contreras, 2012b, Murtagh, 2017]. Dans ces espaces, plusieurs symétries permettent de simplifier la représentation des données avec des retombées computationnelles intéressantes. La complexité de la recherche du plus proche voisin est alors drastiquement réduite ce qui a donc des implications avantageuses en *big data* pour la recherche d'information et la découverte de connaissances. En amont, il est nécessaire de plonger les données dans un espace ultramétrique et dans cette perspective, SNK-AHC pourrait être profitable à la fois du point de vue empirique et théorique.

Méthodes d'apprentissage en analyse de données fonctionnelles

Sommaire du chapitre

6.1	Introduction	136
6.1.1	Contexte	136
6.1.2	Travaux antérieurs	139
	Pré-traitements classiques en ADF	139
	Clustering de données fonctionnelles	141
	Catégorisation de données fonctionnelles	142
	Utilisation des dérivées en analyse de données fonctionnelles	142
6.2	Contributions	143
6.2.1	Représentation de données fonctionnelles avec dérivées	143
6.2.2	<i>k-means</i> à noyaux multiples pour données fonctionnelles avec dérivées	145
6.2.3	SVM à noyaux multiples pour données fonctionnelles avec dérivées	150
6.3	Discussions et perspectives	154

6.1 Introduction

6.1.1 Contexte

Dans ce dernier Chapitre, j'expose des activités de recherche que j'ai entamées plus récemment, à partir de janvier 2019. Je me suis intéressé à l'analyse de données fonctionnelles lors de l'accueil en délégation CNRS que j'ai effectué au sein du Laboratoire Mathématiques Blaise Pascal (LMBP) de l'Université Clermont Auvergne (UCA) au cours du 1er semestre de l'année 2019. Dans ce contexte, je collabore principalement avec Anne-Françoise Yao Professeure du LMBP au sein de l'équipe Probabilités Analyse et Statistiques (PAS).

Cette nouvelle thématique de recherche correspond également à un changement important dans ma vie personnelle. Je suis père de deux enfants, des jumelles, qui sont nées en 2015. Mon épouse, devenue docteur en Science Politique cette même année, a décroché un poste d'ATER au sein de l'UCA à la rentrée 2016. Nous habitons à Lyon à ce moment là. Mon épouse a effectué des allers-retours entre Lyon et Clermont-Ferrand tout au long de l'année universitaire

2016-2017. L'éloignement familial a été difficile et compliqué. Le poste d'ATER de mon épouse a été reconduit l'année d'après. Nous avons pris la décision de vivre à Clermont-Ferrand et nous y sommes installés depuis décembre 2017. J'ai, à mon tour, effectué des allers-retours entre Clermont-Ferrand et Lyon à partir de janvier 2018.

Dans ce contexte, je me suis alors rapproché du LMBP et du Laboratoire Informatique Modélisation et Optimisation des Systèmes (LIMOS) de l'UCA qui deviendra par la suite mon nouveau laboratoire. Le LMBP a accepté de m'accueillir en délégation CNRS sur un projet de recherche en *machine learning*. Anne-Françoise Yao contribue en analyse de données fonctionnelles avec des méthodes statistiques et était intéressée d'investiguer des approches plus répandues au sein de la communauté *machine learning*. Pour ma part, j'étais heureux de pouvoir étudier un nouveau sujet de recherche faisant intervenir des outils mathématiques que je n'avais pas approfondis dans le passé. De plus, l'analyse de données fonctionnelles offre un cadre mathématique englobant de nombreux types de données tels que les séries temporelles, les données de capteurs, les données de télédétection, . . . Elles concernent ainsi de très nombreuses applications en science des données. C'était pour moi l'occasion de continuer à développer mes compétences en analyse de données complexes après, les données relationnelles, les données mixtes, les textes et les images.

Nous avons donc collaboré avec Anne-Françoise sur l'utilisation de **méthodes en apprentissage automatique pour données fonctionnelles**. Une idée centrale que nous développons est l'**utilisation des fonctions dérivées en sus des fonctions initiales pour une information enrichie des données avec un point de vue multi-sources**. Nous nous sommes intéressés aux tâches de *clustering* et de *catégorisation*. Plusieurs communications scientifiques ont été réalisées. Un article de synthèse de ces contributions est actuellement en cours de révision (après relecture) pour le journal *Neurocomputing*.

Dans la section 6.2, je présenterai essentiellement les résultats de cette soumission. Nous nous intéressons en particulier à des fonctions de l'espace de Sobolev \mathbb{H}^q et employons des méthodes à noyaux multiples afin de définir un cadre flexible de représentation et d'analyse de ces données. Nous étendons alors des **méthodes à noyaux multiples** étudiées pour des données vectorielles au cas de données fonctionnelles avec dérivées.

Je m'intéresse également au **problème d'alignement de données fonctionnelles** qui se présente lorsque les observations discrètes des fonctions sont décalées les unes par rapport aux autres. Par exemple, dans le cas de signaux biomédicaux comme les électrocardiogrammes, les courbes peuvent être affectées par des artefacts de mouvement, des interférences électriques, . . . Le problème d'alignement engendre alors des imprécisions lors du calcul de distances entre courbes et peut donc avoir un impact très néfaste sur les résultats d'analyse. Dans ce contexte, les approches qui m'intéressent sont développées en **analyse de données fonctionnelles élastique**. On suppose que les fonctions définies sur $[0, T]$ sont absolument continues et appartiennent à une variété différentiable dotée de la métrique Riemannienne de Fisher-Rao qui a la

singularité d'être invariante vis-à-vis du groupe des re-paramétrisations¹ de $[0, T]$. Une autre propriété attrayante de cette représentation géométrique est qu'il existe une isométrie avec l'espace de Hilbert $\mathbb{L}^2([0, T])$ muni du produit scalaire usuel qui repose sur la transformation dite SRVF pour *Square Root Velocity Functions* (voir par exemple [Srivastava et al., 2011]).

En 2021-2022, j'ai encadré Noé Lebreton qui était alors étudiant en Master DM (*Data Mining*) de l'UL2 dans le cadre d'un TER que j'ai proposé sur l'analyse de données fonctionnelles élastique pour des problèmes de régression. Après avoir aligné les fonctions, on peut définir deux axes d'analyse, l'une centrée sur l'amplitude (distance par rapport à l'axe des ordonnées par comparaison des fonctions alignées), et l'autre fondée sur la phase (distance par rapport à l'axe des abscisses par comparaison des fonctions de re-paramétrisations permettant d'aligner les courbes). Dans [Lee and Jung, 2016, Tucker et al., 2019], les auteurs proposent de combiner ces deux sources d'information par une méthode de *early fusion* que j'ai discuté dans le Chapitre 1. Avec Noé nous avons testé l'approche de *late fusion* et montré qu'elle pouvait aboutir à de meilleures performances. Ce travail de TER a donné lieu à une communication aux Journées de Statistiques (JDS) de 2022 organisées à Lyon [Ah-Pine and Lebreton, 2022]. Je ne détaillerai pas ce travail dans la suite et renvoie le lecteur vers la précédente communication.

Noé Lebreton a poursuivi son cursus par un doctorat. Il est actuellement en thèse CIFRE sous la direction de Julien Jacques Professeur de l'UL2 et actuel directeur du laboratoire ERIC, et moi-même. Son sujet de thèse est le suivant, "Modélisation prédictive ensembliste à l'aide d'approches fonctionnelles".

Les communications/publications qui entrent dans le cadre de ce Chapitre sont les suivantes :

- **J. Ah-Pine** and A-F. Yao. (En cours de révision après relecture). On using derivatives and multiple kernel methods for clustering and classifying functional data. *Neurocomputing*.
- S. Dabo-Niang, **J. Ah-Pine**, P. Llop and A-F. Yao. 2023. Some results on statistics and functional data analysis. *ESAIM : Proceedings and Surveys*, 74, 2-18.
- **J. Ah-Pine** and N. Lebreton. 2022. Fusion tardive an analyse de données fonctionnelles élastique. In *53ème Journées de Statistiques, (JDS 2022)*. [Lien vers les proceedings, <https://jds22.sciencesconf.org/data/pages/LivretJdS22.pdf>].
- **J. Ah-Pine** and A-F. Yao. 2021. Multiple kernel SVM for classifying functional data in Sobolev spaces. *Journées MAS 2020* [Lien vers la conférence, <https://mas2020.sciencesconf.org/>].
- **J. Ah-Pine** and A-F. Yao. 2020. Une approche par noyaux multiples pour l'apprentissage non supervisé de représentation de données fonctionnelles dans des espaces de Sobolev. In *51ème Journées de Statistiques, (JDS 2020)*. [Lien vers les proceedings, https://jds2020.sciencesconf.org/data/pages/book_jds2020_fr_compressed.pdf].

1. Ensemble des difféomorphismes γ de $[0, T]$ dans $[0, T]$ tels que $\gamma(0) = 0$ et $\gamma(T) = T$.

- **J. Ah-Pine** and A-F. Yao. 2019. A study of the manifold hypothesis for functional data by using spectral clustering. *12th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2019)*.

6.1.2 Travaux antérieurs

Je rappelle ci-dessous les bases de l'analyse de données fonctionnelles (ADF). De plus, je donne de brefs états de l'art sur le *clustering* et la classification de données fonctionnelles (DF), ainsi que sur l'utilisation de dérivées en ADF. En effet, les contributions que je présenterai par la suite concernent ces différentes thématiques.

Avant cela, je motive la thématique de ce Chapitre au regard des enjeux actuels en *big data* et en *internet of things* et de leurs applications dans de nombreux domaines scientifiques, technologiques et sociétaux.

Les technologies modernes permettent l'enregistrement massif d'observations de divers phénomènes à des résolutions fines dans l'espace et dans le temps. Par exemple, les changements climatiques et environnementaux peuvent être mesurés grâce à des instruments de télédétection, la santé des machines dans les installations peut être surveillée à l'aide de capteurs, les mouvements humains et les activités physiques peuvent être détectés avec un accéléromètre de smartphone, ... Ces mesures sont associées à des horodatages et/ou des emplacements géographiques et sont enregistrées sous forme de données discrètes. Cependant, elles représentent en réalité des observations discrétisées de courbes ou de surfaces continues. D'un point de vue de l'analyse des données, il peut être avantageux de **considérer la nature continue du phénomène** étudié plutôt que d'analyser uniquement les observations discrètes. En particulier, **travailler avec des fonctions continues permet d'utiliser des outils de l'analyse fonctionnelle tels que les opérateurs différentiels**. L'ADF est la branche de la statistique et de la science des données concernée par ce sujet.

Pré-traitements classiques en ADF

Dans la suite, je me concentrerai exclusivement sur des **données fonctionnelles univariées à l'instar de séries temporelles**. Comme évoqué ci-dessus, en pratique, on n'observe pas directement des courbes entières, mais des échantillons de leurs réalisations à différents points dans l'intervalle $[0, T]$. Par conséquent, avant toute analyse, il est nécessaire de reconstruire une forme fonctionnelle approximative en utilisant l'ensemble fini et discret de valeurs à disposition. Bien que les ensembles des points d'observation pour deux DF distinctes, x_i et $x_{i'}$, puissent être différents, nous supposons que toutes les DF ont été mesurées par rapport à la même grille temporelle $\{t_j\}_{j=1,\dots,p}$. Ainsi, pour toutes les fonctions x_i , $i = 1, \dots, n$, nous avons p observations $\{y_{ij}\}_{j=1,\dots,p}$. Cependant, nous faisons l'hypothèse que ces mesures auraient pu être corrompues par du bruit. Nous supposons donc que :

$$y_{ij} = x_i(t_j) + \epsilon_{ij}, \quad \forall i = 1, \dots, n, \forall j = 1, \dots, p,$$

où $\{\epsilon_{ij}\}_{i=1,\dots,n,j=1,\dots,p}$ sont supposés être indépendants lorsque i ou j varie.

Pour **inférer des expressions fonctionnelles approximatives pour $\{x_i\}_i$ à partir de $\{y_{ij}\}_{i,j}$** , nous supposons que les DF peuvent être représentées comme des combinaisons linéaires d'un ensemble prédéfini de fonctions de base. Dans ce contexte, nous considérons le **système de base B-splines** couramment utilisé, qui est constitué de fonctions polynomiales. Soit alors $\{\phi_k\}_{k=1,\dots,m}$ un ensemble de m B-splines que nous désignons sous forme vectorielle par $\phi = (\phi_k)$. Nous supposons donc que les approximations des DF sont des éléments du sous-espace $\text{Span}(\phi_1, \dots, \phi_m) \subset \mathbb{H}^q$:

$$x_i = \sum_{k=1}^m c_{i,k} \phi_k = \mathbf{c}_i^\top \phi, \quad \forall i = 1, \dots, n,$$

où \mathbf{c}_i est le vecteur $(m \times 1)$ des coefficients de x_i dans la base de fonctions choisie.

Il est important de souligner que l'utilisation d'un ensemble de **fonctions de base lisses** $\{\phi_k\}_k$ facilite la détermination des fonctions dérivées successives de $\{x_i\}_i$. Comme l'opérateur différentiel D est linéaire, il suffit en effet de déterminer les ensembles des dérivées des fonctions de base $\{D^s \phi_k\}_k$ pour $s = 1, \dots, q$.

Pour chaque élément x_i , il est nécessaire d'estimer \mathbf{c}_i en se basant sur les observations $\{y_{ij}\}_j$. En raison de la présence supposée d'un bruit, ce problème est généralement abordé en utilisant une approche par moindres carrés. De plus, pour éviter le sur-ajustement et obtenir un meilleur contrôle sur la régularité des DF, un terme de pénalité, noté R , est appliqué.

Plus formellement, la procédure de *spline smoothing* qui estime \mathbf{c}_i pour la donnée fonctionnelle x_i implique la résolution de :

$$\hat{\mathbf{c}}_i = \arg \min_{\mathbf{c} \in \mathbb{R}^m} \sum_{j=1}^p (y_{ij} - x_i(t_j))^2 + \lambda R(x_i), \quad (6.1)$$

où $x_i(t_j) = \sum_{k=1}^m c_{i,k} \phi_k(t_j)$ et $\lambda > 0$ est un hyperparamètre estimé par une procédure de validation croisée.

Il convient de mentionner qu'il existe d'autres approches pour représenter les DF dans un espace vectoriel de dimension finie. Une alternative à la méthode par projection sur une base de fonctions prédéfinie est l'**Analyse en Composantes Principales (ACP) fonctionnelle**. L'extension de l'ACP du cas vectoriel au cas fonctionnel repose sur le théorème de Karhunen-Loève. Dans ce cadre, on considère que chaque courbe est la réalisation d'un processus stochastique en temps continu qui appartient à $\mathbb{L}^2([0, T])$. La fonction de covariance du processus stochastique peut être vue comme une fonction noyau. Le théorème de Mercer [Mercer, 1909] indique que la décomposition spectrale de la fonction de covariance définit des fonctions propres qui forment une base orthogonale de $\mathbb{L}^2([0, T])$. Le théorème de Karhunen-Loève précise alors que tout processus stochastique en temps continu de $\mathbb{L}^2([0, T])$ peut être décomposé dans cette base de fonctions propres et que les coefficients sont des variables aléatoires centrées et non corrélées. On utilise alors les fonctions propres associées aux

valeurs propres les plus grandes et on peut alors projeter les courbes dans cette base tronquée. Je n'utiliserai pas l'ACP fonctionnelle par la suite mais le lecteur intéressé pourra consulter le survey suivant [Shang, 2014] sur les différentes approches mises en oeuvre dans ce contexte.

Une fois que les données discrètes ont été pré-traitées et que l'on dispose de formes fonctionnelles (approchées) pour les individus à l'étude, l'ADF consiste à étendre les techniques statistiques et de *machine learning* conçues initialement pour des données vectorielles, au cas de données fonctionnelles.

Clustering de données fonctionnelles

De nombreuses méthodes multivariées de *clustering* ont été adaptées ou étendues afin de traiter les DF. Des revues de ces approches peuvent être trouvées dans [Jacques and Preda, 2014, Hitchcock and Greenwood, 2015].

Nous nous sommes particulièrement concentrés sur l'**algorithme k-means fonctionnel**. Les travaux pionniers dans ce domaine incluent [Abraham et al., 2003] et [Tarpey and Kinateder, 2003]. Dans le premier, les DF sont projetées sur un ensemble de B-splines similaires à (6.1), tandis que le second emploie un modèle fondé sur les processus stochastiques et le concept de *principal points*. Dans [Chiou and Li, 2007], une procédure de transfert similaire à celle du k-means a été introduite. Dans ce cas, l'affectation de x_i dans un *cluster* est effectué en comparant les erreurs de reconstruction vis-à-vis des projections sur l'expansion tronquée de Karhunen-Loève associé à chaque *cluster*. Une analyse théorique du problème k-means dans les espaces de Hilbert est présentée dans [Biau et al., 2008]. De plus, dans [García et al., 2015], l'algorithme k-means est appliqué avec des DF représentées dans une base de fonctions appartenant à un *Reproducing Kernel Hilbert Space* (RKHS) associé à un noyau défini sur $[0, T] \times [0, T]$. Un autre article en lien avec notre thématique est [Floriello and Vitelli, 2017], où un k-means partitionne les courbes tandis qu'une fonction de poids définie sur $[0, T]$ est apprise pour sélectionner les sous-intervalles qui favorisent la variance.

Tous ces travaux de recherche appliquent les étapes de base de l'algorithme k-means, les principales différences concernent la représentation utilisée pour les DF. **Notre contribution diverge des approches précédentes en intégrant les fonctions dérivées dans la représentation des DF**. De plus, nous faisons l'hypothèse que les DF et leurs dérivées peuvent appartenir à des sous-espaces non linéaires distincts. Cela est lié à la *manifold hypothesis* déjà évoquée dans le Chapitre précédent page 109. Tout comme les données vectorielles, nous supposons que l'**utilisation de fonctions noyaux** permettant de projeter implicitement les DF dans un autre espace, peut être avantageuse. De plus, l'introduction de **poids non uniformes pour mélanger de manière optimale les informations provenant des fonctions dérivées de différents ordres**, représente un aspect novateur de notre travail.

Catégorisation de données fonctionnelles

D'un point de vue général, soit \mathbb{X} l'espace initial de description des données, \mathbb{C} l'ensemble discret et fini des catégories, $c_i \in \mathbb{C}$ représentant la classe de $x_i \in \mathbb{X}$, et $\{(x_i, c_i)\}_{i=1, \dots, n}$ l'ensemble d'entraînement. Dans la tâche de classification (ou catégorisation), l'objectif est d'apprendre à partir de $\{(x_i, c_i)\}_{i=1, \dots, n}$, une fonction $f : \mathbb{X} \rightarrow \mathbb{C}$ qui prédit correctement $c \in \mathbb{C}$ pour tout $x \in \mathbb{X}$.

Il existe de nombreuses techniques de classification. Dans ce Chapitre, nous nous concentrons sur les modèles paramétriques où la phase d'induction implique la sélection d'une instance optimale dans une classe de fonctions en minimisant une fonction de perte. Dans ce contexte, plusieurs méthodes classiques multivariées ont été étendues aux DF, comme l'Analyse Discriminante Linéaire [Hastie et al., 1995], l'Analyse Discriminante Quadratique [James and Hastie, 2001, Chamroukhi and Nguyen, 2019], la régression logistique et les modèles linéaires généralisés [James, 2002, Müller et al., 2005].

Les méthodes prédictives développées en *machine learning* ont également inspiré les chercheurs et les praticiens travaillant avec des DF. Par exemple, des approches de *random forest* fonctionnelles ont été introduites dans [Fan et al., 2010] et [Möller et al., 2016]. Les réseaux de neurones et les différents paradigmes d'apprentissage ensembliste sont d'autres techniques d'apprentissage automatique qui ont été explorées pour les DF, comme le montrent respectivement [Rossi et al., 2005, Rossi and Conan-Guez, 2006, Hsieh et al., 2021] et [Krämer, 2006, Fuchs et al., 2015].

Dans notre cas, nous nous intéressons spécifiquement aux **méthodes à noyaux**. Mentionnons tout d'abord [Preda, 2007] qui étudie le problème de catégorisation binaire et qui propose de projeter les fonctions dans un RKHS et d'utiliser une régression logistique pénalisée. Un autre travail particulièrement pertinent pour la suite est celui de Fabrice Rossi et Nathalie Vialaneix [Rossi and Villa, 2006], qui étend les *Support Vector Machines* (SVM) aux DF. Les auteurs mettent en avant la nature fonctionnelle des données en discutant des transformations adaptées dans ce contexte qui débouchent sur des fonctions noyaux spécifiques. De plus, [Rossi and Villa, 2006] établit les propriétés de convergence de l'algorithme SVM fonctionnel en adoptant la démarche introduite dans [Biau et al., 2005].

La méthode SVM fonctionnelle mentionnée ci-dessus sert de fondement à notre modèle de classification pour les DF avec dérivées. Notre contribution peut être vue comme une extension de [Rossi and Villa, 2006], où nous considérons explicitement les fonctions et leurs dérivées, et où nous appliquons l'apprentissage de noyaux multiples pour combiner les matrices de noyaux associées aux dérivées d'ordres distincts.

Utilisation des dérivées en analyse de données fonctionnelles

Pour classifier automatiquement ou catégoriser des DF, on peut exploiter **les fonctions dérivées qui apportent des informations discriminantes supplémentaires telles que les pentes ou les courbures**. Dans la communauté ADF, cela a été souligné en tout premier lieu par Philippe Besse et James Ramsay dans [Besse and Ramsay, 1986].

D'un point de vue plus conceptuel, les notions de semi-métriques découlant de l'utilisation des fonctions dérivées ont été mises en valeur par Frédéric Ferraty et Philippe Vieu dans [Ferraty and Vieu, 2003, Ferraty and Vieu, 2006]. Dans la communauté *data mining*, l'utilisation des fonctions dérivées à la place des courbes originales a été proposée pour la première fois par [Keogh and Pazzani, 2001]. Nous mentionnons également [Górecki and Łuczak, 2013] où les auteurs ont été pionniers dans l'utilisation d'une combinaison entre distances entre courbes et distances entre dérivées pour la classification des séries temporelles.

L'utilité des dérivées pour la reconnaissance de motifs an ADF a été démontrée empiriquement dans plusieurs travaux de recherche. En ce qui concerne le *clustering*, [Rossi et al., 2004, Ferraty and Vieu, 2006] montrent que les dérivées secondes peuvent être plus appropriées que les fonctions originales pour l'analyse de données spectrométriques. Dans le cas des courbes d'électrocardiogramme, [Ieva et al., 2013] montre qu'une **mesure de distance composite**, qui agrège distances entre courbes originales et distances entre dérivées premières, améliore les performances de l'algorithme *k-means*. De même, dans [Meng et al., 2018], les auteurs montrent que l'algorithme *k-means* performait mieux avec une distance composite utilisant les dérivées jusqu'à l'ordre deux. Les deux articles précédemment cités appliquent des poids uniformes lors de l'agrégation des mesures de distance. Au contraire, les travaux présentés dans [Villmann, 2007] mettent en avant **l'utilisation de poids non uniformes**. Cependant, la question de l'estimation des poids reste ouverte.

Dans le contexte des problèmes de classification, plusieurs recherches ont également promu l'utilisation de semi-métriques. Dans le contexte de la classification binaire, les auteurs de [Alonso et al., 2012] proposent un cadre basé sur l'Analyse Discriminante Linéaire. D'autres méthodes statistiques multivariées étendues aux DF ont également été examinées avec l'ajout de fonctions dérivées. Par exemple, dans [Ahmedou et al., 2016] le modèle linéaire général fonctionnel est examiné avec des fonctions dérivées incluses en tant que covariables. Concernant les techniques de *machine learning*, nous mentionnons également l'approche basée sur les plus proches voisins introduite dans [Fuchs et al., 2015] et [Rossi and Villa-Vialaneix, 2011] qui étudie d'un point de vue théorique, la tâche d'apprentissage supervisé dans l'espace de Sobolev $\mathbb{H}^q([0, T])$.

À notre connaissance, aucun travail antérieur en ADF n'a proposé un cadre de représentation fonctionnelle qui combine à la fois : une **perspective multi-vue basée sur les dérivées d'ordres variés**, une **utilisation des fonctions noyaux pour la projection des fonctions et des dérivées dans des RKHS**, et un **système de poids non uniforme pour l'agrégation des fonctions noyaux**.

6.2 Contributions

6.2.1 Représentation de données fonctionnelles avec dérivées

Tout comme dans [Rossi and Villa-Vialaneix, 2011], **nous considérons des fonctions de l'espace de Sobolev $\mathbb{H}^q([0, T])$** , c'est à dire l'ensemble des fonctions de $\mathbb{L}^2([0, T])$ dont

les dérivées (au sens faible) jusqu'à l'ordre q sont également des éléments de $\mathbb{L}^2([0, T])$:

$$\mathbb{H}^q([0, T]) = \{x \in \mathbb{L}^2([0, T]) : D^j x \in \mathbb{L}^2([0, T]), \forall j = 1, \dots, q\},$$

où D est l'opérateur différentiel usuel.

Puisque nous supposons que les dérivées jusqu'à l'ordre q sont dans \mathbb{L}^2 , nous travaillons avec le **sous-espace des fonctions engendrées par l'ensemble des B-splines d'ordre $d = q + 2$** , pour garantir un cadre suffisamment flexible pour représenter les DF et leurs dérivées. Par conséquent, le système de base a une dimension de $m = d + p$.

Étant donné l'échantillon $\{x_i\}_{i=1, \dots, n}$, les ensembles de fonctions dérivées jusqu'à l'ordre q sont respectivement désignés par $\{D^1 x_i\}_i, \{D^2 x_i\}_i, \dots, \{D^q x_i\}_i$. **Ces ensembles de fonctions sont interprétés comme des vues distinctes des mêmes objets.** Nous pouvons les utiliser de façon unitaire (mono-vue) ou de façon combinée (multi-vue).

Il est important de noter que, bien que nous supposions que les DF soient des éléments de \mathbb{H}^q , **nous ne nous limitons pas à la métrique de Sobolev classique** suivante :

$$\langle x_i, x_{i'} \rangle_{\mathbb{H}^q} = \sum_{s=0}^q \langle D^s x_i, D^s x_{i'} \rangle_{\mathbb{L}^2}, \quad \forall i, i' = 1, \dots, n,$$

où D^0 est l'identité.

Nous considérons une métrique pour chaque ordre $s = 0, 1, \dots, q$, en utilisant des fonctions noyaux (éventuellement) distinctes $k^s : \mathbb{L}^2 \times \mathbb{L}^2 \rightarrow \mathbb{R}$, à la place du produit scalaire usuel de \mathbb{L}^2 . Par conséquent, pour toute paire $(x_i, x_{i'})$, nous promouvons l'utilisation de :

$$k^s(D^s x_i, D^s x_{i'}), \quad \forall s = 0, \dots, q. \quad (6.2)$$

Dans ce cas, précisons que la notion usuelle d'espace de Sobolev n'est plus valide puisque les représentations des dérivées $\{D^s x_i\}_i$ avec $s > 0$ dans leurs RKHS respectifs associés à k^s , ne correspondent pas, en général, aux dérivées des représentations des fonctions $\{x_i\}_i$ dans son RKHS associé à k^0 .

De plus, nous favorisons l'**usage de poids non uniformes** et supposons ainsi que certains ensembles de dérivées sont plus discriminants que d'autres et doivent être davantage mis en avant. Par conséquent, pour toute paire de fonctions $(x_i, x_{i'})$, **nous adoptons la métrique générale suivante** :

$$k(x_i, x_{i'}) = \sum_{s=0}^q w_s k^s(D^s x_i, D^s x_{i'}), \quad \text{où } w_s \geq 0, \forall s = 0, \dots, q. \quad (6.3)$$

Pour tout $s = 0, \dots, q$, w_s est le poids non négatif attribué à l'information véhiculée par l'ensemble des dérivées d'ordre s , $D^s x_{ii}$.

Nous introduisons à présent, un résultat central de notre approche qui concerne l'**estimation de $\mathbf{w} = (w_s)_{s=1, \dots, q}$, le vecteur de poids des différentes vues.** Les résultats que nous

exposons ci-après ont une portée générale. Nous précisons ultérieurement comment ils s'appliquent dans le contexte de nos modèles de *clustering* et de catégorisation.

Supposons que l'on dispose d'un vecteur de valeurs non négatives $\mathbf{z} = (z_s)_{s=0,\dots,q}$ où z_s est le profit partiel de la vue s . Notre objectif est de déterminer \mathbf{w} qui maximise le profit global donné par la combinaison linéaire $\sum_{s=0}^q w_s z_s = \mathbf{w}^\top \mathbf{z}$. Comme ce problème n'est pas borné, nous ajoutons une contrainte sur \mathbf{w} . L'approche classique consiste à imposer une borne supérieure sur la norme ℓ_r de \mathbf{w} . Ici, on prendra $r > 1$ et une valeur de la borne égale à 1 ce qui donne la contrainte suivante : $\|\mathbf{w}\|_{\ell_r} = (\sum_{s=0}^q w_s^r)^{\frac{1}{r}} \leq 1$. Nous cherchons donc à résoudre le problème d'optimisation :

$$\max_{\mathbf{w} \in \mathbb{R}^{q+1}} \mathbf{w}^\top \mathbf{z} \quad \text{s.l.c.} \quad \begin{cases} \mathbf{w} \geq \mathbf{0}, \\ \|\mathbf{w}\|_{\ell_r} \leq 1, \end{cases} \quad (6.4)$$

où $\mathbf{w} \geq \mathbf{0}$ est un raccourci pour $w_s \geq 0, \forall s = 0, \dots, q$.

Le problème (6.4) est convexe et sa **solution en forme close** est donnée ci-dessous.

Propriété. 1. *Si $\mathbf{z} \geq \mathbf{0}$ et $r > 1$, la solution du Problème (6.4) est donnée par :*

$$w^*_s = \frac{z_s^{\frac{1}{r-1}}}{\left(\sum_{s'=0}^q z_{s'}^{\frac{r}{r-1}}\right)^{\frac{1}{r}}}, \quad \forall s = 0, \dots, q. \quad (6.5)$$

Il convient de mentionner que des problèmes d'optimisation proches du problème (6.4) ont été étudiés en *clustering* dans le contexte de l'approche *fuzzy c-means* [Bezdek, 1973, Bezdek et al., 1984] ou dans le cas du *k-means* à noyaux multiples pour des données multivariées [Tzortzis and Likas, 2012]. En ce qui concerne la catégorisation, notre approche est un cas particulier du modèle SVM à noyaux multiples présenté dans [Kloft et al., 2010, Kloft et al., 2011].

Malgré l'existence de contributions antérieures similaires, à notre connaissance, **l'application de la Proposition 1 pour agréger les informations provenant des fonctions et de leurs dérivées dans le contexte de l'ADF et des machines à noyau multiples est nouvelle.**

6.2.2 *k-means* à noyaux multiples pour données fonctionnelles avec dérivées

Nous proposons d'**étendre l'algorithme *k-means* à noyaux multiples du cas multivarié aux fonctions dans \mathbb{H}^q** , où chaque ensemble de dérivées d'ordre $s = 0, \dots, q$, est considéré comme une vue distincte. Plus formellement, le problème d'optimisation qui nous

intéresse est la minimisation de la variance *intra-cluster* suivante :

$$\begin{aligned} \min_{\mathbb{C}, \mathbf{w}} \frac{1}{n} \sum_{l=1}^k \frac{1}{2|\mathbb{C}_l|} \sum_{i: x_i \in \mathbb{C}_l} \sum_{i': x_{i'} \in \mathbb{C}_l} \sum_{s=0}^q w_s \|\psi^s(D^s x_i) - \psi^s(D^s x_{i'})\|_{\mathbb{F}^s}^2 \quad (6.6) \\ \text{s.l.c.} \quad \begin{cases} \mathbb{C} = \{\mathbb{C}_1, \dots, \mathbb{C}_k\} \text{ est une partition,} \\ \mathbf{w} \geq \mathbf{0}, \\ \|\mathbf{w}\|_{\ell_r} \leq 1, \end{cases} \end{aligned}$$

où pour tout $s = 0, \dots, q$, \mathbb{F}^s est le *feature space* associé au RKHS de noyau k^s dans lequel les fonctions $\{D^s x_i\}_i$ sont projetées au moyen de la *feature map* $\psi^s : \mathbb{L}^2 \rightarrow \mathbb{F}^s$.

En utilisant la **décomposition de la variance totale comme somme de la variance *intra-cluster* et de la variance *inter-cluster***, on peut exprimer de façon équivalente le problème (6.6) en fonction de la variance *inter-cluster*. On obtient le **problème de maximisation** suivant :

$$\begin{aligned} \max_{\mathbb{C}, \mathbf{w}} \sum_{s=0}^q w_s \left(\sum_{l=1}^k \frac{1}{n|\mathbb{C}_l|} \sum_{i: x_i \in \mathbb{C}_l} \sum_{i': x_{i'} \in \mathbb{C}_l} k_{ii'}^s - \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n k_{ii'}^s \right) \quad (6.7) \\ \text{s.l.c.} \quad \begin{cases} \mathbb{C} = \{\mathbb{C}_1, \dots, \mathbb{C}_k\} \text{ est une partition,} \\ \mathbf{w} \geq \mathbf{0}, \\ \|\mathbf{w}\|_{\ell_r} \leq 1. \end{cases} \end{aligned}$$

où pour toutes les vues $s = 0, \dots, q$, nous notons pour tous les couples $\{(x_i, x_{i'})\}_{i, i'=1, \dots, n}$, $\langle \psi^s(D^s x), \psi^s(D^s x') \rangle_{\mathbb{F}^s}$ par $k^s(D^s x_i, D^s x_{i'})$, et nous regroupons toutes ces valeurs dans la matrice de noyaux $\mathbf{K}^s = (k_{ii'}^s) = (k^s(D^s x_i, D^s x_{i'}))$.

Nous employons la **stratégie classique pour résoudre ce type de problèmes d'apprentissage de noyaux multiples**, qui consiste à alterner entre (i) la maximisation par rapport à \mathbb{C} tout en maintenant \mathbf{w} fixé et (ii) la maximisation par rapport à \mathbf{w} tout en maintenant \mathbb{C} fixé. Dans le premier cas, un algorithme usuel de *k-means* à noyau est utilisé pour déterminer \mathbb{C} . Dans le deuxième cas, nous avons une solution analytique en utilisant les résultats exposés dans la sous-section précédente. Dans cette perspective, nous introduisons les quantités suivantes :

$$z_s = \sum_{l=1}^k \frac{1}{n|\mathbb{C}_l|} \sum_{i: x_i \in \mathbb{C}_l} \sum_{i': x_{i'} \in \mathbb{C}_l} k_{ii'}^s - \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n k_{ii'}^s, \quad \forall s = 0, \dots, q. \quad (6.8)$$

Notons que $z_s \geq 0, \forall s = 0, \dots, q$, car ces valeurs correspondent à des mesures de variance *inter-clusters*. Ensuite, grâce à la Proposition 1, nous avons le résultat suivant.

Corollaire. 2. Soit $\mathbb{C} = \{\mathbb{C}_1, \dots, \mathbb{C}_k\}$ fixé et $r > 1$, alors le problème d'optimisation suivant :

$$\begin{aligned} \max_{\mathbf{w}} \sum_{s=0}^q w_s \left(\sum_{l=1}^k \frac{1}{n|\mathbb{C}_l|} \sum_{i:x_i \in \mathbb{C}_l} \sum_{i':x_{i'} \in \mathbb{C}_l} k_{ii'}^s - \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n k_{ii'}^s \right) \\ \text{s.l.c.} \quad \begin{cases} \mathbf{w} \geq \mathbf{0}, \\ \|\mathbf{w}\|_{\ell_r} \leq 1, \end{cases} \end{aligned} \quad (6.9)$$

est convexe et sa solution optimale est donnée par, $\forall s = 0, \dots, q$:

$$w_s^* = \frac{\left(\sum_{l=1}^k \frac{1}{n|\mathbb{C}_l|} \sum_{i:x_i \in \mathbb{C}_l} \sum_{i':x_{i'} \in \mathbb{C}_l} k_{ii'}^s - \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n k_{ii'}^s \right)^{\frac{1}{r-1}}}{\left(\sum_{s'=0}^q \left(\sum_{l=1}^k \frac{1}{n|\mathbb{C}_l|} \sum_{i:x_i \in \mathbb{C}_l} \sum_{i':x_{i'} \in \mathbb{C}_l} \mathbf{K}_{ii'}^{s'} - \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \mathbf{K}_{ii'}^{s'} \right)^{\frac{r}{r-1}} \right)^{\frac{1}{r}}}. \quad (6.10)$$

Nous désignons notre méthode par **MK-KM-FD** pour *Multiple Kernel k-means for Functions with Derivatives*. La procédure est résumée dans Algorithme 1.

Algorithme 1 : Multiple kernel k-means for functions with derivatives (MK-KM-FD).

Input : $\{y_{ij}\}_{i=1, \dots, n; j=1, \dots, p}$ (sampled values of FD), $q \geq 0$ (maximum order of derivative), $r > 1$ (ℓ_r norm, default 2), $\{k^s\}_{s=0, \dots, q}$ (kernel functions, default Gaussian), σ (kernel hyper-parameter if any, default 1), $k \geq 2$ (number of clusters)

Output : \mathbb{C} (partition of FD), \mathbf{w} (weight vector of size $q + 1$)

- 1 Project the sampled FD onto a pre-defined set of $q + 2 + p$ B-splines of order $q + 2$ and determine $\{x_i\}_{i=1, \dots, n}$ by solving (6.1);
- 2 Determine $\{D^s x_i\}_{i=1, \dots, n}, \forall s = 1, \dots, q$;
- 3 Determine $\{\mathbf{K}^s = (k^s(D^s x_i, D^s x_{i'}))_{i, i'=1, \dots, n}\}, \forall s = 0, \dots, q$;
- 4 Normalize the kernel matrices $\mathbf{K}^s, \forall s = 0, \dots, q$ (optional);
- 5 Initialize a uniform weight vector \mathbf{w} ;
- 6 **while** Stopping condition not reached **do**
- 7 Fix \mathbf{w} and apply the kernel k-means algorithm with multiple kernel $\mathbf{K} = \sum_{s=0}^q w_s \mathbf{K}^s$ to determine a new \mathbb{C} (if applicable, use the previous \mathbb{C} as for initialization);
- 8 Fix \mathbb{C} and apply Corollary 2 to determine a new \mathbf{w} ;
- 9 **end**

Puisque la procédure alternée décrite dans Algorithme 1 améliore la fonction objective du Problème (6.7) à chaque itération, elle converge donc vers un **optimum local**.

Nous illustrons l'intérêt de l'approche avec des **données simulées** qui consistent en deux groupes de fonctions Gaussiennes dont les paramètres sont bruités. Je renvoie le lecteur au *preprint* [Ah-Pine and Yao, 2024] pour plus les détails sur les données artificielles. Le jeu de données montré dans la Figure 6.1 est composé de 1000 courbes ainsi que leurs dérivées

premières et secondes. Chaque groupe comporte 500 cas et nous avons également représenté en rouge les courbes moyennes. Notons, que le nombre de points stationnaires pour $\{x_i\}_i$, $\{Dx_i\}_i$ et $\{D^2x_i\}_i$ sont respectivement au nombre de 1, 2 et 3 et qu'autour de ceux-ci les distances entre les courbes des deux groupes sont plus marquées. Par conséquent, ce jeu de données indique qu'il y aurait un intérêt à utiliser les dérivées dans l'analyse.

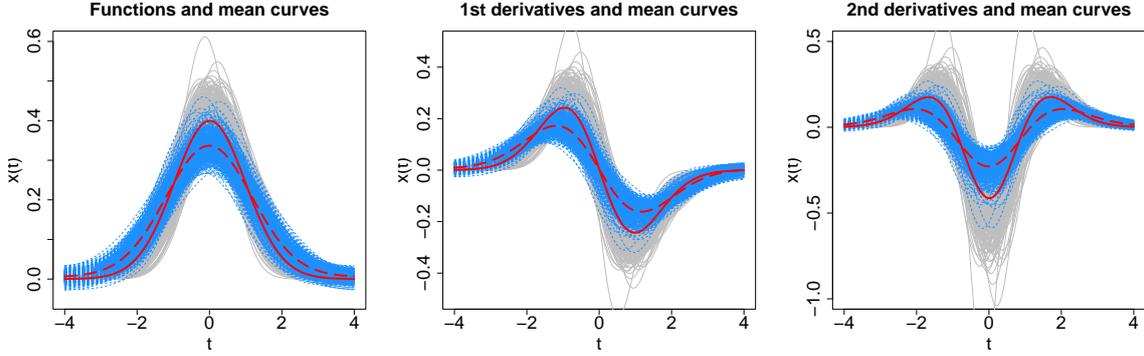


FIGURE 6.1 – De gauche à droite : fonctions initiales, dérivées premières, dérivées secondes. Les fonctions du groupe 1 sont en traits continus gris, les fonctions du groupe 2 sont en traits discontinus bleus. Les fonctions moyennes des groupes sont en rouge : traits continus pour le groupe 1, trait discontinu pour le groupe 2.

Nous avons testé plusieurs représentations des DF utilisant deux types de fonctions noyaux ($\kappa \in \{l, g\}$, avec l pour linéaire et g pour gaussien) et des ensembles variés de dérivées ($s = 0, 1, 2$) avec des approches mono-vues et multi-vues. Nous listons l'ensemble des représentations et leurs acronymes ci-après :

- $\kappa 0 : \mathbf{K}^{\kappa 0} = (\langle \psi^{\kappa 0}(x_i), \psi^{\kappa 0}(x_{i'}) \rangle)_{\mathbb{L}_2}$,
- $\kappa 1 : \mathbf{K}^{\kappa 1} = (\langle \psi^{\kappa 1}(Dx_i), \psi^{\kappa 1}(Dx_{i'}) \rangle)_{\mathbb{L}_2}$,
- $\kappa 2 : \mathbf{K}^{\kappa 2} = (\langle \psi^{\kappa 2}(D^2x_i), \psi^{\kappa 2}(D^2x_{i'}) \rangle)_{\mathbb{L}_2}$,
- $\kappa 01 : \mathbf{K}^{\kappa 01} = \mathbf{K}^{\kappa 0} + \mathbf{K}^{\kappa 1}$,
- $\kappa 012 : \mathbf{K}^{\kappa 012} = \mathbf{K}^{\kappa 0} + \mathbf{K}^{\kappa 1} + \mathbf{K}^{\kappa 2}$.
- $\kappa 01o : \mathbf{K}^{\kappa 01o} = w_0 \mathbf{K}^{\kappa 0} + w_1 \mathbf{K}^{\kappa 1}$,
- $\kappa 012o : \mathbf{K}^{\kappa 012o} = w_0 \mathbf{K}^{\kappa 0} + w_1 \mathbf{K}^{\kappa 1} + w_2 \mathbf{K}^{\kappa 2}$.

Afin d'avoir des résultats d'expériences robustes, nous avons généré 50 échantillons de 1000 courbes. Pour chacun de ces échantillons, nous avons appliqué MK-KM-FD successivement avec les représentations listées ci-dessus. Nous avons comparé les résultats de partitionnement en deux *clusters* donnés par MK-KM-FD avec la vérité terrain et utilisé pour cela la mesure *Normalized Mutual Information* (NMI). Les variations des scores de NMI au sein des 50 échantillons pour chaque cas sont illustrés par des *box plot* dans la Figure 6.4.

Nous avons effectué des tests de Student appariés pour vérifier la significativité statistique des différences entre les valeurs moyennes (triangles rouges). Nous avons trouvé que **le noyau**

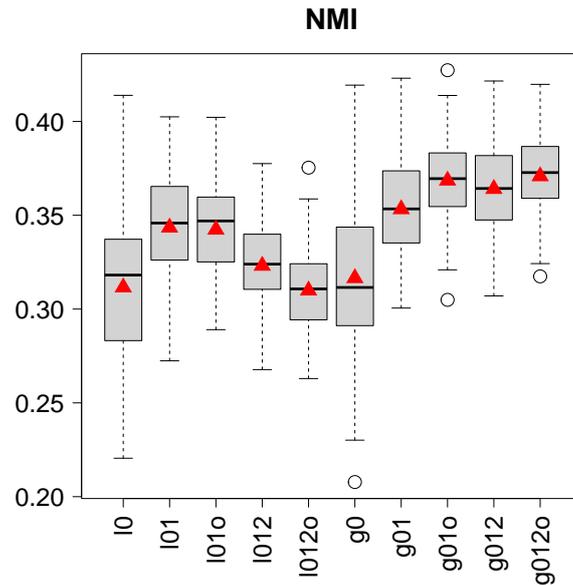


FIGURE 6.2 – *Box plot* des scores NMI mesurés sur 50 échantillons en utilisant MK-KM-FD avec différentes représentations des DF. Les triangles rouges indiquent les moyennes des 50 valeurs.

Gaussien donnait globalement de meilleurs résultats que le noyau linéaire ce qui indique l'**intérêt des fonctions noyaux**. Pour les noyaux Gaussiens, nous observons que l'**ajout des dérivées premières et des dérivées secondes améliore les performances**. Enfin, l'**optimisation des poids permet aussi d'augmenter de façon significative les mesures NMI de $g01$ et de $g012$** . Ainsi, les meilleurs résultats sont atteints par $g01o$ et $g012o$. Cependant, il n'y a pas de différence significative entre ceux deux représentations.

Nous avons testé MK-KM-FD sur des **jeux de données réelles** avec le noyau Gaussien. Étant donné que la procédure *k-means* nécessite une initialisation aléatoire et que les sorties varient selon les cas, pour chaque représentation, nous avons employé MK-KM-FD à 10 reprises et mesuré à chaque fois la valeur NMI. Nous présentons dans la Figure 6.3 les résultats obtenus.

Les performances des approches mono-vues $g0$, $g1$ et $g2$ varient selon les jeux de données : les dérivées peuvent donner des résultats bien meilleurs que les fonctions initiales mais il y a en générale une **incertitude sur l'ordre de dérivation des fonctions dérivées qui donnera la mesure d'évaluation le plus grande**. Nos résultats d'expérience montrent que **les représentations multi-vues $g01$ et $g012$ sont des stratégies averses au risque** : elles peuvent battre la meilleure représentation mono-vue mais lorsque ce n'est pas le cas, elles produisent des performances bien au-dessus de la pire représentation mono-vue.

Par ailleurs, si nous notons par \gtrsim la relation "significativement meilleure que", alors dans la grande majorité des cas nous avons $g01o \gtrsim g01$ et $g012o \gtrsim g012$. Ce constat nous permet de valider l'**intérêt de l'optimisation des poids**.

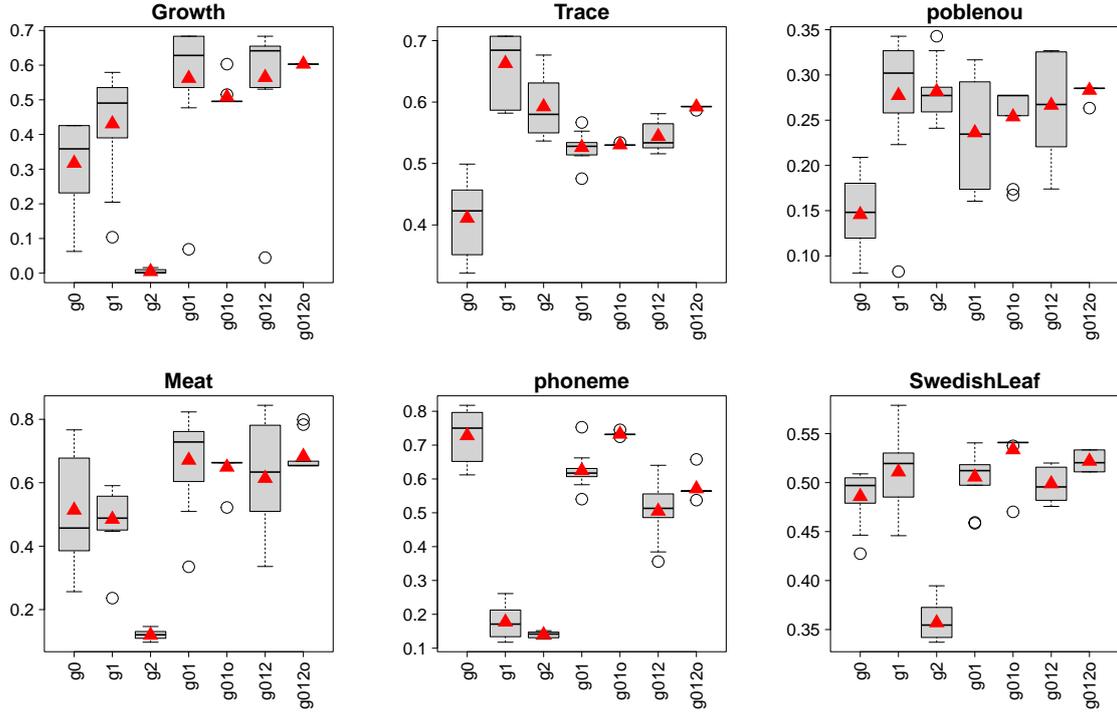


FIGURE 6.3 – Box plots des 10 valeurs de NMI de MK-KM-FD utilisant 10 initialisations différentes pour le *kernel* *k*-means. Les triangles rouges indiquent les moyennes de ces 10 valeurs.

Nos expériences illustrent également une caractéristique intéressante de l’**optimisation des poids** : elle permet à MK-KM-FD d’être **moins sensible à l’initialisation aléatoire de l’algorithme des *k*-means**.

6.2.3 SVM à noyaux multiples pour données fonctionnelles avec dérivées

Rappelons tout d’abord le modèle SVM fonctionnel introduit par Fabrice Rossi et Nathalie Vialaneix dans [Rossi and Villa, 2006]. Dans cet article, les FD $\{x_i\}_i$ sont considérées comme des éléments de \mathbb{L}^2 . Étant donné un ensemble d’entraînement $\{(x_i, c_i)\}_{i=1, \dots, n}$, l’approche SVM pour les FD consiste à résoudre le problème d’optimisation convexe suivant (primal) :

$$\begin{aligned} \min_{a_0 \in \mathbb{R}, a \in \mathbb{L}^2} \quad & \frac{1}{2} \|a\|_{\mathbb{L}^2}^2 + \mu \sum_{i=1}^n \xi_i \\ \text{s.l.c.} \quad & \begin{cases} c_i (a_0 + \langle a, x_i \rangle_{\mathbb{L}^2}) \geq 1 - \xi_i, \forall i = 1, \dots, n; \\ \xi_i \geq 0, \forall i = 1, \dots, n; \end{cases} \end{aligned} \quad (6.11)$$

où $\mu \geq 0$ est un hyperparamètre contrôlant l’équilibre entre la *soft margin*, qui est inversement proportionnelle à $\|a\|_{\mathbb{L}^2}^2$, et la *soft error* $\sum_{i=1}^n \xi_i$.

La problème primal du SVM est équivalent au problème dual qui s'exprime comme suit :

$$\begin{aligned} & \max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} c_i c_{i'} \langle x_i, x_{i'} \rangle_{\mathbb{L}^2} \\ \text{s.l.c.} & \begin{cases} \sum_{i=1}^n \alpha_i c_i = 0; \\ 0 \leq \alpha_i \leq \mu, \forall i = 1, \dots, n. \end{cases} \end{aligned} \quad (6.12)$$

La dualité transforme le problème primal, dont la solution est dans l'espace fonctionnel \mathbb{L}^2 , en un problème dual avec un espace de recherche de dimension finie \mathbb{R}^n . De plus, le problème dual dépend uniquement des produits scalaires entre paires d'objets dans l'échantillon d'entraînement. Cette propriété permet de projeter implicitement les FD dans un RKHS en utilisant le *kernel trick*. Soit $\mathbf{K} = (k_{ii'})$ une matrice carrée d'ordre n dont le terme général $k_{ii'} = \langle \psi(x_i), \psi(x_{i'}) \rangle_{\mathbb{F}} = k(x_i, x_{i'})$, où \mathbb{F} est un RKHS avec une fonction noyau reproduisant k , et ψ une *feature map* qui lui est associée. Alors, l'approche SVM dans son expression duale peut être formulée comme suit :

$$\begin{aligned} & \max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} c_i c_{i'} k_{ii'} \\ \text{s.l.c.} & \begin{cases} \sum_{i=1}^n \alpha_i c_i = 0; \\ 0 \leq \alpha_i \leq \mu, \forall i = 1, \dots, n. \end{cases} \end{aligned} \quad (6.13)$$

Le modèle SVM fonctionnel précédent s'applique aux éléments de \mathbb{L}^2 . Nous définissons ci-après un **nouveau modèle SVM fonctionnel qui traite les DF avec dérivées de \mathbb{H}^q** . Nous proposons pour cela d'utiliser une **matrice de noyaux multiples \mathbf{K}** donnée par une **combinaison linéaire de matrices de noyaux unitaires $\mathbf{K}^s, s = 0, \dots, q$** :

$$\mathbf{K} = \sum_{s=0}^q w_s \mathbf{K}^s, \quad (6.14)$$

où $w_s \geq 0, \forall s = 0, \dots, q$, et pour tout couple $(x_i, x_{i'}) \in \mathbb{H}^q \times \mathbb{H}^q$ de l'ensemble d'entraînement, $k_{ii'}^s = \langle \psi^s(D^s x_i), \psi^s(D^s x_{i'}) \rangle_{\mathbb{F}^s} = k^s(D^s x_i, D^s x_{i'})$.

Comme dans le cas non supervisé, nous exploitons le fait que les fonctions dérivées offrent différentes descriptions des objets initiaux et utilisons le paradigme de l'apprentissage à noyaux multiples pour fusionner ces différentes sources d'informations. De plus, pour tout $s = 0, \dots, q$, nous projetons chaque ensemble $\{D^s x_i\}_i$ d'éléments de \mathbb{L}^2 vers un RKHS en utilisant implicitement les *feature maps* $\psi^s : \mathbb{L}^2 \rightarrow \mathbb{F}^s$ via les noyaux k^s . Cet aspect devient crucial lorsque les classes ne sont pas linéairement séparables.

Dans le cas supervisé également, nous supposons ici que **les matrices de noyaux $\{\mathbf{K}^s\}_s$ doivent se compléter plutôt que se concurrencer**. Par conséquent, nous sommes dans la lignée de l'approche générale étudiée dans [Kloft et al., 2009, Kloft et al., 2010] qui promeut la contrainte $\|\mathbf{w}\|_{\ell_r} \leq 1$ avec $r > 1$. Notre méthode peut être interprétée telle une extension

6.2. CONTRIBUTIONS

du modèle de [Kloft et al., 2009, Kloft et al., 2010] des vecteurs dans \mathbb{R}^p aux fonctions avec dérivées appartenant à \mathbb{H}^q . Nous cherchons à résoudre :

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^{q+1}} \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} & \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} c_i c_{i'} \sum_{s=0}^q w_s k_{ii'}^s \\ \text{s.l.c.} & \begin{cases} \sum_{i=1}^n \alpha_i c_i = 0; \\ 0 \leq \alpha_i \leq \mu, \forall i = 1, \dots, n; \\ \mathbf{w} \geq \mathbf{0}, \|\mathbf{w}\|_{\ell_r} \leq 1. \end{cases} \end{aligned} \quad (6.15)$$

La procédure d'optimisation est similaire au cas non supervisé et consiste à alterner entre (i) maximiser par rapport à $\boldsymbol{\alpha}$ avec \mathbf{w} fixé en utilisant l'algorithme SVM classique, et (ii) minimiser par rapport à \mathbf{w} avec $\boldsymbol{\alpha}$ fixé. Le deuxième problème a une solution analytique que l'on déduit de la Proposition 1. Considérons la maximisation en \mathbf{w} de l'opposé de la fonction objectif du Problème (6.15), et introduisons le vecteur $\mathbf{z} \in \mathbb{R}^{q+1}$ défini par :

$$z_s = \sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} c_i c_{i'} k_{ii'}^s, \quad \forall s = 0, \dots, q. \quad (6.16)$$

Comme pour tout $s = 0, \dots, q$, \mathbf{K}^s est semi-définie positive alors z_s est non négatif et nous avons le résultat suivant.

Corollaire. 3. *Soit $\boldsymbol{\alpha}$ fixé et $r > 1$, alors le problème d'optimisation suivant :*

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^{q+1}} & \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} c_i c_{i'} \sum_{s=0}^q w_s k_{ii'}^s \\ \text{s.l.c.} & \begin{cases} \mathbf{w} \geq \mathbf{0}; \\ \|\mathbf{w}\|_{\ell_r} \leq 1; \end{cases} \end{aligned} \quad (6.17)$$

est convexe et sa solution optimale est donnée par, $\forall s = 0, \dots, q$:

$$w_s^* = \frac{\left(\sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} c_i c_{i'} k_{ii'}^s \right)^{\frac{1}{r-1}}}{\left(\sum_{s'=0}^q \left(\sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} c_i c_{i'} k_{ii'}^{s'} \right)^{\frac{r}{r-1}} \right)^{\frac{1}{r}}}. \quad (6.18)$$

Dans Algorithme 2, nous donnons le pseudo-code de notre procédure **SVM à noyaux multiples pour fonctions avec dérivées** que nous désignons par **MK-SVM-FD**. Comme pour le *clustering*, la fonction objectif globale s'améliore à chaque itération, en conséquence Algorithme 2 converge vers un **optimum local**.

Nous illustrons les apports de notre modèle d'apprentissage supervisé avec les mêmes **données simulées** et les mêmes représentations que précédemment. Nous utilisons cette fois-ci le taux de reconnaissance **Accuracy** pour évaluer les différentes approches.

Plus précisément, nous générons aléatoirement 1200 courbes uniformément réparties entre le groupe 1 et le groupe 2. Nous retenons 200 cas pour l'ensemble de test. Les 1000 courbes

Algorithme 2 : Multiple kernel SVM for functions with derivatives (MK-SVM-FD).

Input : $\{y_{ij}\}_{i=1,\dots,n;j=1,\dots,p}$ (sampled values of FD), $q \geq 0$ (maximum order of derivative), $r > 1$ (ℓ_r norm, default 2), $\{k^s\}_{s=0,\dots,q}$ (kernel functions, default Gaussian), σ (kernel hyper-parameter if any)

Output : α (support vectors' weight), \mathbf{w} (weight vector of size $q + 1$)

- 1 Project the sampled FD onto a pre-defined set of $q + 2 + p$ B-splines of order $q + 2$ and determine $\{x_i\}_{i=1,\dots,n}$ by solving (6.1);
- 2 Determine $\{D^s x_i\}_{i=1,\dots,n}, \forall s = 1, \dots, q$;
- 3 Determine $\{\mathbf{K}^s = (k^s(D^s x_i, D^s x_{i'}))_{i,i'=1,\dots,n}\}, \forall s = 0, \dots, q$;
- 4 Normalize the kernel matrices $\mathbf{K}^s, \forall s = 0, \dots, q$ (optional);
- 5 Initialize a uniform weight vector \mathbf{w} ;
- 6 **while** Stopping condition not reached **do**
- 7 Fix \mathbf{w} and apply the SVM algorithm with multiple kernel $\mathbf{K} = \sum_{s=0}^q w_s \mathbf{K}^s$ to determine a new α ;
- 8 Fix α and apply Corollary 3 to determine a new \mathbf{w} ;
- 9 **end**

restantes sont utilisées pour apprendre le modèle et l'hyperparamètre de régularisation μ . Ce dernier est choisi parmi les valeurs $\{0.01, 0.1, 1, 10, 100\}$ suivant une approche *grid search* basée sur une validation croisée à 5 *folds*. Une fois déterminée μ^* , nous apprenons à nouveau le modèle avec cette valeur mais en utilisant l'ensemble des 1000 observations d'entraînement. Nous appliquons ensuite le modèle estimé sur l'ensemble de test et relevons le taux *Accuracy*. Nous répétons cette procédure 50 fois. Dans la Figure 6.4, nous illustrons à l'aide de *box plots*, la variabilité des scores *Accuracy* de test au sein des 50 échantillons et ceci pour chaque représentation de DF considérée.

Comme pour le non supervisé, **le noyau linéaire est globalement moins performant que le noyau Gaussien**. Toutefois, nous observons ici que *g01* et *g02* **ne donne pas de meilleurs résultats que g0**. Les apports de l'optimisation de poids sont également **mitigés**.

Nous avons ensuite testé MK-SVM-FD sur des **données réelles**, les mêmes que pour le *clustering*. Ces *benchmarks* étant de petites tailles, nous n'avons pas évalué les modèles sur un jeu de données test. Nous donnons la distribution des scores de la valeur μ ayant aboutit à la meilleure moyenne de l'*Accuracy* en suivant une procédure de validation croisée à 10 *folds*. La Figure 6.5 expose les résultats obtenus.

Du point de vue **mono-vue**, pour les six cas, **les fonctions originales ont mieux performé que les fonctions dérivées du premier ou du second ordre**.

Concernant le multi-vue, malgré des scores *Accuracy* de validation différents, dans la plupart des cas, **l'approche g01 a donné des performances similaires ou meilleures² que la meilleure vue unitaire g0**. Plus globalement, similairement à la tâche de *clustering*, **les représentations à noyaux multiples sont des stratégies *risk-averse*** pour

2. Différence des moyennes évaluée selon le test de Student appairé à un niveau de significativité de 5%.

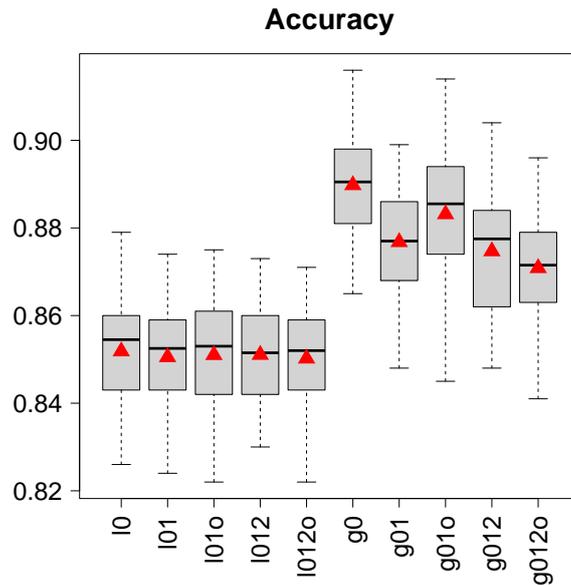


FIGURE 6.4 – *Box plot* des scores Accuracy mesurés sur 50 échantillons en utilisant MK-SVM-FD avec différentes représentations des DF. Les triangles rouges indiquent les moyennes des 50 valeurs.

les problèmes de classification : les scores d'évaluation des représentations multi-vues sont toujours nettement plus élevés que ceux de la représentation mono-vue donnant les moins bonnes performances.

Pour 8 des 12 cas, l'optimisation des poids des vues a permis d'augmenter légèrement ou maintenir les mesures moyennées des Accuracy de validation. Cependant, les différences ne sont généralement pas statistiquement significatives. En conséquence, **sur la base de ces benchmarks, nous ne pouvons pas conclure que l'optimisation des poids permet une amélioration des performances de classification.** *A minima*, la procédure n'a pas tendance à diminuer les scores. En effet, il n'y a qu'un seul cas où l'optimisation des poids était significativement contre-productive. Il s'agit de l'ensemble de données *Trace* pour lequel on a $g01 \gtrsim g010$.

6.3 Discussions et perspectives

L'idée centrale qui nous a guidé tout au long des travaux présentés précédemment est la **prise en compte systématique des dérivées dans l'analyse des DF**. Afin de définir un cadre mathématique riche pour la représentation des fonctions et de leurs dérivées, nous avons proposé l'utilisation de fonctions noyaux. L'agrégation de ces différentes sources d'information que nous supposons être complémentaires, s'effectue par combinaison linéaire. Dans cette perspective, nous avons également suggéré une méthode pour l'apprentissage automatique des coefficients de cette combinaison. Nous avons alors étendu les méthodes de *k-means* à noyaux

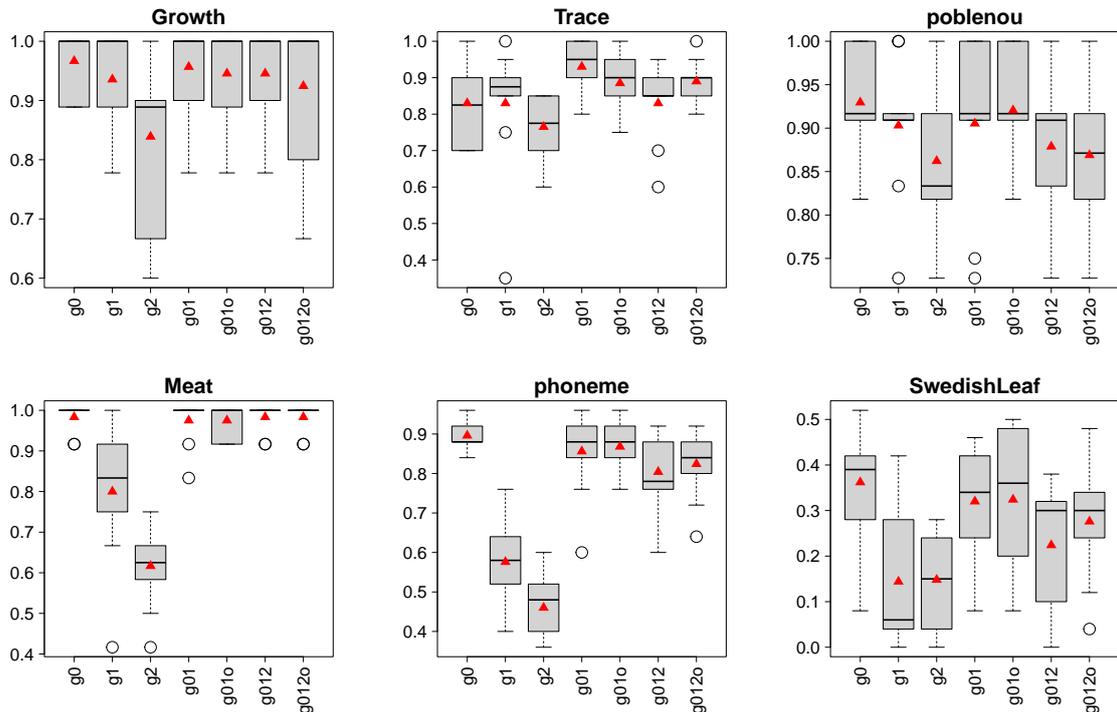


FIGURE 6.5 – Box plots des 10 valeurs d’Accuracy de validation (validation croisée à 10 *folde*). Les triangles rouges indiquent les moyennes des 10 valeurs.

multiples et de SVM à noyaux multiples au cas des fonctions avec dérivées. **Les méthodes MK-KM-FD et MK-SVM-FD ont des propriétés intéressantes.** Spécifiquement, si nous ne disposons pas d’information experte sur l’ordre de dérivation qui permettrait de résoudre la tâche d’apprentissage considérée de façon adéquate, alors nos méthodes donnent des stratégies judicieuses. Dans le cas où le noyau Gaussien est employé pour tous les ordres de dérivation $s = 0, 1, 2$; MK-KM-FD et MK-SVM-FD sont nettement meilleurs que la moins bonne des représentations mono-vues lorsqu’elles ne donnent pas les scores les plus performants.

L’approche multi-vue que nous avons développée avait pour but de combiner les fonctions avec leurs dérivées. Toutefois, **une autre application immédiate de MK-KM-FD et de MK-SVM-FD serait l’analyse de DF multivariées.** Considérons un objet X_i qui est composé de non pas une mais $l > 1$ courbes (x_i^1, \dots, x_i^l) . Chaque dimension $u = 1, \dots, l$ peut être considérée comme une vue unitaire. On peut alors associer une fonction noyau k^u à chaque vue $u = 1, \dots, l$ et employer MK-KM-FD et MK-SVM-FD pour segmenter ou catégoriser un ensemble d’objets. Supposons de plus que pour tout $u = 1, \dots, l$, x_i^u est un élément de \mathbb{H}^q . La représentation de X_i pourrait être étendue en ajoutant les fonctions dérivées des courbes de chaque dimension : $Dx_i^u, \dots, D^q x_i^u$. On aurait alors comme données les ensembles $\{x_i^u\}_i, \{Dx_i^u\}_i, \dots, \{D^q x_i^u\}_i$, avec $i = 1, \dots, n, u = 1, \dots, l$. On pourrait ainsi

utiliser jusqu'à $l \times (q+1)$ fonctions noyaux $k^{u,s}$ avec $u = 1, \dots, l$ et $s = 0, \dots, q$. Les méthodes MK-KM-FD et MK-SVM-FD pourraient être utilisées dans ce contexte également.

Une autre extension de nos modèles que je trouve pertinente serait d'estimer non pas un poids global pour chaque vue, mais d'**inférer une fonction de poids sur le domaine des courbes pour chaque vue**. Dans ce contexte, la focale serait mise sur le **noyau linéaire**. Cette extension s'inspirerait de la méthode *sparse functional clustering* [Floriello and Vitelli, 2017] dans le cas non supervisé, et du modèle *interpretable SVM* [Martin-Barragan et al. dans le cas supervisé. Ces approches permettent d'identifier les régions du domaine les plus importantes pour la reconnaissance de formes. Dans le cas du *clustering* et du *k-means* il s'agit d'estimer une fonction sur $[0, T]$ qui donne plus de poids aux sous-ensembles de valeurs contribuant fortement à la variance. Concernant la catégorisation et le SVM, la fonction objectif étant la *hinge loss*, la fonction de poids vise à préciser les régions de $[0, T]$ permettant de mieux séparer les deux classes. Dans l'une ou l'autre des deux applications, cette fonction de poids permettrait d'**interpréter les règle de décision sous-jacentes aux modèles estimés**. Cette piste de travaux futurs rejoint le **domaine XAI** que j'ai déjà évoqué aux Chapitres 2 et 3.

La méthode MK-SVM-FD s'appuie fortement sur l'article [Rossi and Villa, 2006] de Fabrice Rossi et Nathalie Vialaneix. Dans les années 2000, Fabrice Rossi et ses collègues ont étudié l'extension de diverses autres méthodes de *machine learning* au cas des DF. Dans [Rossi et al., 2005] notamment, les auteurs s'intéressent aux réseaux de neurones. Ils étudient plus spécifiquement les **Radial Basis Function Neural Networks (RBF NN)**, déjà évoqués en sous-section 1.3 d'une part, et les **Multi-Layer Perceptron (MLP)** d'autre part. L'extension de ces approches (et d'autres) du cas vectoriel au cas fonctionnel, peut se faire d'une manière simple en remplaçant le produit scalaire canonique de \mathbb{R}^p par le produit scalaire dans $\mathbb{L}^2([0, T])$. Toutefois, nous avons rappelé en sous-section 6.1.2, qu'en pratique, on ne disposait que d'un ensemble d'observations discrètes des $\{x_i\}_i$. Il est donc au préalable nécessaire de reconstruire de façon approchée la forme fonctionnelle des $\{x_i\}_i$. La procédure classique de *spline smoothing* que nous avons rappelée peut être utilisée à cet effet. Dans ce cas, les DF sont projetées dans un espace de dimension m finie et on se ramène à un produit scalaire dans un espace Euclidien. Toutefois, si la base de fonctions $\{\phi_k\}_{k=1, \dots, m}$ n'est pas orthogonale, comme cela est le cas pour les B-splines, alors le produit scalaire précédent n'est pas canonique mais dépend d'une métrique donnée par la matrice de Gram $\Phi = (\langle \phi_k, \phi_{k'} \rangle_{\mathbb{L}^2})$ d'ordre m . En revanche, il est toujours possible de déterminer la décomposition de Cholesky d'une matrice de Gram et dans ce cas, la factorisation permet de faire un changement de bases et de se ramener à un produit scalaire canonique. Ainsi, [Rossi et al., 2005] montre que des pré-traitements adéquats permettent d'adapter simplement les RBF NN et les MLP à des DF.

Néanmoins, même si du point de vue formel on se ramène à un produit scalaire Euclidien, il ne faut pas perdre de vue le fait que nos vecteurs vivent dans un espace engendré par

une base de fonctions. Prenons le cas du perceptron dans le cas classique. Celui-ci prend un vecteur $\mathbf{x} \in \mathbb{R}^p$ en entrée, combine \mathbf{x} linéairement par un vecteur de poids $\mathbf{a} \in \mathbb{R}^p$, et modifie le signal post-synaptique $\langle \mathbf{x}, \mathbf{a} \rangle_{\mathbb{R}^p} + b$, où $b \in \mathbb{R}$, par une fonction d'activation h . La sortie du perceptron dans le cas vectoriel est donc :

$$h(\langle \mathbf{x}, \mathbf{a} \rangle_{\mathbb{R}^p} + b) = h\left(\sum_{j=1}^p x_j a_j + b\right).$$

Dans [Rossi et al., 2005], les auteurs introduisent la notion de **perceptron fonctionnel** dans l'espace \mathbb{L}^2 . Dans ce cas, le perceptron prend une fonction $x \in \mathbb{L}^2$ en entrée, transforme linéairement x par une fonction de poids $a \in \mathbb{L}^2$, et modifie le signal post-synaptique $\langle \mathbf{x}, \mathbf{a} \rangle_{\mathbb{L}^2} + b$, où $b \in \mathbb{R}$, par une fonction d'activation h . La sortie du perceptron dans le cas fonctionnel est alors :

$$\begin{aligned} h(\langle \mathbf{x}, \mathbf{a} \rangle_{\mathbb{L}^2} + b) &= h\left(\int_0^T x(t)a(t)dt + b\right) \\ &= h(\mathbf{c}^\top \mathbf{\Phi} \mathbf{d} + b). \end{aligned}$$

où $x(t) = \sum_{k=1}^m c_k \phi_k(t) = [\mathbf{c}^\top \boldsymbol{\phi}](t)$; $a(t) = \sum_{k=1}^m d_k \phi_k(t) = [\mathbf{d}^\top \boldsymbol{\phi}](t)$ et $\mathbf{\Phi} = (\langle \phi_k, \phi_{k'} \rangle_{\mathbb{L}^2})$.

Les auteurs de [Rossi et al., 2005] soulignent également la possibilité d'utiliser des dérivées à la place des fonctions initiales ce qui permettrait, par exemple, d'axer la comparaison des fonctions vis-à-vis de leurs formes/silhouettes. Ils précisent que le choix de l'ordre de dérivation à employer pourrait être fait avec l'aide d'un expert métier. Dans la lignée de ce que nous avons proposé précédemment, et en supposant que nous ne disposons pas d'oracle, il serait intéressant d'**exploiter, au sein d'un RBF NN ou MLP fonctionnels, les dérivées de plusieurs ordres de façon collective.**

Ma proposition est motivée par les arguments suivants. Premièrement, dans le cas spécifique du MLP, le modèle fait l'apprentissage d'une représentation des DF à un niveau de granularité plus fin qu'un SVM fonctionnel. Deuxièmement, dans le cas du MLP fonctionnel, la fonction de poids de chaque vue peut être apprise par la procédure de rétro-propagation de l'erreur.

J'illustre cette proposition dans la Figure 6.6 par un réseau de neurones à une couche cachée de nature fonctionnelle et dans le cas d'un problème de classification supervisée à q classes. Il s'agit donc d'un MLP qui prend $q+1$ fonctions $x, Dx, \dots, D^q x$ en entrée, transforme linéairement chaque fonction $D^s x, s = 0, \dots, q$ par un perceptron fonctionnel dont la fonction de poids est notée a_s^1 et modifie chaque signal post-synaptique par la fonction d'activation h^1 . A l'issue de la couche cachée fonctionnelle, la deuxième couche (couche de sortie) du réseau de neurones prend un vecteur $\mathbf{z}^1 = (z_s^1)_{s=1, \dots, q}$ en entrée, transforme \mathbf{z}^1 par une combinaison linéaire pour chaque perceptron de sortie $g_l, l = 1, \dots, q$, dont le vecteur de coefficients est noté \mathbf{a}_l^2 , et modifie chaque signal post-synaptique par la fonction d'activation h^2 . La sortie de ce MLP est un vecteur $\mathbf{g} = \mathbf{z}^2$ dont chaque élément g_l a appris à combiner l'information pré-traitée par la couche fonctionnelle cachée pour chaque fonction d'ordre de dérivation $s = 0, \dots, q$ afin de prédire correctement la classe l . De façon générale, il me semble très

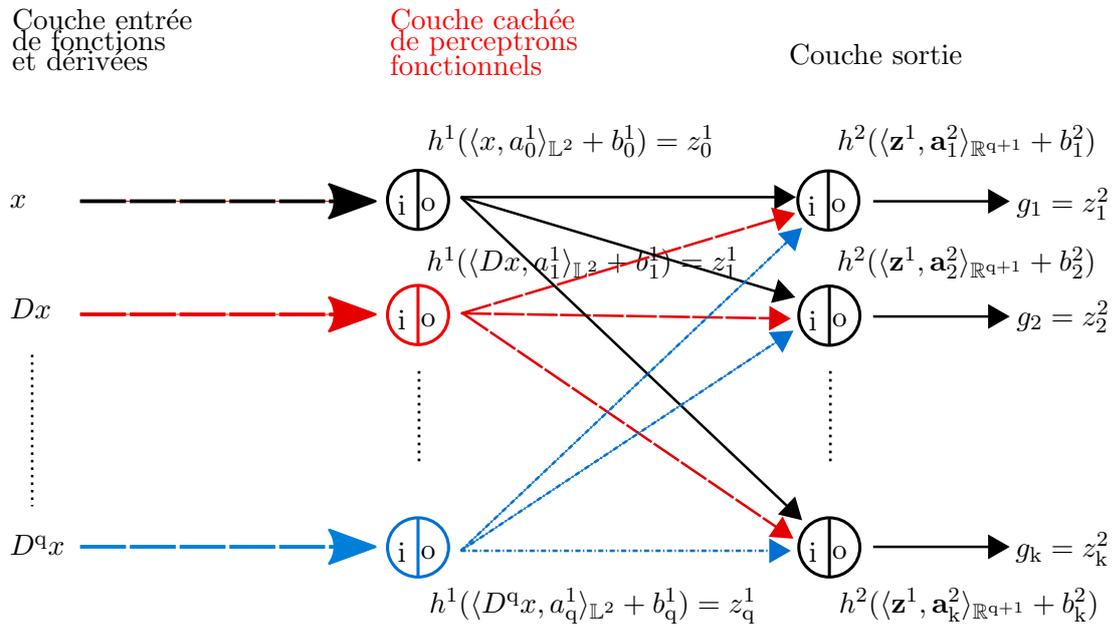


FIGURE 6.6 – Une architecture neuronale MLP avec couche cachée fonctionnelle et apprentissage de fusion entre DF et dérivées dans le cas de la catégorisation à k classes.

intéressant de développer les architectures avancées de *deep learning* (MLP, auto-encodeurs, ...) pour l'analyse de DF et dans cette perspective, le perceptron fonctionnel jouerait un rôle central.



Conclusion générale, travaux actuels et futurs

Sommaire du chapitre

7.1 Conclusion générale	159
7.2 Travaux actuels et futurs	160

7.1 Conclusion générale

Tout au long de ce manuscrit, j'ai décrit les activités scientifiques que j'ai menées au cours de ces seize dernières années. J'ai systématiquement organisé les Chapitres en distinguant une partie sur les travaux antérieurs et une partie sur mes contributions afin de favoriser la bonne compréhension de mes apports. Dans chacun des six thèmes développés, j'ai conclu par une discussion et des pistes d'extension qui me paraissent intéressantes.

En 2024, les méthodes de *deep learning* sont incontournables dans les domaines de la science des données, de l'apprentissage automatique et de l'intelligence artificielle (IA). Les architectures neuronales sont très flexibles et permettent de modéliser des classes de fonction qui peuvent s'adapter à **tout type de données** et tenir compte de leurs **propriétés spécifiques**. La descente de gradient stochastique, les graphes computationnels et l'auto-différentiation forment un écosystème méthodologique et technologique efficace pour l'inférence des paramètres des modèles et la résolution **de tâches diverses en apprentissage automatique**. En raison de l'ensemble de ces arguments, il m'a paru important que je discute et formalise dans la plupart des Chapitres, des liens entre mes travaux et les réseaux de neurones ou des extensions de mes sujets en incluant des modèles neuronaux.

Toutefois, les modèles d'apprentissage profond ou de représentation ont des limites et des inconvénients et il est important de continuer à développer soit des méthodes différentes, soit des architectures plus simples. En effet, les méthodes de *deep learning* atteignent de très bonnes performances à condition que les données soient massives, ce qui n'est pas le cas dans de très nombreux contextes. De plus, l'apprentissage de ces modèles est particulièrement complexe et coûteux en termes d'énergie, ce qui contribue à polluer la planète. Enfin, en raison

de leur complexité, les prédictions données par des réseaux de neurones profonds ne sont pas interprétables, ce qui est un inconvénient majeur dans de nombreuses applications.

Ainsi, dans mes activités de recherche à venir, je souhaite poursuivre les perspectives que j’ai indiquées au cours de ce mémoire en intégrant à bon escient les réseaux de neurones.

7.2 Travaux actuels et futurs

Depuis septembre 2023, je suis MCF à l’Université Clermont Auvergne/Clermont Auvergne INP/SIGMA Clermont et au LIMOS. J’ai été recruté sur ce poste par la voie de la mutation pour rapprochement familial. Cette mutation est un soulagement pour ma famille et moi-même. Comme je l’ai indiqué dans l’introduction du Chapitre 6, j’ai commencé à effectuer des allers-retours entre Clermont-Ferrand et Lyon à partir de janvier 2018. Cette situation a eu un impact sur ma vie professionnelle. Au cours de cette période, par manque de disponibilité en présentiel, j’ai préféré refuser des sollicitations de co-encadrement de thèse et de participation à des projets et ceci a eu des conséquences sur ma production scientifique.

Cette époque est désormais révolue. J’ai de nouveaux projets de recherche qui démarrent et je m’ouvre (à nouveau!) à de nouvelles thématiques de recherche.

Je co-encadre actuellement trois thèses :

- Mohamed Abdillahi Isman (avec Anne-Françoise Yao Université Clermont Auvergne (UCA)/Laboratoire de Mathématiques Blaise Pascal (LMBP), et Paul-Marie Grollemund MCF UCA/LMBP) sur le sujet “Modélisation et prévision par séries temporelles de la consommation d’électricité. Application au cas de Djibouti.”. Thèse Campus France, soutenance prévue pour l’automne 2024.
- Noé Lebreton (avec Julien Jacques Université Lumière Lyon 2/Laboratoire ERIC) sur le sujet “Modélisation prédictive ensembliste à l’aide d’approches fonctionnelles”. Thèse CIFRE débutée en décembre 2022.
- Nicolas Rojas Varela (avec Engelbert Mephu Nguifo UCA/LIMOS) sur le sujet “Étude et conception de modèles interprétables de détection d’anomalies dans les flux de données spectrales. Application à la contamination des chambres de procédés sous vide pour la microélectronique”. Thèse CIFRE débutée en mars 2024.

Ces doctorats portent tous les trois sur des séries temporelles. Dans le travail de thèse de Mohamed, nous utilisons des techniques de *machine learning* et *deep learning*. Dans celui de Noé, nous privilégions l’analyse de données fonctionnelles et les approches statistiques basées sur les plus proches voisins avec diverses métriques et les processus Gaussiens. Le doctorat de Nicolas vient tout juste de débiter, la tâche centrale est la détection d’anomalies dans des flux de données avec des méthodes interprétables. Plusieurs pistes sont envisagées en *data mining* (*matrix profiles*, *shapelet*, ...); en statistiques (modèles Gaussiens, ...); et en apprentissage profond (*auto-encoder*, ...).

Un premier axe de mes activités scientifiques actuelles se situe donc en analyse de séries temporelles. Plus généralement, je souhaite continuer des travaux en analyse de données fonctionnelles à la suite, entre autre, des recherches et perspectives décrites dans les Chapitres 6 et 5. Une thématique qui m'intéresse particulièrement est celle dénommée *physics informed machine learning* où il est question d'intégrer des équations différentielles dans l'apprentissage de modèles afin de mieux spécifier la classe d'hypothèses et de mieux interpréter les résultats. Je souhaite dans ce cas, investiguer des approches à l'intersection de l'analyse de données fonctionnelles et de l'apprentissage profond. Du point de vue des applications, j'envisage d'inscrire mes recherches en **Industrie 4.0** et plus particulièrement en pronostic de défaillance et maintenance prédictive (*Prognostics and Health Management*) comme cela est déjà le cas avec la thèse de Nicolas.

De plus, j'ai décroché en janvier 2024 un financement de l'I-Site CAP 20-25 de Clermont-Ferrand pour un postdoctorat de deux ans, dans le cadre du projet DLISCES (*Deep Learning, Images Satellitaires, et Cartographies d'Indicateurs Économiques et Sociaux*) qui débutera à la rentrée prochaine et dont je suis le *leader*. Il s'agit d'un projet en étroite collaboration avec des collègues économistes du CERDI (Centre d'Études et de Recherches sur le Développement International) de l'UCA. Il implique également des collègues du laboratoire Clermont Recherche Management - Recherche en Management durable (CLERMA) et du LMBP de l'UCA. L'objectif du projet DLISCES est de mettre en place des approches de *deep learning* et de *machine learning* pour analyser des données ouvertes (images satellites, données d'enquêtes, données topographiques et environnementales), afin d'inférer des cartes d'indicateurs de vulnérabilité, dans le but de réduire les risques de catastrophe dans des territoires exposés à des aléas climatiques.

De façon plus générale, je souhaite inscrire mes activités scientifiques dans le domaine **AI for Good**. Dans cette perspective, j'envisage d'élargir mes collaborations avec les **pays du Sud**. Le premier terrain d'expérimentation prévu par le projet DLISCES concerne la ville d'Arequipa au Pérou avec également un objectif d'établir des collaborations locales. Par ailleurs, je développe actuellement des activités de recherche en *data science* avec des membres de l'École Nationale supérieure de Statistiques et d'Économie Appliquée (ENSEA) de la Côte d'Ivoire. Je mène aussi des travaux sur l'analyse des Objectifs de Développement Durable (ODD) avec des approches systémiques où les graphes et les techniques que j'ai détaillées dans les Chapitres 4, 5 et 1, peuvent être mises à contribution. Enfin, un autre axe de recherche que je souhaite développer et qui trouve de très nombreuses applications en sciences des catastrophes, science de la durabilité et sciences économiques, est le développement de modèles d'apprentissage interprétables (*eXplainable Artificial Intelligence* -XAI-) ce qui renvoie notamment aux travaux et perspectives que j'ai détaillées dans le Chapitre 3.

Je serai heureux de poursuivre avec de joyeux.ses collaborateurs.rices, les thématiques évoquées tout au long de ce manuscrit. Les idées que j'ai décrites peuvent être source d'inspiration mais je reste totalement ouvert à explorer de nouvelles pistes, de nouveaux objets,

7.2. TRAVAUX ACTUELS ET FUTURS

de nouvelles théories, de nouvelles applications !

Notations

n, p, m, k	Constantes	(fixées par les données ou fixées manuellement)
i, j, k, l	Itérateurs	
i', i''	Itérateurs	(parcourant le même ensemble que i)
X	Objet quelconque	
x	Réel quelconque	
\mathbf{x}	Vecteur de X	($\mathbf{x} \in \mathbb{R}^p$ en général)
X_i	Objet i	(i indice d'une observation)
\mathbf{x}_i	Vecteur de X_i	($\mathbf{x}_i \in \mathbb{R}^p, \forall i = 1, \dots, n$, en général)
X^j	Variable j	(j id d'une variable)
\mathbf{x}^j	Vecteur de X^j	($\mathbf{x}^j \in \mathbb{R}^n, \forall j = 1, \dots, p$, en général)
\mathbf{X}	Matrice ($\mathbf{x}^1 \dots \mathbf{x}^p$)	($\mathbf{X} \in \mathbb{R}^{n \times p}$, en général)

TABLE 7.1 – Notations génériques

v	Indice indiquant la partie image d'un objet multimédia
t	Indice indiquant la partie texte d'un objet multimédia
u	Paramètre indiquant le type d'information
Q	Requête quelconque
$\mathbf{q} = (\mathbf{q}^v, \mathbf{q}^t)$	Vecteur d'une requête multimédia Q
$X = (X^v, X^t)$	Objet multimédia quelconque formé d'une image X^v et d'un texte X^t
$\mathbf{x} = (\mathbf{x}^v, \mathbf{x}^t)$	Vecteur d'un objet multimédia X
$X_i = (X_i^v, X_i^t)$	Objet multimédia d'une base de données
$\mathbf{x}_i = (\mathbf{x}_i^v, \mathbf{x}_i^t)$	Vecteur d'un objet multimédia X_i d'une base
$\text{Rel}^{\text{early}}(Q, X_i)$	Score de pertinence de X_i étant donné Q basée sur la fusion précoce
F	Fonction d'agrégation entre deux réels
$\text{Rel}^{\text{late}}(Q, X_i)$	Score de pertinence de X_i étant donné Q basée sur la fusion tardive
Sim^u	Fonction de similarité de type u
$\text{Sim}^u(\mathbf{q},)$	Vecteur des similarités de type u entre \mathbf{q} et les objets de la base
\mathbf{S}^u	Matrice de similarités (ou d'adjacence pondérée d'un graphe) de type u
Knn^k	Application de sélection des k plus proches voisins
$\text{MMR}(\mathbf{q}, \mathbf{x}_i)$	Score <i>Maximum Margin Relevance</i> de \mathbf{x}_i étant donnée \mathbf{q}
$\mathbf{e}_n = (1)$	Vecteur rempli de 1 de taille n
\mathbf{P}	Matrice stochastique (valeurs non négatives, lignes somment à 1, carré d'ordre n)
\mathbf{v}	Vecteur stochastique (valeurs non négatives qui somment à 1, de taille n)
$\boldsymbol{\pi}_t$	Vecteur stochastique évalué à l'itération t (de taille n)
$\text{Rel}^{\text{inter}}(Q, X_i)$	Score de pertinence de X_i étant donné Q basée sur une fusion composite

TABLE 7.2 – Notations spécifiques au Chapitre 1

D	Objet de type document ou texte
W	Terme ou mot ou unité lexicale ou <i>chunk</i>
$P(W D)$	Probabilité d'observer W sachant D
\mathbb{C}	Corpus ou collection de documents ou textes
SEN	Type de contexte lexical (<i>SENtence</i> , cooccurrence)
NP	Type de contexte syntaxique (<i>Noun Phrase</i> , <i>shallow parsing</i>)
DEP	Type de contexte syntaxique (<i>DEPendency</i> , <i>dependency tree</i>)
l	Indice indiquant le type d'information lexical
s	Indice indiquant le type d'information syntaxique
f	Indice indiquant le type d'information <i>standard feature</i>
\mathbf{X}^u	Matrice de <i>features</i> (ou d'incidence d'un graphe ou hypergraphe) de type $u \in \{l, s, f\}$
\mathbf{S}^u	Matrice de similarités (ou d'adjacence pondérée d'un graphe) de type $u \in \{l, s, f\}$
\mathbf{M}^u	Matrice générique de type u pouvant être soit de <i>features</i> , soit de similarités
Knn^k	Application de sélection des k plus proches voisins
E	Application de fusion précoce (<i>early</i>) entre matrices
L	Application de fusion tardive (<i>late</i>) entre matrices
P	Application de fusion par propagation (mono ou transmodale) entre matrices
E_i	Entité nommée (EN)
C_j	Contexte ou dépendance syntaxique
\mathbb{K}_l	Clique d'entités nommées
$G(C_j, \mathbb{K}_l)$	Score de pertinence de C_j pour \mathbb{K}_l

TABLE 7.3 – Notations spécifiques au Chapitre 2

$\mathbb{N} = \{1, \dots, N\}$	Ensemble des critères ou votants (identifiés par des entiers)
N	Nombre total de critère ou de votants
$\mathbb{M} = \{1, \dots, M\}$	Ensemble des alternatives ou candidats
M	Nombre total d'alternatives ou de candidats
$X \succ X'$	X "est strictement préféré à" X'
$X \succeq X'$	X "est préféré ou indifférent à" X'
$X \sim X'$	X "est indifférent à" X'
$2^{\mathbb{N}}$	Ensemble des sous-ensembles de \mathbb{N}
F	Fonction d'agrégation
OWA	Opérateur <i>Ordered Weighted Averaging</i>
T	<i>Triangular norm (t-norm)</i>
μ	Mesure floue sur $2^{\mathbb{N}}$ ou capacité
τ	Permutation sur \mathbb{N}
CI_{μ}	Intégrale de Choquet (unipolaire) par rapport à μ
$3^{\mathbb{N}}$	Ensemble des paires de sous-ensembles de \mathbb{N} d'intersection vide
ν	Mesure floue sur $3^{\mathbb{N}}$ ou bicapacité
BCI_{ν}	Intégrale de Choquet bipolaire par rapport à ν
m^{μ}	Transformée de Möbiüs de μ
b^{ν}	Transformée bipolaire de Möbiüs de ν
E_k^N	Sous-ensemble flou "au moins k critères sur N sont satisfaits"
$\mu_{E_k^N}$	Fonction d'appartenance au sous-ensemble flou E_k^N
Σ_k^N	Somme symétrique relative aux sous-ensembles de critères de cardinal k
$\mathbf{w} = (w_k)$	Vecteur de poids non négatifs sommant à 1 (de taille N)
\mathbf{w}^{\uparrow}	Vecteur de poids croissants $(1, 2, \dots, N)/(N(N+1)/2)$
\mathbf{w}^{\downarrow}	Vecteur de poids décroissants $(N, N-1, \dots, 1)/(N(N+1)/2)$
T_{λ}	<i>T-norm</i> paramétrique de paramètre λ
Λ	Matrice triangulaire avec $\forall i < j, \lambda_{ij}$ le paramètre de la <i>t-norm</i> pour $\mu_{S_i \cap S_j}$
$A_{\mathbf{w}, T_{\lambda}, \Lambda}$	Fonction d'agrégation paramétré par $\mathbf{w}, T_{\lambda}, \Lambda$

TABLE 7.4 – Notations spécifiques au Chapitre 3

X_i	Individu ou objet i ($i = 1, \dots, n$)
X^k	Variable k (qualitative ou quantitative) ($k = 1, \dots, p$)
\mathbf{x}^k	Vecteur des valeurs de X^k sur les individus $\{X_i\}_i$
n	Nombre total d'individus ou objets
$\mathbf{N} = (n_{uv})$	Table de contingence croisant X^k et X^l (X^k et X^l qualitatives)
n_{uv}	Nombre d'individus avec modalité u de X^k et v de X^l
p_k	Nombre total de modalités de X^k
n_u^k	Nombre total d'individus avec modalité u de X^k
$\mathbf{X}^k = (x_{ii'}^k)$	Matrice relationnelle de la relation binaire (RB) associée à X^k
$x_{ii'}^k$	Vaut 1 si X_i et $X_{i'}$ en relation pour X^k et 0 sinon
$\mathbf{C} = (c_{ii'})$	Matrice relationnelle collective $\mathbf{C} = \sum_k \mathbf{X}^k$
$c_{ii'}$	Nb. de variables supportant X_i et $X_{i'}$ sont en relation
Δ	Mesure d'association quelconque
\mathbf{M}^t	Moyenne (puissance) généralisée d'ordre t ($t \in \mathbb{R}$)
$\theta_{ii'}$	Angle formé par deux vecteurs \mathbf{x}_i et $\mathbf{x}_{i'}$
$\gamma_{ii'}$	Rapport entre $\max(\ \mathbf{x}_i\ , \ \mathbf{x}_{i'}\)$ et $\min(\ \mathbf{x}_i\ , \ \mathbf{x}_{i'}\)$ ($\gamma_{ii'} > 1$)
$\mathbf{S} = (s_{ii'})$	Matrice de produits scalaires entre individus $\mathbf{S} = \sum_k \mathbf{x}^k [\mathbf{x}^k]^\top$, $s_{ii'} = \langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle$
$\tilde{\mathbf{S}} = (\tilde{s}_{ii'})$	Matrice de similarités transformée par un opérateur
$\mu(\tilde{\mathbf{S}}, i, i')$	Mesure de tendance centrale pour $(X_i, X_{i'})$ (réel)
$\mathbf{A} = (a_{ii'})$	Matrice d'adjacence binaire d'une graphe non orienté (carrée d'ordre n)
$\mathbf{K} = (k_{ii'})$	Matrice de noyaux entre individus (carrée d'ordre n)
$\mathbf{K}^t = (k_{ii'}^t)$	Matrice de noyaux normalisée d'ordre $t > 0$
Γ	Coefficient de corrélation général de Daniels-Kendall
$\mathbf{X}^k = (\check{x}_{ii'}^k)$	Matrice carrée d'ordre n dérivée de \mathbf{x}^k
$\check{\mathbf{X}}^k = (\check{x}_{ii'}^k)$	Matrice relationnelle de la RB inverse de la RB sous-jacente à \mathbf{X}^k
$\mathbf{M}^k = (m_{ii'}^k)$	Matrice de poids associée à \mathbf{X}^k (carrée d'ordre n)
Λ^t	Coefficient de similarité général d'ordre t
\mathbf{L}	Matrice Laplacienne associée à la matrice d'adjacence \mathbf{X} (carrée d'ordre n)
\mathbf{L}_{sym}	Matrice Laplacienne normalisée
\odot	Multiplication matricielle (terme à terme) de Hadamard
$\mathbf{N}^t = (n_{ii'}^t)$	Matrice de normalisation d'ordre $t > 0$
\mathbf{L}_{sym}^t	Matrice Laplacienne normalisée (approche symétrique) d'ordre $t > 0$

TABLE 7.5 – Notations spécifiques au Chapitre 4

X_i	Individu ou objet i ($i = 1, \dots, n$)
\mathbf{x}_i	Vecteur de X_i ($\mathbf{x}_i \in \mathbb{R}^p, \forall i = 1, \dots, n$)
\mathbb{R}^p	<i>Input space</i> (espace initial des \mathbf{x}_i)
\mathbb{F}	<i>Feature space</i> (RKHS en général)
ϕ	Application de $\mathbb{R}^p \rightarrow \mathbb{F}$ ($\phi(\mathbf{x}_i) \in \mathbb{F}$)
$k(\cdot, \cdot)$	Fonction noyau $\mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$
$\mathbb{C} = \{\mathbb{C}_j\}_j$	Partition des ind. en k classes/ <i>clusters</i> ($j = 1, \dots, k$)
\mathbf{c}_j	Vecteur moyen du <i>cluster</i> \mathbb{C}_j dans \mathbb{F} ($j = 1, \dots, k$)
$\mathbf{Y} = (y_{ij})$	Matrice d'affectation ind. \rightarrow <i>clusters</i> (binaire, de taille $n \times k$)
$\mathbf{e}_n = (1)$	Vecteur rempli de 1 ($n \times 1$)
$\mathbf{K} = (k_{ii'})$	Matrice de noyaux (carrée d'ordre n)
$\mathbf{X} = (x_{ii'})$	Matrice d'affinités (carrée d'ordre n)
$\text{Tr}(\mathbf{A})$	Trace d'une matrice carrée \mathbf{A}
$\text{Rk}(\mathbf{B})$	Rang d'une matrice \mathbf{B}
\mathbf{I}_n	Matrice identité (carrée d'ordre n)
\mathbf{J}_n	Matrice remplie de $1/n$ (carrée d'ordre n)
$\mathbf{D} = (d_{ii'})$	Matrice de dissimilarités (initiale) (carrée d'ordre n)
$\mathbf{D}^t = (d_{ii'}^t)$	Matrice de dissim. à l'itération t (carrée d'ordre $n - t$)
α', β', γ	Paramètres dans la formule initiale de LW
α, β, p	Paramètres dans la formule partielle de LW
$\ \cdot, \cdot\ _F$	Distance de Frobenius entre matrices
$\langle \cdot, \cdot \rangle_F$	Produit scalaire de Frobenius entre matrices
$\mathbf{n}_n = (0)$	Vecteur rempli de 0 (vecteur nul) ($n \times 1$)
\mathbf{L}	Matrice Laplacienne d'un graphe d'affinités (carrée d'ordre n)
$\mathbf{L}_\mathbf{X}$	Mat. Lap. d'une mat. bistochastique et idempotence \mathbf{X} ($\mathbf{L}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}$)
\mathbf{S}	Matrice de similarités/affinités observées (carrée d'ordre n)
\mathbf{S}^t	Mat. de sim. à l'itération t de (SN)K-AHC (carrée d'ordre $n - t$)
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	Paramètres des formules de màj. de (SN)K-AHC
\mathbf{p}	Paramètre de la règle de fusion de (SN)K-AHC
$\mathbf{\Lambda}^t$	Mat. de dissimilarité associée à \mathbf{S}^t de (SN)K-AHC
$\mathbf{1}_n$	Matrice remplie de 1 (carrée d'ordre n)
\mathbb{C}^t	Ensemble des <i>clusters</i> à l'itération t de (SN)K-AHC
\mathcal{S}^t	Ens. de paires (i, j) de <i>clusters</i> de $\mathbb{C}^t \times \mathbb{C}^t$ tq. $s_{ij}^t > 0$

TABLE 7.6 – Notations spécifiques au Chapitre 5

X_i	Individu ou objet i ($i = 1, \dots, n$)
x_i	Fonction associée à X_i ($i = 1, \dots, n$)
$\mathbb{L}^2([0, T])$	Ens. des fcts $x : [0, T] \rightarrow \mathbb{R}$ de carré intégrable ($\int_0^T x(t) ^2 dt < \infty$, aussi noté \mathbb{L}^2)
$\{t_i\}_j$	Ens. de points dans $[0, T]$ ($j = 1, \dots, p$)
$\{y_{ij}\}_j$	Ens. des valeurs observées de x_i en les points $\{t_i\}_j$ ($j = 1, \dots, p$)
$\phi = (\phi_k)$	Vecteur de fonctions servant de base ($k = 1, \dots, m$)
$\mathbf{c}_i = (c_{i,k})$	Vecteur des coef. de x_i dans la base de fcts ϕ ($k = 1, \dots, m$)
\mathbb{H}^q	Ens. de fcts de \mathbb{L}^2 avec dérivées (au sens faible) (Espace de Sobolev) jusqu'à l'ordre q (inclus) dans \mathbb{L}^2
D	Opérateur différentiel usuel ($Dx = x'$)
D^j	Opérateur différentiel appliqué j fois ($D^2x = x'', \dots$)
D^0	Opérateur identité ($D^0x = x$)
$\langle \cdot, \cdot \rangle_{\mathbb{L}^2}$	Produit scalaire de \mathbb{L}^2 (espace de Hilbert) ($\langle x_i, x_{i'} \rangle_{\mathbb{L}^2} = \int_0^T x_i(t)x_{i'}(t)dt$)
$k^s(\cdot, \cdot)$	Fct. noyau $\mathbb{L}^2 \times \mathbb{L}^2 \rightarrow \mathbb{R}$ pour comparer les $\{D^s x_i\}_i$ ($s = 0, \dots, q$)
$\mathbf{w} = (w_s)$	Vecteur des poids des vues (dérivées d'ordre s) ($s = 0, \dots, q$)
$\kappa 0, \kappa 1, \kappa 2$	Acronymes, représentations mono-vues $\kappa \in \{l, g\}$ (<i>l</i> inear, <i>g</i> aussian)
$\kappa 01, \kappa 012$	Acronymes, représentations multi-vues $\kappa \in \{l, g\}$ ($\kappa 01(2) \leftrightarrow \kappa 0 + \kappa 1(+\kappa 2)$)
$\kappa 01o, \kappa 012o$	Acronymes, rep. multi-vues avec poids optimisés ($\kappa 01(2)o \leftrightarrow w_0 \kappa 0 + w_1 \kappa 1(+w_2 \kappa 2)$)

TABLE 7.7 – Notations spécifiques au Chapitre 6

Table des figures

1.1	Fusion précoce (<i>early</i>), tardive (<i>late</i>) et intermédiaire ou transmodale (<i>cross-media</i>)	11
1.2	Similarités image-texte en recherche d'information image-texte	12
1.3	Exemples de résultats sans (ligne du haut) et avec (ligne du bas) <i>diversity re-ranking</i> basé sur le <i>clustering</i>	14
1.4	Architecture de notre système d' <i>information seeking</i>	15
1.5	Complémentarité des deux cartes et des deux modes de recherche d'information.	16
1.6	<i>Workflow</i> du modèle général de fusion intermédiaire d'information multimédia basé sur le filtrage sémantique et la diffusion dans des graphes.	23
1.7	Une architecture neuronale de fusion transmodale pour la tâche <i>learning to rank</i> utilisant des RBF NN en parallèle en entrée et en sortie et un MLP intermédiaire.	27
2.1	Architecture et briques technologiques de la proposition <u>RéSoCo</u>	31
2.2	Illustration de contexte lexical du mot " <i>contains</i> ".	34
2.3	Illustration d'un arbre syntaxique simple indiquant les catégories grammaticales et donnant lieu à un premier type de contexte syntaxique.	34
2.4	Illustration d'un arbre syntaxique indiquant les grammaires de dépendance et donnant lieu à un deuxième type de contexte syntaxique.	35
2.5	Illustration de la représentation par hypergraphe des différents contextes.	37
2.6	Matrice d'incidence de l'hypergraphe avec divers types d'hyperarêtes selon les contextes.	38
2.7	Chaîne de traitements pour la tâche de reconnaissance d'entités nommées.	40
2.8	Chaîne de traitements pour la tâche de reconnaissance d'entités nommées.	43
2.9	Exemple de deux annotations possibles de " <i>London</i> " (LOC -capitale- et ORG -université-) représentée par deux cliques maximales contenant ce terme.	44
2.10	<i>Clique Based Clustering (CBC) System</i>	46

2.11	Architecture des <i>Transformers</i> pour la traduction automatique [Vaswani et al., 2017].	49
3.1	Représentation sous forme matricielle de $3^{\mathbb{N}}$ (paires marquées par \checkmark et $*$), des valeurs de \mathbf{b}^{ν} lorsque ν est quelconque (\checkmark et $*$) et lorsqu'elle est 2-additive (\checkmark uniquement).	63
4.1	Ligne du haut : valeurs NMI (axe vertical) <i>versus</i> paramètre de mélange μ (axe horizontal). Ligne du bas : Nombre de <i>clusters</i> trouvés (axe vertical) <i>versus</i> paramètre de mélange μ (axe horizontal). Les courbes pour chaque critère de partitionnement évoqué sont exposées. De gauche à droite, les graphiques correspondent à des graphes comportant 500, 1000 et 2000 sommets.	96
4.2	Courbes représentant les valeurs de $k_{ii'}^t$ pour $t \rightarrow 0, t = 1, t = 10, t = 100, t \rightarrow \infty$; $\gamma_{ii'} \in [1, 2]$ (en abscisse); $\cos(\theta_{ii'}) = -1$ (à gauche) et $\cos(\theta_{ii'}) = 1$ (à droite).	98
4.3	Valeurs ARI sur 5 jeux de données classiques et pour 5 normalisation distinctes : $t \rightarrow 0$ (<i>baseline</i>), $t = 1, 2, 5, 10$.	105
5.1	Jeu de données <i>Compound</i> comportant 399 points de \mathbb{R}^2 répartis en 6 <i>clusters</i> .	131
5.2	Résultat de SNK-AHC sur le jeu de données <i>Compound</i> (les <i>clusters</i> de taille ≤ 3 sont indiqués de façon indifférenciée par des $*$ en gris).	132
5.3	Résultats de SNK-AHC avec un noyau Gaussien pour le jeu de données réelles <i>Landsat</i> . L'axe des abscisses correspond au % de voisins supprimés. L'axe des ordonnées correspond à des valeurs d'indicateurs variant dans $[0, 1]$. Les traits continus verts avec $+$ représentent la mémoire relative employée, les traits discontinus jaunes avec \times indiquent le temps de traitement relatif, les traits pointillés bleu avec Δ montrent la mesure ARI <i>Adjusted Rand Index</i> . Je ne commente pas les traits rouges avec \circ .	133
6.1	De gauche à droite : fonctions initiales, dérivées premières, dérivées secondes. Les fonctions du groupe 1 sont en traits continus gris, les fonctions du groupe 2 sont en traits discontinus bleus. Les fonctions moyennes des groupes sont en rouge : traits continus pour le groupe 1, trait discontinu pour le groupe 2.	148
6.2	<i>Box plot</i> des scores NMI mesurés sur 50 échantillons en utilisant MK-KM-FD avec différentes représentations des DF. Les triangles rouges indiquent les moyennes des 50 valeurs.	149
6.3	<i>Box plots</i> des 10 valeurs de NMI de MK-KM-FD utilisant 10 initialisations différentes pour le <i>kernel</i> k-means. Les triangles rouges indiquent les moyennes de ces 10 valeurs.	150
6.4	<i>Box plot</i> des scores Accuracy mesurés sur 50 échantillons en utilisant MK-SVM-FD avec différentes représentations des DF. Les triangles rouges indiquent les moyennes des 50 valeurs.	154
6.5	<i>Box plots</i> des 10 valeurs d'Accuracy de validation (validation croisée à 10 <i>folds</i>). Les triangles rouges indiquent les moyennes des 10 valeurs.	155

- 6.6 Une architecture neuronale MLP avec couche cachée fonctionnelle et apprentissage de fusion entre DF et dérivées dans le cas de la catégorisation à k classes.158

Liste des tableaux

2.1	Mot “tesgüino” dans quatre phrases distinctes.	33
2.2	F – mesure sur trois tâches de NER en utilisant plusieurs représentations des mots.	41
2.3	Résultats du système CBC seul et en combinaison avec d’autres systèmes de NER.	47
3.1	(a) Table des notes partielles traduites ; (b) Note agrégée traduite et prédictions des différents modèles.	68
3.2	(a) Table des notes partielles traduites ; (b) Note agrégée traduite avec modification pour l’élève g pour un cas d’ <i>inconsistency</i> et prédictions des différents modèles.	69
3.3	Valeurs estimées de b' des le cas des données de la Table 3.2.	70
4.1	Modèles de tendance centrale des 3 critères de partitionnement B, E et J. . .	91
4.2	Matrices de poids \mathbf{M}^k (même formules pour \mathbf{M}^l) et mesures correspondantes : corrélations de rangs (variables quantitatives ou ordinales) en haut et mesures d’association (variables qualitatives) en bas.	101
4.3	Matrices de poids \mathbf{M}^k (même formules pour \mathbf{M}^l) pour r le coefficient de corrélation linéaire de Bravais-Pearson.	107
5.1	Cas particuliers de la formule initiale de LW (5.19).	117
5.2	Cas particuliers de la formule partielle de LW (5.21) (D-AHC).	118
5.3	NMI performances.	125
5.4	Cas particuliers connus dans le cadre de K-AHC défini par (5.54), (5.50) et (5.51).	128
7.1	Notations génériques	163
7.2	Notations spécifiques au Chapitre 1	164

LISTE DES TABLEAUX

7.3	Notations spécifiques au Chapitre 2	165
7.4	Notations spécifiques au Chapitre 3	166
7.5	Notations spécifiques au Chapitre 4	167
7.6	Notations spécifiques au Chapitre 5	168
7.7	Notations spécifiques au Chapitre 6	169

Bibliographie

- [Abraham et al., 2003] Abraham, C., Cornillon, P.-A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using b-splines. Scandinavian journal of statistics, 30(3) :581–595.
- [Ah-Pine, 2007] Ah-Pine, J. (2007). Sur des aspects algébriques et combinatoires de l’analyse relationnelle : applications en classification automatique, en théorie du choix social et en théorie des tresses. PhD thesis, Université Paris 6.
- [Ah-Pine, 2009] Ah-Pine, J. (2009). Cluster analysis based on the central tendency deviation principle. In International Conference on Advanced Data Mining and Applications, pages 5–18. Springer.
- [Ah-Pine, 2010] Ah-Pine, J. (2010). Normalized kernels as similarity indices. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 362–373. Springer.
- [Ah-Pine, 2013a] Ah-Pine, J. (2013a). A general framework for comparing heterogeneous binary relations. In Geometric Science of Information - First International Conference, GSI 2013, Paris, France, August 28-30, 2013. Proceedings, pages 188–195.
- [Ah-Pine, 2013b] Ah-Pine, J. (2013b). Graph clustering by maximizing statistical association measures. In Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013. Proceedings, pages 56–67.
- [Ah-Pine, 2016] Ah-Pine, J. (2016). On aggregation functions based on linguistically quantified propositions and finitely additive set functions. Fuzzy Sets and Systems, 287 :1–21.
- [Ah-Pine, 2017] Ah-Pine, J. (2017). Sur la normalisation de la matrice laplacienne en partitionnement spectral. In Rencontres de la SFC (Société Francophone de Classification).
- [Ah-Pine, 2018] Ah-Pine, J. (2018). An efficient and effective generic agglomerative hierarchical clustering approach. The Journal of Machine Learning Research, 19(1) :1615–1658.
- [Ah-Pine, 2022] Ah-Pine, J. (2022). Learning doubly stochastic and nearly idempotent affinity matrix for graph-based clustering. European Journal of Operational Research, 299(3) :1069–1078.

- [Ah-Pine et al., 2008] Ah-Pine, J., Cifarelli, C., Clinchant, S., Csurka, G., and Renders, J. (2008). Xrce’s participation to imageclef 2008. In Working Notes for CLEF 2008 Workshop co-located with the 12th European Conference on Digital Libraries (ECDL 2008), Aarhus, Denmark, September 17-19, 2008.
- [Ah-Pine et al., 2015] Ah-Pine, J., Csurka, G., and Clinchant, S. (2015). Unsupervised visual and textual information fusion in cbmir using graph-based methods. ACM Transactions on Information Systems (TOIS), 33(2) :1–31.
- [Ah-Pine et al., 2009] Ah-Pine, J., Csurka, G., and Renders, J.-M. (2009). Evaluation of diversity-focused strategies for multimedia retrieval. In Evaluating Systems for Multilingual and Multimodal Information Access : 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers 9, pages 677–684. Springer.
- [Ah-Pine and Jacquet, 2009] Ah-Pine, J. and Jacquet, G. (2009). Clique-based clustering for improving named entity recognition systems. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 51–59.
- [Ah-Pine and Lebreton, 2022] Ah-Pine, J. and Lebreton, N. (2022). Fusion tardive en analyse de données fonctionnelles élastique. In Journées de statistique de la SFdS.
- [Ah-Pine and Marcotorchino, 2010] Ah-Pine, J. and Marcotorchino, J.-F. (2010). Unifying some association criteria between partitions by using relational matrices. Communications in Statistics—Theory and Methods, 39(3) :531–542.
- [Ah-Pine et al., 2013] Ah-Pine, J., Mayag, B., and Rolland, A. (2013). Identification of a 2-additive bi-capacity by using mathematical programming. In Algorithmic Decision Theory - Third International Conference, ADT 2013, Bruxelles, Belgium, November 12-14, 2013, Proceedings, pages 15–29.
- [Ah-Pine et al., 2009] Ah-Pine, J., Renders, J., and Viaud, M. (2009). A continuum between browsing and query-based search for user-centered multimedia information access. In Adaptive Multimedia Retrieval. Understanding Media and Adapting to the User - 7th International Workshop, AMR 2009, Madrid, Spain, September 24-25, 2009, Revised Selected Papers, pages 111–123.
- [Ah-Pine et al., 2012] Ah-Pine, J., Renders, J.-M., and Viaud, M.-l. (2012). System and method for information seeking in a multimedia collection. US Patent 8,171,049.
- [Ah-Pine and Wang, 2016] Ah-Pine, J. and Wang, X. (2016). Similarity based hierarchical clustering with an application to text collections. In International Symposium on Intelligent Data Analysis, pages 320–331. Springer.
- [Ah-Pine and Yao, 2024] Ah-Pine, J. and Yao, A.-F. (2024). Using derivatives and multiple kernel methods for clustering and classifying functional data. Preprint.
- [Ahmedou et al., 2016] Ahmedou, A., Marion, J.-M., and Pumo, B. (2016). Generalized linear model with functional predictors and their derivatives. Journal of Multivariate Analysis, 146 :313–324.

- [Aloise et al., 2009] Aloise, D., Deshpande, A., Hansen, P., and Papat, P. (2009). Np-hardness of euclidean sum-of-squares clustering. Machine learning, 75 :245–248.
- [Alonso et al., 2012] Alonso, A. M., Casado, D., and Romo, J. (2012). Supervised classification for functional data : A weighted distance approach. Computational Statistics & Data Analysis, 56(7) :2334–2346.
- [Bauschke and Borwein, 1996] Bauschke, H. H. and Borwein, J. M. (1996). On projection algorithms for solving convex feasibility problems. SIAM review, 38(3) :367–426.
- [Bellman and Zadeh, 1970] Bellman, R. E. and Zadeh, L. A. (1970). Decision-making in a fuzzy environment. Management science, 17(4) :B–141.
- [Belson, 1959] Belson, W. (1959). Matching and prediction on the principle of biological classification. Applied statistics, 7.
- [Benhadia and Marcotorchino, 1998] Benhadia, H. and Marcotorchino, F. (1998). Introduction à la similarité régularisée en analyse relationnelle. Revue de statistique appliquée, 46(1) :45–69.
- [Bertrand, 2021] Bertrand, P. (2021). Conditions de Monge, Transport Optimal et Pont Relationnel Propriétés, applications et extension du couplage d’indétermination. PhD thesis, Sorbonne Université.
- [Bertrand et al., 2022] Bertrand, P., Broniatowski, M., and Marcotorchino, J.-F. (2022). Independence versus indetermination : basis of two canonical clustering criteria. Advances in Data Analysis and Classification, 16(4) :1069–1093.
- [Besse and Ramsay, 1986] Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. Psychometrika, 51(2) :285–311.
- [Bezdek, 1973] Bezdek, J. C. (1973). Fuzzy Mathematics In Pattern Classification. Cornell University.
- [Bezdek et al., 1984] Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm : The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3) :191–203.
- [Biau et al., 2005] Biau, G., Bunea, F., and Wegkamp, M. H. (2005). Functional classification in hilbert spaces. IEEE Transactions on Information Theory, 51(6) :2163–2172.
- [Biau et al., 2008] Biau, G., Devroye, L., and Lugosi, G. (2008). On the performance of clustering in hilbert spaces. IEEE Transactions on Information Theory, 54(2) :781–790.
- [Boyd et al., 2011] Boyd, S., Parikh, N., and Chu, E. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc.
- [Bron and Kerbosch, 1973] Bron, C. and Kerbosch, J. (1973). Algorithm 457 : finding all cliques of an undirected graph. Communications of the ACM, 16(9) :575–577.
- [Bruynooghe, 1978] Bruynooghe, M. (1978). Classification ascendante hiérarchique des grands ensembles de données : un algorithme rapide fondé sur la construction des voisinages réductibles. Cahiers de l’analyse des données, 3(1) :7–33.

- [Bunescu and Pasca, 2006] Bunescu, R. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation.
- [Cespedes, 2013] Cespedes, P. C. (2013). Modélisations et extensions du formalisme de l'analyse relationnelle mathématique à la modularisation des grands graphes. PhD thesis, Université Paris 6.
- [Chamroukhi and Nguyen, 2019] Chamroukhi, F. and Nguyen, H. D. (2019). Model-based clustering and classification of functional data. WIREs Data Mining and Knowledge Discovery, 9(4).
- [Chen et al., 2022] Chen, L., Liu, C., Zhou, R., Xu, J., and Li, J. (2022). Efficient maximal bi-clique enumeration for large sparse bipartite graphs. Proceedings of the VLDB Endowment, 15(8) :1559–1571.
- [Chen and Van Ness, 1994] Chen, Z. and Van Ness, J. W. (1994). Space-contracting, space-dilating, and positive admissible clustering algorithms. Pattern recognition, 27(6) :853–857.
- [Chiou and Li, 2007] Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 69(4) :679–699.
- [Choquet, 1954] Choquet, G. (1954). Theory of capacities. In Annales de l'institut Fourier, volume 5, pages 131–295.
- [Clark et al., 2019] Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. arXiv preprint arXiv :1906.04341.
- [Clinchant et al., 2011] Clinchant, S., Ah-Pine, J., and Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval. In Proceedings of the 1st ACM international conference on multimedia retrieval, pages 1–8.
- [Clinchant et al., 2007] Clinchant, S., Renders, J., and Csurka, G. (2007). Xrce's participation to imageclef. CLEF Working Notes.
- [Clinchant et al., 2008] Clinchant, S., Renders, J.-M., and Csurka, G. (2008). Trans-media pseudo-relevance feedback methods in multimedia retrieval. In Advances in Multilingual and Multimodal Information Retrieval : 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers 8, pages 569–576. Springer.
- [Coifman and Lafon, 2006] Coifman, R. R. and Lafon, S. (2006). Diffusion maps. Applied and computational harmonic analysis, 21(1) :5–30.
- [Collins, 2002] Collins, M. (2002). Discriminative training methods for hidden markov models : Theory and experiments with perceptron algorithms. In Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002), pages 1–8.
- [Combettes and Pesquet, 2011] Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In Fixed-point algorithms for inverse problems in science and engineering, pages 185–212. Springer.

- [Condorcet, 1785] Condorcet, M. M. d. (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Paris.
- [Cucerzan, 2007] Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pages 708–716.
- [Daniels, 1944] Daniels, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. Biometrika, 33(2) :129–135.
- [Das et al., 2020] Das, R., Sen, S., and Maulik, U. (2020). A survey on fuzzy deep neural networks. ACM Computing Surveys (CSUR), 53(3) :1–25.
- [Daumé III, 2006] Daumé III, H. (2006). Practical structured learning techniques for natural language processing. Citeseer.
- [Day and Edelsbrunner, 1984] Day, W. H. and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. Journal of classification, 1(1) :7–24.
- [Defays, 1977] Defays, D. (1977). An efficient algorithm for a complete link method. The Computer Journal, 20(4) :364–366.
- [Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv :1810.04805.
- [Dhillon et al., 2004] Dhillon, I. S., Guan, Y., and Kulis, B. (2004). A unified view of kernel k-means, spectral clustering and graph cuts. Citeseer.
- [Escoufier, 1970] Escoufier, Y. (1970). Echantillonnage dans une population de variables aléatoires réelles. In Annales de l'ISUP, volume 19, pages 1–47.
- [Escoufier, 1973] Escoufier, Y. (1973). Le traitement des variables vectorielles. Biometrics, pages 751–760.
- [Fallah Tehrani et al., 2012] Fallah Tehrani, A., Cheng, W., Dembczyński, K., and Hüllermeier, E. (2012). Learning monotone nonlinear models using the choquet integral. Machine Learning, 89 :183–211.
- [Fan et al., 2010] Fan, G., Cao, J., and Wang, J. (2010). Functional data classification for temporal gene expression data with kernel-induced random forests. In 2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pages 1–5. IEEE.
- [Ferraty and Vieu, 2003] Ferraty, F. and Vieu, P. (2003). Curves discrimination : a nonparametric functional approach. Computational Statistics & Data Analysis, 44(1-2) :161–173.
- [Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis : theory and practice. Springer Science & Business Media.
- [Firth, 1957] Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis, pages 10–32.

- [Floriello and Vitelli, 2017] Floriello, D. and Vitelli, V. (2017). Sparse clustering of functional data. Journal of Multivariate Analysis, 154 :1–18.
- [Fuchs et al., 2015] Fuchs, K., Gertheiss, J., and Tutz, G. (2015). Nearest neighbor ensembles for functional data with interpretable feature selection. Chemometrics and Intelligent Laboratory Systems, 146 :186–197.
- [Fujimoto, 2004] Fujimoto, K. (2004). New characterizations of k -additivity and k -monotonicity of bi-capacities. In Joint 2nd Int. Conf. on Soft Computing and Intelligent Systems and 5th International Symposium on Advanced Intelligent Systems.
- [Fujimoto and Murofushi, 2005] Fujimoto, K. and Murofushi, T. (2005). Some characterizations of k -monotonicity through the bipolar möbius transform in bi-capacities. JACIII, 9(5) :484–495.
- [García et al., 2015] García, M. L. L., García-Ródenas, R., and Gómez, A. G. (2015). K-means algorithms for functional data. Neurocomputing, 151 :231–245.
- [Gershgorin, 1931] Gershgorin, S. A. (1931). Über die abgrenzung der eigenwerte einer matrix. Proceedings of the Russian Academy of Sciences. Mathematical series., (6) :749–754.
- [Górecki and Łuczak, 2013] Górecki, T. and Łuczak, M. (2013). Using derivatives in time series classification. Data Mining and Knowledge Discovery, 26 :310–331.
- [Gower and Ross, 1969] Gower, J. C. and Ross, G. J. (1969). Minimum spanning trees and single linkage cluster analysis. Applied statistics, pages 54–64.
- [Grabisch et al., 2008] Grabisch, M., Kojadinovic, I., and Meyer, P. (2008). A review of methods for capacity identification in Choquet integral based multi-attribute utility theory : Applications of the kappalab R package. EJOR, 186(2) :766 – 785.
- [Grabisch and Labreuche, 2002a] Grabisch, M. and Labreuche, C. (2002a). Bi-capacities for decision making on bipolar scales. In EUROFUSE Workshop on Informations Systems, pages 185–190. Citeseer.
- [Grabisch and Labreuche, 2002b] Grabisch, M. and Labreuche, C. (2002b). Bi-capacities for decision making on bipolar scales. In Proceedings of the EUROFUSE 02 Workshop on Information Systems, pages 185–190.
- [Grabisch and Labreuche, 2003] Grabisch, M. and Labreuche, C. (2003). The choquet integral for 2-additive bi-capacities. In EUSFLAT Conf., pages 300–303.
- [Grabisch and Labreuche, 2005a] Grabisch, M. and Labreuche, C. (2005a). Bi-capacities—i : definition, möbius transform and interaction. Fuzzy sets and systems, 151(2) :211–236.
- [Grabisch and Labreuche, 2005b] Grabisch, M. and Labreuche, C. (2005b). Bi-capacities—ii : the choquet integral. Fuzzy sets and systems, 151(2) :237–259.
- [Grimm, 1987] Grimm, E. C. (1987). Coniss : a fortran 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. Computers & geosciences, 13(1) :13–35.

- [Hamilton et al., 2022] Hamilton, K., Nayak, A., Božić, B., and Longo, L. (2022). Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. Semantic Web, (Preprint) :1–42.
- [Harris, 1954] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3) :146–162.
- [Hartigan, 1975] Hartigan, J. (1975). Clustering Algorithms. John Wiley and Sons.
- [Hastie et al., 1995] Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. The Annals of Statistics, pages 73–102.
- [Hitchcock and Greenwood, 2015] Hitchcock, D. B. and Greenwood, M. C. (2015). Clustering functional data. In Handbook of Cluster Analysis, pages 286–309. Chapman and Hall/CRC.
- [Hsieh et al., 2021] Hsieh, T.-Y., Sun, Y., Wang, S., and Honavar, V. (2021). Functional autoencoders for functional data representation learning. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), pages 666–674. SIAM.
- [Hsu et al., 2007] Hsu, W. H., Kennedy, L. S., and Chang, S.-F. (2007). Video search reranking through random walk over document-level context graph. In Proceedings of the 15th ACM international conference on Multimedia, pages 971–980.
- [Ieva et al., 2013] Ieva, F., Paganoni, A. M., Pigoli, D., and Vitelli, V. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. Journal of the Royal Statistical Society : Series C (Applied Statistics), 62(3) :401–418.
- [Islam et al., 2019a] Islam, M. A., Anderson, D. T., Petry, F., and Elmore, P. (2019a). An efficient evolutionary algorithm to optimize the choquet integral. International Journal of Intelligent Systems, 34(3) :366–385.
- [Islam et al., 2019b] Islam, M. A., Anderson, D. T., Pinar, A. J., Havens, T. C., Scott, G., and Keller, J. M. (2019b). Enabling explainable fusion in deep learning with fuzzy integral neural networks. IEEE Transactions on Fuzzy Systems, 28(7) :1291–1300.
- [Jacques and Preda, 2014] Jacques, J. and Preda, C. (2014). Functional data clustering : a survey. Advances in Data Analysis and Classification, 8(3) :231–255.
- [James, 2002] James, G. M. (2002). Generalized linear models with functional predictors. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 64(3) :411–432.
- [James and Hastie, 2001] James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 63(3) :533–550.
- [Janson and Vegelius, 1982] Janson, S. and Vegelius, J. (1982). The j-index as a measure of nominal scale response agreement. Applied psychological measurement, 6(1) :111–121.
- [Jelinek and Mercer, 1980] Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of markov source parameters from sparse data, in proceedings of workshop on pattern recognition in practice.

BIBLIOGRAPHIE

- [Jordan, 1926] Jordan, C. (1926). Sur la probabilité des épreuves répétées, le théorème de Bernoulli et son inversion. Bulletin de la S.M.F., 54 :101–137.
- [Jordan, 1927] Jordan, C. (1927). Les coefficients d'intensité relative de körösy. Revue de la société hongroise de statistique, 5.
- [Jordan, 1939] Jordan, C. (1939). Problèmes de la probabilité des épreuves répétées dans le cas général. Bulletin de la S.M.F., 67 :223–242.
- [Joulin et al., 2016] Joulin, A., Van Der Maaten, L., Jabri, A., and Vasilache, N. (2016). Learning visual features from large weakly supervised data. In Computer Vision–ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, pages 67–84. Springer.
- [Kendall, 1948] Kendall, M. G. (1948). Rank correlation methods.
- [Keogh and Pazzani, 2001] Keogh, E. and Pazzani, M. (2001). Dynamic time warping with higher order features. In Proceedings of the 2001 SIAM Intl. Conf. on Data Mining, volume 2.
- [Kloft et al., 2009] Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K.-R., and Zien, A. (2009). Efficient and accurate lp-norm multiple kernel learning. In NIPS, volume 22, pages 997–1005.
- [Kloft et al., 2010] Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. (2010). Non-sparse regularization and efficient training with multiple kernels.
- [Kloft et al., 2011] Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. (2011). Lp-norm multiple kernel learning. The Journal of Machine Learning Research, 12 :953–997.
- [Krämer, 2006] Krämer, N. (2006). Boosting for functional data. arXiv preprint math/0605751.
- [Kuang et al., 2012] Kuang, D., Ding, C., and Park, H. (2012). Symmetric nonnegative matrix factorization for graph clustering. In Proceedings of the 2012 SIAM international conference on data mining, pages 106–117. SIAM.
- [Kulis et al., 2007] Kulis, B., Surendran, A. C., and Platt, J. C. (2007). Fast low-rank semi-definite programming for embedding and clustering. In Artificial Intelligence and Statistics, pages 235–242.
- [Lance and Williams, 1967] Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies : 1. hierarchical systems. The Computer Journal, 9(4) :373–380.
- [Lancichinetti et al., 2008] Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. Physical review E, 78(4) :046110.
- [Landa and Cheng, 2023] Landa, B. and Cheng, X. (2023). Robust inference of manifold density and geometry by doubly stochastic scaling. SIAM Journal on Mathematics of Data Science, 5(3) :589–614.

- [Landa et al., 2021] Landa, B., Coifman, R. R., and Kluger, Y. (2021). Doubly stochastic normalization of the gaussian kernel is robust to heteroskedastic noise. SIAM journal on mathematics of data science, 3(1) :388–413.
- [Langville and Meyer, 2005] Langville, A. N. and Meyer, C. D. (2005). A survey of eigenvector methods for web information retrieval. SIAM review, 47(1) :135–161.
- [Lebart, 1978] Lebart, L. (1978). Programme d’agrégation avec contraintes. Les cahiers de l’analyse des données, 3(3) :275–287.
- [Lee and Jung, 2016] Lee, S. and Jung, S. (2016). Combined analysis of amplitude and phase variations in functional data. arXiv preprint arXiv :1603.01775.
- [Lesot et al., 2009] Lesot, M.-J., Rifqi, M., and Benhadda, H. (2009). Similarity measures for binary and numerical data : a survey. International Journal of Knowledge Engineering and Soft Data Paradigms, 1(1) :63–84.
- [Light and Margolin, 1971] Light, R. J. and Margolin, B. H. (1971). An analysis of variance for categorical data. Journal of the American Statistical Association, 66(335) :pp. 534–544.
- [Manzoor et al., 2023] Manzoor, M. A., Albarri, S., Xian, Z., Meng, Z., Nakov, P., and Liang, S. (2023). Multimodality representation learning : A survey on evolution, pretraining and its applications. arXiv preprint arXiv :2302.00389.
- [Marcotorchino and Michaud, 1981] Marcotorchino, J. and Michaud, P. (1981). Heuristic approach of the similarity aggregation problem. Methods of operations research, 43 :395–404.
- [Marcotorchino, 1984a] Marcotorchino, J.-F. (1984a). Utilisation des comparaisons par paires en statistique des contingences partie I. Technical Report F069, IBM.
- [Marcotorchino, 1984b] Marcotorchino, J.-F. (1984b). Utilisation des comparaisons par paires en statistique des contingences partie II. Technical Report F071, IBM.
- [Marcotorchino and Michaud, 1979] Marcotorchino, J.-F. and Michaud, P. (1979). Optimisation en analyse ordinaire des données. (No Title).
- [Marichal and Roubens, 2000] Marichal, J.-L. and Roubens, M. (2000). Determination of weights of interacting criteria from a reference set. EJOR, 124(3) :641 – 650.
- [Martin-Barragan et al., 2014] Martin-Barragan, B., Lillo, R., and Romo, J. (2014). Interpretable support vector machines for functional data. European Journal of Operational Research, 232(1) :146–155.
- [Mayag et al., 2012] Mayag, B., Rolland, A., and Ah-Pine, J. (2012). Elicitation of a 2-additive bi-capacity through cardinal information on trinary actions. In Advances in Computational Intelligence - 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2012, Catania, Italy, July 9-13, 2012, Proceedings, Part IV, pages 238–247.
- [Meila and Shi, 2000] Meila, M. and Shi, J. (2000). Learning segmentation by random walks. In NIPS, volume 14.

- [Meng et al., 2018] Meng, Y., Liang, J., Cao, F., and He, Y. (2018). A new distance with derivative information for functional k-means clustering algorithm. Information Sciences, 463 :166–185.
- [Menger, 1942] Menger, K. (1942). Statistical metrics. Proceedings of the National Academy of Sciences of the United States of America, 28(12) :535.
- [Mercer, 1909] Mercer, J. (1909). Functions of positive and negative type, and their connection the theory of integral equations. Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character, 209(441-458) :415–446.
- [Michaud, 1987] Michaud, P. (1987). Condorcet—a man of the avant-garde. Applied stochastic models and Data Analysis, 3(3) :173–189.
- [Michaud and Marcotorchino, 1979] Michaud, P. and Marcotorchino, F. (1979). Modèles d’optimisation en analyse des données relationnelles. Mathématiques et Sciences humaines, 67 :7–38.
- [Milligan, 1979] Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. Psychometrika, 44(3) :343–346.
- [Möller et al., 2016] Möller, A., Tutz, G., and Gertheiss, J. (2016). Random forests for functional covariates. Journal of Chemometrics, 30(12) :715–725.
- [Müller et al., 2005] Müller, H.-G., Stadtmüller, U., et al. (2005). Generalized functional linear models. Annals of Statistics, 33(2) :774–805.
- [Müllner et al., 2013] Müllner, D. et al. (2013). fastcluster : Fast hierarchical, agglomerative clustering routines for r and python. Journal of Statistical Software, 53(9) :1–18.
- [Murray et al., 2021] Murray, B., Anderson, D. T., and Havens, T. C. (2021). Actionable xai for the fuzzy integral. In 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pages 1–8. IEEE.
- [Murray et al., 2020] Murray, B. J., Islam, M. A., Pinar, A. J., Anderson, D. T., Scott, G. J., Havens, T. C., and Keller, J. M. (2020). Explainable ai for the choquet integral. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(4) :520–529.
- [Murtagh, 1984] Murtagh, F. (1984). Complexities of hierarchic clustering algorithms : state of the art. Computational Statistics Quarterly, 1(2) :101–113.
- [Murtagh, 2004] Murtagh, F. (2004). On ultrametricity, data coding, and computation. Journal of classification, 21(2) :167–184.
- [Murtagh, 2017] Murtagh, F. (2017). Data Science Foundations : Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics. Chapman and Hall/CRC.
- [Murtagh and Contreras, 2012a] Murtagh, F. and Contreras, P. (2012a). Algorithms for hierarchical clustering : an overview. Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery, 2(1) :86–97.

- [Murtagh and Contreras, 2012b] Murtagh, F. and Contreras, P. (2012b). The future of search and discovery in big data analytics : Ultrametric information spaces. [arXiv preprint arXiv :1202.3451](#).
- [Newman, 2018] Newman, M. (2018). [Networks](#). Oxford university press.
- [Newman, 2006a] Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. [Physical review E](#), 74(3) :036104.
- [Newman and Girvan, 2004] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. [Physical review E](#), 69(2) :026113.
- [Newman, 2006b] Newman, M. E. J. (2006b). Finding community structure in networks using the eigenvectors of matrices. [Physical Review E](#), 74(3) :036104.
- [Ng et al., 2001] Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2001). On spectral clustering : Analysis and an algorithm. In [NIPS](#), volume 14, pages 849–856.
- [Nida, 1979] Nida, E. A. (1979). [A componential analysis of meaning : An introduction to semantic structures](#). De Gruyter.
- [Nie et al., 2016] Nie, F., Wang, X., Jordan, M. I., and Huang, H. (2016). The constrained laplacian rank algorithm for graph-based clustering. In [AAAI](#), pages 1969–1976. Citeseer.
- [Nothman et al., 2009] Nothman, J., Murphy, T., and Curran, J. R. (2009). Analysing wikipedia and gold-standard corpora for ner training. In [Proceedings of the 12th Conference of the European Chapter of the ACL \(EACL 2009\)](#), pages 612–620.
- [Park and Kim, 2017] Park, J. and Kim, T. (2017). Learning doubly stochastic affinity matrix via davis-kahan theorem. In [2017 IEEE International Conference on Data Mining \(ICDM\)](#), pages 377–384. IEEE.
- [Peng and Wei, 2007] Peng, J. and Wei, Y. (2007). Approximating k-means-type clustering via semidefinite programming. [SIAM journal on optimization](#), 18(1) :186–205.
- [Peng and Xia, 2005] Peng, J. and Xia, Y. (2005). A new theoretical framework for k-means-type clustering. In [Foundations and advances in data mining](#), pages 79–96. Springer.
- [Perovic et al., 2011] Perovic, A., Ognjanovic, Z., Raskovic, M., and Radojevic, D. G. (2011). Finitely additive probability measures on classical propositional formulas definable by Gödel’s t-norm and product t-norm. [Fuzzy Sets and Systems](#), 169(1) :65–90.
- [Preda, 2007] Preda, C. (2007). Regression models for functional data by reproducing kernel hilbert spaces methods. [Journal of statistical planning and inference](#), 137(3) :829–840.
- [Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In [International conference on machine learning](#), pages 8748–8763. PMLR.
- [Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

- [Rand, 1971] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336) :846–850.
- [Randriamihamison et al., 2021] Randriamihamison, N., Vialaneix, N., and Neuvial, P. (2021). Applicability and interpretability of ward’s hierarchical agglomerative clustering with or without contiguity constraints. Journal of Classification, 38(2) :363–389.
- [Robert and Escoufier, 1976] Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods : the rv-coefficient. Journal of the Royal Statistical Society Series C : Applied Statistics, 25(3) :257–265.
- [Rocchio Jr, 1971] Rocchio Jr, J. J. (1971). Relevance feedback in information retrieval. The SMART retrieval system : experiments in automatic document processing.
- [Rossi and Conan-Guez, 2006] Rossi, F. and Conan-Guez, B. (2006). Theoretical properties of projection based multilayer perceptrons with functional inputs. Neural Processing Letters, 23(1) :55–70.
- [Rossi et al., 2004] Rossi, F., Conan-Guez, B., and El Golli, A. (2004). Clustering functional data with the som algorithm. In ESANN, pages 305–312.
- [Rossi et al., 2005] Rossi, F., Delannay, N., Conan-Guez, B., and Verleysen, M. (2005). Representation of functional data in neural networks. Neurocomputing, 64 :183–210.
- [Rossi and Villa, 2006] Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. Neurocomputing, 69(7-9) :730–742.
- [Rossi and Villa-Vialaneix, 2011] Rossi, F. and Villa-Vialaneix, N. (2011). Consistency of functional learning methods based on derivatives. Pattern Recognition Letters, 32(8) :1197–1209.
- [Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. science, 290(5500) :2323–2326.
- [Russell and Upton, 1987] Russell, A. M. and Upton, C. J. F. (1987). A class of positive semidefinite matrices. Linear Algebra and its Applications, 93 :121–126.
- [Salton and Buckley, 1990] Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. Journal of the American society for information science, 41(4) :288–297.
- [Sang and De Meulder, 2003] Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. arXiv preprint cs/0306050.
- [Sarkar and Bickel, 2015] Sarkar, P. and Bickel, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. The Annals of Statistics.
- [Schweizer and Sklar, 1958] Schweizer, B. and Sklar, A. (1958). Espaces métriques aléatoires. Comptes Rendus Hebdomadaires des Séances de l’Académie des Sciences, 247(23) :2092–2094.

- [Schweizer and Sklar, 1983] Schweizer, B. and Sklar, A. (1983). Probabilistic metric spaces. North-Holland.
- [Shang, 2014] Shang, H. L. (2014). A survey of functional principal component analysis. AStA Advances in Statistical Analysis, 98 :121–142.
- [Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8) :888–905.
- [Sinkhorn, 1968] Sinkhorn, R. (1968). Two results concerning doubly stochastic matrices. The American Mathematical Monthly, 75(6) :632–634.
- [Sinkhorn and Knopp, 1967] Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics, 21(2) :343–348.
- [Sleeman IV et al., 2022] Sleeman IV, W. C., Kapoor, R., and Ghosh, P. (2022). Multimodal classification : Current landscape, taxonomy and future directions. ACM Computing Surveys, 55(7) :1–31.
- [Soriano-Morales et al., 2017] Soriano-Morales, E., Ah-Pine, J., and Loudcher, S. (2017). Fusion techniques for named entity recognition and word sense induction and disambiguation. In Discovery Science - 20th International Conference, DS 2017, Kyoto, Japan, October 15-17, 2017, Proceedings, pages 340–355.
- [Soriano-Morales, 2018] Soriano-Morales, E.-P. (2018). Hypergraphes et fusion d’information pour l’enrichissement de la représentation de termes. PhD thesis, Université Lumière Lyon 2.
- [Srivastava et al., 2011] Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011). Registration of functional data using fisher-rao metric. arXiv preprint arXiv :1103.3817.
- [Strubell et al., 2018] Strubell, E., Verga, P., Andor, D., Weiss, D., and McCallum, A. (2018). Linguistically-informed self-attention for semantic role labeling. arXiv preprint arXiv :1804.08199.
- [Summaira et al., 2021] Summaira, J., Li, X., Shoib, A. M., Li, S., and Abdul, J. (2021). Recent advances and trends in multimodal deep learning : A review. arXiv preprint arXiv :2105.11087.
- [Takács, 1967] Takács, L. (1967). On the method of inclusion and exclusion. Journal of the American Statistical Association, 62(317) :102–113.
- [Tarpey and Kinateder, 2003] Tarpey, T. and Kinateder, K. K. (2003). Clustering functional data. Journal of classification, 20(1) :093–114.
- [Tenenbaum et al., 2000] Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. science, 290(5500) :2319–2323.
- [Torgerson, 1952] Torgerson, W. (1952). Multidimensional scaling : I. theory and method. Psychometrika, 17 :401–419.

- [Tucker et al., 2019] Tucker, J. D., Lewis, J. R., and Srivastava, A. (2019). Elastic functional principal component regression. Statistical Analysis and Data Mining : The ASA Data Science Journal, 12(2) :101–115.
- [Tversky and Kahneman, 1992] Tversky, A. and Kahneman, D. (1992). Advances in prospect theory : Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5 :297–323.
- [Tzortzis and Likas, 2012] Tzortzis, G. and Likas, A. (2012). Kernel-based weighted multi-view clustering. In 2012 IEEE 12th international conference on data mining, pages 675–684. IEEE.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
- [Villmann, 2007] Villmann, T. (2007). Sobolev metrics for learning of functional data - mathematical and theoretical aspects. Machine Learning Reports, Research group on Computational Intelligence.
- [Von Luxburg, 2007] Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing, 17(4) :395–416.
- [Von Luxburg et al., 2008] Von Luxburg, U., Belkin, M., and Bousquet, O. (2008). Consistency of spectral clustering. The Annals of Statistics, pages 555–586.
- [Wang et al., 2016] Wang, X., Nie, F., and Huang, H. (2016). Structured doubly stochastic matrix for graph based clustering. In Proceedings of the 22nd ACM SIGKDD International conference on Knowledge discovery and data mining, pages 1245–1254.
- [Yager, 1988] Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Transactions on systems, Man, and Cybernetics, 18(1) :183–190.
- [Yager, 2005] Yager, R. R. (2005). Extending multicriteria decision making by mixing t-norms and OWA operators. Int. J. Intell. Syst., 20(4) :453–474.
- [Zass and Shashua, 2005] Zass, R. and Shashua, A. (2005). A unifying approach to hard and probabilistic clustering. In Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, volume 1, pages 294–301. IEEE.
- [Zass and Shashua, 2007] Zass, R. and Shashua, A. (2007). Doubly stochastic normalization for spectral clustering. In Advances in neural information processing systems, pages 1569–1576.
- [Zha et al., 2002] Zha, H., He, X., Ding, C., Gu, M., and Simon, H. D. (2002). Spectral relaxation for k-means clustering. In Advances in neural information processing systems, pages 1057–1064.
- [Zheng et al., 2021] Zheng, Y., Xu, Z., and Wang, X. (2021). The fusion of deep learning and fuzzy systems : A state-of-the-art survey. IEEE Transactions on Fuzzy Systems, 30(8) :2783–2799.

BIBLIOGRAPHIE
