Apports de la théorie des possibilités et des fonctions de croyance à l'analyse de données imprécises

Marie-Hélène Masson

Mémoire présenté en vue de l'obtention du diplôme
d'Habilitation à Diriger des Recherches

Habilitation à diriger des Recherches soutenue le 2 décembre 2005
devant le jury composé de :

Isabelle Bloch, Professeur, Ecole Nat. Sup. des Télécommunication de Paris, (rapporteur)
Didier Dubois, Directeur de Recherches CNRS, Université Paul Sabatier, (rapporteur)
Laurent Foulloy, Professeur, Université de Savoie, (rapporteur)
Thierry Denoeux, Professeur, Université de Technologie de Compiègne, (examinateur)
Bernard Dubuisson, Université de Technologie de Compiègne, (examinateur)

Remerciements

Je tiens à remercier Isabelle Bloch, Professeur à l'ENST Paris, Didier Dubois, directeur de recherches à l'Université Paul Sabatier, et Laurent Foulloy, Professeur à l'Université de Savoie, pour m'avoir fait l'honneur d'accepter de rapporter sur ce travail. Leurs remarques éclairées, toujours constructives, m'incitent à continuer dans la voie de recherche que j'ai choisie.

Je remercie également Bernard Dubuisson, Professeur à l'Université de Technologie de Compiègne, qui a été à l'origine de mes premiers pas dans le monde de la recherche et qui n'a cessé depuis de m'encourager.

J'adresse de chaleureux remerciements à Thierry Denoeux avec qui je collabore depuis de nombreuses années. Il a su, en me témoignant sa confiance et son amitié, m'apprendre mon métier d'enseignant-chercheur.

Je citerai également dans ces remerciements l'équipe Perception et Facteurs Humains du groupe PSA, dirigée par Anne Bardot, qui nous a permis depuis de nombreuses années de nourrir nos recherches théoriques par des applications concrètes originales.

Je n'oublie pas dans ces remerciements toutes les personnes qui facilitent le travail au quotidien dans le laboratoire Heudiasyc : les secrétaires du génie informatique ainsi que la division logistique d'Heudiasyc.

Une petite pensée pour mon compère de bureau Yves ainsi qu'à d'autres collègues croisés au détour des chemins : Véronique, Gérard, Christophe, Stéphane…

Enfin, merci à la famille et les copains pour leur soutien sans faille !

Table des matières

Introduction

La statistique est une branche des mathématiques fondée sur le recueil et l'analyse de données mesurées sur un échantillon d'une population. L'objectif fondamental de la statistique est d'extrapoler des résultats observés sur l'échantillon à l'ensemble de la population. Cette démarche inductive est appelée *l'inférence statistique*. Dans une étape préliminaire, il est souvent nécessaire de simplifier l'échantillon en fournissant des représentations graphiques synthétiques interprétables, en résumant les données grâce à des indicateurs numériques, ou encore en cherchant des structures naturelles dans les données. C'est le rôle de la *statistique descriptive* ou *exploratoire* sur laquelle porte la majeure partie de ce mémoire.

La plupart des méthodes d'analyse de données ont été développées pour traiter des données numériques classiques, validées, complètes, spécialement préparées en vue d'une analyse statistique. En pratique cependant, le statisticien doit de plus en plus faire face à des données dont la nature et le format ne correspondent pas au schéma classique. Les raisons en sont multiples. On assiste d'une part au développement exponentiel des moyens d'enregistrement et de stockage de données. Celles-ci ne sont plus en conséquence systématiquement validées et formatées pour une analyse classique. D'autre part, il semble illusoire de considérer une absolue précision des données. Tout système de mesure a ses limites en termes de précision. Plutôt que de masquer le problème (en ne considérant que la valeur moyenne par exemple), il peut paraître intéressant au contraire d'intégrer la connaissance que l'on a de l'imprécision dans l'analyse. Les descriptions linguistiques, vagues par essence, fournies par des experts sont un autre exemple de données difficilement analysables de façon classique.

On voit ainsi depuis quelques années se développer un intérêt croissant pour des méthodes d'analyse de données imparfaites, exprimées sous forme vague ou imprécise. Nous présentons dans ce mémoire notre contribution dans ce domaine.
Le premier chapitre a pour objectif de passer en revue différents cadres théoriques de représentation et de manipulation de données imparfaites : sous-ensembles flous, probabilités, possibilités et fonctions de croyance. Puis, dans les chapitres qui suivent, seront présentés les outils que nous avons développés pour l'aide à l'analyse de données imprécises. Le chapitre 2 présente une méthode originale de calcul de corrélation basée sur une extension du coefficient de corrélation ordinale de Kendall. Les chapitres 3 et 4 sont consacrés aux méthodes de visualisation de données multidimensionnelles (positionnement multidimensionnel et analyse en composantes principales). Le chapitre 5 propose une méthode de classification automatique, présentée d'abord dans le cas de données classiques, puis étendue à des données floues. Le chapitre 6 s'attaque au problème du lien entre deux cadres théoriques de représentation, les probabilités et les possibilités. Enfin, la conclusion évoque les voies que nous comptons explorer dans les années à venir.

Chapitre

1

Incertitude, imprécision

1 Nulle donnée n'est parfaite !

Dans le domaine de l'analyse de données et de la reconnaissance des formes, nous manipulons des informations, le plus souvent numériques, qui sont censées donner une image aussi fidèle que possible de la réalité. Or, le plus souvent, ces informations sont imparfaites : imprécises, incertaines, vagues, incomplètes,... Plusieurs auteurs, parmi lesquels Bonissone et Tong [10], Bosc et al [12], Smets [93], se sont sont attachés à analyser précisemment les différentes formes d'imperfection dans une donnée. Nous en donnons ici une présentation rapide ainsi qu'un schéma de synthèse en figure 1.1. On peut décomposer l'imperfection dans les données en trois catégories (non exclusives) : l'*incertitude*, l'*inconsistance* et l'*imprécision*, chacune pouvant se décliner en plusieurs sous-catégories.
Pour illustrer ces notions, prenons l'exemple d'un supporter constituant une base de données portant sur le championnat de foot de Ligue 1, avec comme

Fig. 1.1 – Différentes formes d'imperfection.

variables à renseigner, le nom de l'équipe, la dernière équipe rencontrée, le vainqueur du match et la différence de buts. L'*incomplétude* fait référence à l'absence d'information : par exemple, le supporter ne dispose pas du dernier résultat de Lens. L'information sera considérée comme *imprécise* si le supporter sait que l'équipe de Lens a rencontré soit le PSG soit Marseille, ou encore que la différence de but était inférieure à 3. Dans le premier cas, la variable, qualitative, n'est connue que pour appartenir à un ensemble de valeurs possibles, tout autre valeur étant exclue. Le second cas fait référence à une variable quantitative dont la valeur appartient à un certain intervalle. Notons que l'incomplétude peut être considérée comme un cas particulier de l'imprécision. Autre forme d'imprécision, la caractère *vague* ou *flou* d'une donnée, se manifeste généralement lorsque la donnée est exprimée sous forme linguistique. Par exemple, Lens a battu le PSG sur une différence de but "faible". Cette information est considérée comme vague car la signification d'une différence de but faible peut varier d'un individu à l'autre et fait donc référence à un ensemble de valeurs dont les contours ne sont pas précisemment définis.
L'*incertitude* fait référence à la véracité de l'information. Par exemple si l'information manque au supporter, il peut penser à sélectionner un ami au hasard et l'interroger sur le score d'un match, quelle que soit sa compétence dans le domaine du ballon rond. L'information peut alors être complète, précise mais fausse. Certains auteurs choisissent de distinguer deux types d'incertitude : l'incertitude *objective* que l'on peut assimiler à de l'aléa (on dispose d'un système de mesure sujet à une certaine variabilité, par exemple la technique d'échantillonnage précédente), et l'incertitude *subjective* essentiellement liée au crédit que l'on accorde à la source qui fournit l'information (le supporter sait que la personne qui lui a fournit l'information n'est pas fiable). A ces deux termes, on peut préférer les notions d'incertitude *probabiliste* (quelles sont les chances de succès de Lens au prochain match), *possibiliste* (la possibilité que la différence de but excède 5) ou *crédibiliste* (ma propre croyance dans le fait que Lens va gagner).
Enfin, à l'incertitude et l'imprécision, Bosc et al [12] proposent d'ajouter *l'inconsistance* qui survient, en présence de redondance, lorsque plusieurs informations sont en conflit.
On le voit, l'imperfection dans les données peut prendre plusieurs formes non exclusives l'une de l'autre. Pendant longtemps, on a considéré que le cadre probabiliste était le seul cadre adapté à la représentation et à la manipulation de données imparfaites. Dans les trente dernières années, d'autres théories de gestion de l'imprécis et l'incertain on vu le jour, en raison notamment du constat que réduire l'imperfection d'une donnée à son caractère aléatoire était loin d'être satisfaisant. Les paragraphes qui suivent évoquent les bases de ces différentes théories, leurs liens et le traitement de données imprécises dans la littérature statistique.

2 Quantifier et manipuler l'imprécision des mesures

2.1 Sous-ensembles flous

Nous pensons que le cadre le plus adapté à la représentation et la manipulation de mesures imprécises est celui des ensembles flous. Les concepts d'ensembles flous et de logique floue ont été introduits par Zadeh [103]. L'idée de Zadeh était de pouvoir manipuler des informations exprimées en langage naturel. Cet objectif nécessitait d'étendre la théorie des ensembles et la logique propositionnelle classiques. Un sous-ensemble A d'un référentiel Ω est classiquement défini par les objets qui le composent : un objet x appartient ou n'appartient pas à l'ensemble en question. La proposition logique associée "l'objet x appartient à l'ensemble A" est soit vraie, soit fausse. Ce concept d'ensemble classique a été étendu à celui d'ensemble flou grâce à l'idée d'appartenance partielle ou de vérité partielle. Un objet x est alors membre d'un *sous-ensemble flou* A avec un certain *degré d'appartenance* $\mu_A(x)$. La *fonction d'appartenance* $\mu_A : \Omega \rightarrow [0, 1]$ attribue à chaque élément du référentiel Ω un degré d'appartenance. Le *noyau* de A est défini comme l'ensemble classique (ou net) des éléments x appartenant totalement à A (c'est-à-dire pour lesquels $\mu_A(x) = 1$), le *support* comme l'ensemble net des éléments ayant un degré d'appartenance non nul. Par ailleurs, on définit l'*α-coupe* de A (ou la coupe de niveau α), A_α, comme l'ensemble net des éléments ayant une appartenance supérieure à α :

$$A_\alpha = \{x \in \Omega | \mu_A(x) \geq \alpha\}, \tag{1.1}$$

l'α-coupe stricte, notée A_{α^+}, faisant référence à l'ensemble construit à partir de l'inégalité stricte.

Exemple 1.1 Considérons une personne chargée de d'évaluer la température d'une pièce. Plutôt que d'exprimer la température de façon précise, elle peut être tentée de fournir l'information sous une forme naturelle comme "la température est élevée". Pour pouvoir manipuler cette information par la suite, il est nécessaire d'expliciter sur l'échelle des températures ce que représente l'ensemble des températures de niveau élevé. On définit donc sur le référentiel des températures possibles, par exemple $\Omega = [0; 40]$, le sous-ensemble flou des "températures élevées". Sa représentation est donnée en figure 1.2. Pour construire la fonction d'appartenance μ_{Eleve}, on a considéré ici qu'en dessous de 15°C, la température n'est certainement pas élevée, d'où un degré d'appartenance nul, et que les températures au dessus de 25°C sont assurément élevées, d'où un degré d'appartenance maximum. Entre ces deux zones de températures, l'appartenance passe graduellement de 0 à 1. On voit ici qu'une part d'arbitraire est intervenue dans la construction de la fonction d'appartenance. La détermination de fonctions d'appartenance adéquates constitue,

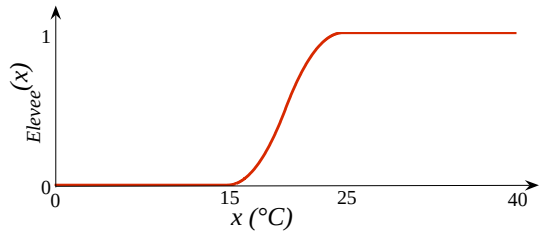

FIG. 1.2 – Fonction d'appartenance de l'ensemble flou des températures élevées.

quelle que soit l'application envisagée, un problème pratique important qui conditionne bien souvent les résultats qu'on peut attendre d'une approche floue. On pourra se reporter à [78, page 19] et [7] pour une description des principales méthodes expérimentales de construction.

La représentation numériques de concepts vagues étant posée, il reste à étendre les opérations classiques d'intersection, d'union et de complémentation sur les ensembles et leur équivalent logique de conjonction, disjonction et complémentation. Soient A et B deux sous-ensembles flous d'un même référentiel. On pose :

$$\mu_{A\cap B}(x) = \min\left(\mu_A(x), \mu_B(x)\right) = \mu_A(x) \wedge \mu_B(x) \quad (1.2)$$

$$\mu_{A\cup B}(x) = \max\left(\mu_A(x), \mu_B(x)\right) = \mu_A(x) \vee \mu_B(x) \quad (1.3)$$

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x) \quad (1.4)$$

REMARQUE 1 Ces définitions ne sont pas uniques. Les opérateurs min et max peuvent être remplacés par des opérateurs faisant respectivement partie de la famille des t-normes (applications t de $[0,1]^2$ dans [0,1] commutatives, associatives, non décroissantes en chaque argument et telles que $t(x,1) = x \quad \forall x \in [0,1]$) et des t-conormes (applications s de $[0,1]^2$ dans [0,1] commutatives, associatives, non décroissantes en chaque argument et telles que $s(x,0) = x \quad \forall x \in [0,1]$).

2.2 Principe d'extension

Le principe d'extension, proposé à l'origine par Zadeh, est un des outils fondamentaux de la théorie des sous-ensembles flous. Il permet d'étendre des relations fonctionnelles classiques à des quantités floues. Soit une application f d'un univers X vers un univers Y. Soit A un sous-ensemble flou défini sur X. Le principe d'extension stipule que l'image par f de A, $f(A)$, est un sous-ensemble flou de Y dont la fonction d'appartenance est définie par :

$$\mu_B(y) = \sup_{x|y=f(x)} \mu_A(x) \quad (1.5)$$

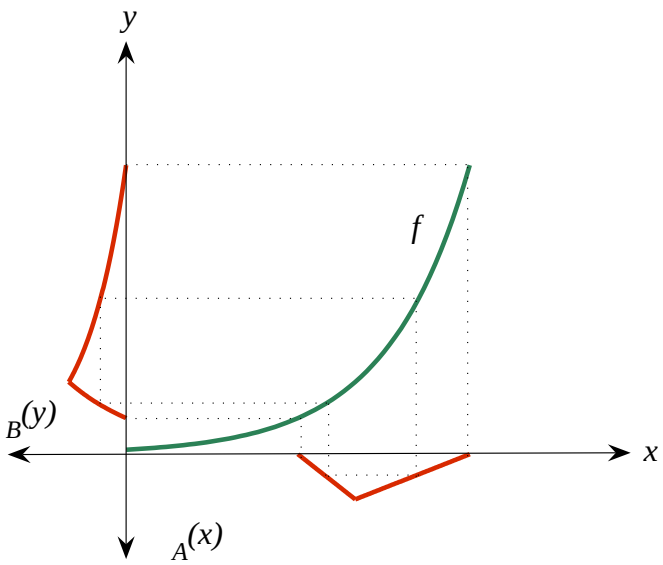

FIG. 1.3 – Principe d'extension.

Ce principe est ilustré en figure 1.3. La généralisation à des fonctions de plusieurs variables est la suivante : soient $X_i \quad i = 1, n$, n univers, X le produit Cartésien des X_i, et f une fonction de X vers Y. L'image des sous-ensembles flous A_1, A_2,...,A_n de X_1,X_2,...,X_n par f est donnée par :

$$\mu_B(y) = \sup_{\substack{(x_1,x_2,...,x_n)\in X_1\times X_2\times...\times X_n \\ y=f(x_1,x_2,...,x_n)}} \min\left[\mu_{A_1}(x_1), \mu_{A_2}(x_2), ..., \mu_{A_n}(x_n)\right] \quad (1.6)$$

Notons que le principe d'extension est compatible avec les α-coupes strictes :

$$A_{\alpha^+} = f\left((A_1)_{\alpha^+}, ..., (A_n)_{\alpha^+}\right), \quad (1.7)$$

et avec les α-coupes non strictes si le sup est atteint dans (1.6).

2.3 Sous-ensembles flous de $\mathbb{R}$: nombres et arithmétique flous

Dubois et Prade définissent dans leur ouvrage [39] différentes notions autour des données floues. En voici les principales :

Une *quantité floue* est un sous-ensemble flou de $\mathbb{R}$. Tout nombre réel appartenant au noyau d'une quantité floue A est appelée une *valeur modale* de A.

Un *intervalle flou* généralise la notion classique d'intervalle. Il s'agit d'une quantité floue convexe. Une quantité floue est convexe si et seulement si ses α-coupes sont des intervalles (fermés ou non). Les intervalles fermés sont étendus

à des intervalles flous dont les fonctions d'appartenance sont semi-continues supérieurement (c'est-à-dire dont les α-coupes sont des intervalles fermés).

Un *nombre flou*, comme l'ont défini Dubois et Prade, est un intervalle flou à support compact ne possédant qu'une seule valeur modale. Par abus de langage, nous emploierons par la suite de terme de nombre flou pour désigner des intervalles flous. Pour rendre plus simple et plus efficace leur manipulation, certaines classes de nombres flous ont été définies à l'aide d'une représentation paramétrique dite "*L-R*". On se donne deux fonctions de forme, L (Left) et R (Right), de $\mathbb{R}^+$ dans [0,1], symétriques, non décroissantes sur $[0;+\infty[$, telles que $L(0) = R(0) = 1$.
Un nombre flou, noté $(a^-, a^+, \gamma, \beta)_{LR}$ est alors défini de la manière suivante :

$$\mu_A(x) = \begin{cases} L\left(\dfrac{a^- - x}{\gamma}\right) & \text{if } x \leq a^- \\ 1 & \text{if } a^- \leq x \leq a^+ \\ R\left(\dfrac{x - a^+}{\beta}\right) & \text{if } x \geq a^+ \end{cases} \tag{1.8}$$

Les fonctions L et R les plus courantes sont des fonctions exponentielles ou linéaires qui permettent de construire des nombres flous gaussiens, triangulaires ou trapézoïdaux. Si $L = R$, on parle alors de nombres LL. L'intérêt de la famille de nombres LL est d'être close vis-à-vis des opérations d'addition, de soustraction et de multiplication par un réel. Soit $A = (a^-, a^+, \gamma_a, \beta_a)_{LL}$ et $B = (b^-, b^+, \gamma_b, \beta_b)_{LL}$ deux nombres flous LL. Sur la base du principe d'extension, on peut établir les règles suivantes :

- *addition :* $A \oplus B = (a^- + b^-, a^+ + b^+, \gamma_a + \gamma_b, \beta_a + \beta_b)_{LL}$
- *soustraction :* $A \ominus B = (a^- - b^+, a^+ - b^-, \gamma_a + \beta_b, \beta_a + \gamma_b)_{LL}$
- *multiplication par* $\lambda \in \mathbb{R}^+$ *:* $\lambda A = (\lambda a^-, \lambda a^+, \lambda \gamma_a, \lambda \beta_a)_{LL}$

3 Quantifier et manipuler l'incertitude

La gestion de l'incertain dans les systèmes d'information se fait classiquement en attachant une mesure de certitude aux éléments manipulés. La manière dont ces valeurs sont ensuite utilisées dépend du cadre théorique choisi. Nous présentons dans ce paragraphe les principaux cadres de représentation de l'incertain ainsi que quelques éléments de comparaison. Les notions de base seront évoquées, des notions plus pointues seront données, lorsqu'elles sont nécessaires, dans les chapitres qui suivent.

3.1 Probabilités

La notion de probabilité est liée à celle de d'expérience aléatoire. Une expérience est aléatoire si l'on ne peut pas prédire avec certitude son résultat. Le résultat d'une expérience aléatoire est un élément ω de l'ensemble Ω de tous les résultats possibles, appelé univers des possibles ou référentiel. On note $\mathcal{P}(\Omega)$ l'ensemble des parties de Ω. Un événement, liée à une expérience aléatoire, est une proposition logique relative au résultat de l'expérience, il est choisi dans un ensemble d'événements $\mathcal{A}$, sous-ensemble de $\mathcal{P}(\Omega)$. Si A et B désigne deux éléments de $\mathcal{A}$, alors :
– $A \cup B$ désigne la réalisation de A *ou* B
– $A \cap B$ désigne la réalisation de A *et* B
– $\bar{A} = \Omega \backslash A$ désigne le contraire de A.
D'autre part,
– Ω est l'événement certain
– $\emptyset$ est l'événement impossible
Lorsque le référentiel Ω est fini, $\mathcal{A}$ regroupe toutes les parties de Ω, noté habituellement 2^{Ω}. Lorsque le référentiel est $\mathbb{R}$ ou un intervalle de $\mathbb{R}$, on fait appel pour définir $\mathcal{A}$ à la notion de tribu. Une tribu est définie de la manière suivante :

Definition 1
$\mathcal{A}$ est une tribu sur Ω si et seulement si $\mathcal{A}$ est un ensemble de parties de Ω contenant l'ensemble vide, stable par passage au complémentaire et par union et intersection d'une suite finie ou dénombrable d'éléments :

1. $\mathcal{A} \subseteq \mathcal{P}(\Omega)$
2. $\emptyset \in \mathcal{A}$
3. *Si l'on dispose d'une suite $A_1, ..., A_n$ d'éléments de $\mathcal{A}$, alors leur réunion et leur intersection $\cup_i A_i$ et $\cap_i A_i$ sont aussi dans $\mathcal{A}$.*
4. *si $A \in \mathcal{A}$ alors $\bar{A}$ est aussi dans $\mathcal{A}$.*

La tribu définie sur $\mathbb{R}$, ou tribu Borélienne, est la tribu engendrée par des intervalles de $\mathbb{R}$. Le couple $(\Omega, \mathcal{A})$ est appelé un espace mesurable.

3.1.1 Les axiomes des probabilités

Soit $(\Omega, \mathcal{A})$ un espace mesurable. Une *mesure de probabilité* est une fonction de $\mathcal{A}$ dans [0,1] telle que :
– $P(\Omega) = 1$;
– quels que soient deux événements A et B incompatibles ($A \cap B = \emptyset$), on a $P(A \cup B) = P(A) + P(B)$.

Le nombre $P(A)$ quantifie dans quelle mesure l'événement $A \subseteq \Omega$ est probable. En lien avec P, dans le cas où Ω est fini, on définit une *distribution de probabilité* comme une fonction p de Ω dans $[0, 1]$ telle que :

$$P(A) = \sum_{\omega \in A} p(\omega) \quad \forall A \subseteq \Omega, \tag{1.9}$$

avec la condition de normalisation :

$$\sum_{\omega \in \Omega} p(\omega) = 1. \tag{1.10}$$

On montre à partir des axiomes de base que

$$P(A) + P(\bar{A}) = 1. \tag{1.11}$$

Cette équation montre que la connaissance de la probabilité de A définit complètement celle de son événement contraire.

3.1.2 Vers d'autres théories de l'incertain

Depuis son avènement au 17ème siècle, on a donné au terme de *probabilité* plusieurs interprétations [93]. Une des premières interprétations, défendue par Laplace, repose sur le principe dit de *raison insuffisante* qui, en l'absence de connaissances, accorde à chaque événement d'une expérience aléatoire une probabilité équivalente. Ce principe conduit parfois à des conclusions contraires au sens commun ou différentes suivant le mode de raisonnement adopté comme le célèbre paradoxe de Bertrand [84]. Une autre interprétation usuelle est l'interprétation *fréquentiste* qui voit une probabilité comme la limite d'une fréquence d'apparition d'un événement lorsque l'expérience est répétée un grand nombre de fois. Ce point de vue est critiqué par les *subjectivistes* qui arguent du fait qu'une expérience aléatoire n'est pas toujours répétable et qu'on peut attacher à un événement, en dehors de toute notion de répétition, une valeur *subjective* quantifiant notre croyance dans le fait que l'événement se produise. La notion de probabilité n'existe pas en elle-même, cette mesure d'incertitude peut varier suivant les circonstances ou l'observateur. Ils proposent d'interpréter la probabilité d'un événement comme le montant qu'un individu serait prêt à payer si l'événement contraire se produisait. Pourvu que le comportement de l'individu soit rationnel, on montre alors que la mesure d'incertitude ainsi définie obéit aux axiomes des probabilités. Dubois et Prade [39] critiquent cette vision des choses tant au plan philosophique (peut on ramener toute situation d'incertitude à un pari ?) que d'un point de vue pratique. Un individu peut avoir des difficultés à décrire de manière précise son état de connaissance. Il peut paraître plus naturel de fournir un ensemble de

valeurs possibles. Cette caractérisation de l'incertitude non pas par un nombre unique mais par plusieurs valeurs ouvre la voie à d'autre théories de gestion de l'incertain qui sont décrites dans ce qui suit.

3.2 Fonctions de croyance

La théorie des fonctions de croyance a été développée par Shafer en 1976 à la suite des travaux de Dempster sur les probabilités inférieures et supérieures. Philippe Smets [89, 95, 91, 92, 94] a ensuite énormément contribué au développement de cette théorie grâce à son modèle des croyances transférables (appelé aussi TBM pour *Transferable Belief Model*), dérivé de celui de Shafer. En voici les grandes lignes. Soit Ω un référentiel fini[1]. Une connaissance imparfaite sur Ω est représentée par une masse de croyance (en anglais *basic belief assignment* ou bba) [87, 95], définie comme une fonction de 2^Ω dans $[0, 1]$, vérifiant :

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1.12}$$

Les éléments A de Ω tels que $m(A) > 0$ sont appelés des *éléments focaux* de m. Le nombre $m(A)$ peut être interprété comme la fraction de la masse unité allouée à A sur la base de notre état de connaissance. A la différence des probabilités, on voit qu'il est possible d'allouer de la masse à des sous-ensembles de Ω et non uniquement à des singletons. Cette possibilité fournit au modèle une grande souplesse de représentation. Il est en effet possible de modéliser des connaissances précises ou imprécises, certaines ou incertaines de manière très naturelle. L'ignorance complète correspond à $m(\Omega) = 1$, alors qu'une connaissance précise et sûre correspond à l'attribution de la totalité de la masse à un singleton de Ω (m est alors appelée une *masse certaine*). Une connaissance imprécise et sûre se traduira par l'allocation de la masse unité à un élément focal non singleton. Une connaissance incertaine correspondra à l'allocation de fractions de la masse à plusieurs éléments focaux.

Un autre intérêt de la modélisation réside dans la notion de *monde ouvert* qui est mal appréhendée par la théorie classique des probabilités. Une bba m telle que $m(\emptyset) = 0$ est dite normale. Cette condition a été à l'origine imposée par Shafer [87], mais elle peut être relâchée si l'on accepte l'hypothèse de monde ouvert qui pose que Ω peut être incomplet [89]. La quantité $m(\emptyset)$ est alors interprétée comme la part de croyance dans le fait que la vérité se trouve ailleurs que dans Ω. On verra dans le chapitre 5 l'intérêt de cette modélisation. Une bba m peut être de façon équivalente représentée par deux mesures non

[1] Même si la théorie existe dans le cas continu, elle est plus difficilement manipulable et sera donc présentée uniquement dans le cas fini

additives : une fonction de *crédibilité* $\text{bel} : 2^\Omega \mapsto [0,1]$, définie par

$$\text{bel}(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \quad \forall A \subseteq \Omega \ , \tag{1.13}$$

et une fonction de *plausibilité* $\text{pl} : 2^\Omega \mapsto [0,1]$, définie par

$$\text{pl}(A) = \text{bel}(\Omega) - \text{bel}(\overline{A}) \quad \forall A \subseteq \Omega \ . \tag{1.14}$$

Chaque masse $m(A)$ quantifie la part de croyance allouée exactement à A et non la croyance totale en A représentée par $\text{bel}(A)$. Celle-ci est calculée comme la somme des masses attribués aux éléments *impliquant* A. La plausibilité de A est quant à elle calculée comme la somme des masses des éléments ne *contredisant* pas A. Le couple $(\text{bel}(A), \text{pl}(A))$ caractérise la connaissance que l'on a de A.

Une des pierres angulaires de la théorie des fonctions de croyance est la possibilité de combiner deux sources de croyance m_1 et m_2. Une manière standard de combiner ces deux masses est l'opération de somme conjonctive définie par :

$$(m_1 m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \ , \tag{1.15}$$

for all $A \subseteq \Omega$. La quantité

$$K = (m_1 m_2)(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{1.16}$$

est appelée le *degré de conflit* entre m_1 et m_2. Elle peut être vue comme de degré de désagrément entre les deux sources. Si nécessaire, la condition de normalité $m(\emptyset) = 0$ peut être rétablie en divisant chaque masses $(m_1 m_2)(A)$ par $1 - K$ (cette opération est appelée la normalisation de Dempster). L'opération résultante, appelée règle de combinaison de Dempster [87] est alors :

$$(m_1 \oplus m_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C) \ . \tag{1.17}$$

Les deux opérations et $\oplus$ sont commutative et associative, deux propriétés désirables pour un opérateur de combinaison. Notons cependant que l'usage de ces règles de combinaison nécessite que les sources soient distinctes et fiables[2]. Une règle plus prudente est celle de la somme disjonctive $\textcircled{\cup}$ définie par :

$$(m_1 \textcircled{\cup} m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C) \ , \tag{1.18}$$

[2]se reporter à [90] pour une discussion approfondie sur ces notions.

qui est justifiée, en particulier, lorsque l'on sait qu'au moins une des sources est fiable.
Dans le modèle des croyances transférables, Smets choisit de distinguer deux niveaux : le niveau crédal où les masses sont affectées et combinées et le niveau pignistique où les décisions sont prises. On revient alors aux probabilités en recourant au concept de *probabilité pignistique* [95] défini, pour une masse m normalisée, par :

$$BetP(A) = \sum_{\emptyset \neq B \subseteq \Omega} m(B) \frac{|A \cap B|}{|B|}. \tag{1.19}$$

3.3 Possibilités

La théorie des possibilités [39, 106] est présentée comme un cadre alternatif pour représenter des informations incertaines. Elle est étroitement liée à la théorie des sous-ensembles flous présentée précédemment. Soit $(\Omega, \mathcal{A})$ un espace mesurable. Une *mesure de possibilité* Π est une fonction de $\mathcal{A}$ dans [0,1] telle que :

$$\Pi(\emptyset) = 0 \tag{1.20}$$

$$\Pi(A \cup B) = \max(\Pi(A), \Pi(B)). \tag{1.21}$$

De plus, si $\Pi(\Omega) = 1$, la mesure de possibilité est dite *normale*. Le nombre $\Pi(A)$ quantifie dans quelle mesure l'événement $A \subseteq \Omega$ est possible. Notons que les mesures de possibilité satisfont la relation :

$$\max(\Pi(A), \Pi(\bar{A})) = 1, \tag{1.22}$$

où $\bar{A}$ dénote le complément de A. Cette équation traduit le fait que, de deux événements contraires, l'un au moins est possible, la possibilité de l'un n'impliquant pas l'impossibilité de l'autre. Dans le cas où Ω est fini, une mesure de possibilité Π est caractérisée par une *distribution de possibilité*, c'est-à-dire une fonction $\pi : \Omega \to [0, 1]$ telle que :

$$\Pi(A) = \sup_{\omega \in A} \pi(\omega) \quad \forall A \subseteq \Omega. \tag{1.23}$$

Notons qu'une distribution de possibilité peut être vue comme la fonction d'appartenance de l'ensemble flou des éléments possibles pour la solution cherchée. Cette distribution de possibilité définit une mesure duale dite de *nécessité* :

$$N(A) = \inf_{\omega \notin A} (1 - \pi(\omega)) \quad \forall A \subseteq \Omega, \tag{1.24}$$

qui est liée à la mesure de possibilité par la relation suivante :

$$\Pi(A) = 1 - N(\bar{A}). \tag{1.25}$$

ω	1	2	3	4	5	6	7	8
$\pi(\omega)$	1	1	1	1	0.8	0.6	0.4	0.2
$p(\omega)$	0.1	0.8	0.1	0	0	0	0	0

TAB. 1.1 – Possibilités et probabilités associées aux nombres d'oeufs

Les mesures de nécessité satisfont la relation suivante :

$$\min(N(A), N(\bar{A})) = 0. \tag{1.26}$$

De plus, on a :

$$N(A) > 0 \Rightarrow \Pi(A) = 1 \tag{1.27}$$

$$\Pi(A) < 1 \Rightarrow N(A) = 0 \tag{1.28}$$

L'incertitude d'un événement A, au contraire des probabilités, est donc caractérisée par deux valeurs : sa possibilité $\Pi(A)$ et sa nécessité $N(A)$. L'interprétation d'un degré de possibilité est très différente de celle d'une probabilité. A titre illustratif, reprenons l'exemple célèbre de Zadeh [106] du petit déjeuner de Hans. On suppose connues les valeurs de possibilité et de probabilité concernant le nombre d'oeufs que Hans mangera demain. Elles sont données dans le tableau 1.1. On observe que la possibilité que Hans mange trois oeufs est de 1 alors que la probabilité n'est que de 0.1. On voit donc qu'un fort de degré de possibilité n'implique pas un fort degré de probabilité, et qu'un faible degré de probabilité n'est pas synonyme d'un faible degré de possibilité. Seulement peut-on dire qu'un degré de possibilité nul implique une probabilité nulle. Ces principes intuitifs de consistance entre possibilités et probabilités seront repris en détails au chapitre 6.

Le contenu informationnel d'une distribution de possibilité peut être caractérisé en terme de *spécificité*. Une distribution π_1 est dite plus spécifique qu'une distribution π_2 si :

$$\pi_1(\omega) \leq \pi_2(\omega) \quad \forall \omega \in \Omega. \tag{1.29}$$

En particulier, la distribution de possibilité la moins spécifique définie sur Ω, qui représente un état d'ignorance total, est donnée par :

$$\pi(\omega) = 1 \quad \forall \omega \in \Omega, \tag{1.30}$$

tandis que la distribution la plus spécifique est de la forme :

$$\pi(\omega) = \begin{cases} 1 & \text{if } \omega = \omega_0 \\ 0 & \forall\ \omega \neq \omega_0 \end{cases}, \tag{1.31}$$

pour un $\omega_0 \in \Omega$.

Tout comme pour les fonctions de croyance, il est possible de définir une distribution de possibilité sur le produit cartésien de plusieurs référentiels, de marginaliser des distributions ou encore de les étendre (par une opération dite d'extension cylindrique) [39]. D'autres part, des opérateurs conjonctifs, disjonctifs ou adaptatifs sont disponibles pour combiner plusieurs distributions de possibilité [9, 8, 41, 42]. Tous ces détails, non utilisés dans la suite de ce document, ne seront pas développés ici.

3.4 Quel cadre ?

Une littérature abondante a été consacrée à la comparaison des différentes théories de l'incertain. Au delà des différences parfois subtiles d'interprétation, il apparaît pourtant bien difficile de conclure sur la supériorité de l'une ou l'autre des théories présentées. Il est clair en tout cas que les objets mathématiques manipulés sont proches et que la théorie des fonctions de croyance peut être considérée comme plus générale que celle des probabilités ou des possibilités puisque l'on retrouve celles-ci comme des cas particuliers : en effet,

– si la masse est attribuée à des singletons uniquement, la masse est dite *Bayesienne*, la construction de la fonction de croyance correspondante donne une unique mesure de probabilité $\text{bel} = P = \text{pl}$;
– si la masse est attribuée à des éléments focaux A_i, $i = 1 \in I$ consonants (ou emboîtés $i \leq j \Rightarrow A_i \subseteq A_j$) , alors la fonction de croyance correspondante bel est une fonction de Nécessité.

Cependant les opérateurs de combinaison diffèrent et le résultat de la combinaison de deux masses de croyances consonantes est rarement consonant. Autre analogie, il faut souligner que l'inférence statistique classique se fonde sur la notion de vraisemblance (qui n'est pas une probabilité), qui, une fois normalisée, s'apparente clairement à une distribution de possibilité.

Ces quelques réflexions nous incitent non pas à considérer ces différentes théories comme rivales, mais comme proposant des représentations complémentaires de l'incertitude. Un travail important de comparaison et de mise en relation des différentes théories est encore nécessaire à ce jour. Constatant l'intérêt de manipuler conjointement différents formalismes notamment pour manipuler des données hétérogènes, certains auteurs ont d'ores et déjà proposé des transformations permettant le passage d'un formalisme à l'autre [37, 19, 23, 67, 70, 33]. Ce travail mérite d'être encore approfondi. Nous présenterons au chapitre 6 une contribution dans ce domaine.
Il nous apparaît en tout cas vain à l'heure actuelle de choisir de façon dogmatique un cadre unique de représentation. C'est la raison pour laquelle, tout

au long de ce mémoire, suivant l'application visée, des briques provenant de différents cadres théoriques seront utilisées.

4 L'analyse de données imprécises dans la littérature

L'analyse de données vagues ou imprécises a suscité ces dernières années un grand nombre de travaux comme en témoigne le nombre d'ouvrages parus sur le sujet [68, 3, 98, 4, 31].

Le cadre théorique des sous-ensembles flous apparaissant comme le plus adapté à la représentation de l'imprécision, la majorité de ces travaux, sous le terme anglais de *fuzzy data analysis*, cherchent à mêler statistiques classiques et logique floue. Avant d'aller plus loin, il faut noter que le terme de "fuzzy data analysis" est trompeur car il confond deux démarches complètement différentes. L'analyse de données *floue* consiste à utiliser des techniques floues pour analyser des données classiques. C'est le cas par exemple du célèbre algorithme de classification automatique des c-moyennes floues [5]. La démarche qui nous intéresse ici est celle de l'analyse de données *floues* qui cherche des techniques d'analyse spécifiques d'échantillons de données imprécises. Dans ce cas, on fait l'hypothèse de l'existence d'une variable aléatoire réelle classique dont les réalisations ne sont pas directement accessibles mais rapportées par un observateur avec une certaine imprécision. Comme le soulignent Gebhardt et al. [49, page 317], deux voies principales ont été explorées qui dépendent essentiellement de la manière dont l'observateur rapporte l'information.

La première est basée sur la notion de variable aléatoire floue qui généralise la notion classique de variable aléatoire réelle. On définit en effet une fonction qui, plutôt que d'associer un nombre réel à toute issue possible d'une expérience aléatoire, associe un nombre flou de l'ensemble des réels. Cette approche a suscité de nombreux développements mathématiques qui ont notamment étendu d'importants théorèmes limites. Néanmoins peu d'applications pratiques sur des jeux de données réels sont rapportées dans la littérature. Comme le font encore remarquer Gebhardt et al. [49, page 317], les variables aléatoires floues décrivent des situations où l'incertitude et l'imprécision sur la réalisation d'une variable aléatoire sont fonctionnellement dépendantes de l'issue de l'expérience aléatoire. A titre illustratif, prenons l'exemple d'une application traitée par Montenegro et al. [74] qui concerne l'analyse de données météorologiques relevées en Espagne. Une des variables observées porte sur le caractère nuageux du temps, les valeurs observables (brillant, clair, couvert, nuageux, et sombre) ayant été traduites par des experts comme des nombres flous *fixes* d'un référentiel continu allant de 0 à 100. Il est clair que le résultat de l'expérience aléatoire conditionne ici l'imprécision de l'observation et que dans

ce cas, les variables aléatoires floues sont, d'après Gebhardt et al., adaptées. On peut penser également aux résultats d'un questionnaire avec des réponses linguistiques codées sous forme de nombres flous fixes.

Si, au contraire, les conditions d'observation ne sont pas influencées par l'expérience aléatoire, alors on peut se passer de tout l'arsenal mathématique des variables aléatoires floues. Il suffit de généraliser les opérations effectuées en inférence statistique classique sur des données nettes à des opérations sur des données floues grâce au principe d'extension exposé plus haut. On suppose simplement disposer d'une vision imparfaite $\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}$ (où chaque $\tilde{x_i}$ est un nombre flou) d'un échantillon $x_1, ..., x_n$ d'une variable aléatoire parente. C'est l'approche suivie notamment par Viertl [98, 46]. Les travaux présentés dans la suite de ce mémoire se situent clairement dans cette lignée.

Pour finir, il faut citer les travaux de l'équipe de Diday [31] sur *l'Analyse de Données Symboliques*. Au travers notamment du projet européen SODAS (et du logiciel public du même nom) a été développé un grand nombre de méthodes d'analyse de données dites symboliques. Partant du constat que les bases de données actuelles deviennent trop volumineuses pour être traitées directement par des méthodes classiques, la proposition est de résumer les données initiales à l'aide de concepts sous-jacents. Par exemple, si les données portent sur des individus, on peut choisir de les regrouper par ville, par catégorie socio-professionnelle, etc. On construit alors des objets d'étude de deuxième niveau (les villes, les catégories socio-professionelles) que l'on va soumettre à l'analyse. Ces objets sont décrits par des variables plus complexes que celles habituellement rencontrées en statistique. Les valeurs des variables peuvent être des histogrammes, des intervalles, des valeurs uniques, des valeurs associés à des probabilités ou des degrés de confiance, etc. Une théorie complète autour de l'analyse de données symboliques a été développée et des méthodes spécifiques permettant la description, la visualisation et la classification de ces données ont été proposées. Elles se fondent pour l'essentiel sur la recherche de métriques ou de mesures de similarité adaptées. Outre que les concepts manipulés sont d'une complexité parfois peu justifiée, et bien que la forme des données traitées s'apparente à celle abordée dans ce mémoire (par exemple des données de type intervalle), cette démarche s'écarte conceptuellement de la nôtre. En effet, notre but n'est pas de traiter des "méta-données" ou des données de deuxième niveau mais bien d'intégrer la notion de flou, d'imprécision dans notre analyse.

5 Conclusion

Dans ce chapitre, nous avons présenté le cadre général de ce mémoire que constitue l'analyse de données imprécises. Après être rapidement revenu sur la notion de donnée imparfaite, nous avons ensuite rappelé les différents cadres

classiques de représentation et de manipulation de l'incertitude et l'imprécision :

- la théorie des sous-ensembles flous qui fournit des outils simples et bien adaptés à la représentation d'informations imprécises ;
- la théorie des probabilités, des possibilités et des fonctions de croyance pour la gestion de l'incertitude.

Enfin, nous avons évoqué comment le problème de l'analyse de données imprécises était abordé dans la littérature et comment notre approche se situait par rapport à ces travaux. Dans la suite de ce mémoire, nous verrons comment les outils théoriques que nous avons présentés sont mis en oeuvre pour aboutir à une proposition cohérente d'outils d'analyse de données imprécises.

Chapitre

2

Corrélation

1 Introduction

Lorsque l'on dispose de deux séries de mesures X et Y prélevées simultanément sur une population d'individus, une première manière de décrire ces données est d'étudier leur lien grâce à un calcul de corrélation. Si les variables sont continues, la mesure de corrélation usuelle est celle de Bravais-Pearson [84]. On suppose implicitement dans ce cas l'existence d'une relation linéaire entre les quantités mesurées. Une autre manière de procéder consiste à s'affranchir de l'hypothèse de linéarité en ne considérant que les ordres induits par chacune des variables. On mesure alors la concordance des ordres grâce à un coefficient de corrélation de rang comme celui du τ de Kendall [66]. De nombreux travaux ont cherché à généraliser la notion de corrélation dans un contexte flou [50, 102, 18, 17]. Dans la plupart des cas, le coefficient proposé est un nombre net. Cette formulation ne nous semble pas appropriée. En effet, si les données sont imprécisément décrites, il nous paraît plus naturel de considérer qu'il existe un ensemble (net ou flou) de valeurs possibles pour le coefficient de corrélation et non une valeur unique. C'est la raison pour laquelle les travaux de Liu et Kao [71], qui ont proposé une version floue du coefficient de Bravais-Pearson basée sur le principe d'extension de Zadeh, ont attiré notre attention. Leur approche est décrite dans le paragraphe 2. Le paragraphe 3 est, quant à lui, consacré à l'extension du tau de Kendall à des variables imprécises. Nous montrons comment tester la significativité de la valeur floue obtenue en introduisant la notion de degré de signification flou (en anglais p-value) proposée indépendamment par Filzmoser et Viertl [46].

2 Approche de Liu et Kao pour des variables quantitatives

Soit $(x^1, y^1), ...(x^n, y^n)$, n paires d'observations issues de deux variables X et Y. Le coefficient de corrélation de Bravais-Pearson, entre les deux séries d'observations X et Y est défini par

$$r_{\mathrm{XY}} = \frac{\sum_{p=1}^{n}(x^p - \bar{x})(y^p - \bar{y})}{\sqrt{\sum_{p=1}^{n}(x^p - \bar{x})^2 \sum_{p=1}^{n}(y^p - \bar{y})^2}}, \tag{2.1}$$

où $\bar{x}$ et $\bar{y}$ désignent respectivement la moyenne arithmétique des valeurs x^p et y^p. Ce coefficient mesure le caractère linéaire du lien entre les variables X et Y. Les propriétés de r_{XY} sont bien connues :

– $0 \leq |r_{\mathrm{XY}}| \leq 1$;
– plus $|r_{\mathrm{XY}}|$ est proche de 1, plus forte est la dépendance linéaire entre X et Y.

Supposons maintenant que les valeurs des couples (x^p, y^p) ne soient pas connues, mais seulement décrites au travers d'observations floues $(\tilde{x}^p, \tilde{y}^p)$ de fonctions d'appartenance respectives $\mu_{\tilde{x}^p}$ et $\mu_{\tilde{y}^p}$. Il est alors naturel d'envisager la corrélation entre X et Y non plus comme une valeur précise mais comme un nombre flou $\tilde{r}$. Le principe d'extension de Zadeh permet à Liu et Kao [71] d'écrire

$$\forall r \in \mathbb{R} \quad \mu_{\tilde{r}}(r) = \sup_{\{x^1,...,x^n,y^1,...,y^n / r = r_{\mathrm{XY}}\}} \min_p \left(\mu_{\tilde{x}^p}(x^p) \wedge \mu_{\tilde{y}^p}(y^p)\right), \tag{2.2}$$

où $\mu_{\tilde{r}}$ désigne la fonction d'appartenance de $\tilde{r}$, et $\wedge$ désigne l'opérateur minimum. Plus précisément, si $[(\tilde{x}^p)_\alpha^-; (\tilde{x}^p)_\alpha^+]$ et $[(\tilde{y}^p)_\alpha^-; (\tilde{y}^p)_\alpha^+]$ désigne les intervalles fermés obtenus par α-coupes de $\tilde{x}^p$ et $\tilde{y}^p$, alors, chaque α-coupe de $\tilde{r}$ est un intervalle fermé $[(\tilde{r})_\alpha^-; (\tilde{r})_\alpha^+]$ dont on trouve les bornes minimales et maximales en résolvant les programmes non-linéaires suivants :

$$\begin{aligned} (\tilde{r})_\alpha^- \quad = \quad & \min_{x^1,...,x^n,y^1,...,y^n} \frac{\sum_{p=1}^{n}(x^p - \bar{x})(y^p - \bar{y})}{\sqrt{\sum_{p=1}^{n}(x^p - \bar{x})^2 \sum_{p=1}^{n}(y^p - \bar{y})^2}} \\ \text{avec} \quad & (\tilde{x}^p)_\alpha^- \leq x^p \leq (\tilde{x}^p)_\alpha^+ \quad \forall p, \\ & (\tilde{y}^p)_\alpha^- \leq y^p \leq (\tilde{y}^p)_\alpha^+ \quad \forall p, \end{aligned} \tag{2.3}$$

$$\begin{aligned} (\tilde{r})_\alpha^+ \quad = \quad & \max_{x^1,...,x^n,y^1,...,y^n} \frac{\sum_{p=1}^{n}(x^p - \bar{x})(y^p - \bar{y})}{\sqrt{\sum_{p=1}^{n}(x^p - \bar{x})^2 \sum_{p=1}^{n}(y^p - \bar{y})^2}} \\ \text{avec} \quad & (\tilde{x}^p)_\alpha^- \leq x^p \leq (\tilde{x}^p)_\alpha^+ \quad \forall p, \\ & (\tilde{y}^p)_\alpha^- \leq y^p \leq (\tilde{y}^p)_\alpha^+ \quad \forall p. \end{aligned} \tag{2.4}$$

Pour une coupe de niveau α donné, la résolution de (2.3) et (2.4) peut être menée à bien en utilisant une routine classique de résolution de programmes non linéaires. Notons qu'en raison de la non linéarité entre les observations et le coefficient de corrélation, même si $\tilde{x}$ et $\tilde{y}$ sont des nombres flous trapézoïdaux, le coefficient de corrélation n'est pas un nombre flou trapézoïdal. En pratique, les deux programmes linéaires sont résolus pour un petit nombre d'alpha-coupes ce qui donne une vue raisonnable de $\tilde{r}$.

EXEMPLE 2.1 (*Notes*) Soit un ensemble de neuf étudiants dont on a relevé les notes dans 5 disciplines : les MATHématiques, la PHYsique, la LITTérature, le LATin et le DESsin. Les notes sont soit précises soit imprécises, codées sous forme de nombres flous trapézoïdaux, comme le montre la figure 2.1. Les corrélations entre les différentes matières sont données en figure 2.2. Les traits verticaux correspondent à la corrélation classique calculée avec les centres des noyaux des notes. On observe logiquement une corrélation élevée entre les disciplines MATHématiques et PHYsique d'une part, et LITTérature, et LATin d'autre part. De plus, on constate qu'une corrélation forte est complètement possible entre le LATin et les MATHématiques ou la PHYsique, mais que des valeurs faibles de corrélation sont également possibles, ce qui permet de nuancer le résultat obtenu sur les notes centrales nettes. Sur la diagonale, les auto-corrélations floues renseignent sur la quantité d'imprécision attachée à une variable : la valeur 1 est toujours complètement possible, mais des valeurs plus faibles le sont aussi si la variable considérée est imprécise.

3 Liaison entre variables ordinales

3.1 Tau de Kendall

3.1.1 Principe

Soit $U = \{u_1, u_2, ..., u_n\}$ un ensemble d'individus. Un ordre strict total sur U est une relation binaire $L \subseteq U \times U$ qui est [104] :
– transitive : $u_i L u_j, u_j L u_k \Rightarrow u_i L u_k \; \forall\, u_i, u_j, u_k \in U$;
– asymétrique[1] : $\forall\, u_i, u_j \in U$, non ($u_i L u_j$ et $u_j L u_i$) ;
– complète : $\forall\, u_i \neq u_j \in U \; u_i L u_j$ ou $u_j L u_i$.
Supposons que les individus de U soient décrits par deux variables continues X et Y associant à chaque individu u_i deux valeurs x_i et y_i. Ces deux variables définissent deux ordres stricts totaux (ou ordres stricts linéaires), L_{X} and L_{Y}

[1]Notons que la propriété d'asymétrie est équivalente aux propriétés conjointes d'antiréflexivité : $\forall u_i \in U, non(u_i L u_i)$ et d'antisymétrie : $\forall u_i, u_j \in U, u_i L u_j$ et $u_j L u_i \Rightarrow u_i = u_j$.

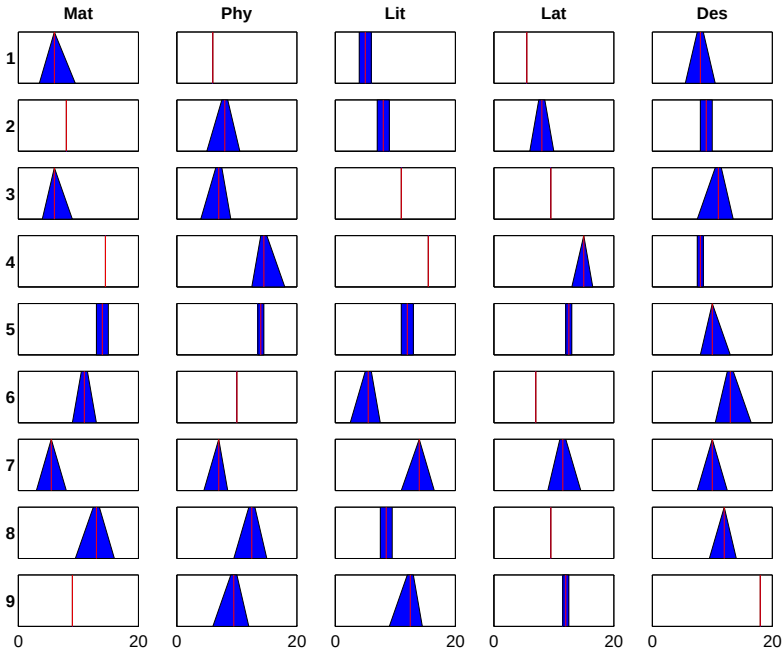

FIG. 2.1 – Jeu de données des notes.

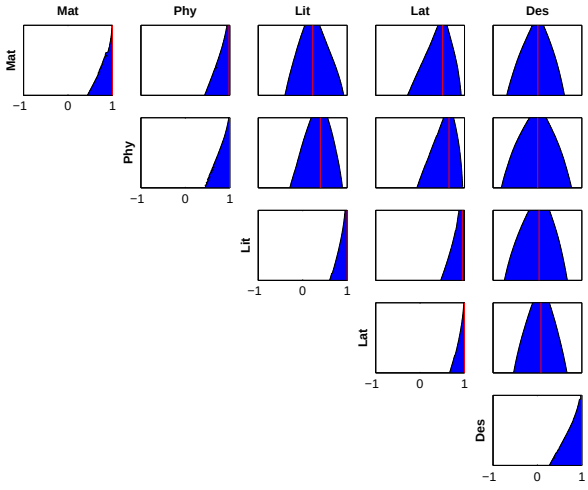

FIG. 2.2 – Jeu de données des notes ; Corrélations floues entres les matières.

sur U :

$$x_i > x_j \Leftrightarrow u_i \succ_{L_X} u_j \Leftrightarrow u_i L_X u_j \tag{2.5}$$

$$y_i > y_j \Leftrightarrow u_i \succ_{L_Y} u_j \Leftrightarrow u_i L_Y u_j \tag{2.6}$$

Soit N le nombre de paires d'individus ($N = n(n-1)/2$) et D le nombre de paires discordantes entre L_X et L_Y : une paire est dite discordante si elle appartient à un ordre mais pas à un autre. Le τ de Kendall est alors défini par [66] :

$$\tau(L_X, L_Y) = 1 - \frac{2D}{N}, \tag{2.7}$$

ou, de façon equivalente, en notation ensembliste :

$$\tau(L_X, L_Y) = \frac{2 \times |L_X \cap L_Y|}{N} - 1. \tag{2.8}$$

Comme le coefficient de Bravais-Pearson, le τ de Kendall est compris entre -1 et 1, une valeur égale à 1 indiquant que les classements sont identiques, et une valeur de -1 indiquant que les classements sont complètement inversés.

3.1.2 Caractère significatif de la corrélation

On peut utiliser le τ de Kendall pour tester l'indépendance (hypothèse nulle H_0) contre la dépendance entre X et Y (l'hypothèse alternative H_1). Sous l'hypothèse d'indépendance H_0, la loi de τ est asymptotiquement une loi normale de moyenne nulle et d'écart-type égal à :

$$\sigma_n = \sqrt{\frac{2(2n+5)}{9n(n-1)}}.$$

L'approximation est très bonne à partir de $n \geq 8$. A l'aide cette loi, on construit, par exemple, un test bilatéral. On rejette alors l'indépendance entre U et V au niveau θ si la valeur de τ constatée vérifie :

$$|\tau| > \Phi^{-1}(1 - \frac{\theta}{2})\sqrt{\frac{2(2n+5)}{9n(n-1)}}, \tag{2.9}$$

où Φ est la fonction de répartition de la loi normale centrée réduite. Le choix du seuil de signification θ est parfois jugé arbitraire. C'est la raison pour laquelle certains préfèrent garder l'information contenue dans la valeur de la statistique de test, en retournant le seuil de significativité limite auquel H_0 aurait été rejetée, compte tenu de l'observation : c'est le degré de signification p qui est ici égal à :

$$p = 2\left[1 - \Phi\left(|\tau| / \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)\right]. \tag{2.10}$$

Le test concernant les classements induits par X et Y se reformule alors de la manière suivante : on rejette l'hypothèse H_0 au niveau θ si le degré de signification p est inférieur à θ, d'où la décision $\mathcal{D}_\theta(U,V)$:

$$\mathcal{D}_\theta(U,V) = \begin{cases} 1 & \text{si } p < \theta \\ 0 & \text{sinon} \end{cases} \tag{2.11}$$

3.2 Liaison ordinale entre variables de type intervalle

3.2.1 Ordres partiels

On considère maintenant que les mesures recueillies sur les individus sont fournies sous forme d'intervalles. On cherche dans cette partie à généraliser le τ de Kendall. Cet objectif nous amène à considérer non plus des ordres totaux mais des ordres *partiels*. Notre travail rejoint ici celui de Ha et Haddawy [54] qui s'intéressent aux mesures de similarité entre des préférences partiellement spécifiées.
Soit $[X]$ une variable associant un intervalle $[x_i] = [x_i^-, x_i^+]$ à chaque individu $u_i \in U$. $[x_i]$ représente l'ensemble des valeurs possibles pour la quantité x_i inconnue.
Soit P la relation binaire définie sur U qui mesure la supériorité stricte induite par $[X]$ entre chaque paire d'individus (i,j). Une définition naturelle de la supériorité stricte d'intervalles est : $u_i P u_j \Leftrightarrow x_i^- > x_j^+$. Il est simple de montrer que P est transitive et asymétrique et donc définit une relation d'ordre strict. Cependant P n'est pas nécessairement totale, certains individus peuvent ne pas être classés les uns par rapport aux autres : il s'agit d'une relation d'ordre partiel.
Exemple. Supposons que les mesures dont on dispose soient les suivantes : $[x_1] = [1;2]$; $[x_2] = [1.5;3]$, et $[x_3] = [4;5]$. La relation P n'est constituée que des couples suivants : $(u_3,u_1),(u_3,u_2)$, le classement entre u_1 et u_2 étant indéterminé.

3.2.2 Notion d'extension linéaire

Pour un ordre partiel donné, on peut montrer (cf [47]) qu'il existe au moins un ordre total tel que $u_i P u_j \Rightarrow\ u_i L u_j$, i.e., $P \subseteq L$. Un tel ordre total est appelé une *extension linéaire* de P.
Exemple. Reprenons l'exemple précédent où P est constitué des couples (u_3,u_2) et (u_3,u_1). Il existe deux extensions linéaires correspondant respectivement à $u_3 \succ_P u_2 \succ_P u_1$ et $u_3 \succ_P u_1 \succ_P u_2$, seuls classements compatibles avec P.
De facon générale, on associera à tout ordre partiel P, l'ensemble $E(P)$ contenant toutes les extensions linéaires de P.

3.2.3 Corrélation entre ordres partiels

On suppose que les individus sont décrits simultanément par deux variables de type intervalle $[X]$ et $[Y]$. Soient P_{X} la relation d'ordre partiel associée à $[X]$ et P_{Y} son homologue associée à $[Y]$. La corrélation entre $[X]$ et $[Y]$ est équivalente à la corrélation entre P_{X} and P_{Y}. Pour définir la corrélation entre deux ordres partiels, on assimile chacun à son ensemble d'extensions linéaires. Le coefficient de corrélation de rang entre P_{X} et P_{Y} peut alors être défini comme l'intervalle le plus petit contenant toutes les valeurs possibles de corrélation :

$$\bar{\tau}(P_{\mathrm{X}}, P_{\mathrm{Y}}) = [\min_{L_{\mathrm{X}} \in E(P_{\mathrm{X}}),\, L_{\mathrm{Y}} \in E(P_{\mathrm{Y}})} \tau(L_{\mathrm{X}}, L_{\mathrm{Y}})\,;\, \max_{L_{\mathrm{X}} \in E(P_{\mathrm{X}}),\, L_{\mathrm{Y}} \in E(P_{\mathrm{Y}})} \tau(L_{\mathrm{X}}, L_{\mathrm{Y}})] \quad (2.12)$$

Notons que si P_{X} and P_{Y} sont des ordres linéaires stricts, cette définition coïncide avec celle de Kendall. Dans le cas contraire, la taille de l'intervalle de corrélation reflète la quantité d'information contenue dans les ordres partiels.

3.2.4 Calcul pratique

L'idée la plus naturelle pour calculer la corrélation de rang entre deux ordres partiels P_{X} et P_{Y} est de calculer le τ de Kendall $\tau(L_{\mathrm{X}}, L_{\mathrm{Y}})$ entre toutes les paires d'extensions linéaires de P_{X} et P_{Y} et d'en retenir le minimum et le maximum. Cette approche se révèle toutefois très coûteuse en temps de calcul, le nombre d'extensions linéaires d'un ordre partiel étant potentiellement très grand (le cas limite d'un ordre partiel vide conduit à $n\,!$ extensions linéaires). Un algorithme permettant de résoudre le problème sans générer toutes les extensions linéaires a été proposé par Hébert et al [58]. Cette technique est néanmoins limitée à un faible nombre d'individus. Par ailleurs, des techniques approchées ont été proposées. Elles consistent, non plus à dresser une liste exhaustive de toutes les extensions linéaires, mais à en générer aléatoirement un sous-ensemble, échantillonné de manière quasi-uniforme [14, 54], grâce à des techniques de simulation de Monte Carlo à base de chaînes de Markov. L'algorithme le plus efficace connu à ce jour est celui de Bubley et Dyer [14]. Il peut être utilisé pour générer de façon répétée $L_{\mathrm{X}} \in E(P_{\mathrm{X}})$, $L_{\mathrm{Y}} \in E(P_{\mathrm{Y}})$, calculer $\tau(L_{\mathrm{X}}, L_{\mathrm{Y}})$, et stocker cette valeur si elle plus faible que le minimum courant ou plus élevée que le maximum courant. L'algorithme s'arrête lorsque le minimum et le maximum n'ont pas évolué durant les η dernières itérations.

3.2.5 Test de significativité

On cherche à construire un test portant sur l'indépendance des ordres induits par $[X]$ et $[Y]$ à l'aide de la statistique proposée. Le coefficient de corrélation $\bar{\tau}$ étant maintenant défini comme un intervalle $[\tau^-; \tau^+]$, le degré de signification

associé au test d'indépendance des classements devient tout naturellement imprécis. Grâce aux propriétés suivantes :

$$\min_{\tau^- \leq \tau \leq \tau^+} |\tau| = \max(0, \tau^-, -\tau^+),$$

$$\max_{\tau^- \leq \tau \leq \tau^+} |\tau| = \max(\tau^+, -\tau^-),$$

on en déduit que p varie dans l'intervalle $[p^-; p^+]$ dont les bornes sont définies par :

$$p^- = 2\left[1 - \Phi\left(\max(\tau^+, -\tau^-)\sigma_n^{-1}\right)\right]$$
$$p^+ = 2\left[1 - \Phi\left(\max(0, \tau^-, -\tau^+)\sigma_n^{-1}\right)\right].$$

Le résultat du test est obtenu par comparaison de p au seuil de signification θ. Puisque p est imprécis, trois cas peuvent se produire :

1. si $p^+ < \theta$, alors p est assurément plus faible que θ, il faut donc rejeter l'hypothèse H_0 ;
2. si $p^- > \theta$, alors p est assurément plus élevée que θ, il faut donc retenir l'hypothèse H_0 ;
3. si $p^- < \theta < p^-$, alors il y a indétermination, on ne peut pas conclure.

La décision du test au niveau de signification θ concernant les classements induits par $[X]$ et $[Y]$ peut se formuler de la manière suivante :

$$\mathcal{D}_\theta([X],[Y]) = \begin{cases} 1 & \text{si } p^+ < \theta \\ 0 & \text{si } p^- > \theta \\ \{0,1\} & \text{sinon} \end{cases} \tag{2.13}$$

3.3 Liaison ordinale de variables floues

3.3.1 Ordres partiels flous

On suppose maintenant que les mesures recueillies sont des nombres flous. Pour définir la notion d'ordre entre ces nombres, on fait appel au concept de relation floue de Zadeh [104], dont une présentation récente est donnée dans [75]. Une relation floue sur un ensemble U est un sous-ensemble flou de U^2 qui quantifie le degré de mise en relation de deux éléments u_i et u_j de U par une valeur $\tilde{R}(u_i, u_j)$ dans l'intervalle [0 ;1]. Les propriétés classiques telles que la symétrie, la réflexivité, ou la transitivité sont aisément étendues pour ce type de relations, en s'appuyant sur les définitions d'inclusion d'ensembles flous et sur les opérateurs classiques de la logique floue. Un *ordre partiel flou* est une relation floue particulière $\widetilde{P} \in [0,1]^{U \times U}$ qui possède les propriétés suivantes :

– max-min transitivité :
$\forall\ u_i, u_j, u_k \in U,\ \widetilde{P}(u_i, u_k) \geq \bigvee_j \widetilde{P}(u_i, u_j) \wedge \widetilde{P}(u_j, u_k)$;

– asymétrie : $\forall\ u_i, u_j \in U,\ \widetilde{P}(u_i, u_j) \wedge \widetilde{P}(u_j, u_i) = 0$.

Propriété remarquable, chaque α-coupe $\widetilde{P}_\alpha$ d'un ordre partiel flou est un ordre partiel classique [104].

3.3.2 Ordres partiels flous induits par des nombres flous

Soit $\widetilde{X}$ une variable associant un nombre flou $\widetilde{x}_i$ à chaque individu $u_i \in U$. Une généralisation de la mesure de supériorité stricte introduite pour les intervalles est donnée par la relation $\widetilde{P}$ définie par :

$$\widetilde{P}(u_i, u_j) = 1 - \sup_{a \leq b} \min(\mu_{\widetilde{x}_i}(a), \mu_{\widetilde{x}_j}(b)) \tag{2.14}$$

Cette mesure s'interprète comme la nécessité de l'événement $x_i > x_j$ [38]. La relation $\tilde{P}$ ainsi définie est antisymétrique et transitive, c'est une relation d'ordre partiel flou sur U.

Exemple. Soit les nombres flous trapézoïdaux représentés sur la figure 2.3. La

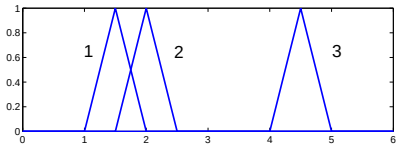

FIG. 2.3 – Trois nombres flous trapézoïdaux.

relation de supériorité stricte P calculée grâce à l'équation (2.14) est donnée dans la table 2.1.

$\widetilde{P}\ \nearrow$	u_1	u_2	u_3
u_1	0	0	0
u_2	0.5	0	0
u_3	1	1	0

TAB. 2.1 – Relation floue de supériorité stricte.

3.3.3 Corrélation entre ordres partiels flous

On a vu précédemment que l'on pouvait associer à tout ordre partiel P l'ensemble $E(P)$ constitué de toutes ses extensions linéaires. Il s'agit maintenant d'étendre cela aux ordres partiels flous. Soit $\mathcal{L}(U) \subseteq 2^{U \times U}$ l'ensemble des ordres totaux sur U. L'ensemble $E(\widetilde{P}) \in [0,1]^{\mathcal{L}(U)}$ désigne l'ensemble flou des

extensions linéaires de $\tilde{P}$ dont la fonction d'appartenance peut être définie par :

$$\mu_{E(\widetilde{P})}(L) = I(\widetilde{P}, L) \quad \forall L \in \mathcal{L}(U),$$

où I désigne une mesure d'inclusion. En utilisant un choix classique [43], on obtient :

$$\begin{aligned} \mu_{E(\widetilde{P})}(L) &= 1 - ht(\widetilde{P} \cap \overline{L}) \\ &= 1 - \sup_{u_i,u_j} \min\Big(\widetilde{P}(u_i,u_j), 1 - L(u_i,u_j)\Big) \end{aligned}$$

Pour définir le coefficient de corrélation de rang entre deux ordres partiels flous, nous nous appuyons sur le résultat fondamental suivant :

PROPOSITION 1
L'α-coupe de l'ensemble flou des extension linéaires de $\tilde{P}$ est l'ensemble net des extensions linéaires des coupes strictes de niveau $(1-\alpha)$ de $\tilde{P}$.

Démonstration. Soit $\alpha \in (0,1]$. L'α-coupe de $E(\widetilde{P})$ est l'ensemble net $E(\widetilde{P})^\alpha$ défini par tout $L \in \mathcal{L}(U)$ qui vérifie :

$$1 - \sup_{u_i,u_j} \min\Big(\widetilde{P}(u_i,u_j), 1 - L(u_i,u_j)\Big) \geq \alpha.$$

On a donc :

$$\begin{aligned} & L \in E(\widetilde{P})^\alpha \\ \Leftrightarrow\ & \sup_{u_i,u_j} \min\Big(\widetilde{P}(u_i,u_j), 1 - L(u_i,u_j)\Big) \leq 1 - \alpha \\ \Leftrightarrow\ & \forall (u_i,u_j)\ \min\Big(\widetilde{P}(u_i,u_j), 1 - L(u_i,u_j)\Big) \leq 1 - \alpha \\ \Leftrightarrow\ & \forall (u_i,u_j)\ \widetilde{P}(u_i,u_j) > 1 - \alpha \Rightarrow 1 - L(u_i,u_j) = 0 \\ \Leftrightarrow\ & \forall (u_i,u_j) \in \widetilde{P}^{(1-\alpha)+},\ (u_i,u_j) \in L \\ \Leftrightarrow\ & L \in E(\widetilde{P}^{(1-\alpha)+}) \square \end{aligned}$$

Grâce au résultat précédent, on peut maintenant étendre la notion de corrélation de rang pour deux ordres partiels nets $\bar{\tau}(P_{\mathrm{X}}, P_{\mathrm{Y}})$ à celle de corrélation de rang pour deux ordres partiels flous : $\tilde{\tau}(\widetilde{P}_{\mathrm{X}}, \widetilde{P}_{\mathrm{Y}})$ est un intervalle flou dont les α-coupes sont des intervalles fermés définis par :

$$\tilde{\tau}(\widetilde{P}_{\mathrm{X}}, \widetilde{P}_{\mathrm{Y}})^\alpha = \bar{\tau}(P_{\mathrm{X}}^{(1-\alpha)+}, P_{\mathrm{Y}}^{(1-\alpha)+}). \tag{2.15}$$

En pratique, $\tilde{\tau}$ peut être approximé en estimant les bornes minimales et maximales d'un nombre limité d'α-coupes grâce à la technique de Monte-Carlo décrite dans le paragraphe 3.2.4, par échantillonnage de $E(\widetilde{P}_{\mathrm{X}}^{(1-\alpha)+})$ et $E(\widetilde{P}_{\mathrm{Y}}^{(1-\alpha)+})$ suivant une loi uniforme.

3.3.4 Test de significativité

Soit $\mu_{\tilde{\tau}}$ la fonction d'appartenance du τ de Kendall flou calculé entre deux ordres partiels flous. Le principe d'extension [39] permet de définir la valeur p associée au test de significativité de τ comme un nombre flou $\tilde{p}$ dont la fonction d'appartenance est :

$$\mu_{\tilde{p}}(p) = \sup_{\tau/p=2[1-\Phi(|\tau|\sigma_\tau)]} \mu_{\tilde{\tau}}(\tau). \tag{2.16}$$

Plus précisément, chaque α-coupe de $\tilde{p}$ est un intervalle fermé $[p_\alpha^-, p_\alpha^+]$ défini par :

$$p_\alpha^- = 2\left[1 - \Phi\left(\max(\tau_\alpha^+, -\tau_\alpha^-)\sigma_n^{-1}\right)\right]$$
$$p_\alpha^+ = 2\left[1 - \Phi\left(\max(0, \tau_\alpha^-, -\tau_\alpha^+)\sigma_n^{-1}\right)\right],$$

où τ_α^- et τ_α^+ désignent les bornes inférieure et supérieure de l'α-coupe du τ de Kendall flou. Pour obtenir le test associé, il suffit d'appliquer de nouveau le principe d'extension à l'équation (2.11). La décision concernant $\tilde{X}$ et $\tilde{Y}$ est alors un ensemble flou de $\{0, 1\}$ défini par :

$$\begin{cases} \mu_{\tilde{\mathcal{D}}_\theta}(1) = \sup_{p \leq \theta} \mu_{\tilde{p}}(p) \\ \mu_{\tilde{\mathcal{D}}_\theta}(0) = \sup_{p > \theta} \mu_{\tilde{p}}(p) \end{cases} \tag{2.17}$$

La quantité $\mu_{\tilde{\mathcal{D}}_\theta}(1)$ peut s'interpréter comme la possibilité de l'événement $p \leq \theta$ compte tenu de la distribution de possibilité de $\tilde{p}$. C'est donc la possibilité qu'on aurait de rejeter H_0 (accepter la non indépendance) si on observait des valeurs précises sur l'échantillon. De façon similaire, la quantité $\mu_{\tilde{\mathcal{D}}_\theta}(0)$ peut s'interpréter comme la possibilité de l'événement $p > \theta$, c'est-à-dire la possibilité de retenir H_0 (hypothèse d'indépendance) en présence de données précises.

Remarque 2 On pourrait de la même façon tester la significativité du coefficient de corrélation de Liu et Kao. En particulier, si n est grand, la loi de r est approximativement une loi normale et le même type de raisonnement s'applique.

Exemple 2.2 Nous reprenons l'exemple des notes des étudiants. Les valeurs de tau de Kendall obtenues sont représentées figure 2.4. Les traits verticaux indiquent les bornes d'acceptation de H_0 ($\pm 1.96\sigma_n$) compte tenu d'un seuil de signification $\theta = 5\%$. Les résultats sont assez semblables à ceux trouvés avec le coefficient de Liu et Kao. On observe sans surprise la corrélation entre d'une part les disciplines scientifiques (Mathématiques et Physique), et d'autre

part, les disciplines littéraires (Latin et Littérature). Sur la diagonale, les autocorrélations floues renseigne sur la complétude des ordres partiels. La figure 2.5 présente les degrés de signification flous, les traits verticaux indiquant le degré de signification $\theta = 5\%$. A titre d'exemple, voici les valeurs des décisions concernant les MAThématiques et la PHYsique :

$$\mu_{\tilde{\mathcal{D}}_\theta(\mathrm{MAT,PHY})}(1) = 1,$$

$$\mu_{\tilde{\mathcal{D}}_\theta(\mathrm{MAT,PHY})}(0) = 0.75,$$

ce qui montre que la possibilité de rejeter l'hypothèse d'indépendance est maximale, alors qu'elle n'est que de 0.75 d'accepter l'indépendance. On dispose alors d'une décision nuancée, la décision finale, si elle est nécessaire, étant laissée à l'appréciation de l'utilisateur.

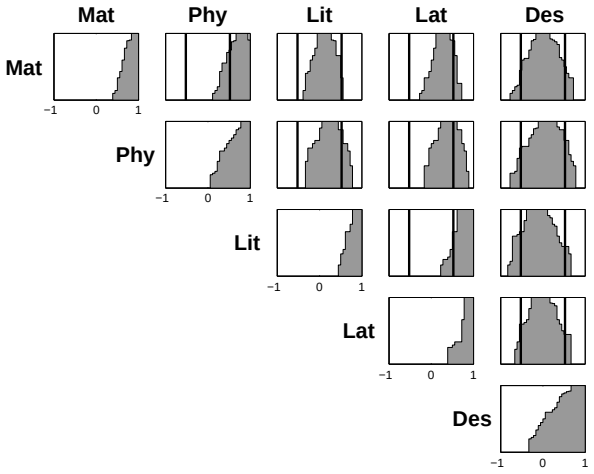

FIG. 2.4 – Taux de Kendall flous.

4 Conclusion

Nous avons exposé dans ce chapitre des méthodes permettant d'étendre la notion de corrélation à des variables imprécises. Deux points sont à souligner :

- Nous défendons le point de vue selon lequel la corrélation entre deux séries d'observations imprécises doit elle-même être imprécise. Dans ce cadre, nous avons d'abord présenté le coefficient de Pearson pour des variables continues étendu par Liu et Kao. Nous avons ensuite présenté notre approche pour étendre le coefficient de corrélation de rang de Kendall (appelé τ de Kendall) au cas des intervalles et des nombres flous. Nous avons montré que les ordres

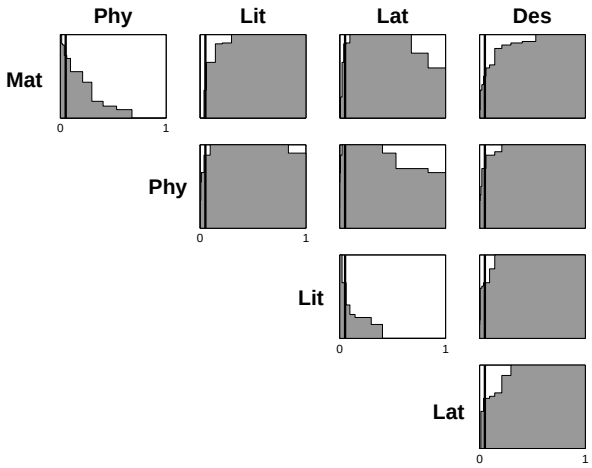

FIG. 2.5 – Degrés de signification flous.

induits par des observations imprécises sont des ordres partiels flous. En utilisant la notion d'extension linéaire d'un ordre partiel, nous avons défini le coefficient de corrélation de rang comme étant lui-même un nombre flou.

- De la même manière, nous pensons que le test d'une hypothèse précise à l'aide de données imprécisemment décrites, doit conduire à envisager une troisième voie aux situations classiques d'acceptation ou de rejet de l'hypothèse nulle : celle ou aucune décision ne peut être prise, compte tenu du caractère imprécis des données. Un test statistique permettant de juger du caractère significatif de la corrélation a été proposé dans ce cadre. La notion de degré de signification attaché au test classique a été étendue à celle de degré de signification flou conduisant à une décision floue.

L'approche proposée ici est suffisamment générale pour être mise en oeuvre sur d'autres statistiques non paramétriques. Nous pensons qu'elle ouvre des perspectives intéressantes dans le domaine de l'analyse de données statistiques.

Publications

M. Masson, P.-A. Hébert, et T. Denoeux. Corrélation de rang entre nombres flous. *Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'03)*, 137-142, Tours, France, Novembre 2003.

P.-A. Hébert, M. Masson et T. Denoeux. Fuzzy rank correlation between fuzzy numbers. *10th IFSA World congress*, pages 224-227, 29 Juin-2 Juillet, Istanbul,Turquie, 2003.

T. Denoeux, M. Masson, P.-A. Hébert. Nonparametric rank-based statistics and significance tests for fuzzy data. *Fuzzy sets and systems*, 153(1), 1-28, 2005.

Chapitre

3

Positionnement multidimensionnel

1 Introduction et contexte

Les méthodes de positionnement multidimensionnel [86, 21, 11] sont des techniques désormais classiques d'analyse de mesures de dissimilarité entre objets. L'idée est de représenter les dissimilarités comme des *distances* entre points dans un espace, le plus souvent euclidien, chaque point représentant un objet. On obtient des représentations facilement interprétables sous forme de cartes dans lesquelles deux objets seront d'autant plus proches qu'ils sont associés à des dissimilarités faibles.

Initialement développées comme techniques de réduction de dimension pour des données complexes, les méthodes MDS ont suscité un grand intérêt dans des disciplines relevant de la psychologie. En analyse sensorielle, en particulier, on cherche à comprendre comment des produits sont perçus par des consommateurs, un produit pouvant être un yaourt, une voiture, une couleur, un parfum, etc... A cette fin, il est courant de faire appel à des panels de sujets (experts ou naïfs), en les interrogeant sur les différences ressenties entre produits. Le but est ensuite de visualiser ces différences et de déterminer quelles sont les dimensions sous-jacentes structurant leur perception. Nous avons engagé avec PSA depuis plusieurs années des travaux portant sur l'analyse de données recueillies au cours de tests d'évaluation sensorielle. Les données concernaient initialement le confort acoustique dans un habitacle automobile, mais progressivement d'autres sens comme le toucher avec l'étude de textiles pour les sièges ou de matières plastiques pour les tableaux de bord, ou encore l'odorat avec des parfums d'intérieur ont été abordés. Les difficultés pour un sujet hu-

main d'évaluer de manière précise ses sensations nous ont conduit à envisager le recueil de dissimilarités sous forme imprécise comme des intervalles ou encore des nombres flous. Nous avons donc été amenés naturellement à proposer l'extension des méthodes MDS à ce type de données.

2 Positionnement multidimensionnel classique

Dans ce paragraphe, nous donnons une rapide description des principes de base du positionnement multidimensionnel. Une description complète peut être trouvée dans plusieurs ouvrages comme [86, 21, 11].

2.1 Généralités

On suppose que l'on dispose d'une matrice de dissimilarités $\Delta = (\delta_{ij})$, où δ_{ij} désigne la dissimilarité entre les objets i and j. Le but est de déterminer les coordonnées de n objets dans un espace de dimension p, sous la forme d'une matrice $\mathbf{X} = (x_{il})$ de taille $n \times p$, de telle sorte que la matrice des distances inter-points $D(\mathbf{X}) = (d_{ij}(\mathbf{X}))$ soit la plus proche possible de Δ.
La distance euclidienne est généralement choisie comme mesure de proximité dans l'espace de représentation :

$$d_{ij}(\mathbf{X}) = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{jl})^2}. \tag{3.1}$$

La détermination de $\mathbf{X}$ se fonde sur la minimisation itérative d'un critère mesurant la proximité entre $D(\mathbf{X})$ et Δ, connu dans la littérature sous le nom de fonction de "*stress*" [69]. Plusieurs variantes ont été proposées parmi lesquelles :

$$\sigma(\mathbf{X}) = \sum_{i<j} \left(d_{ij}(\mathbf{X}) - \delta_{ij}\right)^2. \tag{3.2}$$

Exemple 3.1 (*Jeu de données des couleurs*) Nous considérons ici une expérience de Helm [60] rapportée dans [11, p360] sur la perception des couleurs par des êtres humains. Dix objets colorés ont été présentés à différents sujets à qui il a été demandé d'évaluer la similarité perçue. Dans ce premier exemple, on considère seulement la réponse du premier sujet. Une méthode d'échelonnement multidimensionnel classique appliquée à ces données conduit à la représentation donnée en figure 3.1. Notons que le critère de stress (3.2) étant invariant par toute transformation isométrique (rotation, translation, dilatation), l'orientation a été choisie de manière arbitraire. On voit que les proximités des couleurs dans le plan sont conformes au sens commun et, comme

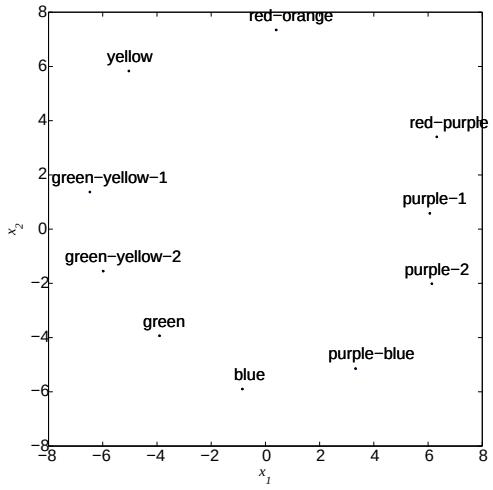

FIG. 3.1 – Jeu de données des couleurs ; Configuration bidimensionnelle.

l'avait déjà montré Eckman [45], qu'elles se positionnent autour d'un cercle imaginaire. La qualité d'approximation des dissimilarités peut se juger au travers d'un diagramme donné en figure 3.2, appelé le diagramme de Shepard, qui croise les dissimilarités d'entrée avec les distances reconstruites.

Sous le terme de méthodes MDS, on trouve en fait une grande variété de modèles et d'algorithmes. Les méthodes MDS varient principalement suivant :
– la façon, quantitative ou qualitative, dont les dissimilarités sont prises en compte : on parle alors d'approches métriques ou non métriques ;
– le modèle de distance choisi ;
– le nombre de tableaux échelonnés simultanément.

2.2 Approches métriques et non métriques

Le modèle décrit précédemment cherche à imposer l'égalité entre dissimilarités et distances. On peut cependant relâcher cette contrainte en imposant seulement que les dissimilarités soient dépendantes d'une fonction croissante, paramétrée, des données d'entrée. Le critère de stress est alors modifié en :

$$\sigma(\mathbf{X}) = \sum_{i<j} \left(d_{ij}(\mathbf{X}) - f(\delta_{ij})\right)^2 . \tag{3.3}$$

La fonction affine est un choix usuel pour f. Les paramètres de la fonction sont optimisés conjointement avec la configuration de points. Quelle que soit la forme paramétrique de f (affine, logarithmique, exponentielle), cette approche est qualifiée de *métrique*. Parfois, spécialement dans des disciplines

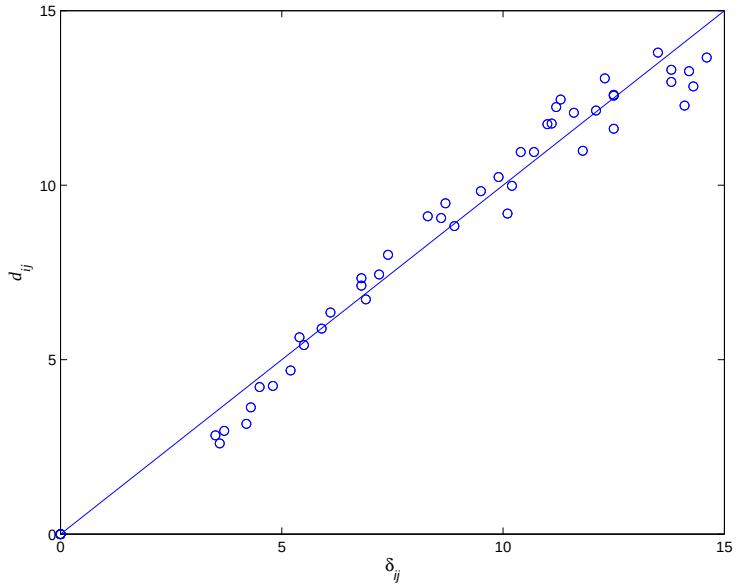

FIG. 3.2 – Shepard diagram for the color data.

relevant des sciences humaines, seul l'ordre induit par les dissimilarités est porteur de sens. Les approches *non métriques* permettent de ne pas imposer de forme particulière à f. La seule contrainte imposée est la monotonie de f. Dans ce cadre Kruskal a proposé un algorithme de régression isotonique [69] qui permet d'imposer la contrainte $d_{ij} \leq d_{kl}$ lorsque $\delta_{ij} \leq \delta_{kl}$. En pratique, nous avons constaté que les méthodes non métriques, utilisant un principe d'optimisation alternée, étaient très lourdes à mettre en oeuvre et, bien que nous ayons développé un algorithme non métrique pour les données intervalles [26], il ne sera pas détaillé ici.

2.3 Modèle de distance sphérique

Le positionnement multidimensionnel sphérique a été proposé par Cox et Cox [20] comme alternative au positionnement Euclidien. Il est préconisé pour trouver des configurations d'objets dans lesquelles la notion de points extrêmes n'a pas de sens. Cette méthode est donc particulièrement adaptée pour représenter des mesures de corrélation entre des variables statistiques. Nous commençons la description de la méthode par le positionnement sur un cercle avant de généraliser à la sphère. Supposons que les données disponibles consistent en une matrice $\mathcal{T} = (\tau_{ij})$ de dimension $n \times n$, où τ_{ij} désigne la corrélation entre deux variables i et j. L'idée est de représenter chaque variable par un vecteur de norme unité de telle sorte que le cosinus de l'angle formé par les

vecteurs associés aux deux variables soit lié à la corrélation fournie en entrée. Le problème peut se formaliser de la manière suivante : on note $(1, \theta_i)$ les coordonnées sphériques de la variable i. Le cosinus de l'angle ϕ_{ij} entre deux variables i and j, qui est le produit scalaire des coordonnées cartésiennes, est donné par :

$$\cos \phi_{ij} = cos\theta_i \cos \theta_j + \sin \theta_i \sin \theta_j \tag{3.4}$$

L'ensemble des vecteurs $\Theta = (\theta_1, \theta_2, ..., \theta_n)$, en partant d'une configuration initiale aléatoire, peut être déterminé par minimisation itérative du critère suivant :

$$\sigma(\Theta) = \sum_{i<j} (\cos \phi_{ij} - \tau_{ij})^2, \tag{3.5}$$

La généralisation au positionnement sur la surface bi-dimensionnelle d'une sphère est immédiate : chaque variable est représentée par des coordonnées sphériques $(1, \theta_{i1}, \theta_{i2})$ équivalentes en coordonnées Cartésiennes à :
$\mathbf{x_i} = (\cos \theta_{i1} \sin \theta_{i2}, \sin \theta_{i1} \sin \theta_{i2}, \cos \theta_{i1})$. Le cosinus de l'angle ϕ_{ij} entre deux variables i and j peut être calculé grâce au produit scalaire $< \mathbf{x_i}, \mathbf{x_j} >$ et le critère (3.5) est minimisé par rapport à $\Theta = (\theta_{11}, \theta_{12}, ..., \theta_{n1}, \theta_{n2})$.

3 Positionnement Euclidien de données imprécises

3.1 Dissimilarités de type intervalle

On suppose maintenant que les données disponibles consistent en une matrice $\Delta = ([\delta_{ij}])$ de dissimilarités exprimées sous forme d'intervalles. Chaque intervalle $[\delta_{ij}] = [\delta_{ij}^-; \delta_{ij}^+]$ s'interprète comme l'ensemble des valeurs possibles pour la dissimilarité δ_{ij}, inconnue, entre l'objet i et l'objet j.

3.1.1 Modèle général

Puisque la position relative des objets n'est pas décrite de manière précise, on choisit d'associer à un objet, non plus un point dans l'espace, mais une région R_i, et l'on définit les distances minimales et maximales entre objets de la façon suivante :

$$d_{ij}^- = \min_{\mathbf{x}_i \in R_i, \mathbf{x}_j \in R_j} \|\mathbf{x}_i - \mathbf{x}_j\| \tag{3.6}$$

$$d_{ij}^+ = \max_{\mathbf{x}_i \in R_i, \mathbf{x}_j \in R_j} \|\mathbf{x}_i - \mathbf{x}_j\|. \tag{3.7}$$

L'approche pratique la plus simple consiste à définir chaque région R_i comme une sphère dans l'espace, paramétrée par un centre $\mathbf{c}_i \in \mathbb{R}^p$ et un rayon r_i. On obtient alors un modèle à $n(p+1)$ paramètres (n centres définis par p coordonnées, et n rayons).

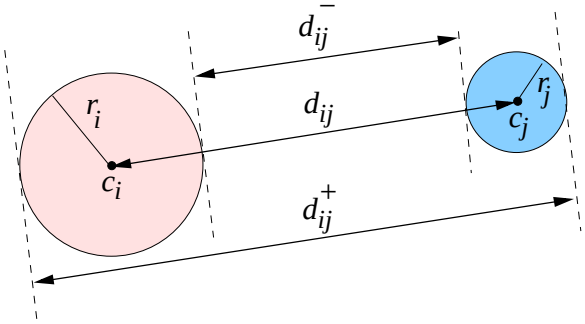

FIG. 3.3 – Distances minimales et maximales entre deux régions sphériques.

Comme le montre la figure 3.3, les distances entre les région se calculent alors très facilement par les équations suivantes :

$$d_{ij}^- = \max(0, d_{ij} - r_i - r_j) \tag{3.8}$$
$$d_{ij}^+ = d_{ij} + r_i + r_j, \tag{3.9}$$

où $d_{ij} = \|\mathbf{c}_i - \mathbf{c}_j\|$ désigne la distance euclidienne entre les centres $\mathbf{c}_i$ et $\mathbf{c}_j$.
Pour que le modèle soit complet, il reste à définir comment juger de la proximité entre les données d'entrée et les distances minimales et maximales produites par le modèle. Nous avons proposé deux variantes qui sont décrites dans ce qui suit.

3.1.2 Ajustement par les moindres carrés

Une première idée consiste à chercher à approcher au mieux, au sens des moindres carrés, les bornes minimales et maximales des dissimilarités. Le critère de stress (3.2) se généralise de manière naturelle :

$$\sigma'(\mathcal{R}) = \sum_{i<j} (d_{ij}^- - \delta_{ij}^-)^2 + \sum_{i<j} (d_{ij}^+ - \delta_{ij}^+)^2, \tag{3.10}$$

où $\mathcal{R}$ désigne l'ensemble des n régions $\{R_1, \ldots, R_n\}$. Les paramètres du modèle peuvent alors être déterminés simplement en minimisant $\sigma'(\mathcal{R})$ par rapport à $\mathcal{R}$ par une technique de descente du gradient.
Notons que la minimisation de $\sigma'(\mathcal{R})$ est un problème d'optimisation sous contraintes car les valeurs des rayons doivent être positifs. On peut s'affranchir d'utiliser une procédure d'optimisation sous contrainte en posant $r_i = \rho_i^2$ et en optimisant les ρ_i.

Le modèle obtenu à l'optimum a des propriétés intéressantes. On peut en effet montrer en premier lieu que lorsque les dissimilarités d'entrée sont précises ($\delta_{ij}^- = \delta_{ij}^+$) alors les rayons convergent vers une valeur nulle, la méthode généralise la MDS classique. D'autre part, on peut également montrer [26] que chaque rayon r_k dépend linéairement de la quantité

$$s_k = \sum_{i \neq k} (\delta_{ik}^+ - \delta_{ik}^-), \tag{3.11}$$

qui est une mesure globale de l'imprécision concernant le positionnement de l'objet k relativement aux autres objets. Cette remarque est importante car elle donne des pistes d'interprétation des résultats obtenus en reliant la taille des régions R_k à l'imprécision concernant l'objet k.

3.1.3 Ajustement possibiliste

Le modèle précédent est sous doute l'extension la plus naturelle du modèle MDS classique. Cependant, la carte qu'il fournit n'est qu'une représentation approchée des données. En d'inspirant des travaux de Tanaka dans le domaine de la régression possibiliste, nous avons proposé un second modèle fournissant une représentation plus fidèle des données.
Supposons que les centres des régions aient été déterminés auparavant (on peut par exemple utiliser le modèle des moindres carrés pour les calculer). Dans ce cas, les distances d_{ij} entre les centres des régions sont fixées. On peut alors chercher les rayons les plus faibles possibles qui respectent la contrainte suivante :

$$[\delta_{ij}^-, \delta_{ij}^+] \subseteq [d_{ij}^-, d_{ij}^+] \quad \forall i, j. \tag{3.12}$$

L'idée est de rendre compte de façon exacte de l'imprécision contenue dans les dissimilarités. Il s'avère que le problème ainsi posé conduit à la résolution d'un programme linéaire très simple. En effet, on pose :

$$\min_{\mathbf{r}} \sum_{i=1}^{n} r_i \tag{3.13}$$

sous les contraintes :

$$d_{ij}^- \leq \delta_{ij}^- \quad \forall i, j \tag{3.14}$$

$$d_{ij}^+ \geq \delta_{ij}^+ \quad \forall i, j \tag{3.15}$$

$$r_i \geq 0 \quad \forall i = 1, n\ , \tag{3.16}$$

Dans (3.13), $\mathbf{r}$ désigne le vecteur des rayons $(r_1, r_2, ..., r_n)^t$. En utilisant les expressions de d_{ij}^- et d_{ij}^- données par (3.8) et (3.9), les contraintes (3.14) et (3.15) peuvent se réécrire de la manière suivante :

$$\max(0, d_{ij} - r_i - r_j) \leq \delta_{ij}^- \tag{3.17}$$

$$r_i + r_j \geq \delta_{ij}^+ - d_{ij}, \tag{3.18}$$

ce qui peut se formuler de manière plus compacte sous la forme :

$$r_i + r_j \geq \max(d_{ij} - \delta_{ij}^-, \delta_{ij}^+ - d_{ij}) \quad \forall i, j. \tag{3.19}$$

La minimisation de (3.13) sous les constraintes (3.16) et (3.19) est un programme linéaire. On peut observer que le problème a toujours une solution réalisable, puisque $d_{ij}^- \to 0$ et $d_{ij}^+ \to \infty$ quand r_i et $r_j \to \infty$. Les paramètres du modèle peuvent donc toujours être obtenus quelles que soient les dissimilarités d'entrée.

Remarque 3 Contrairement à l'ajustement par moindres carrés, l'ajustement possibiliste ne conduit pas à des rayons nuls lorsque toutes les dissimilarités d'entrée sont précises ($\delta_{ij}^- = \delta_{ij}^+$) mais erronées. En effet, le modèle obtenu représente à la fois l'*imprécision* dans les données (la taille des intervalles) et l'*adéquation* du modèle aux données (i.e., le choix du modèle Euclidien, la dimension de la configuration cherchée, et les erreurs d'estimation).

Exemple 3.2 *(Jeu de données des villes européennes)* On a demandé à un sujet humain d'évaluer les distance entre plusieurs villes européennes. Vue la difficulté supposée de la tâche, l'évaluateur était autorisé à fournir ses estimations sous forme d'intervalles de distances. Ces intervalles sont donnés en table 3.1. Les résultats obtenus avec les deux méthodes sont donnés en figure 3.4 et 3.5. Notons encore que l'orientation nord/sud et est/ouest résulte encore d'un choix subjectif. Les centres dans le modèle possibiliste ont été initialisés en utilisant les valeurs obtenus avec l'ajustement des moindres carrés. La figure 3.4 suggère que l'évaluateur a eu plus de difficultés à estimer les grandes distances, les cercles en effet étant plus larges pour les villes situées à la périphérie de la carte (Dublin, Berlin, Madrid et Rome). L'ajustement par les moindres carrés est donc capable de rendre compte de l'imprécision globale dans les données d'entrée. Dans la représentation obtenue par l'ajustement possibiliste, les cercles sont plus larges. Ceci est conforme à ce que l'on attendait, car le modèle possibiliste rend compte à la fois de l'imprécision et de l'incertitude des données. La figure 3.6 montre un diagramme de Shepard modifié, dans lequel les distances hautes et basses sont représentées en fonction des dissimilarités maximales et minimales. On constate que les contraintes d'inclusion sont bien respectées.

Remarque 4 Pour être complet, il faut noter que, dans le même ordre d'idée, on pourrait tenter de résoudre le problème dual consistant à *maximiser* les rayons des hypersphères sous les contraintes :

$$[d_{ij}^-, d_{ij}^+] \subseteq [\delta_{ij}^-, \delta_{ij}^+] \quad \forall i, j. \tag{3.20}$$

	Paris	Dublin	London	Frankfort	Berlin	Marseille	Rome
Paris	0						
Dublin	[850 ;1050]	0					
London	[250 ;450]	[450 ;650]	0				
Frankfort	[500 ;700]	[1300 ;1700]	[600 ;800]	0			
Berlin	[900 ;1100]	[1700 ;2300]	[1000 ;1400]	[450 ;650]	0		
Marseille	[800 ;1000]	[1800-2400[	[1100 ;1400]	[1000 ;1200]	[1600 ;2000]	0	
Rome	[1400 ;1800]	[2200 ;2800]	[1800 ;2100]	[1000 ;1200]	[1700 ;2300]	[700 ;900]	0
Madrid	[1500 ;1900]	[1700 ;2300]	[1700 ;2000]	[1500 ;2500]	[2100 ;2800]	[900 ;1100]	[1200 ;1800]

TAB. 3.1 – Intervalles de distance estimés par l'évaluateur.

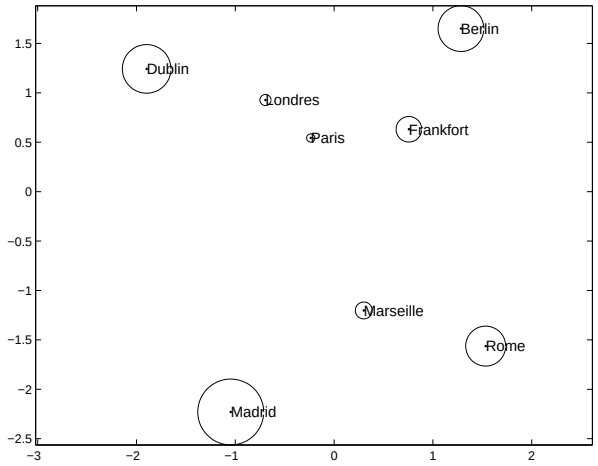

FIG. 3.4 – Jeu de données des villes : configuration obtenue par l'ajustement des moindres carrés.

Cette nouvelle forme d'ajustement, que l'on pourrait appeler *modèle de nécessité*, en référence à Tanaka, amène encore à la résolution d'un programme linéaire. Ce programme n'a cependant pas toujours de solution. De plus, les expériences que nous avons menées nous ont montré que les représentations obtenues sont parfois difficilement interprétables.

3.2 Extension à des dissimilarités floues

On suppose maintenant que les dissimilarités sont fournies sous forme de nombres flous. Ces données peuvent provenir d'une évaluation linguistique d'un unique sujet humain (en utilisant des termes comme "très proches", "peu différents", etc...) ou d'une synthèse de plusieurs réponses fournies par un ensemble de sujets. Le modèle ainsi que les algorithmes proposés pour les intervalles s'étendent très facilement, comme l'expliquent les paragraphes qui suivent.

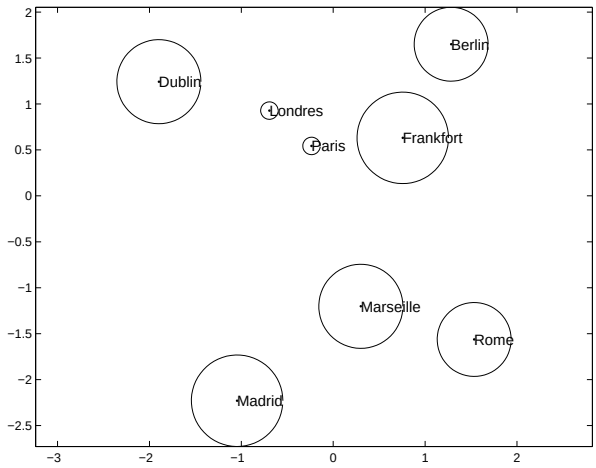

FIG. 3.5 – Jeu de données des villes : configuration obtenue par l'ajustement possibiliste.

3.2.1 Modèle

Il est maintenant naturel de représenter chaque objet par une *région floue* $\widetilde{R}_i$ dans $\mathbb{R}^p$ définie par une fonction d'appartenance $\mu_{\widetilde{R}_i}$. En appliquant le principe d'extension [105], la distance floue entre deux régions $\widetilde{R}_i$ et $\widetilde{R}_j$ peut être définie par :

$$\mu_{\widetilde{d}_{ij}}(w) = \sup_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \min(\mu_{\widetilde{R}_i}(\mathbf{x}), \mu_{\widetilde{R}_j}(\mathbf{y})), \tag{3.21}$$

où le supremum est calculé sous la contrainte $\|\mathbf{x}-\mathbf{y}\| = w$. Si $\widetilde{R}_i$ et $\widetilde{R}_j$ sont des nombres flous multidimensionnels [65, p.146], alors chaque α-coupe de $\widetilde{d}_{ij}$ est un intervalle fermé ${}^{\alpha}\widetilde{d}_{ij} = [{}^{\alpha}\widetilde{d}^{-}_{ij}, {}^{\alpha}\widetilde{d}^{+}_{ij}]$, dont les bornes sont respectivement le minimum et le maximum des distances entre les α-coupes de $\widetilde{R}_i$ et $\widetilde{R}_j$. Comme précédemment, nous avons opté pour une représentation simple des objets, dans laquelle les α-coupes sont des hypershères imbriquées de rayon ${}^{\alpha}r_i$ et de centre $\mathbf{c}_i$, de telle sorte que

$$\begin{aligned} {}^{\alpha}\widetilde{d}^{-}_{ij} &= \max(0, d_{ij} - {}^{\alpha}r_i - {}^{\alpha}r_j) && (3.22)\\ {}^{\alpha}\widetilde{d}^{+}_{ij} &= d_{ij} + {}^{\alpha}r_i + {}^{\alpha}r_j, && (3.23) \end{aligned}$$

où d_{ij} désigne, comme auparavant, la distance Euclidienne distance entre les centres $\mathbf{c}_i$ et $\mathbf{c}_j$.

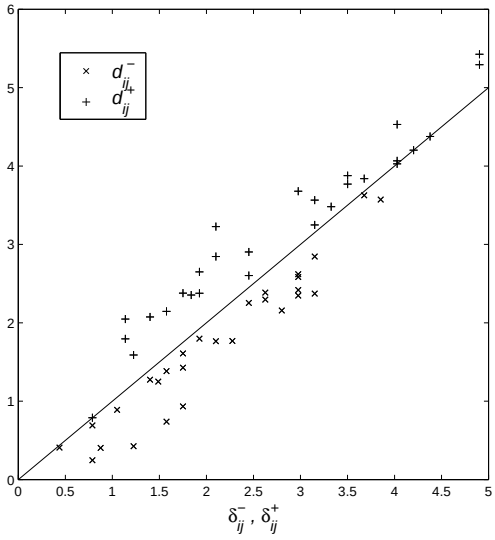

FIG. 3.6 – Jeu de données des villes : distances minimales et maximales entre deux régions après ajustement possibiliste.

3.2.2 Ajustement par les moindres carrés

Pour ajuster le modèle, un ensemble prédéfini de niveaux $\{\alpha_i\}_{i=1,c}$ doit être choisi, avec la convention :

$$1 = \alpha_1 > \ldots > \alpha_c = 0 \tag{3.24}$$

Ensuite, il suffit d'étendre la fonction de stress (5.15) de la façon suivante :

$$\sigma''(\widetilde{\mathcal{R}}) = \sum_{k=1}^{c} \sum_{i<j} ({}^{\alpha_k}\widetilde{d}_{ij}^- - {}^{\alpha_k}\widetilde{\delta}_{ij}^-)^2 + \sum_{k=1}^{c} \sum_{i<j} ({}^{\alpha_k}\widetilde{d}_{ij}^+ - {}^{\alpha_k}\widetilde{\delta}_{ij}^+)^2, \tag{3.25}$$

où $\widetilde{\mathcal{R}}$ désigne l'ensemble des régions floues $\widetilde{R}_i$, et ${}^0\widetilde{x}$ représente, par convention, le support d'un nombre flou $\widetilde{x}$. Notons que la fonction de stress est équivalente au critère des moindres carrés flous proposé par Diamond [29, 30] et étendu par Ming and al. [73]. Le nombre de paramètres du modèle est de $n(p+c)$: n centres définis par p coordonnées c_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$ et $n \times c$ rayons ${}^{\alpha_k}r_i$, $i = 1, \ldots, n$, $k = 1, \ldots, c$. Pour imposer la contrainte d'imbrication entre les hypersphères, une nouvelle paramétrisation de la forme

$${}^{\alpha_k}r_i = \sum_{h=1}^{k} \alpha_h \rho_i^2, \tag{3.26}$$

permet de transformer un problème d'optimisation sous contraintes en un problème sans contrainte.

3.2.3 Ajustement possibiliste

En suivant la même idée que celle développée au paragraphe 3.1.3, nous généralisons la condition (3.12) par

$$\widetilde{\delta}_{ij} \subseteq \widetilde{d}_{ij}, \quad \forall i, j \tag{3.27}$$

où $\subseteq$ désigne l'inclusion floue standard, i.e.

$$\mu_{\widetilde{\delta}_{ij}} \leq \mu_{\widetilde{d}_{ij}}, \quad \forall i, j. \tag{3.28}$$

Puisque $\widetilde{\delta}_{ij}$ et $\widetilde{d}_{ij}$ sont des nombres flous, cette condition peut s'exprimer sous la forme

$$[{}^{\alpha_k}\widetilde{\delta}_{ij}^{-}, {}^{\alpha_k}\widetilde{\delta}_{ij}^{+}] \subseteq [{}^{\alpha_k}\widetilde{d}_{ij}^{-}, {}^{\alpha_k}\widetilde{d}_{ij}^{+}] \quad \forall i, j, k. \tag{3.29}$$

Comme dans le paragraphe 3.1.3, on suppose que les centres $\mathbf{c}_i$, $i = 1, \ldots, n$ ont été déterminés auparavant en utilisant, par exemple, la procédure des moindres carrés décrite au paragraphe précédent. Le problème consiste alors à trouver les rayons les plus faibles satisfaisant la condition (3.29). La solution du problème est trouvée en résolvant, successivement, les programmes linéaires suivants, en commençant avec $k = 1$ jusqu'à $k = c$:

$$\min_{{}^{\alpha_k}\mathbf{r}} \sum_{i=1}^{n} {}^{\alpha_k}r_i \tag{3.30}$$

sous les contraintes :

$${}^{\alpha_k}r_i + {}^{\alpha_k}r_j \geq \max(d_{ij} - {}^{\alpha_k}\delta_{ij}^{-}, {}^{\alpha_k}\delta_{ij}^{+} - d_{ij}) \quad \forall i, j \tag{3.31}$$

$${}^{\alpha_k}r_i \geq \begin{cases} 0 & \text{if } k = 1 \\ {}^{\alpha_{k-1}}r_i & \text{if } k > 1 \end{cases} \quad \forall i = 1, n. \tag{3.32}$$

avec ${}^{\alpha_k}\mathbf{r} = ({}^{\alpha_k}r_1, \ldots, {}^{\alpha_k}r_n)^t$.

Exemple 3.3 Nous revenons à l'exemple sur la perception des couleurs en considérant maintenant les réponses de plusieurs sujets. Ceux-ci étaient classés en deux groupes distincts : certains avaient une vision normale, alors que d'autres souffraient de daltonisme. La perception des couleurs dans chaque groupe a été résumée en utilisant un nombre flou triangulaire de support défini par les réponses minimale et maximale des sujets et de noyau égal à la moyenne des réponses. La partie supérieure de la figure 3.7 présente les résultats obtenus pour deux α-coupes (support et noyau) par ajustement possibiliste. On voit que la configuration circulaire des couleurs est bien retrouvée dans le groupe de sujets sains, et la taille faible des cercles les plus foncés indique une bonne adéquation du modèle Euclidien ainsi qu'une très bonne

précision des réponses moyennes. On note, comme c'était attendu, une plus grande confusion des couleurs dans le groupe de sujets daltoniens. Comme dans les résultats de Helm, le cercle des couleurs est légèrement déformé. Il apparaît de façon évidente, au travers de supports et de noyaux plus larges que pour le premier groupe, que les réponses du second groupe sont plus confuses. L'ajustement par moindres carrés a également été appliqué à ces mêmes données. Comme auparavant, seules $c = 2$ α-coupes ont été retenues. Les configurations obtenues pour les deux groupes sont données en partie inférieure de la figure 3.7. Encore une fois, on retrouve la configuration annulaire des couleurs, avec une légère déformation pour les sujets daltoniens. Les régions représentant les couleurs sont plus précises que celles obtenues avec l'ajustement possibiliste, et leur noyau est réduit à un point en raison du choix d'un nombre triangulaire pour représenter les dissimilarités. De façon surprenante à première vue, les configurations obtenues pour les deux groupes sont assez similaires. Cela peut s'expliquer par le fait que l'imprécision obtenue grâce à l'ajustement par les moindres carrés ne reflète pas les erreurs d'estimation contenues dans les données mais seulement l'imprécision des dissimilarités (qui est comparable dans les deux groupes).

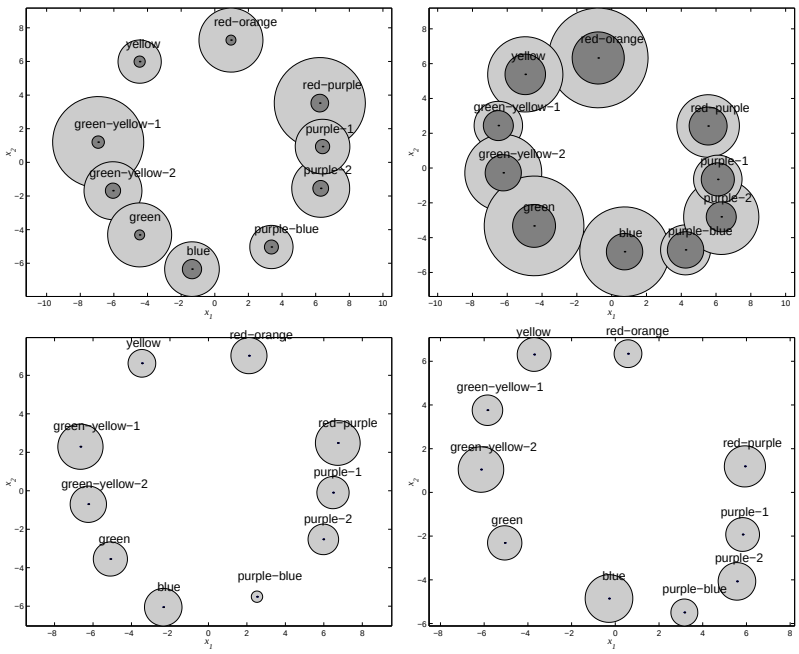

FIG. 3.7 – Jeu de données des couleurs ; (en haut) : ajustement possibiliste avec (à gauche) les sujets normaux ; (à droite) les sujets daltoniens. (en bas) : ajustement des moindres carrés. Les supports sont représentés en gris clair, et les noyaux en gris foncés.

4 Positionnement multidimensionnel sphérique

4.1 Données de type intervalle

On suppose maintenant que les données consistent en des corrélations exprimées sous forme d'intervalles. Chaque intervalle $[\tau_{ij}] = [\tau_{ij}^-; \tau_{ij}^+]$ est interprété comme l'ensemble des valeurs possibles pour la corrélation entre deux variables ou attributs i et j. De telles corrélations imprécises sont disponibles par exemple lorsque des objets sont décrits par des variables intervalles. Plusieurs coefficients de corrélation classiques ont été étendus pour prendre en compte ce type de données (par exemple le coefficient de corrélation de rang de Kendall [58, 25] que nous avons étendu ou encore le coefficient le plus courant de Bravais-Pearson [71]).

4.1.1 Modèle

Comme dans le modèle sphérique classique, les objets sont représentés sur une hypersphère $\mathbf{S}^p$ de centre O et de rayon 1 dans un espace de dimension p : $\mathbf{S}^2$ est un cercle et $\mathbf{S}^3$ est une sphère.
Lorsque les corrélations sont précises, on a vu dans le paragraphe 2.3 que le modèle classique permet de représenter τ_{ij} par le cosinus de l'angle $\widehat{(\mathbf{x}_i, \mathbf{x}_j)}$. Cet angle, défini dans $[0, \pi]$, est celui que forment les segments les points $\overline{O\mathbf{x}_i}$ et $\overline{O\mathbf{x}_j}$ dans le plan engendré par les trois points $\mathbf{x}_i$, $\mathbf{x}_j$ et O de $\mathbf{S}^p$. Comme dans le modèle Euclidien, il est naturel de représenter l'imprécision relative à une variable par une région S_i de $\mathbf{S}^p$. Par conséquent, une paire de régions, S_i et S_j, situées sur $\mathbf{S}^p$ induisent un ensemble de cosinus que l'on peut caractériser par leur minimum et leur maximum :

$$\min_{\mathbf{x}_i \in S_i, \mathbf{x}_j \in S_j} \cos\left(\widehat{(\mathbf{x}_i, \mathbf{x}_j)}\right) = \cos\left(\max_{\mathbf{x}_i \in S_i, \mathbf{x}_j \in S_j} \left(\widehat{(\mathbf{x}_i, \mathbf{x}_j)}\right)\right) \tag{3.33}$$

$$\max_{\mathbf{x}_i \in S_i, \mathbf{x}_j \in S_j} \cos\left(\widehat{(\mathbf{x}_i, \mathbf{x}_j)}\right) = \cos\left(\min_{\mathbf{x}_i \in S_i, \mathbf{x}_j \in S_j} \left(\widehat{(\mathbf{x}_i, \mathbf{x}_j)}\right)\right). \tag{3.34}$$

Les angles maximaux et minimaux correspondant sont donc :

$$\phi_{ij}^- = \max_{\mathbf{x}_i \in S_i, \mathbf{x}_j \in S_j} \left(\widehat{(\mathbf{x}_i, \mathbf{x}_j)}\right) \tag{3.35}$$

$$\phi_{ij}^+ = \min_{\mathbf{x}_i \in S_i, \mathbf{x}_j \in S_j} \left(\widehat{(\mathbf{x}_i, \mathbf{x}_j)}\right). \tag{3.36}$$

En pratique, il suffit de décider d'une forme paramétrée pour chaque région pour que le modèle soit complet. On choisit d'associer à chaque région un centre $\mathbf{c}_i$ sur $\mathbf{S}^p$ et un angle d'imprécision $\beta_i \in [0, \pi]$ de sorte que :

$$S_i = \left\{ \mathbf{x} \in \mathbf{S}^p \ / \ \widehat{(\mathbf{x}, \mathbf{c}_i)} \leq \beta_i \right\}. \tag{3.37}$$

Chaque région S_i est donc un arc circulaire quand $p = 2$, et une calotte sphérique lorsque $p = 3$. L'intervalle $[\phi_{ij}^+, \phi_{ij}^-]$ définit l'ensemble des angles $\widehat{(\mathbf{x}_i, \mathbf{x}_j)}$ tels que $\mathbf{x}_i \in S_i$ et $\mathbf{x}_j \in S_j$. La fonction cosinus étant décroissante, les intervalles $[\cos(\phi_{ij}^-), \cos(\phi_{ij}^+)]$ définissent donc l'ensemble des valeurs possibles de cosinus que l'on peut obtenir pour toute paire $(\mathbf{x}_i, \mathbf{x}_j) \in (S_i, S_j)$.
Soit ϕ_{ij} l'angle $\widehat{(\mathbf{c}_i, \mathbf{c}_j)}$. Tout comme dans le cas net, il peut être calculé comme le produit scalaire des centres $\mathbf{c}_i$ et $\mathbf{c}_j$:

$$\phi_{ij} = \arccos \langle \mathbf{c}_i, \mathbf{c}_j \rangle . \tag{3.38}$$

Quelle que soit la dimension de l'espace de représentation, les angles $(\phi_{ij}^+, \phi_{ij}^-)$ peuvent se mesurer dans le plan $(O, \mathbf{c}_i, \mathbf{c}_j)$ dont une représentation est donnée en figure 3.8.

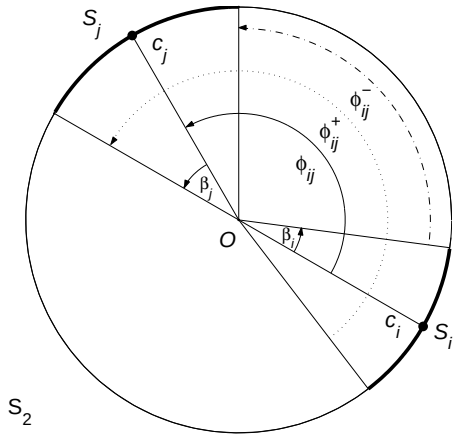

FIG. 3.8 – Angles minimal et maximal entre deux régions.

On voit que les angles sont définis par les équations suivantes :

$$\phi_{ij}^- = \min(\pi, \phi_{ij} + \beta_i + \beta_j) \tag{3.39}$$

$$\phi_{ij}^+ = \max(0, \phi_{ij} - \beta_i - \beta_j) . \tag{3.40}$$

Les fonctions min et max sont nécessaires pour tenir compte de deux situations particulières :

- $\exists (\mathbf{x}_i, \mathbf{x}_j) \in (S_i, S_j)$ / $\widehat{(\mathbf{x}_i, \mathbf{x}_j)} = \pi$: l'angle maximal π correspondant à des oppositions complètes des variables peut être atteint, dans ce cas ϕ_{ij}^- est égal à π ;
- $\exists (\mathbf{x}_i, \mathbf{x}_j) \in (S_i, S_j)$ / $\widehat{(\mathbf{x}_i, \mathbf{x}_j)} = 0$: l'angle minimal 0 correspondant à des positions identiques est atteint de telle sorte que ϕ_{ij}^+ est égal à 0.

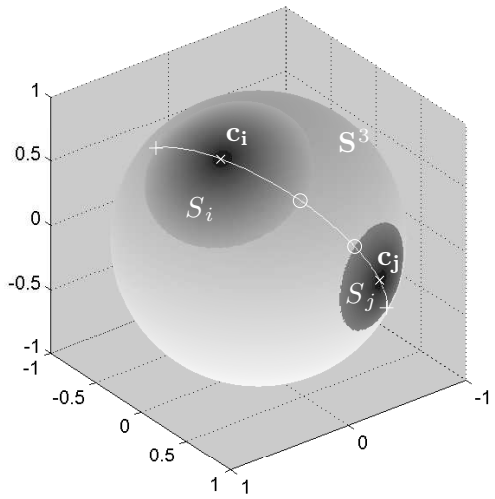

FIG. 3.9 – Représentation des deux régions sur la sphère S^3.

Sur la figure 3.9 sont représentés deux objets (ou variables) sur la sphère $\mathbf{S}^3$. Les régions S_i et S_j sont donc des calottes sphériques. Chaque centre $\mathbf{c}_i$ and $\mathbf{c}_j$ est représenté par le signe $\times$. Inclus dans le plan $(O, \mathbf{c}_i, \mathbf{c}_j)$, l'arc de cercle blanc délimité par le signe $+$ connecte les points les plus éloignés respectivement contenus dans S_i et S_j, alors que l'arc délimité par les symboles o connecte les points les plus proches : ces deux arcs définissent respectivement les angles ϕ_{ij}^- and ϕ_{ij}^+.

Le modèle étant posé, il reste à montrer comment déterminer les centres $\mathbf{c}_i$ et les angles d'imprécision β_i pour chaque objet i, de manière à représenter au mieux les corrélations d'entrée. Chaque centre $\mathbf{c}_i$ situé sur la sphère $\mathbf{S}^p$ sera caractérisé par des coordonnées sphériques $(1, \theta_{i1}, \ldots, \theta_{i(p-1)})$, où $\theta_{i(p-1)} \in [0, 2\pi]$ et $\theta_{iq} \in [0, \pi]$ pour tout $q < p-1$. La démarche proposée pour le positionnement Euclidien peut être intégralement reproduite : nous présentons d'abord une méthode d'ajustement par moindres carrés puis une méthode possibiliste.

4.1.2 Ajustement par les moindres carrés

On propose de minimiser le critère suivant :

$$\sigma'(\mathcal{S}) = \sum_{i<j} \left(\cos(\phi_{ij}^-) - \tau_{ij}^-\right)^2 + \sum_{i<j} \left(\cos(\phi_{ij}^+) - \tau_{ij}^+\right)^2, \qquad (3.41)$$

où $\mathcal{S}$ désigne l'ensemble des n régions $\{S_1, \ldots, S_n\}$. Les np paramètres du modèle(n centres definis par $p-1$ coordonnées θ_{iq} et n angles β_i) peuvent, encore une fois, être déterminés en minimisant $\sigma'(\mathcal{S})$ par rapport à $\mathcal{S}$ en utilisant une technique itérative de descente de gradient. Pour éviter d'utiliser des routines d'optimisation sous contraintes, chaque paramètre β_i est remplacé par par un paramètre $b_i \in \mathbb{R}$ transformé par une fonction monotone dérivable $e(x) : \mathbb{R} \to [0, \pi]$:

$$e(x) = \frac{\pi}{1 + \exp(-x)} \ .$$

4.1.3 Ajustement possibiliste

Tout comme précédemment, on suppose que les centres des régions ont été déterminés au préalable, en minimisant, par exemple, le critère (3.41). Ensuite, on cherche les angles d'imprécision les plus faibles vérifiant :

$$[\tau_{ij}^-, \tau_{ij}^+] \subseteq \left[\cos(\phi_{ij}^-), \cos(\phi_{ij}^+)\right], \ \forall i, j. \tag{3.42}$$

Les distances angulaires distances $[\phi_{ij}^+, \phi_{ij}^-]$ peuvent alors être interprétées comme des vues "pessimistes" des intervalles de corrélation $[\tau_{ij}^-, \tau_{ij}^+]$. Pour les déterminer, on résout le problème d'optimisation suivant :

$$\text{Minimiser} \sum_{i=1}^{n} \beta_i, \tag{3.43}$$

sous les contraintes :

$$\cos(\phi_{ij}^-) \leq \tau_{ij}^-, \ \forall i, j \tag{3.44}$$

$$\cos(\phi_{ij}^+) \geq \tau_{ij}^+, \ \forall i, j \tag{3.45}$$

$$\beta_i \geq 0, \ \forall i \tag{3.46}$$

$$\beta_i \leq \pi, \ \forall i. \tag{3.47}$$

En utilisant (3.39) et (3.40), les contraintes (3.44) et (3.45) se simplifient en :

$$\begin{aligned}
\cos(\phi_{ij}^-) \leq \tau_{ij}^- \quad &\Leftrightarrow \quad \cos(\min(\pi, \phi_{ij} + \beta_i + \beta_j)) \leq \tau_{ij}^- && (3.48)\\
&\Leftrightarrow \quad \min(\pi, \phi_{ij} + \beta_i + \beta_j) \geq \arccos(\tau_{ij}^-) && (3.49)\\
&\Leftrightarrow \quad \phi_{ij} + \beta_i + \beta_j \geq \arccos(\tau_{ij}^-) && (3.50)\\
&\Leftrightarrow \quad \beta_i + \beta_j \geq \arccos(\tau_{ij}^-) - \phi_{ij} \ , && (3.51)
\end{aligned}$$

et

$$\begin{aligned}
\cos(\phi_{ij}^{+}) \geq \tau_{ij}^{+} &\Leftrightarrow \cos(\max(0, \phi_{ij} - \beta_i - \beta_j)) \geq \tau_{ij}^{+} && (3.52)\\
&\Leftrightarrow \max(0, \phi_{ij} - \beta_i - \beta_j) \leq \arccos(\tau_{ij}^{+}) && (3.53)\\
&\Leftrightarrow \phi_{ij} - \beta_i - \beta_j \leq \arccos(\tau_{ij}^{+}) && (3.54)\\
&\Leftrightarrow \beta_i + \beta_j \geq \phi_{ij} - \arccos(\tau_{ij}^{+}). && (3.55)
\end{aligned}$$

Au final, chaque contrainte (3.44) et (3.45) peut donc s'exprimer sous la forme suivante :

$$\beta_i + \beta_j \geq \max\left(\arccos(\tau_{ij}^{-}) - \phi_{ij}, \phi_{ij} - \arccos(\tau_{ij}^{+})\right), \ \forall i, j. \tag{3.56}$$

La minimisation de (3.43) sous les contraintes (3.46), (3.47) and (3.56) est un un programme linéaire qui a toujours une solution : en effet, si $\beta_i \to \pi$ et $\beta_j \to \pi$ alors $\phi_{ij}^{-} \to \pi$ et $\phi_{ij}^{+} \to 0$ et donc $\cos(\phi_{ij}^{-}) \to -1$ and $\cos(\phi_{ij}^{+}) \to 1$.

4.2 Données floues

L'extension à des données floues ne présente pas de difficulté particulière et sera donc peu développée : nous en donnons seulement les grandes lignes. On suppose ici que les données disponibles consistent en une matrice carrée de corrélations exprimées sous forme de nombres flous $\widetilde{\tau}_{ij}$. Le concept de corrélation floue a été suggéré dans plusieurs travaux comme ceux de [71] ou [58, 25] pour mesurer le degré de dépendance entre des attributs flous. La même démarche que dans le cas Euclidien est suivie : les données d'entrée étant floues, il est naturel de représenter les objets (ou ici les variables) sous forme de régions floues $\widetilde{S}_i$ dans $\mathbf{S}^p$. Chaque région floue est paramétrée par son centre et des angles d'imprécision croissants générant ainsi des calottes hypersphériques imbriquées. Le principe d'extension permet de définir la distance angulaire floue entre deux régions $\widetilde{S}_i$ and $\widetilde{S}_j$ comme :

$$\mu_{\widetilde{\phi}_{ij}}(w) = \sup_{\mathbf{x},\mathbf{y}\in\mathbf{S}^p \ / \ \widehat{(\mathbf{x},\mathbf{y})}=w} \min(\mu_{\widetilde{S}_i}(\mathbf{x}), \mu_{\widetilde{S}_j}(\mathbf{y})). \tag{3.57}$$

Ensuite un ajustement du type moindres carrés ou possibiliste permet de déterminer les angles d'imprécision tels que les α-coupes des coefficients de corrélation flous soient le plus en accord avec celles de la distance angulaire floue.

4.3 Exemple d'application : données sensorielles

En analyse sensorielle, on utilise des évaluateurs humains pour comprendre comment sont perçus des produits. On cherche par exemple à établir un *profil sensoriel* de chaque produit en décrivant la sensation perçue (qu'il s'agisse

Produits	Opaque	Brillant	Granuleux	Clair/Foncé	Nacré
1	(96,99.9,99.9)	(97.7,100,100)	(96,99.9,99.9)	(96.8,100,100)	(26.7,30,53.4)
2	(19.4,28,31.4)	(60.5,69.4,75.1)	(26.1,37.4,43.7)	(64.4,74.1,80.8)	(59.5,68.1,78)
3	(36,48.3,68.8)	(3.2,10.3,14.6)	(17.7,25.7,31.9)	(42.7,54.4,67.5)	(66.9,78,86)
4	(94.8,96.2,99.1)	(43.9,56.5,63.6)	(84.9,92.1,96.4)	(92.3,94,100)	(58.9,71,77.1)
5	(48.1,58.1,65.5)	(26.7,42.3,48.9)	(18.5,35.3,46.8)	(50.9,59.7,71)	(42.3,58.5,64.9)
6	(46.8,56,65.7)	(0,0.3,4.2)	(48.5,60.3,68.8)	(31.9,44.8,62.8)	(51.7,68.1,72.8)
7	(0,0,0)	(44,50.9,62.6)	(0.3,0.3,0.3)	(0.1,0.1,0.1)	(54.2,71.2,83.5)
8	(6.8,12.4,22.4)	(81.5,85.4,96.6)	(79.2,85.8,95.2)	(16.1,20.6,40.1)	(82.1,91.9,99.9)

TAB. 3.2 – Scores flous attribués par l'expert aux 8 produits suivant 5 descripteurs.

de l'odorat, du goût (!), de l'audition, ou du toucher) grâce à un score, exprimé sur une échelle bornée, attribué à plusieurs variables ou *descripteurs*. Pour tenir compte des difficultés de notation et de l'imprécision inhérente au processus d'évaluation, les notations sont en général répétées plusieurs fois et moyennées. Or les répétitions ont un coût elevé pour l'utilisateur. Il peut être plus judicieux de procéder à une répétition unique en autorisant l'évaluateur à fournir une réponse imprécise, sous forme d'un nombre flou triangulaire par exemple. Nous rapportons ici une étude effectuée en collaboration avec le laboratoire sensoriel de PSA. Les produits étaient 8 plaques de plastique translucides décrites suivant 5 variables : *Opaque*, *Brillant*, *Granuleux*,*Clair/Foncé* et *Nacré*. Les notes floues fournies par un des experts ayant participé à l'étude sont données dans la table 3.2 et représentées en figure 3.10.

Un des résultats attendus de l'étude concernait la compréhension des relations entre les différents descripteurs. Classiquement, la similarité entre deux variables est mesurée par un coefficient de corrélation. Lorsque l'on estime que l'information pertinente réside uniquement dans le classement des produits suivant chaque descripteur, on utilise une coefficient de corrélation ordinale comme celui du tau de Kendall. Ce coefficient a été étendu à des données floues [58, 25] et a donc été retenu pour la suite. Les coefficients de corrélation flous obtenus à partir des données du tableau 3.2 sont représentés en figure 3.11. La représentation MDS obtenue sur une sphère avec l'ajustement des moindres carrés est donné en figure 3.12 ainsi que le diagramme de Shepard qui montre une assez bonne reconstruction des données d'entrée. L'examen de cette figure révèle une forte corrélation entre les descripteurs *Opaque, Clair/foncé*, et *Granuleux*. Ensuite, on constate un degré d'imprécision élevé pour le descripteur *Nacré* qui peut s'interpréter comme un manque de pouvoir discriminant de ce descripteur. En effet (cf table 3.2), beaucoup de produits ont des scores flous qui se superposent. Enfin, il faut noter que ce descripteur s'oppose clairement aux 4 autres descripteurs car il est possible de tracer sur la sphère un diamètre complet. Ceci est conforme aux corrélations d'entrée puisque la possibilité

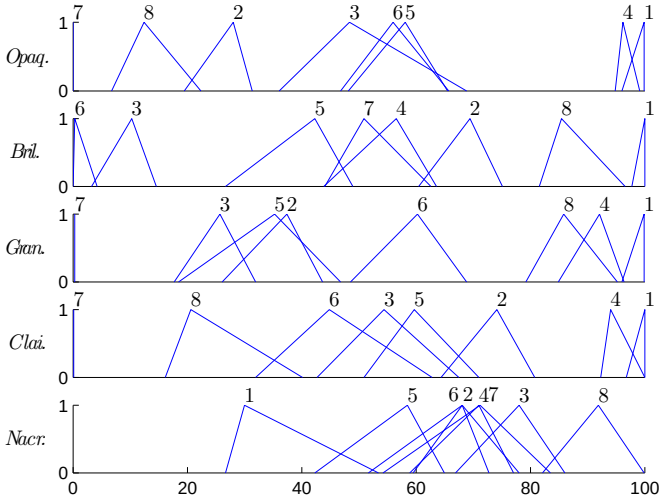

FIG. 3.10 – Jeu de données sensorielles.

d'une corrélation fortement négative est non nulle. Comme c'était attendu, la représentation obtenue avec le modèle possibiliste est plus imprécise (cf figure 3.13) mais conduit aux mêmes interprétations.

5 Conclusion

Les méthodes MDS sont des méthodes de visualisation de données très employées en analyse de données. Curieusement, la prise en compte de données imprécises a été jusqu'à ce jour assez peu abordée dans la littérature. Le principe de base qui sous-tend notre approche, quels que soient le modèle et le type d'ajustement, est que des dissimilarités imprécises doivent être représentées par des régions imprécises dans l'espace choisi. Le modèle Euclidien est bien adapté à la représentation de dissimilarités qui correspondent à des distances (perceptives par exemple), alors que le modèle sphérique est pertinent pour des mesures de corrélations. Les deux types d'ajustement proposés nous apparaissent complémentaires car ils apportent des éclairages différents sur les données : alors que l'ajustement par les moindres carrés fournit une représentation "compromis" en général assez claire, le modèle possibiliste fournit une représentation, exacte, plus imprécise, qui renseigne sur l'adéquation du modèle de distance avec les données.

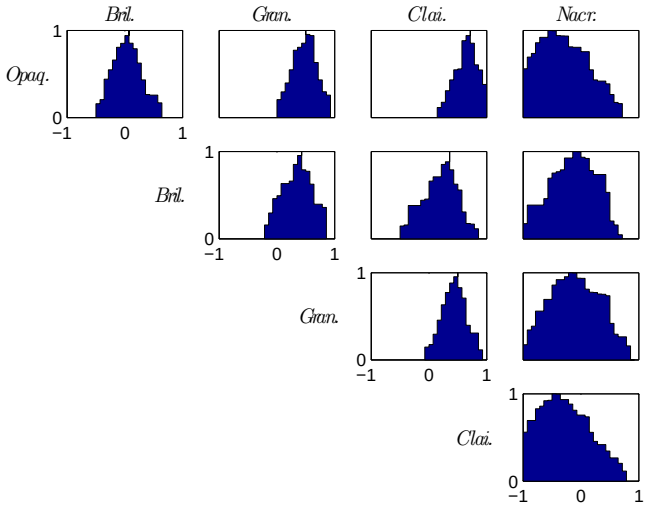

FIG. 3.11 – Jeu de données sensorielles ; Fonctions d'appartenance des coefficients de Kendall flous entre descripteurs.

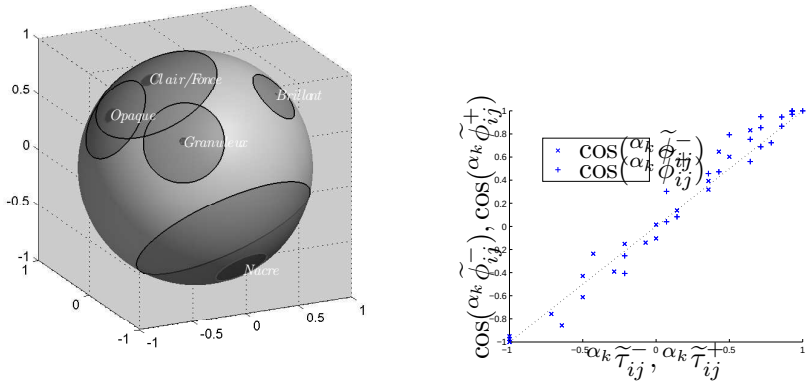

FIG. 3.12 – Jeu de données sensorielles ; MDS sphérique : ajustement par les moindres carrés. Représentation des descripteurs (α-coupes $\{0^+, 0.9\}$) et diagramme de Shepard.

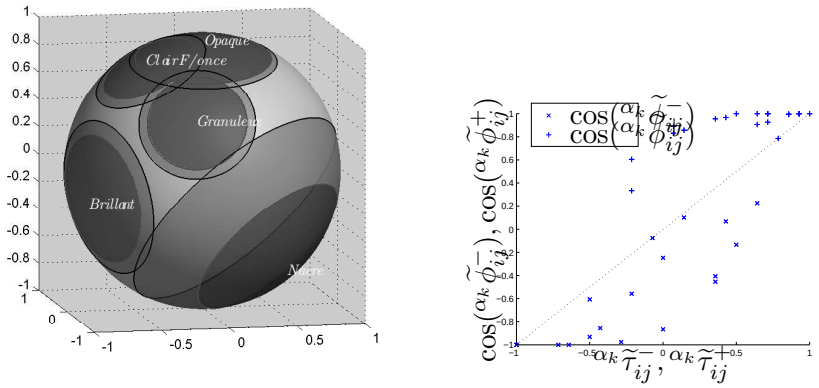

FIG. 3.13 – Jeu de données sensoriel ; MDS sphérique : ajustement possibiliste. Représentation des descripteurs (α-coupes $\{0^+, 0.9\}$) et diagramme de Shepard.

Publications

M. Masson et T. Denoeux. Positionnement multidimensionnel de données de dissimilarités floues. *Actes des Rencontres Francophones sur la Logique Flou et ses Applications (LFA'99)*, 39-45, Valenciennes, France, novembre 1999.

T. Denoeux et M. Masson. Multidimensional scaling of interval-valued dissimilarity data *Pattern Recognition Letters*, 21, 83-92, 2000.

M. Masson et T. Denoeux. Multidimensional scaling of fuzzy dissimilarity data. *Fuzzy sets and Systems*, 128(3), 339-352, 2002.

P.-A. Hébert, M. Masson et T. Denoeux. Fuzzy multidimensional scaling. *Computational Statistics and Data Analysis*, Special Issue on Fuzzy Statistical Analysis, à paraître.

Chapitre

4

Analyse en composantes principales

1 Position du problème

L'Analyse en Composantes Principales (ACP), initiée par Pearson en 1901 [77] et développée par la suite par Hotelling en 1933 [63], est l'une des techniques actuellement les plus couramment utilisées en analyse de données exploratoire. Elle permet l'analyse de données multivariées présentées sous forme d'un tableau dit *individus-variables* $\mathbf{X} = (x_{ij})$ de taille $n \times p$ où chaque individu (une ligne du tableau) est décrit par p variables numériques. L'objectif de l'ACP est d'une part d'étudier les similarités entre les individus (quels sont les individus proches dans l'espace ? existe-t-il des groupes de points homogènes ?) et d'autre part, d'étudier les liens éventuels entre les variables (y-a-t-il des variables corrélées positivement ? Quelles sont les variables qui s'opposent ?). Cet objectif passe par la création d'un petit nombre de variables synthétiques qui permettent de représenter les individus dans un espace de faible dimension retenant les caractéristiques majeures de l'espace originel. Dans l'interprétation géométrique de l'ACP, le tableau initial est vu comme un nuage de points dans $\mathbb{R}^p$ et l'on cherche les directions de l'espace dans lesquelles la dispersion des données est maximale. Ces directions sont appelées les axes principaux d'inertie. Si q ($q < d$) dimensions sont trouvées, alors la projection des n individus $\mathbf{x}^1, \ldots, \mathbf{x}^n$ sur le sous-espace linéaire $\mathcal{L}$ engendré fournit une représentation compressée des données d'entrée sous forme de vecteurs $\mathbf{y}^1, \ldots, \mathbf{y}^n \in \mathbb{R}^q$. Le principe de l'ACP est illustré en figure 4.1. Dans la continuité des travaux sur le positionnement des données imprécises, nous avons cherché à développer une méthode d'ACP pour des données imprécises c'est-à-dire des individus décrits

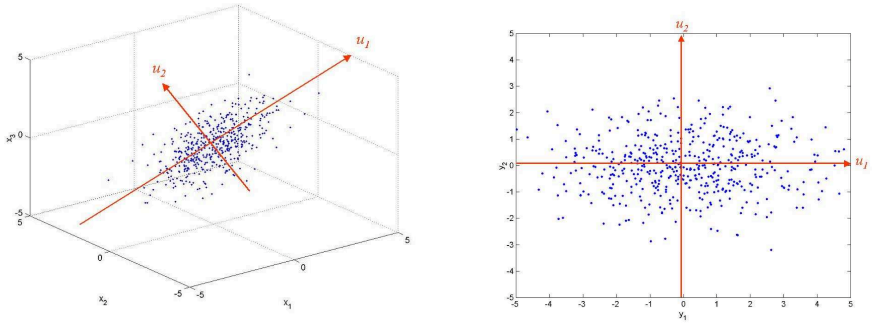

FIG. 4.1 – Principe de l'ACP : (à gauche) deux premiers axes d'inertie ; (à droite) projection sur ces deux axes : premier plan factoriel.

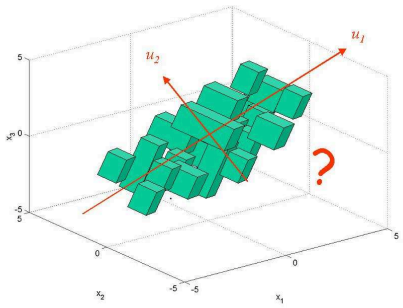

FIG. 4.2 – Principe de l'ACP sur données intervalles.

par un vecteur dont les composantes sont des intervalles ou plus généralement des nombres flous. La figure 4.2 donne la représentation schématique d'un ensemble de données de type intervalle. Dans ce cas, on peut considérer chaque individu peut être vu non plus comme un point dans l'espace mais comme un hypercube. Se pose alors le problème de déterminer les meilleures axes de représentation. L'algorithme proposé se fonde sur des résultats récents concernant la capacité des réseaux de neurones autoassociatifs à réaliser une compression des informations d'entrée. On utilise pour cela des réseaux dits *diabolos*, en référence à leur forme, avec un apprentissage *associatif* : on force le réseau à reproduire en sortie l'entrée qui lui a été présentée. La partie centrale, resserrée, du réseau constitue une version comprimée de l'entrée. Ce principe peut facilement être réutilisé avec des nombres flous en appliquant d'une part le principe d'extension de Zadeh et d'autre part des règles simples d'arithmétique d'intervalles.

2 ACP par réseaux de neurones autoassociatifs

2.1 Rappels sur l'ACP classique

Comme indiqué dans l'introduction, l'ACP cherche un sous-espace de projection qui minimise la perte d'information subie. Il est bien connu que le sous-espace optimal de dimension q est celui engendré par les q premiers vecteurs propres $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q$ de la matrice de variance-covariance des données, associés aux q premières valeurs propres $\lambda_1, ..., \lambda_q$. On appelle *composantes principales* les vecteurs de dimension n des projections des individus sur les axes principaux. Les coordonnées de la i-ème composante principale sont données par :

$$y_i^p = (\mathbf{x}^p)^t \mathbf{u}_i \quad p = 1, n. \tag{4.1}$$

Elle permettent de représenter les individus dans différents plans factoriels. Chaque composante principale i a une variance égale à λ_i. On mesure sa capacité de représentation par le pourcentage d'inertie expliqué calculé comme :

$$\frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} \tag{4.2}$$

On dispose d'autre part de formules de reconstruction qui permettent de retrouver le tableau initial au moyen des composantes principales et des axes principaux, de façon exacte ou de façon approchée suivant le nombre de composantes principales retenues. Si on se limite à q composantes, on a dans ce cas la formule approchée de reconstitution :

$$\mathbf{x}^p \approxeq \mathbf{z}^p = \sum_{i=1}^{q} y_i^p \mathbf{u}_i^t \tag{4.3}$$

On peut montrer, ce qui donne un autre point de vue sur l'ACP, que les vecteurs propres de la matrice de variance-covariance sont solutions du problème de minimisation quadratique suivant :

$$\min_{U_q} = \sum_{i=1}^{p} \|\mathbf{x}^p - \mathbf{z}^p\|^2, \tag{4.4}$$

où U_q désigne la matrice des vecteurs propres $[\mathbf{u}_1, ..., \mathbf{u}_q]$. La matrice Z formée par les vecteurs $\mathbf{z}^p$ constitue la meilleure approximation de rang q du tableau initial X au sens des moindres carrés (theorème d'Eckart-Young). Cette propriété intéressante est à la base de l'approche présentée dans le paragraphe suivant.

2.2 ACP par réseau auto-associatif

Considérons le réseau à trois couches représenté en figure 4.3 constitué de d unités d'entrée, q ($q < d$) unités cachées et d unités de sortie. Soit A la matrice de dimension $q \times d$ des poids de la couche d'entrée vers la couche cachée et B la matrice de dimension $d \times q$ des poids de la couche cachée vers la couche de sortie. Toutes les fonctions d'activation sont linéaires, de telle sorte que la sortie $\mathbf{z}$ calculée pour une entrée $\mathbf{x}$ s'écrit :

$$\mathbf{z} = BA\mathbf{x}. \tag{4.5}$$

Si l'on suppose maintenant que le réseau est entraîné sur un mode *associatif*, c'est-à-dire que l'on force le réseau durant l'apprentissage à reproduire en sortie la forme qui lui a été présentée, alors le réseau cherche à approximer la fonction identité. Puisque la partie centrale du réseau comporte moins de cellules qu'il n'y a en entrée, les sorties de cette couche constituent une version compressée de l'entrée. Cette idée a d'abord été suggérée par Rumelhart et al. [82]. Elle a ensuite été analysée formellement par Bourlard et Kamp [13] puis Baldi et Hornik [1, 2]. La partie qui suit donne un résumé des principaux résultats de ces derniers. Soit $E(A, B)$ l'erreur quadratique définie par :

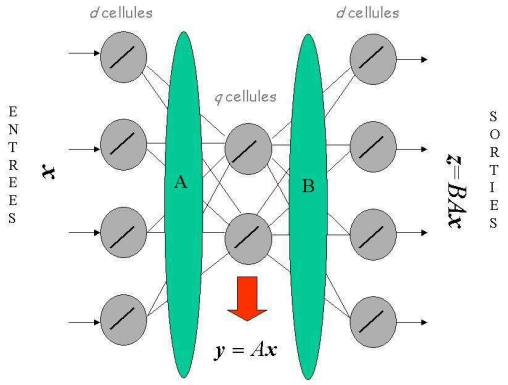

FIG. 4.3 – Schéma d'un réseau diabolo.

$$E(A, B) = \sum_{p=1}^{n} e(\mathbf{x}^p, \mathbf{z}^p), \tag{4.6}$$

où $e(\mathbf{x}^p, \mathbf{z}^p)$ désigne l'erreur de reconstruction pour la forme p :

$$e(\mathbf{x}^p, \mathbf{z}^p) = \|\mathbf{x}^p - \mathbf{z}^p\|^2 = \sum_{k=1}^{d} (x_k^p - z_k^p)^2, \quad p = 1, \ldots, n. \tag{4.7}$$

L'erreur totale peut aussi s'exprimer comme une fonction de la transformation globale $W = BA$ contrainte à être au plus de rang q. Il est clair que $E(A, B) = E(CA, BC^{-1})$ quelle que soit la matrice C de dimension $q \times q$. Comme dans la paragraphe précédent, on note $\mathbf{u}_1, ..., \mathbf{u}_q$ les vecteurs propres de la matrice de variance-covariance des données associés aux valeurs propres $\lambda_1 > ... > \lambda_q$. Baldi et Hornik démontrent la proposition suivante [2] :

PROPOSITION 2
L'erreur E exprimée en fonction de la transformation globale W a un unique minimum de la forme $W = BA$ avec

$$A = CU_q^t \tag{4.8}$$
$$B = U_q C^{-1}, \tag{4.9}$$

où U_q désigne, comme précédemment, la matrice $[\mathbf{u}_1, \ldots, \mathbf{u}_q]$, et C est une matrice quelconque inversible de dimension $q \times q$.

On retrouve donc l'ACP comme correspondant au cas particulier où C est égale à l'identité. Les sorties de la couche cachée sont dans ce cas identiques aux composantes principales des données. Cette solution n'est cependant pas systématiquement obtenue, et les sorties de la couche cachée sont les composantes principales définies à une transformation linéaire près. Bien qu'efficace en terme de compression, la solution générale n'est pas complètement satisfaisante en terme de visualisation des données d'entrée car les axes sont arbitrairement dilatés. Une façon de contourner la difficulté est d'introduire la contrainte $A' = B$. Cette condition impose que $CC' = I$, c'est-à-dire que C soit une matrice orthogonale. Dans ce cas, les composantes principales sont les sorties de la couche cachée à une transformation *isométrique* près (symétrie et rotation), ce qui est plus satisfaisant en termes de pouvoir de représentation des caractéristiques extraites. L'équation (4.5) se réécrit dans ce cas sous la forme suivante :

$$\mathbf{z} = BB^t\mathbf{x} \tag{4.10}$$

ou encore, sous forme scalaire,

$$z_k = \sum_{j=1}^{q} B_{kj} \sum_{i=1}^{d} B_{ij} x_i \quad k = 1, \ldots, d. \tag{4.11}$$

En pratique, les poids du réseau (la matrice B) sont obtenus en minimisant l'erreur de reconstruction moyenne par une technique classique de descente de gradient.

3 Extension à des données floues

3.1 Principe général

Soit $\mathcal{F}(\mathbb{R})$ l'ensemble des nombres flous définis sur $\mathbb{R}$. On suppose maintenant que les données consistent en un ensemble de vecteurs $\widetilde{\mathbf{x}}^1, \ldots, \widetilde{\mathbf{x}}^n$, où chaque $\widetilde{\mathbf{x}}^p \in \mathcal{F}(\mathbb{R})^d$ est un vecteur de d nombres flous notés $(\widetilde{x}_i^p)_{1 \leq i \leq d}$. Le but est de compresser ces données sous la forme de vecteurs $\widetilde{\mathbf{y}}^1, \ldots, \widetilde{\mathbf{y}}^n$ de nombres flous de dimension inférieure, avec $\widetilde{\mathbf{y}}^p \in \mathcal{F}(\mathbb{R})^q$, $p = 1, \ldots, n$, et $q < d$. Dans le cas de données classiques, ce problème peut être résolu par une simple ACP. L'implémentation de celle-ci par réseau autoassociatif, décrite au paragraphe précédent, donne une solution pour la généralisation au cas flou. On considère un réseau dont la structure, représentée en figure 4.4, est la même que précédemment et l'on suppose qu'un vecteur $\widetilde{\mathbf{x}}$ de d nombres flous est placé en entrée du réseau.

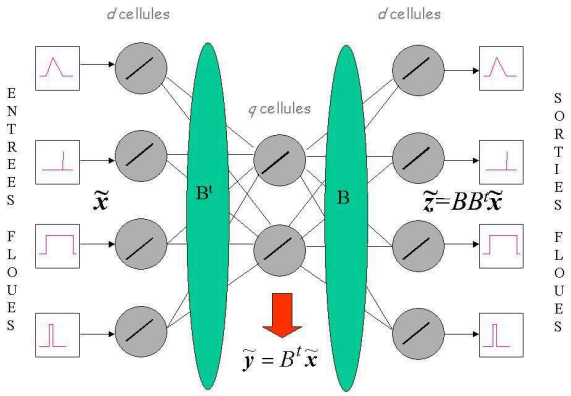

FIG. 4.4 – Schéma du réseau diabolo flou.

La sortie du réseau peut être calculée en appliquant le principe d'extension de Zadeh à l'équation (4.10). La $k^{\text{ième}}$ composante $\widetilde{z}_k$ du vecteur flou de sortie $\widetilde{\mathbf{z}}$ pour l'entrée $\widetilde{\mathbf{x}}$ est donc définie comme :

$$\mu_{\widetilde{z}_k}(u) = \sup \min_{1 \leq i \leq d} \mu_{\widetilde{x}_i}(v_i), \tag{4.12}$$

le supremum étant pris sous la contrainte :

$$u = \sum_{j=1}^{q} B_{kj} \sum_{i=1}^{d} B_{ij} v_i.$$

Une forme plus compacte est donnée par :

$$\widetilde{z}_k = \sum_{j=1}^{q} B_{kj} \sum_{i=1}^{d} B_{ij} \widetilde{x}_i \quad k = 1, \ldots, d, \tag{4.13}$$

où l'addition et la multiplication par un réel sont remplacées par les opérations usuelles d'arithmétique floue [38], ou encore, sous une forme analogue à celle de l'équation (4.10) :

$$\widetilde{\mathbf{z}} = BB'\widetilde{\mathbf{x}}. \tag{4.14}$$

Si le réseau est correctement entraîné, le vecteur $\widetilde{\mathbf{y}} = B'\widetilde{\mathbf{x}}$ constitue, comme dans le cas net, une version compressée de la donnée floue d'entrée $\widetilde{\mathbf{x}}$.

3.2 Application à des nombres flous trapézoïdaux

Pour l'implémentation pratique de la méthode, nous avons choisi la manipulation de nombres flous trapézoïdaux. Cette classe de nombres flous étant close vis-à-vis des opérations d'addition, de soustraction et de multiplication par un réel (cf chapitre 1), si les entrées sont des nombres flous trapézoïdaux, alors les sorties de la couche cachée ainsi que les sorties du réseau sont donc aussi des nombres flous trapézoïdaux. On suppose donc que les composantes $\tilde{x}_i$ de chaque vecteur d'entrée $\tilde{\mathbf{x}} \in \mathcal{F}(\mathbb{R})$ sont des nombres flous trapézoïdaux $\widetilde{x}_i = (x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, x_i^{(4)})$ avec $x_i^{(1)}$ et $x_i^{(4)}$ désignant les bornes minimale et maximale du support du nombre flou et $x_i^{(2)}$ et $x_i^{(3)}$ les bornes minimales et maximales du noyau. On calcule dans un premier temps les sorties de la couche cachée. Par définition, la sortie $\widetilde{y}_j$ de la $j^{\text{ième}}$ cellule cachée s'écrit :

$$\widetilde{y}_j = \sum_{i=1}^{d} B_{ij} \widetilde{x}_i = (y_j^{(1)}, y_j^{(2)}, y_j^{(3)}, y_j^{(4)}) \quad j = 1, \ldots, q. \tag{4.15}$$

En utlisant des formules classiques d'arithmétique floue [39], il vient :

$$y_j^{(1)} = \sum_{\substack{i=1 \\ B_{ij}>0}}^{d} B_{ij} x_i^{(1)} + \sum_{\substack{i=1 \\ B_{ij}<0}}^{d} B_{ij} x_i^{(4)}, \tag{4.16}$$

$$y_j^{(2)} = \sum_{\substack{i=1 \\ B_{ij}>0}}^{d} B_{ij} x_i^{(2)} + \sum_{\substack{i=1 \\ B_{ij}<0}}^{d} B_{ij} x_i^{(3)}, \tag{4.17}$$

$$y_j^{(3)} = \sum_{\substack{i=1 \\ B_{ij}>0}}^{d} B_{ij} x_i^{(3)} + \sum_{\substack{i=1 \\ B_{ij}<0}}^{d} B_{ij} x_i^{(2)}, \tag{4.18}$$

$$y_j^{(4)} = \sum_{\substack{i=1 \\ B_{ij}>0}}^{d} B_{ij} x_i^{(4)} + \sum_{\substack{i=1 \\ B_{ij}<0}}^{d} B_{ij} x_i^{(1)}. \tag{4.19}$$

De même, la $k^{\text{ième}}$ sortie est par définition :

$$\widetilde{z}_k = \sum_{j=1}^{p} B_{kj} \widetilde{y}_j = (z_k^{(1)}, z_k^{(2)}, z_k^{(3)}, z_k^{(4)}) \quad k = 1, \ldots, d, \tag{4.20}$$

et donc :

$$z_k^{(1)} = \sum_{\substack{k=1 \\ B_{kj}>0}}^{q} B_{kj} y_j^{(1)} + \sum_{\substack{k=1 \\ B_{kj}<0}}^{q} B_{kj} y_j^{(4)}, \tag{4.21}$$

$$z_k^{(2)} = \sum_{\substack{k=1 \\ B_{kj}>0}}^{q} B_{kj} y_j^{(2)} + \sum_{\substack{k=1 \\ B_{kj}<0}}^{q} B_{kj} y_j^{(3)}, \tag{4.22}$$

$$z_k^{(3)} = \sum_{\substack{k=1 \\ B_{kj}>0}}^{q} B_{kj} y_j^{(3)} + \sum_{\substack{k=1 \\ B_{kj}<0}}^{q} B_{kj} y_j^{(2)}, \tag{4.23}$$

$$z_k^{(4)} = \sum_{\substack{k=1 \\ B_{kj}>0}}^{q} B_{kj} y_j^{(4)} + \sum_{\substack{k=1 \\ B_{kj}<0}}^{q} B_{kj} y_j^{(1)}. \tag{4.24}$$

Un expression analytique de la fonction d'erreur à optimiser peut être obtenue en choisissant une métrique adaptée. En considérant chaque nombre flou trapézoïdal comme un point dans un espace à quatre dimensions comme proposé dans [59] et [30], on peut généraliser la fonction d'erreur (4.6) par le critère suivant :

$$E(B) = \sum_{p=1}^{n} \sum_{k=1}^{d} e(\widetilde{x}_k^p, \widetilde{z}_k^p)$$

avec,

$$e(\widetilde{x}_k, \widetilde{z}_k) = \sum_{t=1}^{4} (z_k^{(t)} - x_k^{(t)})^2, \quad k = 1, \ldots, d. \tag{4.25}$$

Les expressions explicites des dérivées partielles de $E(B)$ par rapport aux poids B_{ij} du réseau sont données dans [27].

3.3 Corrélation entre les composantes principales et les variables initiales

L'un des intérêts primordiaux de l'ACP réside dans les nombreux outils d'interprétation fournis à l'analyste. Parmi, ceux-ci, la corrélation entre les composantes principales construites et les variables initiales donne des indications précieuses pour comprendre les représentations obtenues. Nous avons choisi d'utiliser le coefficient de corrélation flou proposé par Lui et Kao [71] pour aider à l'interprétation des axes. Son principe est expliqué dans le chapitre 2.

4 Autres approches

Contrairement aux methodes MDS qui n'ont pas, à notre connaissance, été abordées sous cet angle, on trouve plusieurs approches concurrentes dans la littérature pour réaliser une ACP de données imprécises. On trouve les premiers travaux portant sur l'Analyse en Composantes Principales de données imprécises dans ceux développpés par l'équipe de Diday autour de l'analyse de données symboliques [31]. Considérant les données intervalles comme expliqué en introduction de ce chapitre comme des hyperrectangles dans l'espace, Cazes et al. [16] proposent deux méthodes pour réaliser l'ACP de ces données. La première méthode, dite *méthode des sommets* consiste simplement à réaliser une ACP sur les sommets des hyperrectangles. On passe alors d'une matrice initiale de données de dimension $n \times p$ à une matrice de dimension $n2^p \times p$ dans laquelle chaque ligne représente un des 2^p sommets associés à un individu. La projection des sommets des hyperrectangles dans les différents plans factoriels ne définissant pas un rectangle, la représentation d'un individu s'obtient en recherchant le rectangle de taille minimum englobant toutes les projections des sommets de son hypercube. Des formules analytiques très simples permettent son calcul. Cette méthode s'avère cependant très coûteuse en temps de calcul spécialement lorsque le nombre de variables est grand, le nombre de lignes de la matrice dépendant, comme on l'a vu précédemment, exponentiellement de p.

Conçue pour contourner ce problème, la *méthode des centres* consiste à chercher les axes principaux d'inertie sur la matrice constituée par les centres des hyperrectangles. On projette ensuite les sommets des hyperrectangles dans ce sous-espace et l'on détermine comme dans la méthode des sommets les rectangles englobants. Cette dernière approche présente à nos yeux deux inconvénients. Les axes principaux d'inertie sont déterminés indépendamment de l'imprécision résidant dans les données or parfois, c'est justement l'information que l'on cherche à représenter. D'autre part, dans les différentes expériences que nous avons réalisées, nous avons constaté que la méthode avait tendance à surestimer l'imprécision des données en produisant des représentations beau-

coup plus confuses que celles obtenues avec notre approche.

5 Deux exemples d'application

Exemple 4.1 (*Jeu de données des étudiants)* Pour illustrer la capacité de la méthode à fournir une représentation condensée adéquate des données, nous reprenons un exemple adapté de [38, page 237]. Les données traitées sont reportées dans le tableau 4.1. Elles regroupent les notes obtenues par six étudiants en mathématiques et en physique, lors de deux semestres consécutifs (M1, M2, P1, et P2). Certaines notes sont précisément connues, d'autres sont seulement connues sous forme d'intervalle, ou sous forme linguistique, ou encore absentes.

	M1	M2	P1	P2
Tom	15	fairly good	unknown	[14,16]
David	9	good	fairly good	10
Bob	6	[10,11]	[13,20]	good
Jane	fairly good	very good	19	[10,12]
Joe	very bad	fairly bad	[10,14]	[14]
Jack	1	[4,6]	9	[6,9]

Tab. 4.1 – Jeu de données des étudiants : tableau de notes.

On choisit de représenter chaque note sous forme d'un nombre flou trapezoïdal. Les appréciations linguistiques sont caractérisées par les fonctions d'appartenance représentées figure 4.5.
Un réseau comprenant deux cellules cachées ($q = 2$) a permis d'obtenir une représentation plane des données. L'examen de la matrice de poids (tableau 4.2) permet une interprétation des axes : le premier axe est lié aux performances en mathématiques alors que le second est plutot lié à la physique. Ces remarques sont confirmées par l'étude des coefficients de corrélation donnés en figure 4.6. Le premier axe est clairement corrélé avec M1 et M2. Pour le deuxième axe, les corrélations sont moins claires, en raison de la note très imprécise de Tom en P1.
Sur le plan de projection représenté figure 4.7, on retrouve de nombreuses caractéristiques importantes contenues dans les données originales. Par exemple, Jack, avec des notes dans l'ensemble assez mauvaises et précises, est bien localisé dans la partie basse de la figure. La grande dispersion de Tom le long du second axe s'explique par le fait qu'une de ses notes en physique est inconnue. Bob est très précisément situé sur l'axe lié aux mathématiques et de façon moins précise sur le second.
Une comparaison des données d'entrée et des données reconstruites est fournie

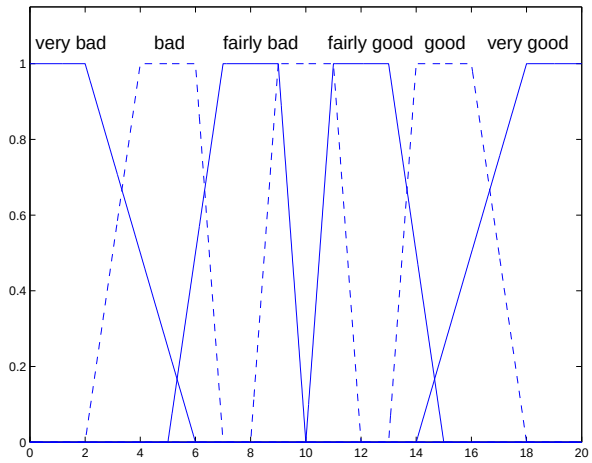

FIG. 4.5 – Fonctions d'appartenance des notes floues.

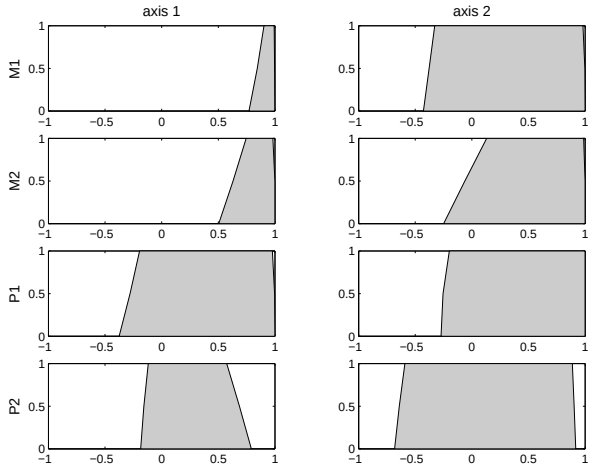

FIG. 4.6 – Coefficients de corrélation flous.

matière	axe 1	axe 2
M1	0.80	-0.07
M2	0.57	0.17
P1	0.00	0.95
P2	0.08	0.13

TAB. 4.2 – Valeurs des poids B_{ij} pour le jeu de données des étudiants.

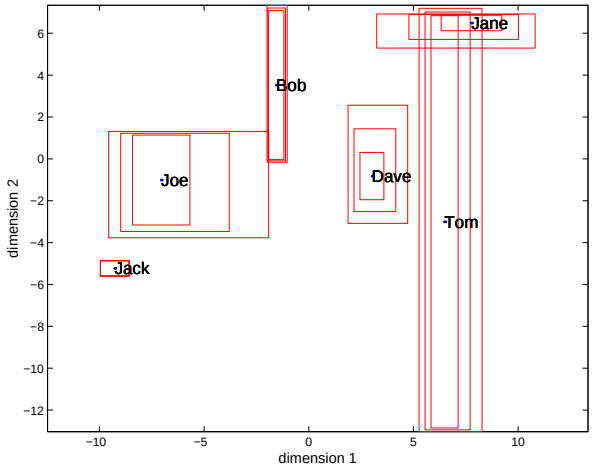

FIG. 4.7 – Représentation des étudiants dans le plan (supports, noyaux, et coupes de niveau 0.5).

en figure 4.8. Elle révèle quels sont les aspects qui sont bien préservés dans la représentation compressée et ceux qui ne le sont pas : par exemple, la note de Bob dans la matière P1 est bien reconstruite alors que celle de Jack dans la matière P2 ne l'est pas. Les erreurs totales de reconstruction pour les six étudiants sont reportés table 4.3. Elles renseignent sur la confiance que l'on peut avoir dans la représentation des différents étudiants.

Tom	David	Bob	Jane	Joe	Jack
20.97	13.00	10.93	10.66	7.65	14.22

TAB. 4.3 – Erreurs de reconstruction pour le jeu de données des étudiants

EXEMPLE 4.2 (*Jeu de données sensorielles)* La méthode proposée a été appliquée à des données d'évaluation sensorielle dans le cadre d'un projet mené en collaboration avec PSA Peugeot-Citröen. Les objets étudiés sont des sons

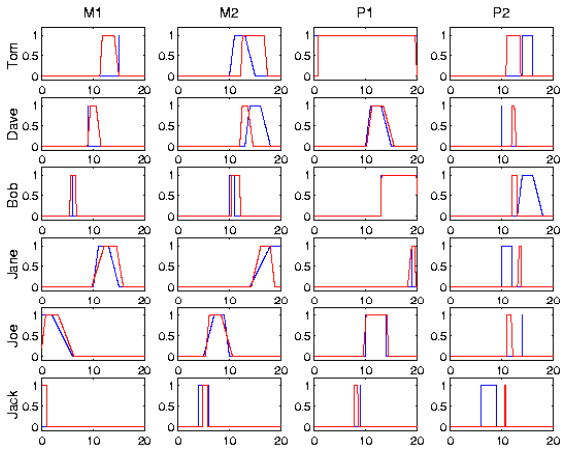

FIG. 4.8 – Jeu de données des étudiants. Données d'entrée (traits bleus) et données reconstruites (traits rouges).

enregistrés à l'intérieur de véhicules. Les données consistent en des scores attribués par un panel de 12 juges décrivant la perception de 21 sons suivant 5 descripteurs. Chaque son a été présenté 3 fois à chaque sujet de telle sorte que l'on dispose d'une matrice à quatre entrées : sons $\times$ descripteurs $\times$ sujets $\times$ répétitions. Le but de l'étude était d'étudier la variabilité des réponses entre les différents panelistes et la répétabilité de chaque sujet au cours des répétitions. Pour cela, chacun des 21 $\times$ 12 couples (son, sujet) a été considéré comme un objet dans l'analyse. Pour chaque descripteur, les trois valeurs issues des répétitions ont servi à la construction de nombres flous triangulaires (qui est un cas particulier de nombre flou trapézoïdal) définis par une valeur minimale, maximale et centrale. L'ensemble de données initial consiste donc en 21 $\times$ 12 vecteurs composés de 5 nombres flous triangulaires. Une représentation en deux dimensions a été obtenue grâce à un réseau comprenant $q = 2$ cellules cachées. Les poids, reportés dans le tableau 4.4, et les coeffients de corrélation en figure 5.1 montrent que les axes 1 et 2 sont fortement liés, respectivement, aux descripteurs 2 et 4.

Une première partie des résultats est présentée sur la figure 4.10. Pour plus de clarté, les réponses des 12 sujets pour 4 sons différents sont présentées sur 4 figures différentes. Les quatre sons ont été choisis car ils permettent de mettre en évidence quatre comportements différents du panel. Le premier son est perçu de manière similaire par les juges et avec une variabilité faible au cours des répétitions. Les jugements pour le son 5 sont en assez bon accord mais avec une variabilité plus importante. Le son 16 présente les mêmes caractéristiques, mis à part un juge, qui est clairement en désaccord avec l'en-

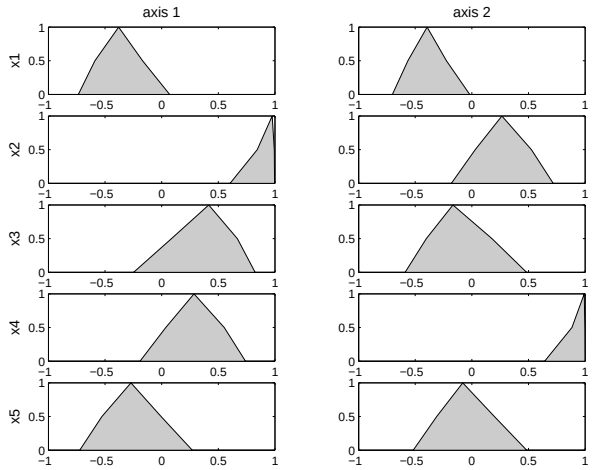

FIG. 4.9 – Jeu de données sensorielles. Corrélations floues entre les composantes principales et les descripteurs initiaux.

semble du groupe. Enfin, le son 21 semble être difficile à noter avec une variabilité inter et intra-juges importante. Une autre façon de présenter les résultats consiste à montrer l'intégralité des réponses fournies par un même juge. Sur la figure 4.11 sont présentés deux juges aux comportements extrêmes : le juge 4 utilise une échelle de notation de faible amplitude avec une bonne répétabilité. Le juge 12 au contraire utilise une plage de notation beaucoup large mais apparaît très variable dans ses évaluations. En étudiant attentivement toutes les représentations, il est donc possible de répondre à un certain nombre de questions qui se posent en analyse sensorielle : est-ce que les produits sont correctement distingués les uns des autres, quels sont les juges répétables, l'échelle de notation est-elle utilisée de façon pertinente, etc.

	axe 1	axe 2
x_1	-0.17	-0.16
x_2	0.92	0.00
x_3	0.32	-0.17
x_4	0.01	0.97
x_5	-0.13	-0.01

TAB. 4.4 – Jeu de données sensorielles. Poids B_{ij}.

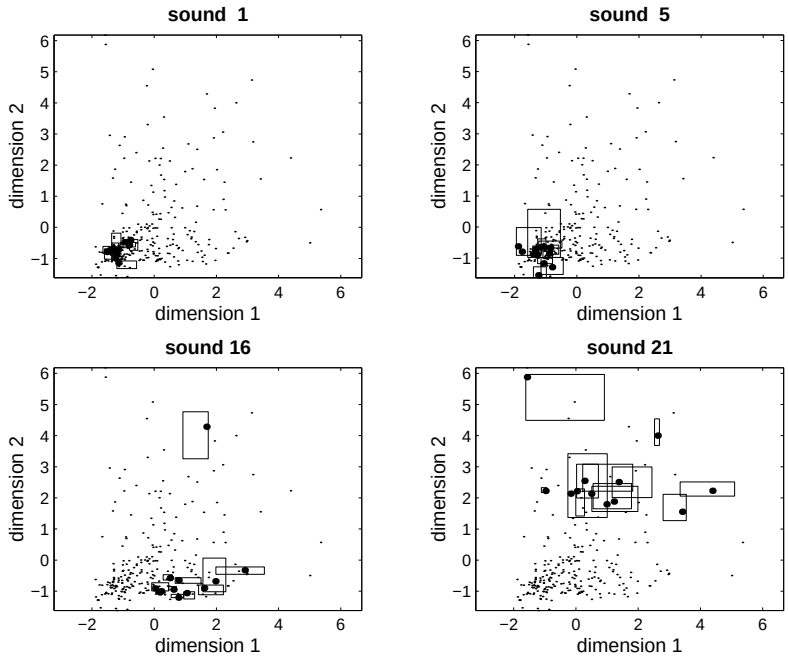

FIG. 4.10 – Jeu de données sensorielles. Chaque figure montre comment un son est perçu par les douze juges. Les cercles et les traits pleins représentent respectivement le noyau et l'alpha-coupe de niveau 0.5 des projections. Les points représentent le centre de gravité des projections de toutes les autres paires (son,sujet) qui ne concernent pas le son considéré.

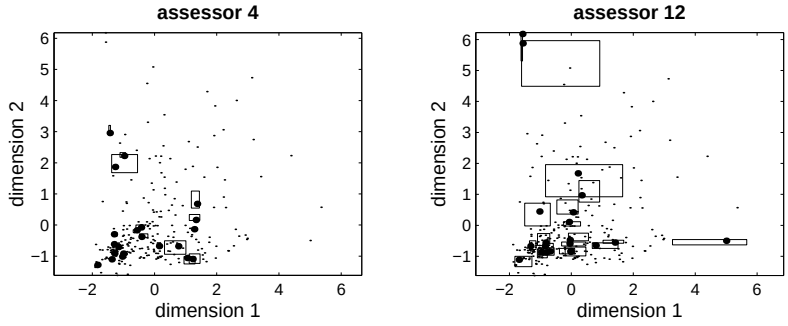

FIG. 4.11 – Jeu de données sensorielles. Chaque figure montre comment un juge particulier perçoit les 21 sons. Les cercles et les traits pleins représentent respectivement le noyau et l'alpha-coupe de niveau 0.5 des projections. Les points représentent le centre de gravité des projections de toutes les autres paires (son,sujet) qui ne concernent pas le juge considéré.

6 Conclusion

Les données floues apparaissent naturellement dans un certain nombre de situations pour lesquelles l'incertitude ou l'imprécision des observations ne peuvent être ignorée. Par exemple, dans l'application à l'évaluation sensorielle décrite dans cet article, l'écart des évaluations entre répétitions est aussi important que la valeur centrale. De nouvelles méthodes d'analyse exploratoire de ce type de données doivent être développées pour prendre en compte cette complexité supplémentaire. La technique présentée ici est une extension de l'ACP classique. Elle exploite des développements récents concernant la capacité d'un réseau autoassociatif à réaliser une compression de l'information exactement comme l'ACP, mais sans diagonalisation explicite de la matrice de variance-covariance. Les expériences menées tant sur données simulées que sur données réelles ont démontré la capacité de la méthode à fournir des représentation concises pertinentes de données multidimensionnelles complexes.

Publications

M. Masson et T. Denoeux. Analyse en composantes principales de données floues par réseau de neurones autoassociatif. *Actes des Rencontres Francophones sur la Logique Flou et ses Applications (LFA'01)*, 89-95, Mons, Belgique, novembre 2001.

T. Denoeux et M. Masson. Principal component analysis of fuzzy data using autoassociative neural networks. *IEEE Transactions on Fuzzy Systems*, 12(3), 336-349, 2004.

Chapitre

5

Classification automatique

1 Introduction à la classification automatique

1.1 Objectifs, méthodes, données

On reconnaît aux êtres vivants, et particulièrement aux êtres humains, la faculté de regrouper un grand nombre d'objets en catégories d'objets *similaires*. Cette faculté peut être par exemple considérée comme étant à la base du développement du langage, un nom commun représentant simplement une famille d'objets, de choses, d'animaux, etc, ayant en commun certaines caractéristiques.
Une des branches de l'analyse de données s'est naturellement intéressée à l'automatisation de ce processus de classification. On rassemble sous le terme de *classification automatique*[1] un grand nombre d'algorithmes qui visent à découvrir dans un ensemble de données des groupes ou classes d'observations homogènes. L'objectif est d'obtenir un résumé des données ou de vérifier une structure existante, sur la base d'informations expertes par exemple. Suivant la méthode employée, le résultat de l'application d'un algorithme de classification automatique sera soit :

- *une partition*, c'est-à-dire un ensemble de classes distinctes, tel qu'un objet appartient à une et une seule classe de la partition ;
- *une hiérarchie*, c'est-à-dire une suite de partitions imbriquées (*dendrogramme*) ;
- *une partition empiétante*, dans ce cas, on autorise un même objet à appar-

[1] En anglais *clustering*. Notons que l'on trouve parfois dans la littérature française le terme de *coalescence* qui désigne l'obtention d'une partition.

tenir à plusieurs classes différentes ;
– *une partition floue*, dans ce cas chaque objet, appartient à toutes les classes avec un certain degré d'appartenance.

La nature des données disponibles est un second mode suivant lequel varient les méthodes de classification automatique. Le plus souvent, on dispose d'un tableau individus-variables dans lequel chaque objet est décrit par un ensemble d'attributs ou de caractéristiques. Deux objets sont alors jugés semblables s'ils sont proches (au sens d'une métrique à définir, comme exemple la distance Euclidienne) dans l'espace des caractéristiques. Parfois, on ne dispose pas des valeurs d'attributs, mais directement d'une mesure par paire de similarité ou de dissimilarité entre objets : on parle alors de *données relationnelles*. Bien que moins communes, ces données peuvent être considérées comme plus générales que les données individus-variables puisqu'il est toujours possible de convertir des données attributs en données relationnelles. D'autre part, de nombreux jeux de données réels, dans des domaines comme par exemple la psychologie, la biologie, l'économie, ou la chimie, sont par nature relationnels. Enfin, notons qu'il peut s'agir d'une stratégie pertinente lorsque les données sont hétérogènes (qualitatives, quantitatives, structurées, etc) ou encore si l'on veut incorporer une connaissance *a priori* dans le calcul de la dissimilarité.

1.2 Partition de données relationnelles

On trouvera dans [6, chapter 3] une description détaillée des modèles classiques et flous de classification automatique de données relationnelles. Nous en donnons ici une brève introduction. Ces méthodes peuvent être classées en trois catégories : les méthodes hiérarchiques, les méthodes basées sur la décomposition de relations floues, et celles fondées sur l'optimisation d'une fonction coût. Dans cette dernière catégorie, étant donné un ensemble de n objets à classer en c classes, on cherche une matrice de partition floue $U = (u_{ik})$ de taille $n \times c$ telle que

$$\sum_{k=1}^{c} u_{ik} = 1 \quad \forall\, i \in \{1, \ldots, n\} \tag{5.1}$$

et

$$\sum_{i=1}^{n} u_{ik} > 0 \quad \forall\, k \in \{1, \ldots, c\}\ . \tag{5.2}$$

Chaque nombre $u_{ik} \in [0, 1]$ peut s'interpréter comme le *degré d'appartenance* de l'objet i à la classe k. La matrice de partition floue est déterminée en optimisant un crière qui mesure la compacité et la séparation des classes. Parmi les méthodes les plus connues, citons le modèle FNM (*Fuzzy Non Metric*) de Roubens [81], le modèle AP (*Assignment-Prototype*) de Windham [100], et le

modèle des c-moyennes relationnel (RFCM) de Hathaway [57] (dont une version similaire peut être trouvée dans [64]). Ce dernier modèle a été par la suite étendu sous le nom de NERF par Hathaway and Bezdek [56] pour prendre en compte des données de dissimilarité non Euclidiennes. Enfin, signalons que des versions "robustes" de FNM et RFCM ont été proposées par Davé pour résister aux perturbations induites par des points aberrants [22].

2 Quelques outils de la théorie des fonctions de croyance

Dans cette partie nous donnons quelques éléments supplémentaires par rapport au chapitre introductif de ce mémoire sur la théorie des fonctions de croyance. Ces éléments sont nécessaires à la mise en place de la méthode de classification proposée. Considérons une masse de croyance m^{Ω} définie sur le produit Cartésien $\Omega = \Omega_1 \times \Omega_2$ (afin d'éviter toute ambiguïté, il est d'usage de noter le domaine comme un exposant). La bba marginale $m^{\Omega\downarrow\Omega_1}$ sur Ω_1 est définie, pour tout $A \subseteq \Omega_1$, comme

$$m^{\Omega\downarrow\Omega_1}(A) \triangleq \sum_{\{B\subseteq\Omega \mid \mathrm{Proj}(B\downarrow\Omega_1)=A\}} m^{\Omega}(B) \ , \tag{5.3}$$

où $\mathrm{Proj}(B \downarrow \Omega_1)$ dénote la projection de B sur Ω_1, définie comme

$$\mathrm{Proj}(B \downarrow \Omega_1) \triangleq \{\omega_1 \in \Omega_1 \mid \exists \omega_2 \in \Omega_2, (\omega_1, \omega_2) \in B\} \ . \tag{5.4}$$

Cette opération est l'analogue de la marginalisation d'une loi de probabilité. La théorie des fonctions de croyance dispose d'une autre opération qui cette fois n'existe pas en probabilité, appelée *l'extension vide*. Soit une bba m^{Ω_1} sur Ω_1, son *extension vide* [87, 90] sur $\Omega = \Omega_1 \times \Omega_2$ est définie pour tout $B \subseteq \Omega$ comme :

$$m^{\Omega_1\uparrow\Omega}(B) \triangleq \begin{cases} m^{\Omega_1}(A) & \text{if } B = A \times \Omega_2 \text{ pour un } A \subseteq \Omega_1 \\ 0 & \text{sinon} \end{cases} \tag{5.5}$$

Cette définition de l'extension vide résulte du Principe d'Engagement Minimal [90], qui formalise l'idée selon laquelle on ne doit jamais donner plus de crédit que justifié à une proposition. Etant données deux bba m^{Ω_1} et m^{Ω_2}, leur somme conjonctive sur $\Omega = \Omega_1 \times \Omega_2$ peut être obtenue en combinant leur extension vide sur Ω, en utilisant (1.15). On obtient alors :

$$(m^{\Omega_1} m^{\Omega_2})(A \times B) = m^{\Omega_1}(A) m^{\Omega_2}(B) \ , \tag{5.6}$$

pour tous sous-ensembles non vides $A \subseteq \Omega_1$ et $B \subseteq \Omega_2$.

3 Classification automatique dans le cadre de la théorie des fonctions de croyance

3.1 Partition crédale

On considère que l'on dispose d'un ensemble $O = \{o_1, \ldots, o_n\}$ de n objets, et l'on cherche une partition $\Omega = \{\omega_1, \ldots, \omega_c\}$ en c classes de O. Pour cela, on dispose comme données d'entrée d'une matrice carrée $\Delta = (\delta_{ij})$, où $\delta_{ij} \geq 0$ mesure le degré de dissimilarité entre les objets o_i et o_j. La matrice Δ sera supposée symétrique avec une diagonale nulle. La valeur de dissimilarité peut provenir directement de l'évaluation d'un expert ou d'un calcul de distance entre attributs, non nécessairement homogènes [64]. Notons que des valeurs de similarité σ_{ij} peuvent être convertie en dissimilarités grâce à la transformation $\delta_{ij} = g(\sigma_{ij})$, où g est une fonction décroissante.

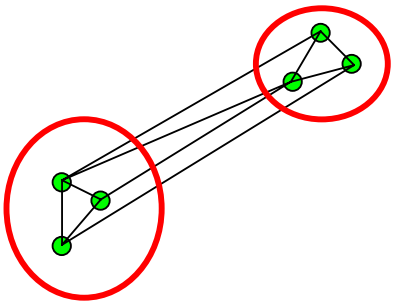

FIG. 5.1 – Classification de données relationnelles : connaissant toutes les dissimilarités inter-points, est-il possible de déterminer une partition des données en un nombre fixé de classes ?

Se plaçant d'emblée dans le cadre des fonctions de croyances, et plus particulièrement dans celui du modèle des croyances transférables, on choisit de représenter l'incertitude sur l'appartenance d'un objet o_i aux différentes classes par une fonction de croyance m_i^Ω définie sur Ω. Cette représentation permet de coder toutes les situations allant de la certitude totale à l'ignorance complète comme l'illustre l'exemple suivant.

EXEMPLE 5.1 Cet exemple fictif constitué de 5 objets, dont les masses associées sont présentées dans le tableau 3.1, permet de passer en revue les différentes formes que peut prendre l'incertitude que l'on a sur la partition :

- l'appartenance de l'objet 2 est connue avec certitude puisque toute la masse est portée sur une classe unique ;
- au contraire, l'incertitude concernant l'objet 5 est totale puisque la masse est portée sur Ω.

A	$m_1^\Omega(A)$	$m_2^\Omega(A)$	$m_3^\Omega(A)$	$m_4^\Omega(A)$	$m_5^\Omega(A)$
$\emptyset$	0	0	0	0	0
$\{\omega_1\}$	0	0	0	0.2	0
$\{\omega_2\}$	0	1	0	0.4	0
$\{\omega_1, \omega_2\}$	0.7	0	0	0	0
$\{\omega_3\}$	0.3	0	0.2	0.4	0
$\{\omega_1, \omega_3\}$	0	0	0.5	0	0
$\{\omega_2, \omega_3\}$	0	0	0	0	0
Ω	0	0	0.3	0	1

TAB. 5.1 – Exemple de partition crédale avec 5 objets.

- avec les autres objets, on observe des situations d'incertitude intermédiaires avec deux cas particuliers : la masse associée à l'objet 4 est *Bayesienne* car les éléments focaux sont des singletons, la masse associée à l'objet 3 est possibiliste car les éléments focaux sont emboîtés.

L'ensemble des vecteurs de masse $M^\Omega = (m_1^\Omega, m_2^\Omega, ..., m_n^\Omega)$ constitue ce que l'on appelle une *partition crédale*. Cette notion généralise les notions de partition usuelles. En effet,
– si toutes les masses sont certaines, on retrouve une partition stricte de Ω ;
– si toutes les masses sont Bayesiennes, on retrouve une partition floue de Ω.

Une *partition crédale de taille c* sera définie comme une partition crédale $M^\Omega = (m_1^\Omega, \ldots, m_n^\Omega)$ sur un cadre de discernement Ω constitué de c éléments si

$$\mathrm{pl}_i^\Omega(\{\omega\}) > 0 \tag{5.7}$$

pour un $i \in \{1, \ldots, n\}$, pl_i^Ω étant la fonction de plausibilité associée à m_i^Ω. Notons que cette condition qui énonce le fait que chaque classe a un degré strictement positif de plausibilité pour au moins un des objets, est l'équivalent de (5.2) dans la définition d'une partition floue de taille c.

EXEMPLE 5.2 Les plausibilités de chaque singleton pour l'exemple précédent sont données en table 5.2. Puisque chaque classe est plausible pour au moins un des objets, M^Ω constitue une partition crédale de taille 3.

3.2 Partition crédale et dissimilarités

Le cadre de travail étant fixé, il reste à déterminer comment inférer une partition crédale à partir des données de dissimilarité d'entrée. Comme évoqué dans l'introduction, il parait naturel de considérer que *plus deux objets sont*

k	$\mathrm{pl}_1^\Omega(\{\omega_k\})$	$\mathrm{pl}_2^\Omega(\{\omega_k\})$	$\mathrm{pl}_3^\Omega(\{\omega_k\})$	$\mathrm{pl}_4^\Omega(\{\omega_k\})$	$\mathrm{pl}_5^\Omega(\{\omega_k\})$
1	0.7	0	0.8	0.2	1
2	0.7	1	0.3	0.4	1
3	0.3	0	1	0.4	1

TAB. 5.2 – Plausibilités des singletons.

similaires, plus il est plausible qu'ils appartiennent à la même classe. Ce principe simple se modélise parfaitement dans le cadre du modèle des croyances transférables.
Considérons deux objets o_i et o_j dont l'appartenance aux classes est caractérisée par les fonctions de masse m_i^Ω et m_j^Ω. Pour calculer la plausibilité qu'ils appartiennent à la même classe, on va construire à partir de m_i^Ω et m_j^Ω, une nouvelle fonction de croyance, $m_{i\times j}^{\Omega^2}$, définie sur le produit cartésien Ω^2, qui quantifie notre croyance sur l'appartenance simultanée des objets. Pour cela, on procède à l'extension vide des fonctions de masses individuelles puis on les combine dans l'espace commun, de sorte que

$$m_{i\times j}^{\Omega^2}(A\times B) = m_i^\Omega(A)\cdot m_j^\Omega(B), \qquad \forall A,B\subseteq\Omega, A\neq\emptyset, B\neq\emptyset\ . \tag{5.8}$$

Dans Ω^2, l'événement "les objects o_i et o_j appartiennent à la même classe" correspond au sous-ensemble de Ω^2 suivant :

$$S = \{(\omega_1,\omega_1),(\omega_2,\omega_2),\ldots,(\omega_c,\omega_c)\} \tag{5.9}$$

Soit $\mathrm{pl}_{i\times j}$ la plausibilité associée à la fonction de masse $m_{i\times j}^{\Omega^2}$. On a

$$\begin{aligned}
\mathrm{pl}_{i\times j}(S) &= \sum_{\{A\times B\subseteq\Omega^2 \mid (A\times B)\cap S\neq\emptyset\}} m_{i\times j}^{\Omega^2}(A\times B) \\
&= \sum_{A\cap B\neq\emptyset} m_i^\Omega(A)\cdot m_j^\Omega(B) \\
&= 1-\sum_{A\cap B=\emptyset} m_i^\Omega(A)\cdot m_j^\Omega(B) \\
&= 1-L_{ij}\ ,
\end{aligned} \tag{5.10}$$

où L_{ij} est le degré de conflit entre m_i^Ω and m_j^Ω.
La plausibilité que deux objets o_i et o_j appartiennent à la même classe est donc simplement égale au degré de conflit entre leurs fonctions de masse associées. Etant donné deux couples d'objets (o_i,o_j) et $(o_{i'},o_{j'})$, il est donc naturel d'imposer la condition suivante :

$$d_{ij} > d_{i'j'} \Rightarrow \mathrm{pl}_{i\times j}(S) \leq \mathrm{pl}_{i'\times j'}(S) \tag{5.11}$$

ou, de façon équivalente :

$$d_{ij} > d_{i'j'} \Rightarrow L_{ij} \geq L_{i'j'} \ , \tag{5.12}$$

i.e., plus les objets sont dissemblables, moins il est plausible qu'ils appartiennent à la même classe, et plus le conflit est fort entre les bba's. Une partition M^Ω satisfaisant cette condition sera dite *compatible* avec Δ.

3.3 Inférer une partition crédale

Inférer une partition crédale à partir des données, c'est déterminer une matrice M^Ω compatible (ou le plus compatible) avec la matrice de dissimilarités Δ. Ce problème s'avère en fait similaire à celui évoqué dans le chapitre consacré aux méthodes MDS (cf chapitre 3). Les masses recherchées ici jouent le rôle de coordonnées des objets dans l'espace dans la MDS classique, le conflit entre les masses joue le rôle de distance interpoint et l'on cherche à approximer au mieux les dissimilarités d'entrée par ces distances données par le modèle.
L'approche MDS qui correspond le plus à notre objectif (équation (5.12)) est l'approche *ordinale* ou *non métrique* (cf paragraphe 2.2 du chapitre 3)). On cherche alors à déterminer les masses m_i^Ω, $i = 1, n$ en minimisant une fonction de stress définie par :

$$I_{nm}(M^\Omega, f) \triangleq \frac{\sum_{i<j}[L_{ij} - f(\delta_{ij})]^2}{\sum_{i<j}[L_{ij} - \overline{L}]^2} \ , \tag{5.13}$$

où $\overline{L}$ désigne le degré moyen de conflit, et f est une fonction croissante. Classiquement, $I_{nm}(M^\Omega, f)$ est minimisé de façon alternée, d'abord par rapport à M^Ω en utilisant une technique itérative de descente de gradient, puis, par rapport à f en utilisant une régression monotone [11]. Cette approche, bien que puissante, est coûteuse en temps de calcul et on peut lui préférer l'approche *métrique*. Il s'agit dans ce cas d'imposer une forme paramétrée à la relation entre les degrés de conflit et les dissimilarités. Une fonction de stress doit être définie. Il en existe plusieurs formes. Le critère le plus simple peut se formuler de la manière suivante :

$$I(M^\Omega, a, b) \triangleq \frac{1}{C} \sum_{i<j} w_{ij}(aL_{ij} + b - \delta_{ij})^2 \ , \tag{5.14}$$

où a et b sont deux coefficients et C est une constante de normalisation égale à $\sum_{i<j} d_{ij}$. Les poids w_{ij} sont fixés à 1 ou 0 selon que la dissimilarité δ_{ij} est disponible ou non. Il s'agit d'un moyen classique et efficace pour traiter des jeux de données avec des données manquantes ou pour s'attaquer à des ensembles d'objets de taille importante [11]. En effet, dans ce dernier cas, calculer ou faire

évaluer par un sujet humain la totalité des paires de dissimilarité est souvent irréalisable. Or, de nombreuses études MDS ont montré qu'il était possible de laisser de côté une partie des dissimilarités lors de la détermination du modèle sans pour autant sacrifier la qualité de la solution [96, 97].
Nous avons choisi ici une version de stress plus sophistiquée qui permet d'accorder plus d'importance aux faibles dissimilarités. Il s'agit du stress normalisé de Sammon [83] défini par

$$I(M^{\Omega}, a, b) \triangleq \frac{1}{C} \sum_{i<j} \frac{w_{ij}(aL_{ij} + b - \delta_{ij})^2}{\delta_{ij}} \ , \tag{5.15}$$

Le critère I peut être minimisé par rapport à M^{Ω}, a et b en utilisant une procédure de descente de gradient. Notons que le critère de Sammon donne plus de poids aux faibles dissimilarités, ce qui nous parait une bonne stratégie dans un objectif de classification automatique.

Remarque 5 Chaque bba m_i^{Ω} doit prendre ses valeurs dans $[0, 1]$ et satisfaire $\sum_{A\subseteq\Omega} m(A) = 1$. Il s'agit donc d'un problème d'optimisation sous contraintes. On peut cependant relâcher les contraintes en adoptant le changement de variables suivant :

$$m_i^{\Omega}(A_k) = \frac{\exp(\alpha_{ik})}{\sum_{l=1}^{f} \exp(\alpha_{il})}, \tag{5.16}$$

où $A_k,\ k = 1, \ldots, f$ sont les f éléments focaux ($f = 2^c$ dans le cas général), et les α_{ik} pour $i = 1, \ldots, n$ and $k = 1, \ldots, f$ sont les nf paramètres réels représentant la partition crédale.

Exemple 5.3 *(Les données papillon)* Nous illustrons la méthode sur un jeu de données synthétique inspiré de l'exemple classique de Windham [100]. On considère une matrice de dimension 13×13 dont une image est donnée en figure 5.2. Les dissimilarités entre les objets 2 à 13 ont été calculées comme la distance au carré entre douze points d'un espace de dimension 2 qui sont représentés en figure 5.3. Les onze premiers points correspondent à l'exemple de Windham, le douzième, un "outlier", a été ajouté pour tester la robustesse de la méthode par rapport aux points aberrants. De plus, un 13ème objet (l'objet 1) a été incorporé à l'exemple. Cet objet, qui n'a pas de représentation dans le plan, est similaire à tous les autres objets (cf figure 5.2). Cet objet est censé représenter des situations où les données sont bruitées, incohérentes ou encore provenant d'évaluations subjectives.
Nous comparons le résultat de notre algorithme, intitulé evclus, avec ceux obtenus par les méthodes traditionnelles de classification de données relationnelles citées en introduction : ap [100], fnm [81], rfcm [57], et sa version dérivée nrfcm, [22], et nerf [56]. nrfcm, par l'utilisation d'une classe de

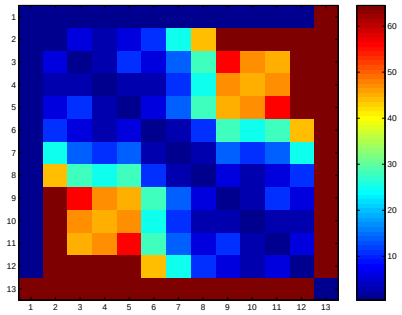

FIG. 5.2 – Données papillon : image du tableau de dissimilarités.

bruit, est bien adapté à la détection d'outliers, alors que NERF a été conçu pour traiter des données non Euclidiennes. L'objectif est d'obtenir une partition raisonnable des objets 2 à 12 et de détecter la particularité des objets 1 ("inlier") et 13 (outlier).

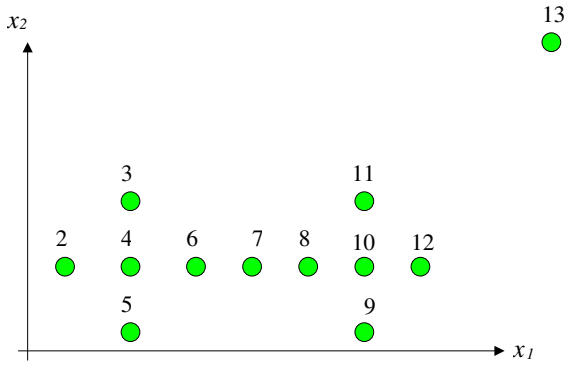

FIG. 5.3 – Données papillon : représentation des points 2 à 13.

Parmi les cinq algorithmes évoqués plus haut, trois (FNM, RFCM et sa version bruitée NRFCM) ont une constante de "fuzzification" h qui contrôle le degré de "dureté" de la partition résultante. Une valeur $h = 2$ est communement adoptée dans la littérature[2]. D'autre part, NRFCM possède un autre paramètre α qui sert à définir la distance de la classe de bruit par rapport aux autres objets.
La figure 5.4 montre les fonctions d'appartenance obtenues avec les 5 algorithmes (avec $h = 2$ et $\alpha = 50$) et les bbas obtenues avec EVCLUS ($m_i(\{\omega_1\})$, $m_i(\{\omega_2\})$, $m_i(\Omega)$ and $m_i(\emptyset)$ sont représentées en fonction de i). L'object 13

[2]le code Matlab de NERF disponible à l'adresse `http ://www2.gasou.edu/facstaff/hathaway` utilise exclusivement cette valeur.

(l'outlier) est associé à un fort degré d'appartenance pour la classe 2 par toutes les méthodes sauf par NRFCM et EVCLUS, qui détecte l'atypicalité de cet objet de façon satisfaisante (en le classant dans la classe de bruit pour NRFCM et en assignant une masse importante à l'ensemble vide pour EVCLUS). La principale différence entre NRFCM et EVCLUS est constatée pour l'objet 1, pour lequel NRFCM donne arbitrairement une valeur élevée d'appartenance à la classe 1, alors que EVCLUS choisit pour ce point d'allouer toute la masse à Ω.

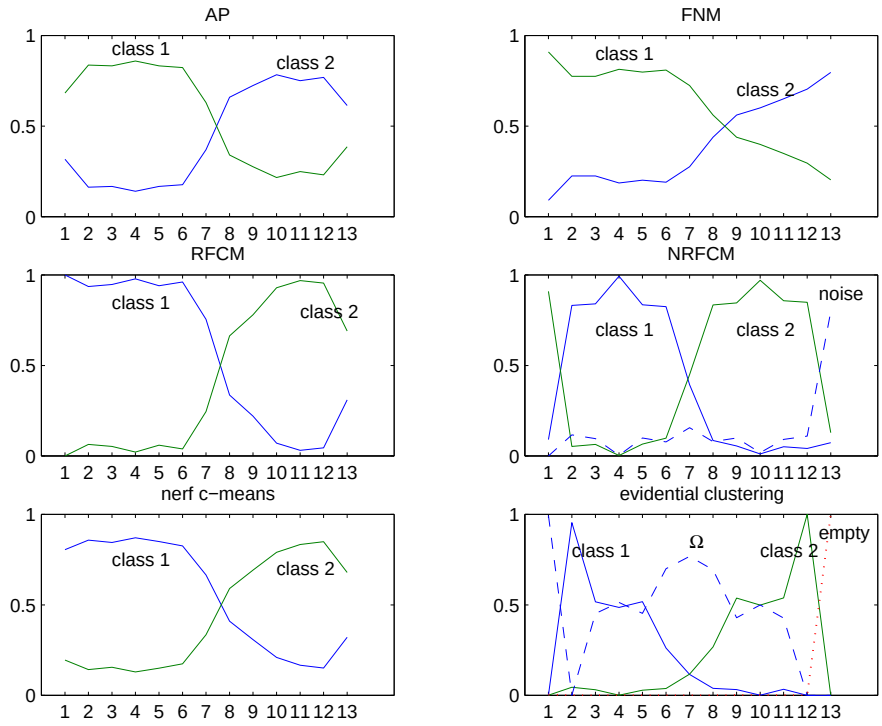

FIG. 5.4 – Données papillon : partitions obtenues.

Autre manière de présenter les résultats obtenus avec EVCLUS, sur la figure 5.5 sont représentées les plausibilités des classes 1 et 2 pour les 13 objets. On constate que les deux classes sont toutes deux complètement plausibles pour l'objet 1 ($\text{pl}_1^\Omega(\{\omega_1\}) = \text{pl}_1^\Omega(\{\omega_2\}) \approx 1$) alors qu'aucune des classes n'est plausible pour l'objet 13 ($\text{pl}_{13}^\Omega(\{\omega_1\}) = \text{pl}_{13}^\Omega(\{\omega_2\}) \approx 0$). Notons que le report de la quasi totalité de la masse sur l'ensemble vide pour l'objet 13 est conforme à l'interprétation de cet ensemble sous l'hypothèse de monde ouvert [89], aucune des hypothèses en présence n'étant satisfaisante pour cet objet.

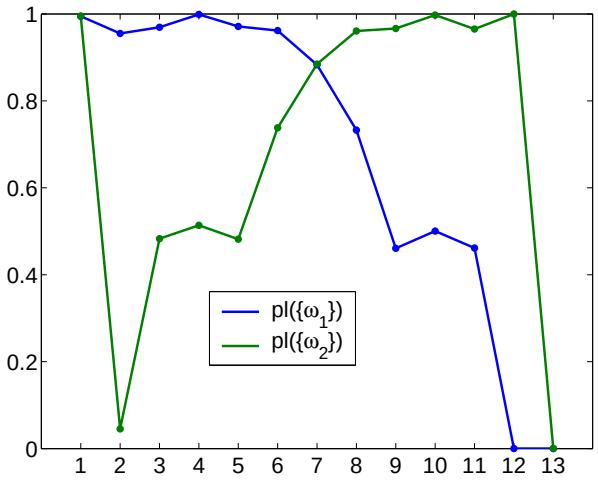

FIG. 5.5 – Données papillon : plausibilités des singletons.

3.4 Limiter la complexité

Comme dans tout problème d'apprentissage, un écueil réside dans le nombre de paramètres du modèle à optimiser. Si l'on considère tous les éléments focaux sur un ensemble Ω constitué de c classes, il y a $n \times 2^c$ paramètres à déterminer. Ce nombre est donc linéaire par rapport au nombre d'objets mais exponentiel par rapport au nombre de classes. Si c est grand, le nombre de paramètres libres du modèle doit être contrôlé. Deux voies peuvent alors être envisagées :

- premièrement, on peut simplement limiter le nombre d'élements focaux, par exemple en contraignant les éléments focaux à être des singletons, l'ensemble vide et Ω. Le nombre de paramètres est alors réduit à $n(c+2)$ tout en gardant une certaine richesse de description qui permet de détecter des points éloignés ou des valeurs incohérentes.
- on peut ajouter à la fonction de stress un terme de pénalisation contrôlant le contenu informationnel des masses obtenues. Cette approche ne réduit pas le nombre de paramètres à optimiser mais limite le nombre de paramètres *effectifs*. Il s'agit donc d'un moyen efficace pour contrôler la complexité du modèle de classification.

Dans notre cas, on cherche à extraire des masses aussi "informatives" que possible des données.

La définition de la quantité d'information contenue dans une fonction de croyance est encore un sujet de débat à l'heure actuelle. Cependant, plusieurs mesures d'entropie ont été proposées, parmi elles l'incertitude totale, introduite par Pal et al. [76]. Elle est définie pour une masse de croyance

normalisée m par :

$$H(m) = \sum_{A \in \mathcal{F}(m)} m(A) \log_2 \left(\frac{|A|}{m(A)} \right) , \tag{5.17}$$

où $\mathcal{F}(m)$ désigne l'ensemble des éléments focaux de m. $H(m)$ est minimale lorsque la masse est affectée à peu d'éléments focaux de faible cardinalité (il est prouvé dans [76] que $H(m) = 0$ ssi $m(\{\omega\}) = 1$ pour un $\omega \in \Omega$). Lorsque la masse attribuée à l'ensemble vide est non nulle, une procédure de normalisation doit être appliquée auparavant. Deux procédures sont communément employées :

– celle de Dempster [87] dans laquelle la masse du $\emptyset$ est annulée et toutes les autres masses sont divisées par $1 - m(\emptyset)$;
– celle de Yager [101] dans laquelle la masse du $\emptyset$ est transferée à Ω.

Nous avons retenu le critère d'entropie (5.17) avec la normalisation de Yager, considérant que la masse attribuée au vide doit être pénalisée de la même façon que les autres masses ce qui n'est pas le cas avec la normalisation de Dempster. L'expression de l'incertitude totale devient alors :

$$\begin{aligned} H(m) &= \sum_{A \in \mathcal{F}(m) \setminus \{\emptyset\}} m(A) \log_2 \left(\frac{|A|}{m(A)} \right) \\ &+ m(\emptyset) \log_2 \left(\frac{|\Omega|}{m(\emptyset)} \right) . \end{aligned} \tag{5.18}$$

La fonction coût optimisée s'écrit finalement :

$$J(M^\Omega, a, b) \triangleq I(M^\Omega, a, b) + \lambda \sum_{i=1}^{n} H(m_i^\Omega), \tag{5.19}$$

où λ règle le compromis entre l'adéquation du modèle aux données et sa complexité.

3.5 D'une partition crédale à une partition nette ou floue

Bien la partition crédale soit porteuse de beaucoup plus d'information, il est toujours possible de la transformer en une partition nette ou floue. Cette conversion se fait au moyen du concept de *probabilité pignistique* [95] défini (cf chapitre 1), pour une masse m^Ω normalisée, par :

$$BetP(A) \triangleq \sum_{\emptyset \neq B \subseteq \Omega} m^\Omega(B) \frac{|A \cap B|}{|B|} \tag{5.20}$$

Pour obtenir une partition floue, il suffit de calculer la probabilité pignistique de chaque classe ω_k. Dans le cas où seuls les singletons, l'ensemble vide et Ω ont

été choisis comme éléments focaux, l'expression des probabilités pignistiques est la suivante :

$$BetP(\{\omega_k\}) = m^{\Omega}(\{\omega_k\}) + \frac{m^{\Omega}(\Omega) + m^{\Omega}(\emptyset)}{c}, \tag{5.21}$$

pour tout $k = 1, c$ (en supposant que la normalisation de Yager est utilisée). Une partition nette peut facilement s'en déduire. La partition crédale apparaît donc comme un modèle général de partitionnement incluant comme cas particuliers les partitions nettes et floues.

3.6 Deux jeux de données réelles

EXEMPLE 5.4 *(Les données de cortex du chat)* Ce jeu de données réelles consiste en une matrice décrivant les forces de connections entre 65 régions corticales du chat. Collecté par Scannell [85], il a été utilisé par de nombreux auteurs pour illustrer des méthodes de visualisation, de discrimination ou de classification automatique de données de proximité [52, 53, 61]. Les valeurs de proximité vont de 0 (auto-connection), à 4 (absence de connection) avec des valeurs intermédiaires : 1 (connection dense), 2 (intermediate connection moyenne) et 3 (connection faible). D'autre part, on sait que le cortex peut être divisé en quatre zones fonctionnelles : le cortex auditif (A), le cortex visuel (V), le cortex somatosensoriel (S) et le cortex frontolimbique (F). Chacune des 65 zones dispose d'une étiquette fonctionnelle. L'idée était d'étudier s'il était possible de retrouver ces zones fonctionnelles à l'aide des données de proximités en recherchant une partition en 4 classes.

L'algorithme EVCLUS a été lancé 50 fois avec différentes initialisations aléatoires des paramètres (avec $\lambda = 0.01$, et 6 élements : $\{\omega_i\}, i = 1, \ldots, 4$, Ω et $\emptyset$), et la solution de coût minimum (équation (5.19))a été retenue. Une carte bidimensionnelle des données a été obtenue avec un algorithme MDS classique. La figure 5.6 montre les partitions nette et floue obtenues par la transformation pignistique. Sur la figure, un symbole différent est utilisé pour chacune des classes et la taille du symbole est proportionnelle au maximum de probabilité pignistique. On voit que les classes fonctionnelles sont presque parfaitement retrouvées avec un taux d'erreur de 4.6 % (seules trois zones sont mal classées). Ce chiffre est conforme à l'erreur de type "leave-one-out" rapportée par Graepel [52, 53] dans un contexte supervisé.
Nous avons également appliqué les 5 algorithmes classiques en utilisant le même dispositif expérimental : pour différentes valeurs de h comprises entre 1 et 2, chaque algorithme a été lancé 50 fois, puis la solution de moindre coût a été retenue. Toutes les méthodes convergent vers une solution dégénérée avec des valeurs d'appartenance de $1/c$ pour tous les points lorsque h est fixé à 2.

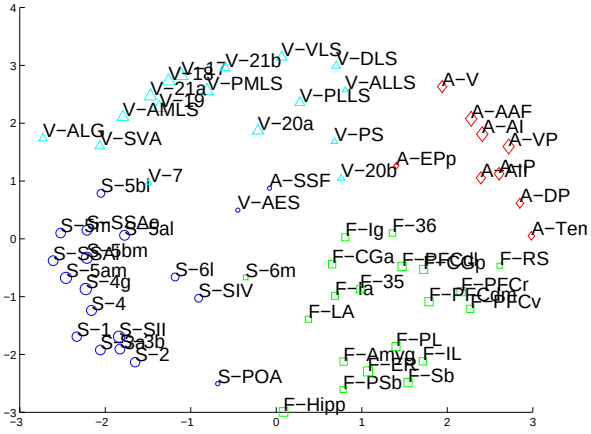

FIG. 5.6 – Données du cortex du chat : partition obtenue.

FNM, RFCM, FCM sont capables de donner une solution raisonnable pour des valeurs faibles de h ($h < 1.4$). La meilleure solution est obtenue avec RFCM avec un taux de 4.6% pour une valeur de $h = 1.2$. On constate donc que les performance de EVCLUS sont équivalentes à la meilleure méthode sur ce type de données.

EXEMPLE 5.5 *(Les données protéines)* Ce jeu de données consiste en une matrice de proximités dérivée de la comparaison structurelle de 213 séquence de protéines (se reporter à [62] et [52] pour des exemples antérieurs d'expériences avec ces données). Les protéines sont classées en quatre catégories appelées hemoglobine-α (HA), hemoglobine-β (HB), myoglobine (M) et autres globines (G). De même que précédemment, on a cherché à vérifier si cette partition pouvait être retrouvée à partir des données de proximité. EVCLUS fournit la partition en 4 classes représentée figure 5.7. Cette solution a été obtenue en utlisant le même dispositif expérimental que dans l'exemple précédent : 50 exécutions, $\lambda = 0.005$, et 6 éléments focaux ($\{\omega_i\}, i = 1, \ldots, 4$, Ω and $\emptyset$). La partition nette trouvée est très proche de celle connue, puisqu'une seule protéine sur les 213 est mal classée.

Nous avons appliqué sur ce même jeu de données les autres algorithmes. Avec une constante de fuzzification $h = 2$, tous convergent vers une solution triviale d'égalité des appartenances à toutes les classes pour chauqe protéine. Les meilleurs résultats sont obtenus avec RFCM (ou de façon équivalente NRFCM) avec une constante $h = 1.05$: 5 protéines sur 213 sont mal classées.
Outre un taux d'erreur inférieur, il est intéressant de noter que EVCLUS fournit des informations supplémentaires sur les données grâce à l'analyse des masses allouées à l'ensemble vide. De façon surprenante à première vue, les masses

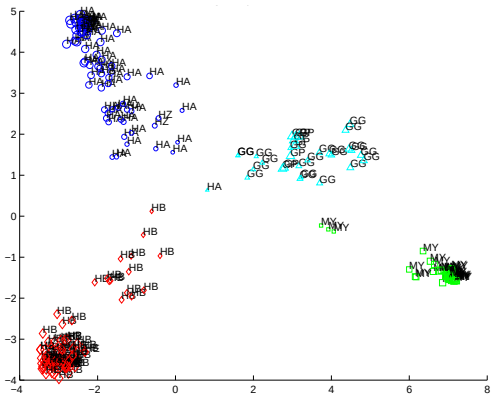

FIG. 5.7 – Données protéines : partition obtenue.

les plus élevées sont attribuées à la classe G (cf figure 5.8) alors que ces points ne peuvent pas être considérés comme des "outliers". La figure 5.9 donne des pistes d'explications à ce phénomène grâce à une représentation des distances inter et intra-classes pour les 4 catégories de protéines. On voit que la classes G se caractérise par des distances intra-classe très elevées et par des distances inter-classes du même ordre. La masse de l'ensembe vide permet donc de détecter particularité de cette classe. NERFCM parvient à isoler cette classe pour certaines valeurs de α : suivant sa valeur, les éléments de la classe G sont soient regroupés dans une classe, soit rejetés dans une classe de bruit. EVCLUS, en opérant à deux niveaux (le niveau crédal avec l'affectation des masses et le niveau décisionnel avec la probabilité pignistique), évite cet écueil : la particularité de la classe G est représentée au niveau crédal alors qu'une partition satisfaisante est déterminée au niveau décisionnel.

4 Extension à des données relationnelles de type intervalle

4.1 Objectifs et principe

Jusqu'à présent nous avons supposé que les dissimilarités étaient exprimées sous forme nette. On suppose maintenant que chaque dissimilarité est seulement connue pour appartenir à un intervalle $[\delta_{ij}^-; \delta_{ij}^+]$. Les principes exposés dans le paragraphe 3 peuvent s'étendre de manière assez naturelle, comme nous l'expliquons dans ce qui suit.

Les bornes minimales et maximales de dissimilarité constituent deux sources d'information sur la proximité entre les objets qui ont des significations complètement différentes :

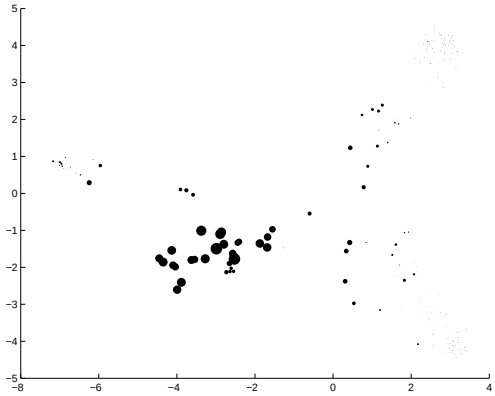

FIG. 5.8 – Données protéines : masse allouée à l'ensemble vide (la taille des disques est proportionnelle à $m(\emptyset)$).

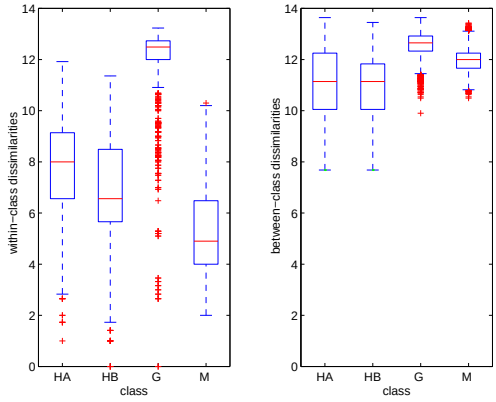

FIG. 5.9 – Données protéines : dissimilarités inter et intra-classes.

- si δ_{ij}^{-} est voisin de 0, alors il est possible que les objets i et j soit proches l'un de l'autre, et donc l'appartenance à la même classe est plausible ;
- dans le même temps, si δ_{ij}^{+} est grand, il est aussi plausible que les objets n'appartiennent pas à la même classe, ou, de façon équivalente, en raison de la dualité bel/pl, il est peu crédible que les objets appartiennent à la même classe.

Ces principes intuitifs peuvent se résumer en deux points :

- plus δ_{ij}^{-} est faible, plus la proposition S (équation (5.9)) est plausible ;
- plus δ_{ij}^{+} est forte, moins la proposition S est crédible ;

Etant donné deux couples d'objets (o_i, o_j) et $(o_{i'}, o_{j'})$, il est donc naturel d'im-

poser les conditions :

$$\begin{cases} \delta^-_{ij} > \delta^-_{i'j'} \Rightarrow \mathrm{pl}_{i\times j}(S) \leq \mathrm{pl}_{i'\times j'}(S) \\ \delta^+_{ij} > \delta^+_{i'j'} \Rightarrow \mathrm{bel}_{i\times j}(S) \leq \mathrm{bel}_{i'\times j'}(S) \end{cases} \tag{5.22}$$

où $\mathrm{pl}_{i\times j}(S)$ est définie par l'équation (5.10) et $\mathrm{bel}_{i\times j}(S)$ est la crédibilité de l'événement S que l'on peut calculer de la façon suivante :

$$\begin{aligned} \mathrm{bel}_{i\times j}(S) &= \sum_{\{A\times B\subseteq\Omega^2 \mid \emptyset\neq(A\times B)\subseteq S\}} m^{\Omega^2}_{i\times j}(A\times B) \\ &= \sum_{k=1}^{c} m^{\Omega}_i(\{\omega_k\}) m^{\Omega}_j(\{\omega_k\}) \\ &= 1 - U_{ij}, \end{aligned} \tag{5.23}$$

avec

$$U_{ij} = 1 - \sum_{k=1}^{c} m^{\Omega}_i(\{\omega_k\}) m^{\Omega}_j(\{\omega_k\}) \tag{5.24}$$

Les conditions (5.22) s'énoncent donc de manière équivalente par :

$$\begin{cases} \delta^-_{ij} > \delta^-_{i'j'} \Rightarrow L_{ij} \geq L_{i'j'} \\ \delta^+_{ij} > \delta^+_{i'j'} \Rightarrow U_{ij} \geq U_{i'j'}. \end{cases} \tag{5.25}$$

4.2 Inférer les masses à partir des dissimilarités

Il s'agit, à partir des principes évoqués ci-dessus de compatibilité entre les bornes minimales et maximales des dissimilarités et les quantités U et L, de déterminer une partition crédale des données $M^\Omega = (m^\Omega_1, m^\Omega_2, ..., m^\Omega_n)$. Tout comme dans le cas net, on choisit une approche métrique en posant le critère suivant :

$$J(M^\Omega, a_1, b_1, a_2, b_2) \triangleq \frac{1}{C_1}\sum_{i<j} w_{ij}(a_1 L_{ij} + b_1 - \delta^-_{ij})^2 + \frac{1}{C_2}\sum_{i<j} w_{ij}(a_2 U_{ij} + b_2 - \delta^+_{ij})^2 \tag{5.26}$$

où C_1 et C_2 sont deux constantes de normalisation définies par :

$$C_1 = \sum_{i<j} {\delta^-_{ij}}^2, \tag{5.27}$$

et

$$C_2 = \sum_{i<j} {\delta^+_{ij}}^2. \tag{5.28}$$

Les paramètres C_1 et C_2 sont utilisés pour égaliser l'influence des deux termes dans J. Pour fournir au modèle une plus grande flexibilité, deux relations

affines différentes plutôt qu'une sont supposées (paramétrées par a_1, b_1, a_2 et b_2) : une pour les bornes basses de dissimilarités, l'autre pour les bornes hautes. Il est également possible d'ajouter à J, tout comme dans le cas net, un terme d'entropie (équation [5.18]) permettant de pénaliser la complexité du modèle.
En utilisant la même reparamétrisation que dans le cas net (cf équation (5.16)), la minimisation de J par rapport à M^{Ω}, a_i and b_i $(i = 1, 2)$ est un problème d'optimisation non linéaire sans contrainte.

REMARQUE 6 Si l'on veut accorder moins d'influence aux données les plus imprécises durant la phase d'optimisation, la valeur de w_{ij} peut être choisie dans l'intervalle [0 ;1] plutôt que dans l'ensemble $\{0, 1\}$. Une expression possible pour w_{ij} serait :

$$w_{ij} = \begin{cases} \dfrac{1}{\beta(\delta_{ij}^{+} - \delta_{ij}^{-}) + 1} & \text{si la dissimilarité est connue} \\ 0 & \text{sinon,} \end{cases} \tag{5.29}$$

où le paramètre β contrôle globalement l'influence de l'imprécision dans la pondération. S'il est fixé à zéro, on retrouve le critère classique.

4.3 Exemples

EXEMPLE 5.6 *Données synthétiques.* Pour illustrer notre approche, nous avons généré un jeu de données artificielles. Il est composé de 24 vecteurs $\tilde{\mathbf{x}}_\mathbf{i}$, dont chaque composante est un intervalle $[x_{i\ell}^{-}, x_{i\ell}^{+}]$ $\ell = 1, 2$. Chaque vecteur peut donc se représenter dans le plan par un rectangle comme le montre la figure 5.10. Une matrice de dissimilarités de type intervalle a été calculée de la façon suivante : δ_{ij}^{-} et δ_{ij}^{+} sont définies, respectivement, comme le minimum et le maximum de :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|^2},$$

sous les contraintes

$$x_{i\ell}^{-} \leq x_{i\ell} \leq x_{i\ell}^{+} \qquad \ell = 1, 2$$
$$x_{j\ell}^{-} \leq x_{j\ell} \leq x_{j\ell}^{+} \qquad \ell = 1, 2.$$

L'algorithme proposé a été appliqué avec $c = 2$ classes, en limitant les éléments focaux aux deux classes, l'ensemble vide et Ω. Aucune pénalisation par un terme d'entropie n'est employée. Aucune différence majeure n'ayant été constatée en faisant varier la valeur de β, seuls les résultats obtenus avec une valeur nulle sont présentés. La qualité de la partition peut être jugée au travers du digramme de Shepard fourni en figure 5.11. On voit qu'un bon ajustement

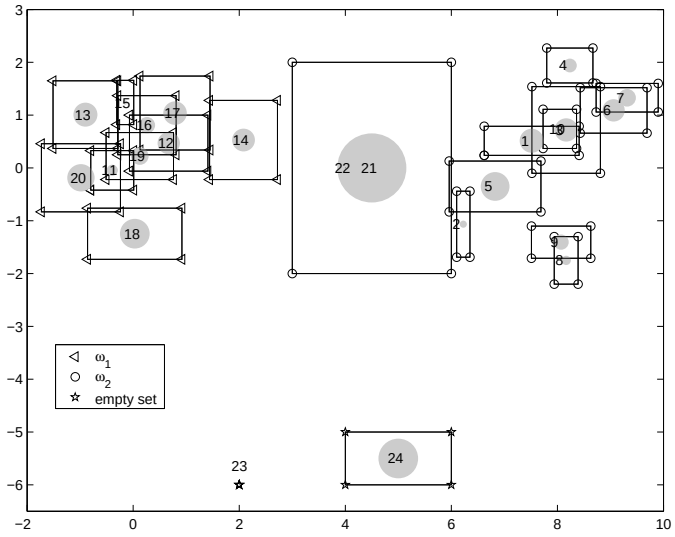

FIG. 5.10 – Jeu de données synthétiques. Partition en deux classes avec rejet. Chaque point est affecté soit à une classe (triangle ou cercle), soit rejeté (étoile] en fonction du maximum de masse alloué. La masse allouée à Ω est proportionnelle au rayon du disque gris sur chaque point. La taille des symboles est proportionnelle à la masse de la classes gagnante.

linéaire est obtenu tant pour les dissimilarités hautes que basses. La figure 5.10 montre les données avec la partition, dans laquelle chaque point est affecté à une classe ou rejeté suivant la valeur de la masse la plus importante. La partition est conforme à ce que l'on attendait, les deux classes étant correctement isolées. La méthode est donc capable de fournir une partition correcte sur des données imprécises. Un deuxième intŕêt de la méthode est de pouvoir déceler des observations atypiques. C'est le cas des points 23 et 24 pour lesquelles la masse la plus importante est allouée à l'ensemble vide. Celui-ci joue le rôle de classe de "bruit" comme dans la méthode Davé [22] permettant la détection de points éloignés. D'autre part, l'algorithme apporte des informations complémentaires grâce à la masse allouée à Ω. Celle-ci est représentée sur la même figure par des disques pleins dont le diamètres est proportionnel à $m^{\Omega}(\Omega)$. On voit clairement une dépendance entre la masse de Ω et l'imprécision des données. Par exemple, les points 22 et 23, dont les composantes sont toutes nettes, ont une masse nulle sur Ω alors que le point 21, qui est le point le plus imprécis du jeu de données, reçoit la masse la plus importante sur Ω.

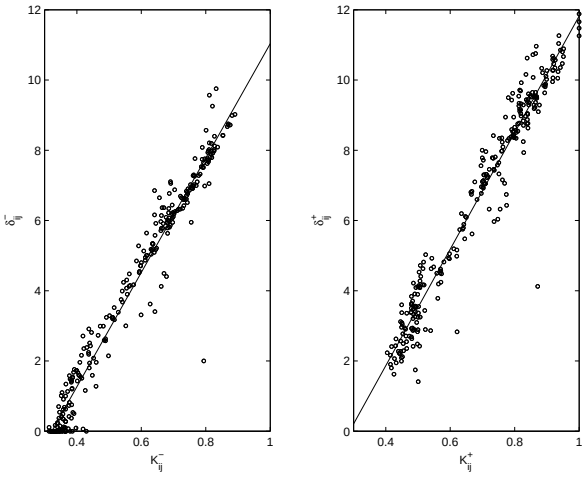

FIG. 5.11 – Dissimilarités et degrés de conflit pour le jeu de données synthétiques (gauche : dissimilarités basses ; droite : dissimilarités hautes). Les lignes représentent les relations affines définies par les coefficients a_1, b_1 (gauche) et a_2, b_2 (droite).

EXEMPLE 5.7 *(Données sensorielles)* Ce jeu de données réel est repris d'une étude sensorielle rapportée dans [86]. Dix variétés de boisson à base de cola ont été présentées à 10 sujets, à qui l'on a demandé de noter la différence de perception entre chaque paire de boissons sur une échelle de 0 à 100 (les valeurs des 45 dissimilarités pour les 10 sujets peuvent être trouvées dans [86]). Chaque dissimilarité est donc caractérisée par uen distribution empirique de réponses, que nous avons résumée par l'intervalle interquartile. Les variété de colas utilisés dans cette expérience sont connues pour se diviser en deux catégories : les boissons normales (Pepsi, Coca, Pepper, Shasta, RC-cola, Yukon) et les boissons diététiques (D-pepper, D-pepsi, D-rite, Tab). Nous avons cherché à retrouver cette partition à partir des données de dissimilarité. L'algorithme a été appliqué avec $c = 2$ classes, en limitant les éléments focaux aux deux classes, l'ensemble vide et Ω et aucune pénalisation. La méthode MDS pour données intervalle développée au chapitre 3 a été employée pour déterminer une configuration bidimensionnelle des points et permettre ainsi une visualisation des résultats. La partition obtenue est présentée en figure 5.12. L'affectation aux classes est faite suivant le maximum de masse. On voit que la dichotomie dietétique/non dietétique est bien retrouvée. De plus, trois boissons (Yukon, Pepper, D-pepper) sont détectées comme atypiques.

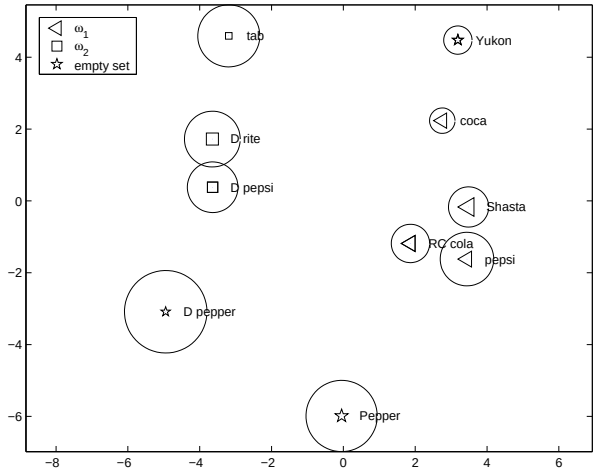

FIG. 5.12 – Données sensorielles. Partition en deux classes avec rejet. Chaque point est affecté soit à une classe, soit rejeté en fonction du maximum de masse alloué. La taille des symboles est proportionnelle à la masse de la classes gagnante.

5 Conclusion

Dans ce chapitre, nous avons présenté une méthode originale de classification de données relationnelles. La méthode a été développée dans un premier temps pour traiter des données classiques, puis étendue pour prendre en compte des données de type intervalle. Elle se fonde sur la théorie des fonctions de croyance et plus précisément sur la notion de partition crédale qui généralise les notions classiques de partitions probabilistes ou floues. Le cadre des fonctions de croyance fournit une grande richesse de description de problème. La possibilité d'attribuer de la masse à l'ensemble vide permet d'intégrer de façon très naturelle le problème classique de la détection de points de points aberrants. L'allocation d'une fraction de la masse à Ω permet quant à elle de gérer l'imprécision et l'inconsistance dans les données. En s'appuyant sur la notion de structure de croyances imprécises déjà développée par certains auteur(cf [24], la généralisation à des données floues pourrait passer par l'attribution aux éléments focaux de masses de croyance de type intervalle. Cette voie n'a pas encore été explorée.

Publications

M. Masson et T. Denoeux. Clustering of proximity data using belief functions. *IPMU'2002*, Vol I, 609-616, Annecy, France, juillet 2002.

T. Denoeux et M. Masson. Clustering of proximity data using belief functions. B. Bouchon-Meunier, L. Foulloy and R. R. Yager, Eds, *Intelligent Systems for Information Processing : From representation to applications*, Elsevier, Amsterdam, pages 291-302, 2003.

T. Denoeux et M. Masson. EVCLUS : Evidential Clustering of Proximity Data *IEEE Transactions on Systems, Man and Cybernetics*, 34(1), 95-109, 2004.

M. Masson et T. Denoeux. Clustering Interval-valued Data using Belief Functions *Pattern Recognition Letters*, 25(2), 163-171, 2004.

Chapitre 6

Inférer une distribution de possibilité à partir de données

1 Introduction

Dans le premier chapitre de ce mémoire, nous avons présenté différentes théories de gestion de l'incertitude et de l'imprécision et notamment la théorie des possibilités et celle des probabilités. Dans certaines applications, il est parfois utile de passer d'un cadre théorique à l'autre. Plusieurs transformations, dans le cas continu ou discret, ont été proposées dans la littérature [37, 19, 23, 67, 70, 33], en se basant sur des principes de consistance ("ce qui est probable est possible") ou d'invariance de l'information. Inférer une distribution de possibilités à partir de données est un problème différent. Une façon de procéder est de supposer que les données ont été générées par une distribution de probabilité inconnue. Si la taille de l'échantillon est suffisamment grande, l'histogramme des données peut être considéré comme une bonne approximation de la distribution sous-jacente et les transformations citées peuvent s'appliquer. Cette approche est clairement insatisfaisante lorsque la taille de l'échantillon est limitée puisque la distribution empirique des données peut différer notablement de la distribution réelle, or ce qui est d'intérêt, c'est bien la distribution réelle. Il s'agit donc d'un problème typique d'inférence pour lequel nous avons utilisé une méthode classique, les régions de confiance, que avons interprétées dans le cadre possibiliste. Un *intervalle de confiance* pour un paramètre scalaire ou plus généralement *une région de confiance* pour un paramètre vectoriel d'une distribution de probabilité est une région aléatoire dans l'espace de paramètres, définie comme une fonction de l'échantillon, qui contient la vraie valeur du paramètre avec un certain niveau de probabilité ou de confiance $1 - \alpha$. L'idée ici

est de considérer l'histogramme observé comme la réalisation d'une variable aléatoire multinomiale de paramètre inconnu $\mathbf{p}$. On construit alors une région de confiance sur $\mathbf{p}$. Cette région de confiance peut être considérée comme spécifiant un ensemble de distributions de probabilité. Nous proposons ensuite une procédure pour inférer la distribution de possibilité la plus spécifique qui soit consistante avec l'ensemble des distributions de probabilité dans cet ensemble. La procédure garantit asymptotiquement que celle-ci soit consistante avec la distribution réelle dans $100(1-\alpha)\%$ des cas. La première partie de ce chapitre est consacrée à la présentation de la méthode de transformation de probabilités en possibilités de Dubois et Prade, sur laquelle se fonde notre approche. Puis nous exposons comment construire les intervalles de confiance et en déduire une distribution de possibilité. Quelques expériences viennent illustrer la méthode.

2 La transformation de Dubois et Prade

Le problème de transformer des probabilités en possibilités a suscité de nombreux travaux [37, 19, 23, 67, 70, 33]. Un principe de consistance entre probabilités et possibilités a été pour la première fois posé par Zadeh [106] de façon informelle : *ce qui est probable devrait être possible*. Dubois et Prade [34, 36] ont ensuite traduit ce principe par l'inégalité :

$$\mathbb{P}(A) \leq \Pi(A) \quad \forall A \subseteq \Omega, \tag{6.1}$$

où $\mathbb{P}$ and Π désigne, respectivement, une mesure de probabilité et de possibilité sur un domaine $\Omega = \{\omega_1, \ldots, \omega_K\}$. Dans ce cas, on dit que Π *domine* $\mathbb{P}$. Transformer une mesure de probabilité en une mesure de possibilité revient à choisir une mesure de possibilité dans l'ensemble $\mathcal{F}(\mathbb{P})$ des mesures de possibilité dominant $\mathbb{P}$. Dubois et al. [44, 32] ont proposé d'ajouter les contraintes suivantes, qui assurent la préservation de la forme de la distribution (contraintes de préservation d'ordre strict) :

$$p_i < p_j \Leftrightarrow \pi_i < \pi_j \quad \forall i, j \in \{1, \ldots, K\}, \tag{6.2}$$

où $p_i = \mathbb{P}(\{\omega_i\})$ et $\pi_i = \Pi(\{\omega_i\})$, pour tout $i \in \{1, \ldots, K\}$. Il est alors naturel de chercher la distribution *la plus spécifique* vérifiant (6.1) et (6.2) (on rappelle qu'une distribution de possibilité π est *plus spécifique* que π' si $\pi_i \leq \pi'_i, \forall i$). Dubois and Prade [44, 32] montre que la solution à ce problème existe et est unique. Elle peut être décrite de la façon suivante. En supposant que $p_i \neq p_j$ quel que soit i, il est possible de définir une relation d'ordre stricte $\mathcal{L}$ sur $\Omega = \{\omega_1, ..., \omega_K\}$ telle que :

$$(\omega_i, \omega_j) \in \mathcal{L} \Leftrightarrow p_i < p_j \ . \tag{6.3}$$

Soit σ une permutation des indices $\{1, 2, ..., K\}$ associée à cet ordre strict telle que $p_{\sigma(1)} < p_{\sigma(2)} < ... < p_{\sigma(K)}$ or, de façon équivalente :

$$\sigma(i) < \sigma(j) \Leftrightarrow (\omega_{\sigma(i)}, \omega_{\sigma(j)}) \in \mathcal{L} \ . \tag{6.4}$$

La permutation σ est une bijection et la transformation inverse σ^{-1} donne le rang de chaque p_i dans la liste des probabilités triées par ordre croissant. On peut alors exprimer la transformation de Dubois et Prade sous la forme suivante :

$$\pi_i = \sum_{\{j|\sigma^{-1}(j) \leq \sigma^{-1}(i)\}} p_j \quad \forall i. \tag{6.5}$$

EXEMPLE 6.1 Soit $p_1 = 0.2$, $p_2 = 0.35$, $p_3 = 0.4$, et $p_4 = 0.05$. Alors on a $\sigma(1) = 4$, $\sigma(2) = 1$, $\sigma(3) = 2$, $\sigma(4) = 3$ and $\sigma^{-1}(1) = 2$, $\sigma^{-1}(2) = 3$, $\sigma^{-1}(3) = 4$, $\sigma^{-1}(4) = 1$. La transformation (6.5) donne donc la distribution de possibilité suivante :

$$\begin{aligned}
\pi_1 &= p_1 + p_4 = 0.2 + 0.05 = 0.25 \\
\pi_2 &= p_2 + p_1 + p_4 = 0.35 + 0.2 + 0.05 = 0.6 \\
\pi_3 &= p_3 + p_2 + p_1 + p_4 = 0.4 + 0.35 + 0.2 + 0.05 = 1 \\
\pi_4 &= p_4 = 0.05.
\end{aligned}$$

REMARQUE 7 La formulation (6.5) suppose que les valeurs de p_i soient toutes différentes. Si au moins deux valeurs sont égales, (6.3) n'induit pas un ordre strict, mais un ordre partiel $\mathcal{P}$ sur Ω (cf chapitre 2 pour des rappels sur les relations d'ordre). Cet ordre partiel peut être représenté par l'ensemble de ses extensions linéaires $\Lambda(\mathcal{P}) = \{\mathcal{L}_l, l = 1, L\}$. A chaque extension linéaire possible $\mathcal{L}_l$ de $\Lambda(\mathcal{P})$, correspond une permutation σ_l de l'ensemble $\{1, .., K\}$ telle que :

$$\sigma_l(i) < \sigma_l(j) \Leftrightarrow (\omega_{\sigma_l(i)}, \omega_{\sigma_l(j)}) \in \mathcal{L}_l. \tag{6.6}$$

Dans ce cas, la distribution la plus spécifique compatible avec $\mathbf{p} = (p_1, ..., p_K)$ est obtenue en retenant le maximum sur toutes les permutations possibles :

$$\pi_i = \max_{l=1,L} \sum_{\{j|\sigma_l^{-1}(j) \leq \sigma_l^{-1}(i)\}} p_j \quad \forall i. \tag{6.7}$$

EXEMPLE 6.2 Soient $p_1 = 0.2$, $p_2 = 0.5$, $p_3 = 0.2$, et $p_4 = 0.1$. Il y a donc deux permutations possibles $\sigma_1(1) = 4$, $\sigma_1(2) = 1$, $\sigma_1(3) = 3$, $\sigma_1(4) = 2$ et $\sigma_2(1) = 4$, $\sigma_2(2) = 3$, $\sigma_2(3) = 1$, $\sigma_2(4) = 2$. En appliquant la transformation (6.7), on obtient :

$$\begin{aligned}
\pi_1 &= \max(p_4 + p_1, p_4 + p_3 + p_1) = \max(0.3, 0.5) = 0.5 \\
\pi_2 &= p_4 + p_1 + p_3 + p_2 = 1 \\
\pi_3 &= \max(p_4 + p_1 + p_3, p_4 + p_3) = \max(0.5, 0.3) = 0.5 \\
\pi_4 &= p_4 = 0.1.
\end{aligned}$$

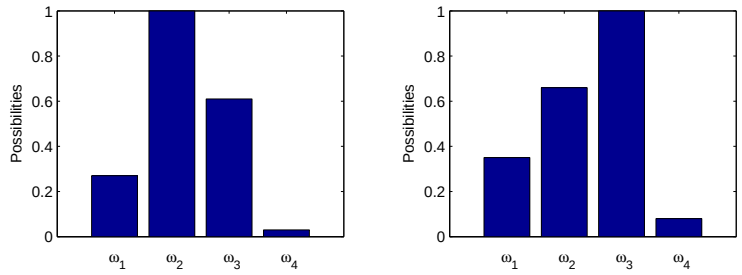

FIG. 6.1 – Deux distributions de possibilité calculées à partir de deux échantillons de même loi.

On voit que $p_1 = p_3$ implique que $\pi_1 = \pi_3$, une condition imposé par la préservation des ordres stricts.

3 Inférer une distribution de possibilité à partir de données expérimentales

3.1 Position du problème

On suppose que les données disponibles consistent en N observations réparties en K classes de $\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$ suivant une distribution de probabilité $\mathbb{P}_X$. Soit n_i le nombre d'observations se répartissant dans la i-ème classe ω_i. Le vecteur $\mathbf{n} = (n_1, n_2, ..., n_K)$ est la réalisation d'une variable aléatoire multinomiale de paramètres $\mathbf{p} = (p_1, p_2, ..., p_K)$, où chaque $p_i = \mathbb{P}_X(\{\omega_i\}) > 0$ est la probabilité d'apparition de la i-ème classe (ou *proportion* de la classe i), avec $\sum_{i=1}^{K} p_i = 1$. L'approche classique pour construire une distribution de possibilité serait d'assimiler le vecteur des fréquences observées dans chaque classe ($\mathbf{f} = (f_1, f_2, ..., f_K)$ avec $f_i = n_i/N$) au vecteur des probabilités $\mathbf{p}$ et d'appliquer la transformation de Dubois et Prade. Cependant, cette approche n'est pas satisfaisante car elle ne tient pas compte du processus d'échantillonnage, comme le montre l'exemple suivant.

EXEMPLE 6.3 Sur la figure 6.1 sont représentées les distributions de possibilité calculées avec (6.5) en partant de deux échantillons tirés suivant la même loi multinomiale de paramètres $\mathbf{p} = [0.2; 0.35; 0.4; 0.05]^t$ et $N = 100$. La première remarque, immédiate, tient dans la forme très différente des deux distributions, alors que les échantillons sont pourtant de même loi parente. Ensuite, on observe qu'une des distributions (celle de gauche) n'est pas consistante avec la distribution de probabilité réelle (voir l'exemple 6.1).

L'inférence statistique classique fournit des outils pour étendre des résultats observés sur un échantillon à une population plus générale. En particulier, les *intervalles ou régions de confiance* sont un des moyens usuels pour estimer les paramètres inconnus d'une distribution de probabilité. Un intervalle de confiance à un niveau de confiance α (par exemple $\alpha = 5\%$) est un intervalle, aléatoire, fonction des observations, qui contient la vraie valeur du paramètre avec une probabilité $1 - \alpha$ (si l'on disposait d'un grand nombre d'échantillons et que l'on calculait un intervalle de confiance sur chaque échantillon, alors cet intervalle contiendrait la vraie valeur du paramètre dans $100(1 - \alpha)\%$ des cas).
L'idée développée dans ce chapitre consiste à d'abord estimer les p_i à l'aide d'intervalles de confiance sur les proportions d'une loi multinomiale, puis à en déduire une distribution de possibilités. La procédure sera construite pour assurer que la distribution de possibilité domine la vraie distribution de probabilité dans au moins $100(1 - \alpha)\%$ des cas, ce qui peut se traduire sous la forme :

$$\mathbb{P}\left(\Pi(A) \geq \mathbb{P}_X(A), \forall A \subseteq \Omega\right) \geq 1 - \alpha \ , \tag{6.8}$$

où $\mathbb{P}_X(A)$ est la probabilité inconnue mais constante de l'événement A, et $\Pi(A)$ est une variable aléatoire fonction des observations. Notons que la proposition (6.8) est équivalente à

$$\mathbb{P}\left(N(A) \leq \mathbb{P}_X(A), \forall A \subseteq \Omega\right) \geq 1 - \alpha \ , \tag{6.9}$$

où N est la mesure de nécessité associée à Π, définie comme $N(A) = 1 - \Pi(\overline{A})$, $\forall A \subseteq \Omega$.

3.2 Intervalles de confiance pour proportions de loi multinomiale

Pour construire un intervalle de confiance sur les proportions d'une loi multinomiale, plusieurs voies sont envisageables. La première consiste à considérer chaque effectif n_i contre tous les autres comme la réalisation d'une loi binomiale et de se servir de cette hypothèse pour construire des intervalles de confiance indépendamment les uns des autres. Cette approche ne permet malheureusement pas de contrôler le degré de confiance global de l'ensemble des intervalles. Une meilleure approche est de construire des *intervalles de confiance simultanés* avec un degré de confiance joint $1 - \alpha$, d'où le nom d'intervalles de confiance simultanés. La recherche d'intervalles de confiance simultanés est un problème ancien et de nombreuses méthodes ont été proposées dans la littérature [80, 51, 48, 88]. Toutes ces méthodes cherchant une région de confiance $\mathcal{C}_{\mathbf{n}}$ dans l'espace des paramètres $\{\mathbf{p} = (p_1, ...p_K) \in [0; 1]^K \,|\, \sum_{i=1}^{K} p_i = 1\}$ comme le produit cartésien de K intervalles $[p_1^-, p_1^+] \times ... \times [p_K^-, p_K^+]$ tels que

$$\mathbb{P}\left(\mathbf{p} \in \mathcal{C}_{\mathbf{n}}\right) \geq 1 - \alpha \tag{6.10}$$

i	1	2	3	4
p_i^-	0.10	0.34	0.25	0
p_i^+	0.28	0.56	0.46	0.08

TAB. 6.1 – Intervalles de confiance pour l'exemple 6.4.

Cette probabilité est appelée le taux de couverture de l'estimation. A taux de couverture équivalent, le meilleur estimateur est celui de volume le plus faible. Nous avons retenu la solution proposée par [51] qui a été testée sur différents jeux de données simulés et a montré des performances satisfaisantes dans la plupart des cas envisagés [72]. Nous ne développerons pas ici comment obtenir la formulation de Goodman, seules les formules principales seront rappelées. Soient

$$A = \chi^2(1 - \alpha/K, 1) + N, \tag{6.11}$$

où $\chi^2(1 - \alpha/K, 1)$ désigne le quantile de niveau $1 - \alpha/K$ d'une distribution du chi-2 à 1 degré de liberté, et $N = \sum_{i=1}^{K} n_i$ la taille de l'échantillon. Soient de plus

$$B_i = \chi^2(1 - \alpha/K, 1) + 2n_i, \tag{6.12}$$

$$C_i = \frac{n_i^2}{N}, \tag{6.13}$$

$$\Delta_i = B_i^2 - 4AC_i. \tag{6.14}$$

Alors on obtient les bornes des intervalles de confiance par la formule suivante :

$$[p_i^-, p_i^+] = \left[\frac{B_i - \Delta_i^{\frac{1}{2}}}{2A}, \frac{B_i + \Delta_i^{\frac{1}{2}}}{2A} \right]. \tag{6.15}$$

Notons que cette formule s'appuie sur des approximations asymptotiques. Se basant sur le résultat de leurs nombreuses simulations, May and Johnson [72] ont montré que les intervalles de Goodman se comportent plutôt bien en termes de taux de couverture et de volume de la région de confiance, pourvu que le nombre de classes soit supérieur à 2 et que les effectifs dnas chaque classe soient supérieurs à 5. Si ces conditions ne sont pas vérifiées, alors il vaut mieux se tourner vers d'autres méthodes comme par exemple celle de Sison et Glaz [88].

EXEMPLE 6.4 Soient les mêmes valeurs de probabilité que précédemment $\mathbf{p} = [0.2; 0.35; 0.4; 0.05]^t$. On suppose qu'on observe un échantillon de taille 100 avec la distribution suivante dans les classes : 18, 45, 35, et 2. En fixant $\alpha = 0.1$, on trouve les intervalles de confiance simultanés donnés dans le tableau 3.2 et représentés sur la figure 6.2.

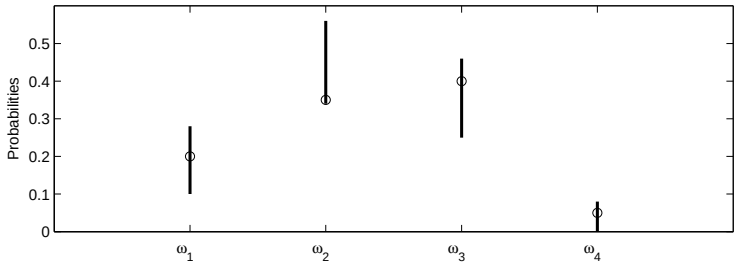

FIG. 6.2 – Intervalles de confiance de l'exemple 6.4 ; les cercles représentent les valeurs de probabilité.

3.3 Mesure de probabilité inférieure induite

Une région de confiance $\mathcal{C}_{\mathbf{n}}$ pour les proportions d'une loi multinomiale comme celle décrite au paragraphe précédent est interprétée généralement comme un ensemble de valeurs plausibles pour le vecteur de paramètres $\mathbf{p}$. Cependant, puisque chaque valeur de $\mathbf{p}$ définit une unique mesure de probabilité, il est clair que $\mathcal{C}_{\mathbf{n}}$ peut aussi bien être vue comme une famille de mesures de probabilité. Pour garder des notations aussi simples que possibles, $\mathcal{C}_{\mathbf{n}}$ sera utilisé par la suite pour désigner à la fois l'ensemble des valeurs possibles pour $\mathbf{p}$ et l'ensemble des mesures de probabilité. Soient P^- and P^+ désignant, respectivement, l'enveloppe basse et l'enveloppe haute de $\mathcal{C}_{\mathbf{n}}$, définies par $P^-(A) = \min_{P \in \mathcal{C}_{\mathbf{n}}} P(A)$ and $P^+(A) = \max_{P \in \mathcal{C}_{\mathbf{n}}} P(A)$. On peut les calculer à l'aide de la proposition suivante :

PROPOSITION 3
Pour tout sous-ensemble non vide A de Ω,

$$P^-(A) = \max\left(\sum_{\omega_i \in A} p_i^-, 1 - \sum_{\omega_i \notin A} p_i^+\right) \tag{6.16}$$

$$P^+(A) = \min\left(\sum_{\omega_i \in A} p_i^+, 1 - \sum_{\omega_i \notin A} p_i^-\right). \tag{6.17}$$

Démonstration. Pour tout $A \subset \Omega$, $P^-(A)$ est solution du programme linéaire suivant :

$$\min_{p_1,\ldots,p_K} \sum_{\omega_i \in A} p_i,$$

sous les contraintes $\sum_{i=1}^K p_i = 1$ et $p_i^- \leq p_i \leq p_i^+$, $i = 1, \ldots, K$. Ce problème est un cas particulier d'une famille de problèmes d'optimisation linéaires étudiés

par Dubois et Prade dans le contexte de l'arithmétique floue [35, 38]. L'équation (6.16) est dérivée de la formule générale donnée dans [38, page 55]. L'équation (6.17) s'obtient de la même manière. □

Notons que, conséquence directe de la proposition 3, on a :

$$P^+(A) = 1 - P^-(\overline{A}), \quad \forall A \subseteq \Omega.$$

Donc, la mesure de probabilité inférieure P^- est suffisante pour caractériser $\mathcal{C}_\mathbf{n}$:

$$\mathcal{C}_\mathbf{n} = \{P \mid P^- \leq P\}.$$

Par construction, on a

$$\mathbb{P}(\mathbb{P}_X \in \mathcal{C}_\mathbf{n}) = \mathbb{P}(P^- \leq \mathbb{P}_X) \geq 1 - \alpha \tag{6.18}$$

et, de façon équivalente,

$$\mathbb{P}(P^+ \geq \mathbb{P}_X) \geq 1 - \alpha \ . \tag{6.19}$$

Les équations (6.18) et (6.19) s'apparentent à (6.9) et (6.8), respectivement. Cependant, P^- n'est pas une mesure de nécessité. Il ne s'agit même pas, en général, d'une fonction de croyance [87] lorsque $K > 2$, comme le montre l'exemple 6.5 ci-dessous. Par contre, on peut montrer qu'il s'agit d'une capacité monotone d'ordre 2, c'est-à-dire que l'on a

$$P^-(A \cup B) \geq P^-(A) + P^-(B) - P^-(A \cap B), \quad \forall A, B \subseteq \Omega.$$

En conséquence, P^- est une *mesure de probabilité inférieure cohérente* [99].

Exemple 6.5 Reprenons l'exemple de la région de confiance calculée dans l'exemple 6.4. Les probabilités inférieures correspondantes sont données dans le tableau 6.2. Comme l'a énoncé Shafer [87], une application $f : 2^\Omega \to [0, 1]$ est une fonction de croyance si et seulement si sa transformée inverse de Möbius, définie par :

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} f(B), \quad \forall A \subseteq \Omega,$$

est une masse de croyance (i.e., si $m(A) \geq 0$ pour tout A, et $\sum_{A \subseteq \Omega} m(A) = 1$). La transformée inverse de Möbius de P^-, donnée dans le tableau 6.2, alloue une valeur négative à Ω, donc P^- n'est pas une fonction de croyance.

A	$P^-(A)$	$P^+(A)$	$m(A)$
$\{\omega_1\}$	0.1038	0.2938	0.1038
$\{\omega_2\}$	0.3323	0.5735	0.3323
$\{\omega_1, \omega_2\}$	0.4362	0.7530	0
$\{\omega_3\}$	0.2429	0.4747	0.2429
$\{\omega_1, \omega_3\}$	0.3467	0.6636	0.0000
$\{\omega_2, \omega_3\}$	0.6139	0.8921	0.0387
$\{\omega_1, \omega_2, \omega_3\}$	0.9077	0.9959	0.1900
$\{\omega_4\}$	0.0041	0.0923	0.0041
$\{\omega_1, \omega_4\}$	0.1079	0.3861	0
$\{\omega_2, \omega_4\}$	0.3364	0.6533	0
$\{\omega_1, \omega_2, \omega_4\}$	0.5253	0.7571	0.0850
$\{\omega_3, \omega_4\}$	0.2470	0.5638	0
$\{\omega_1, \omega_3, \omega_4\}$	0.4265	0.6677	0.0757
$\{\omega_2, \omega_3, \omega_4\}$	0.7062	0.8962	0.0882
Ω	1.0000	1.0000	-0.1607

TAB. 6.2 – Probabilités inférieures et supérieures induites par les intervalles de confiance du tableau 3.2, et transformée inverse de Möbius correspondante.

3.4 Générer une distribution de possibilité à partir de probabilités de type intervalle

On a vu précédemment qu'un ensemble d'intervalles de confiance simultanés à un niveau de confiance $1 - \alpha$ peut être vu comme une famille de distributions de probabilité, décrit de façon exacte par son enveloppe inférieure P^- (ou de façon équivalente par son enveloppe supérieure P^+).
Rappelons que le but est de trouver la mesure de probabilité vérifiant (6.8). Il est maintenant clair que cette propriété est satisfaite pour tout possibilité Π dominant P^+, puisque $\Pi(A) \geq P^+(A)$, $\forall A \subseteq \Omega$ et (6.19) implique (6.8).
Trouver la distribution la plus spécifique dominant une fonction de plausibilité a déjà été abordé dans [40], dans lequel un algorithme de calcul d'une solution a été proposé. Cependant cet algorithme n'est pas applicable, puisque, comme nous l'avons montré dans le paragraphe précédent, P^- n'est pas une fonction de croyance (ou de façon équivalente, P^+ n'est pas une fonction de plausibilité).
Il est clair que Π domine la mesure de probabilité supérieure P^+ si et seulement si elle domine tous les éléments dans la famille correspondante de mesures de probabilité, c'est-à-dire, toutes les mesures de probabilité P telles que $P \leq P^+$, ou de façon équivalente, toutes les distributions de probabilité vérifiant $p_i^- \leq p_i \leq p_i^+$, $i = 1, \ldots, K$.

Nous reformulons donc notre but en cherchant la distribution de possibilité sur Ω la plus spécifique dominant toute distribution de probabilité définie par $p_i \in [p_i^-, p_i^+] \quad \forall i$, ou de façon équivalente toute distribution de possibilité induite par les p_i, quelle que soit leur valeur dans $[p_i^-, p_i^+]$.

Pour atteindre ce but, notre approche sera d'utiliser la transformation décrite dans le paragraphe 2, qui permet de calculer la distribution la plus spécifique dominant une mesure de probabilité particulière.
Soit $\mathcal{P}$ l'ordre partiel induit par les intervalles $[p_i] = [p_i^-, p_i^+]$:

$$(\omega_i, \omega_j) \in \mathcal{P} \Leftrightarrow p_i^+ < p_j^- . \tag{6.20}$$

Comme nous l'avons déjà expliqué dans le paragraphe 2, cet ordre partiel peut être assimilé à l'ensemble de ses extensions linéaires $\Lambda(\mathcal{P}) = \{\mathcal{L}_l, l = 1, L\}$, ou, de façon équivalente, par l'ensemble des permutations correspondantes $\{\sigma_l, l = 1, L\}$.
Formellement, la solution de notre problème peut donc se calculer de la manière suivante :

1. Pour toute permutation possible σ_l associée à une extension linéaire dans $\Lambda(\mathcal{P})$, et chaque classe ω_i, résoudre le programme linéaire suivant :

$$\pi_i^{\sigma_l} = \max_{p_1,\ldots,p_K} \sum_{\{j|\sigma_l^{-1}(j) \leq \sigma_l^{-1}(i)\}} p_j \tag{6.21}$$

sous les contraintes

$$\begin{cases} \sum_{k=1}^{K} p_k = 1 \\ p_k^- \leq p_k \leq p_k^+ \quad \forall k \in \{1, ..., K\} \\ p_{\sigma_l(1)} \leq p_{\sigma_l(2)} \leq ... \leq p_{\sigma_l(K)} \end{cases} \tag{6.22}$$

2. Ensuite, retenir la distribution la plus spécifique dominant toutes les distributions π^{σ_l} :

$$\pi_i = \max_{l=1,L} \pi_i^{\sigma_l} \quad \forall i \in \{1, ..., K\}. \tag{6.23}$$

Si les $[p_i]$ sont des intervalles de confiance simultanés calculés en utilisant (6.15) avec un niveau de confiance $1-\alpha$, cette procédure assure que la distribution de possibilité résultante soit la plus spécifique qui domine toutes les probabilités compatibles et donc la probabilité supérieure P^+. Elle vérifie donc, et c'était le but fixé, la propriété (6.8).

EXEMPLE 6.6 Considérons à nouveau les 4 classes caractérisées par les intervalles de probabilité donnés dans le tableau 3.2 et représentés sur la figure 6.2. L'ordre partiel sur les classes correspondant est :

$$\mathcal{P} = \{(\omega_4, \omega_1), (\omega_1, \omega_2), (\omega_4, \omega_2), (\omega_4, \omega_3)\}.$$

Il y a trois permutations compatibles avec $\mathcal{P}$: $\sigma_1 = (4, 1, 3, 2)$, $\sigma_2 = (4, 1, 2, 3)$, $\sigma_3 = (4, 3, 1, 2)$. Les rangs correspondants sont $\sigma_1^{-1} = (2, 4, 3, 1)$, $\sigma_2^{-1} = (2, 3, 4, 1)$ et $\sigma_3^{-1} = (3, 4, 2, 1)$. Le tableau 6.6 donne la solution des différents programmes linéaires ainsi que la distribution de possibilité finale. Les classes ω_2 et ω_3 reçoivent logiquement un degré de possibilité maximal, puisque chacune des deux peut être classée en dernière position (dans la liste des probabilités triées par ordre croissant). La classe ω_4, toujours classée en première position, reçoit un degré de possibilité correspondant à la borne maximale de son intervalle associé. La valeur de 0.64 pour la première classe est obtenue avec la troisième permutation avec une valeur optimale de $\mathbf{p}$ égale à [0.28,0.36,0.28,0.08].

l	$\pi_1^{\sigma_l}$	$\pi_2^{\sigma_l}$	$\pi_3^{\sigma_l}$	$\pi_4^{\sigma_l}$
1	0.36	1	0.66	0.08
2	0.32	0.66	1	0.08
3	0.64	1	0.36	0.08
$\max_l$	0.64	1	1	0.08

TAB. 6.3 – Solutions des programmes linéaires et distribution de possibilité optimale (Exemple 6.6).

REMARQUE 8 Il est très important de noter que le problème (6.21)-(6.22) n'a pas toujours de solution. A titre illustratif, considérons les intervalles de probabilité suivants : $p_1 = [0; 0.9]$, $p_2 = [0.1; 0.3]$, $p_3 = [0; 0.8]$. Une permutation possible est $\sigma(1) = 1$, $\sigma(2) = 3$, $\sigma(3) = 2$. Cependant, il est impossible de satisfaire la contrainte d'égalité, p_1 et p_3 étant limités par la valeur maximale de p_2 qui est de 0.3. La solution finale doit être trouvée en cherchant le maximum parmi les solutions *réalisables*, ce qui donne ici $\pi_1 = 1$, $\pi_2 = 0.6$ and $\pi_3 = 1$.

3.5 Procédure de calcul

Pour calculer les degrés de possibilité associés à chaque classe, l'approche la plus simple consiste à générer toutes les extensions linéaires compatibles avec l'ordre partiel induit par les intervalles de probabilité, et ensuite à résoudre les programmes linéaires associés. Cette approche est malheureusement limitée à

des valeurs faibles de K (disons $K < 10$) en raison de la complexité des algorithmes de génération des extensions linéaires : l'algorithme le plus rapide est à notre connaissance celui de Pruesse et Ruskey [79], dont la complexité est en $O(L)$, où L désigne le nombre d'extensions linéaires. Même pour des valeurs faibles de K, L peut être très élevé ($K!$ dans le pire des cas) et générer l'ensemble des extensions linéaires devient infaisable. Pour cette raison, nous avons proposé une solution, intitulée `Prob2poss`, que nous avons montrée équivalente à la précédente et qui permet de réduire les calculs de façon très importante. Cette solution ne sera pas explicitée ici, considérant que les démonstrations, assez longues et techniques, noieraient un peu le discours principal de ce chapitre. On pourra se reporter à la publication dans Fuzzy Sets and Systems qui détaille comment cette procédure a été obtenue. L'idée directrice est qu'en fait il n'est pas nécessaire d'évaluer la solution pour toutes les extensions linéaires possibles. Au contraire, il est possible de construire une recherche arborescente de la solution, partant de la racine vers les feuilles, qui est stoppée dès qu'une solution a été trouvée à un certain niveau de l'arbre. Cette procédure a montré son efficacité dans les expériences qui sont décrites dans le paragraphe qui suit.

4 Quelques expériences

4.1 Convergence vers la distribution de Dubois et Prade

Considérons un histogramme de données continues composé de K=11 classes uniformément espacées de -2 à 2. Les fréquences observées dans les différentes classes sont supposées être égales à

$$f = [0.04, 0.02, 0.05, 0.09, 0.14, 0.27, 0.14, 0.14, 0.07, 0.02, 0.02],$$

conduisant à la distribution de possibilité calculée avec la transformation de Dubois et Prade représentée en noir sur la figure 6.3. On suppose que plusieurs échantillons, de taille variable (N=100,500,1000 and 10000) ont conduit aux mêmes valeurs de fréquence. Les intervalles de Goodman, calculés avec $\alpha = 0.05$, sont représentés sur la figure 6.4. Les résultats obtenus avec la transformation proposée sont donnés sur la figure 6.3 en grisé. Ils confirment ce qui était attendu : notre transformation est toujours moins spécifique (car plus prudente) que celle de Dubois et Prade et converge asymptotiquement vers elle lorsque N tend vers l'infini.

4.2 Taux de couverture

Dans un second temps, nous nous sommes intéressés à la probabilité de couverture de la transformation proposée (c'est-à-dire la probabilité que la distri-

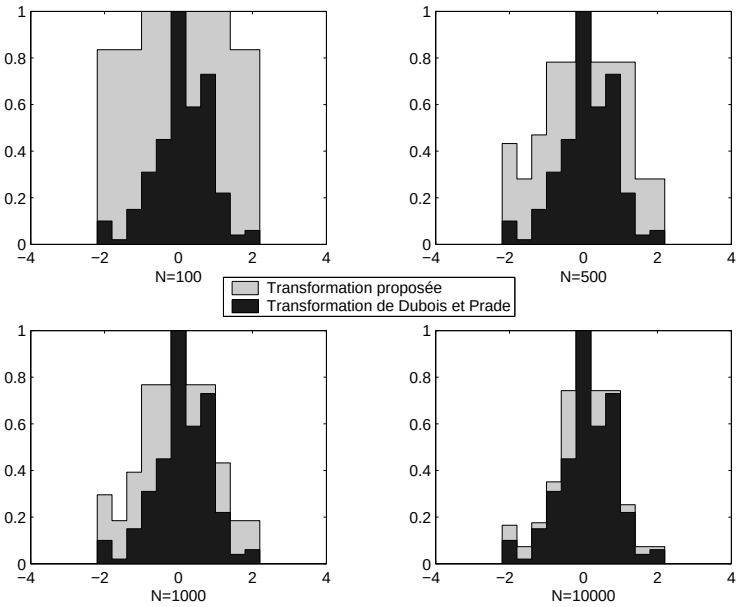

FIG. 6.3 – Convergence de la transformation proposé vers celle de Dubois et Prade lorsque $N \to \infty$.

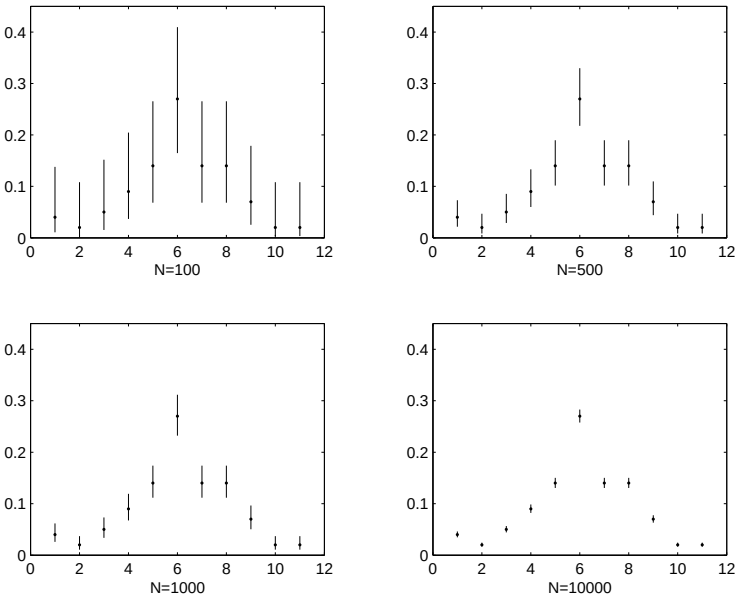

FIG. 6.4 – Intervalles de confiance de Goodman ; les points representent les fréquences des classes.

bution de possibilité résultante domine la distribution de probabilité réelle). Pour cela, la procédure suivante a été observée :

- étape 1 : on considère que Ω est composé de $K = 5$ classes. Cinq proportions p_i ont été uniformément choisies dans l'intervalle [0,1] avec la contrainte $\sum_{i=1}^{K} p_i = 1$;
- étape 2 : la probabilité de couverture a été estimée en utilisant 100 échantillons de taille N fixe, tirés suivant une loi multinomiale de paramètres p_i ; cette estimation est obtenue en calculant, pour chaque échantillon, un ensemble d'intervalles de confiance avec une valeur fixe de α, en appliquant notre transformation et en vérifiant que :

$$\Pi(A) \geq \mathbb{P}_X(A) \quad \forall A \subseteq \Omega$$

  (pour chaque sous-ensemble A de Ω, on vérifie si le maximum des degrés de possibilité des singletons inclus dans A est supérieur ou égal à la somme de leur probabilités) ;
- cette probabilité de couverture est moyennée sur $nrep = 100$ répétitions de l'expérience précédente (étapes 1 and 2) ;
- l'expérience globale est répétée en utilisant différentes valeurs de α (0.01,0.05,0.1,0.2,0.3,0.4) et de N (100, 1000, 10000).

On montre sur la partie gauche de la figure 6.5 le taux de couverture de la transformation proposée. On voit que la solution domine en fait la distribution réelle avec un taux de couverture bien supérieur à $100(1-\alpha)\%$. Notre approche est donc très conservatrice de ce point de vue. A titre de comparaison, le taux de couverture des intervalles de Goodman est représenté sur la partie droite de la figure précédente. Pour notre transformation, le choix de α n'est pas très critique, et, quelle que soit sa valeur, un très bon taux de couverture est assuré. Ce choix doit résulter d'un compromis entre le taux de couverture désiré et la spécificité de la distribution résultante.

5 Conclusion

Nous avons présenté dans ce chapitre les principes généraux qui nous semblent devoir guider la construction d'une distribution de possibilités à partir de données expérimentales. Il s'agit de rester dans le cadre de l'inférence statistique et donc de tenir explicitement compte de la taille de l'échantillon et des fluctuations d'échantillonnage. Nous avons supposé que les données étaient générées suivant une distribution de probabilité inconnue et réparties dans différentes classes. Sur la base de cette hypothèse, des intervalles de confiances simultanés sur les proportions d'une loi multinomiale ont été construits. Ils définissent un ensemble de distributions de probabilités. Nous avons ensuite

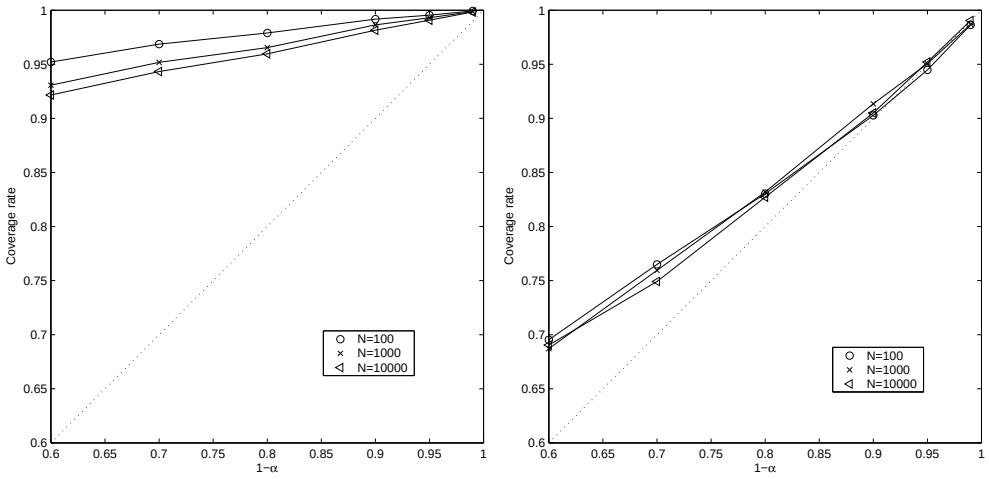

FIG. 6.5 – Expérience 2. Taux de couverture la transformation proposée (à gauche) et des intervalles de Goodman (à droite).

proposé de construire la distribution de possibilité la plus spécifique qui domine l'ensemble de ces distributions de probabilité. Cette procédure garantit que la distribution de possibilité domine la distribution de probabilité réelle dans au moins $100(1-\alpha)\%$. Nous pensons que cette approche peut ouvrir des voies intéressantes pour faire collaborer Inférence Statistique classique et Théorie des Possibilités.

Publications

M. Masson et T. Denoeux. Inferring a possibility distribution from empirical data. *Fuzzy Sets and Systems*, 157(3), 319-340, 2006.

Conclusion générale et perspectives

Nous avons évoqué dans ce mémoire une partie des travaux que nous avons menés autour du thème de l'analyse de données "imprécises". Dans un premier chapitre, nous avons rappelé les différents cadres théoriques classiques pour représenter et manipuler l'imprécis et l'incertain dans les systèmes d'informations. Ensuite, nous avons présenté les outils que nous avons développés. Certains relèvent des statistiques exploratoires, d'autres des statistiques inférentielles même si l'aspect descriptif est le plus présent dans notre travail : corrélation et test flou associé, visualisation (analyse en composantes principales, ou positionnement multidimensionnel), classification automatique. Nous avons tenu à illustrer les méthodes à l'aide de nombreux jeux de données, pour une grande part issue d'applications réelles. Le dernier chapitre, tout en restant dans le cadre des statistiques et de l'analyse de données, s'écarte du thème de l'analyse de données imprécises. Les données traitées sont précises, c'est la caractérisation de leur fonction génératrice qui est imprécise et qui conduit à une proposition de transformation probabilités/possibilités qui s'intègre de manière naturelle dans le cadre inférentiel classique.

Dans un souci d'homogénéité du document, nous avons passé sous silence d'autres études plus ou moins connexes menées grâce au travail de doctorants ou d'étudiants de DEA sous notre responsabilité. Ces études portent notamment sur le diagnostic (thèses de Nassim Boudaoud et Léopold Tsogo), sur l'utilisation de relations floues pour l'analyse de données sensorielles (thèse de Pierre-Alexandre Hébert) ou encore le développement de méthodes de fusion de classifieurs dans le cadre du modèle des croyances transférables (thèse de Benjamin Quost et David Mercier).

A court terme, le travail présenté ouvre plusieurs perspectives intéressantes. En termes d'analyse de données multidimensionnelles, il est clair que de nombreuses méthodes factorielles peuvent être étendues pour prendre en compte des données floues. Nous pensons notamment aux méthodes de traitement de tableaux à trois entrées (ou plus) comme PARAFAC [55] ou STATIS [28]. Ces méthodes sont couramment appliquées dans le domaine de l'analyse sensorielle pour analyser simultanément plusieurs tableaux individus-variables fournis par un panel de consommateurs. Il en va de même de INDSCAL (INDividual Difference SCALing) [15], algorithme de positionnement multidimensionnel de plusieurs tableaux de dissimilarités, qui pourrait facilement être étendu à des données floues. Notons qu'au delà du développement proprement dit d'algorithmes, il semble qu'un travail de diffusion des méthodes soit nécessaire auprès des utilisateurs potentiels, notamment de la communauté d'analyse sensorielle. En termes de classification automatique, nous avons d'ores et déjà évoqué comment on pouvait envisager l'extension de l'algorithme proposé à des données floues, par l'emploi de structures de croyance imprécises.

L'aspect inférentiel n'a été qu'effleuré dans le deuxième et le dernier chapitre. Nous pensons qu'un effort important est à mener sur ce point. Nous avons assez peu travaillé sur la discrimination alors qu'il s'agit d'un problème central dans le domaine de la reconnaissance des formes. Nous disposons du modèle d'ACP de données floues par réseaux autoassociatif qui s'étend de manière immédiate à l'analyse discriminante. Nous pensons également que la méthode des k-plus proches voisins développée dans le cadre du modèle des croyances transférables par Denoeux [107] peut facilement être étendue à des données de type intervalle, et par la suite à des données floues. L'idée serait de considérer, tout comme dans la classification automatique, que les distances minimales et maximales d'un objet à son voisin sont deux sources d'information à modéliser et à combiner de façon adéquate. Pour ce qui est du lien entre probabilités et possibilités, notre travail s'est limité jusque là au cas discret. Il est clair qu'il serait intéressant d'étendre les transformations proposées dans le cas continu [32] à des spécifications imprécises d'une densité de probabilité, ou d'une fonction de répartition.

A plus long terme, il nous semble qu'un travail doit être mené pour mieux comprendre les liens, les domaines d'application privilégiés des différents cadres de gestion de l'incertitude, et encourager leur utilisation conjointe. Ce travail ne pourra aboutir que si des chercheurs issus de communautés différentes (intelligence artificielle, statistiques) collaborent. Même si le cadre des fonctions de croyance nous parait a priori séduisant car très général (et un contexte local nous y pousse...), notre volonté n'est pas de nous y limiter mais plutôt de faire collaborer différents modes de représentation ou de raisonnement. Pourquoi ne

pas penser par exemple à mettre à profit la théorie des sous-ensembles flous et la manipulation de concepts linguistiques pour produire des sorties d'une analyse en composantes principales facilement *interprétables* pour un utilisateur non averti ?

Pour conclure, il faut souligner que nos recherches ont toujours été nourries par des applications, notamment au travers des liens étroits que nous avons tissés avec PSA-Peugeot Citroën. Nous tenterons à l'avenir, dans la mesure du possible, de conserver cette démarche.

Bibliographie

[1] P. Baldi and K. Hornik. Neural networks and principal component analysis : Learning from examples without local minima. *Neural Networks*, 2 :53–58, 1989.

[2] P. Baldi and K. Hornik. Learning in linear neural networks : A survey. *IEEE Transactions on Neural Networks*, 6(4) :837–858, 1995.

[3] H. Bandemer and W. Näther. *Fuzzy data analysis*. Kluwer Academic Publishers, Dordrecht, 1992.

[4] C. Bertoluzza, M. A. Gil, and D. A. Ralescu, editors. *Statistical modeling, analysis and management of fuzzy data*. Physica-Verlag, Heidelberg, 2002.

[5] J.C Bezdek. *Pattern Recognition with fuzzy objective function algorithm*. Plenum Press, New-York, 1981.

[6] J.C. Bezdek, J. Keller, R. Krishnapuram, and N.R. Pal. *Fuzzy models and algorithms for pattern recognition and image processing*. Kluwer Academic Publishers, Boston, 1999.

[7] T. Bilgic and I. Türksen. Measurement of membership functions : theoretical and empirical work. In D. Dubois and H. Prade, editors, *Fundamentals of Fuzzy sets*, pages 195–228. Kluwer Academic Publishers, Boston, 2000.

[8] I. Bloch. Information Combination Operators for Data Fusion : A Comparative Review with Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 26(1) :52–67, 1996.

[9] I. Bloch and H. Maître. Fusion de données en traitement d'images : modèles d'information et décisions. *Traitement du Signal*, 11(6) :435–446, 1994.

[10] P. Bonissone and R.M. Tong. Editorial : reasoning with uncertainty in expert systems. *International journal of Man Machine Studies*, 22 :241–250, 1985.

[11] I. Borg and P. Groenen. *Modern multidimensional scaling*. Springer, New-York, 1997.

[12] P. Bosc, D. Dubois, and H. Prade. An introduction to fuzzy sets and possibility theory based approaches to the treatment of uncertainty and imprecision in database management systems. In *Proceedings of the 2nd Workshop on Uncertainty Management in Information Systems : From needs to solutions, Catalina, CA, USA*, 1993.

[13] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59 :291–294, 1988.

[14] R. Bubley and D.Dyer. Faster random generation of linear extensions. In *Proceedings of the Ninth Annual ACM-SIAM Symposyum on Discrete Algorithms*, pages 175–186, 1998.

[15] J. D. Carrol and J. J. Chang. Analysis of individual differences in multidimensional scaling via n-way generalization of eckart-young decomposition. *Psychometrika*, 35 :283–319, 1970.

[16] P. Cazes, A. Chouakria, E. Diday, and Y. Schektman. Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique Appliquée*, 14(3) :5–24, 1997.

[17] B. B. Chaudhuri and A. Bhattacharya. On correlation between two fuzzy sets. *Fuzzy sets and systems*, 118 :15–35, 2001.

[18] D.A. Chiang and N.P. Lin. Correlation of fuzzy sets. *Fuzzy sets and systems*, 102 :221–226, 1999.

[19] M.R. Civanlar and H. J. Trussel. Constructing membership functions using statistical data. *Fuzzy Sets and Systems*, 18 :1–13, 1986.

[20] T.F. Cox and M.A.A. Cox. Multidimensional scaling on a sphere. *Communications in Statistics*, 20 :2943–2953, 1991.

[21] T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman and Hall, London, 1994.

[22] R.N. Davé. Clustering relational data containing noise and outliers. In *FUZZ'IEEE 98*, pages 1411–1416, Anchorage, 1998.

[23] M. Delgado and S. Moral. On the concept of possibility-probability consistency. *Fuzzy Sets and Systems*, 21 :311–318, 1987.

[24] T. Denoeux. Modeling vague beliefs using fuzzy-valued belief structures. *Fuzzy Sets and Systems*, 116(2) :167–199, 2000.

[25] T. Denœux, M.-H. Masson, and P.-A. Hébert. Nonparametric rank-based statistics and significance tests for fuzzy data. *Fuzzy Set and Systems (To appear)*, 2005.

[26] T. Denoeux and M.H. Masson. Multidimensional scaling of interval-valued dissimilarity data. *Pattern Recognition Letters*, 21 :83–92, 2000.

[27] T. Denoeux and M.H. Masson. Principal component analysis of fuzzy data using autoassociative neural networks. *IEEE Transactions on Fuzzy Systems*, 12(3) :336–349, 2004.

[28] H. L'Hermier des Plantes. *Structuration des tableaux à trois indices de la statistique*. PhD thesis, Université de Montpellier, Montpellier, 1976.

[29] P. Diamond. Fuzzy least squares. *Information Science*, 46 :141–157, 1988.

[30] P. Diamond and H. Tanaka. Fuzzy regression analysis. In R. Slowinski, editor, *Fuzzy sets in decision analysis, operations research and statistics*, pages 349–387. Kluwer Academic Publishers, Boston, 1998.

[31] E. Diday and H. Bock. *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag, Heidelberg, 2000.

[32] D. Dubois, L. Foulloy, G. Mauris, and H. Prade. Probability-possibility transformations, triangular fuzzy sets and probabilistic inequalities. *Reliable Computing*, 10 :273–297, 2004.

[33] D. Dubois, H. T. Nguyen, and H. Prade. Possibility theory, probability and fuzzy sets : Misunderstandings, bridges and gaps. In D. Dubois and H. Prade, editors, *Fundamentals of Fuzzy sets*, pages 343–438. Kluwer Academic Publishers, Boston, 2000.

[34] D. Dubois and H. Prade. *Fuzzy Sets and Systems : Theory and Applications*. Academic Press, New York, 1980.

[35] D. Dubois and H. Prade. Addition of interactive fuzzy numbers. *IEEE Transactions on Automatic Control*, 26(4) :926–936, 1981.

[36] D. Dubois and H. Prade. Unfair coins and necessity measures, towards a possibilistic interpretation of histograms. *Fuzzy sets and Systems*, 10 :15–20, 1983.

[37] D. Dubois and H. Prade. A set-theoretic view of belief functions : logical operations and approximations by fuzzy sets. *International Journal of General Systems*, 12 :193–226, 1986.

[38] D. Dubois and H. Prade. *Possibility Theory : An approach to computerized processing of uncertainty*. Plenum Press, New-York, 1988.

[39] D. Dubois and H. Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Comput. Intell.*, 4 :244–264, 1988.

[40] D. Dubois and H. Prade. Consonant approximations of belief measures. *International Journal of Approximate Reasoning*, 4 :419–449, 1990.

[41] D. Dubois and H. Prade. Data fusion in robotics and machine intelligence. In M. Al Abidi et al, editor, *Combination of information in the framework of possibility theory.* Academic, New York, 1992.

[42] D. Dubois and H. Prade. La fusion d'informations imprécises. *Traitement du Signal*, 11(6) :447–458, 1994.

[43] D. Dubois and H. Prade. Fuzzy sets : history and basic notions. In D. Dubois and H. Prade, editors, *Fundamentals of Fuzzy sets*, pages 21–124. Kluwer Academic Publishers, Boston, 2000.

[44] D. Dubois, H. Prade, and S. Sandri. On possibility/probability transformations. In *Proceedings of the Fourth Int. Fuzzy Systems Association World Congress (IFSA'91), Brussels, Belgium*, pages 50–53, 1991.

[45] G. Eckman. Dimension of color vision. *Journal of Psychology*, 38 :367–474, 1954.

[46] P. Filzmoser and R. Viertl. Testing hypothesis with fuzzy data : the fuzzy p-value. *Metrika*, 59 :21–29, 2004.

[47] P.C. Fishburn. *Interval orders and interval graphs.* Wiley, New-York, 1985.

[48] S. Fitzpatrick and A. Scott. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82(399) :875–879, 1987.

[49] J. Gebhardt, M. A. Gil, and R. Kruse. Fuzzy set-theoretic methods in statistics. In R. Slowinski, editor, *Fuzzy sets in decision analysis, operations research and statistics*, pages 311–347. Kluwer Academic Publishers, Boston, 1998.

[50] T. Gerstenkorn and J. Manko. Correlation of intuitionistic fuzzy sets. *Fuzzy sets and systems*, 44 :39–43, 1991.

[51] L.A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7(2) :247–254, 1965.

[52] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer. Classification on pairwise proximity data. In M. Kearns, S. Solla, , and eds. D. Kohn, editors, *Advances in Neural Information Processing Systems 11*, pages 438–444. MIT Press, Cambridge, 1999.

[53] Thore Graepel, Ralf Herbrich, Bernhard Schölkopf, Alex Smola, Peter Bartlett, Klaus Robert-Maller, Klaus Obermayer, and Robert Williamson. Classification on proximity data with lp–machines. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, pages 304–309, 1999.

[54] V. Ha and P. Haddawy. Similarity of personal preferences : theoretical foundations and empirical analysis. *Artificial Intelligence*, 146(2) :149–173, 2003.

[55] R.A. Harshman and M.E. Lundy. Parafac : Parallel factor analysis. *Computational Statistics and Data Analysis*, 18 :39–72, 1994.

[56] R.J. Hathaway and J.C. Bezdek. Nerf c-means : Non-euclidean relational fuzzy clustering. *Pattern Recognition*, 27 :429–437, 1994.

[57] R.J. Hathaway, J.W. Davenport, and J.C. Bezdek. Relational duals of the c-means clustering algorithms. *Pattern Recognition*, 22(2) :205–211, 1989.

[58] P.-A. Hébert, M.-H. Masson, and T. Denœux. Fuzzy rank correlation between fuzzy numbers. In *Proceedings of the 10th IFSA World congress*, pages 224–227, Istanbul, Turkey, 2003.

[59] S. Heilpern. Representation and application of fuzzy numbers. *Fuzzy sets and systems*, 91 :259–268, 1997.

[60] C.E. Helm. A multidimensional ration scaling analysis of perceived color relations. *Journal of the Optical Society of America*, 54 :256–262, 1964.

[61] T. Hofmann and J. Buhmann. Multidimensional scaling and data clustering. In D. Touretzky G. Tesauro and eds. T. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 459–466. MIT Press, Cambridge, 1995.

[62] T. Hofmann and J. Buhmann. Pairwise data clustering by deterministic annealing. *Transactions on Pattern Analysis and Machine Intelligence*, 19(1) :1–14, 1997.

[63] H. Hotelling. Analysis of a complex of statistical variables with principal components. *Journal of Educational Psychology*, 24 :498–520, 1933.

[64] L. Kaufman and P. J. Rousseeuw. *Findinf groups in data*. Wiley, New-York, 1990.

[65] A. Kaufmann and M. M. Gupta. *Introduction to fuzzy arithmetic. Theory and applications*. International Thomson Computer Press, London, 1991.

[66] M.G. Kendall. *Rank correlation methods*. Hafner, New York, 1970.

[67] G.J. Klir. A principle of uncertainty and information invariance. *Int. J. of General Systems*, 17(2-3) :249–275, 1990.

[68] R. Kruse and K.D. Meyer. *Statistics with vague data*. Reidel, Dordrecht, 1987.

[69] J.B. Kruskal. Nonmetric multidimensional scaling : a numerical method. *Psychometrika*, 29 :115–129, 1964.

[70] V. Lasserre, G. Mauris, and L. Foulloy. A simple possibilistic modelisation of measurement uncertainty. In L.A. Zadeh Eds. B. Bouchon-Meunier, R.R. Yager, editor, *Uncertainty in Intelligent and Information Systems*, pages 58–69. World Scientific, 2000.

[71] S.T. Liu and C. Kao. Fuzzy measures for correlation coefficient of fuzzy numbers. *Fuzzy Sets and Systems*, 128 :267–275, 2002.

[72] W.L. May and W.D. Johnson. A SAS macro for constructing simultaneous confidence intervals for multinomial proportions. *Computer Methods and Programs in Biomedecine*, 53 :153–162, 1997.

[73] M. Ming, M. Friedman, and A. Kandel. General fuzzy least squares. *Fuzzy sets and systems*, 88 :107–118, 1997.

[74] M. Montenegro, M. R. Casals, M. A. Lubiano, , and M. A. Gil. Two-sample hypothesis tests of means of a fuzzy random variable. *Information Sciences*, 133 :89–100, 2001.

[75] S. Ovchinnikov. Fundamentals of fuzzy sets. In D. Dubois and H. Prade, editors, *An introduction to fuzzy relations*, pages 233–259. Kluwer, Boston, 2000.

[76] N. R. Pal, J. C. Bezdek, and R. Hemasinha. Uncertainty measures for evidential reasoning II : New measure of total uncertainty. *8*, pages 1–16, 1993.

[77] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 :559–572, 1901.

[78] W. Pedrycz and F. Gomide. *An introduction to fuzzy sets*. MIT Press, Cambridge, 1998.

[79] G. Pruesse and F. Ruskey. Generating linear extensions fast. *SIAM Journal on Computing*, 23(2) :373–386, 1994.

[80] C.P. Quesenberry and D.C. Hurst. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6(2) :191–195, 1964.

[81] M. Roubens. Pattern classfication problems and fuzzy sets. *Fuzzy sets and systems*, 1 :239–253, 1978.

[82] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing :*

Explorations in the microstructure of Cognition, volume 1, pages 318–362. MIT Press, Cambridge, 1988.

[83] J. W. Sammon. A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, 18 :401–409, 1969.

[84] G. Saporta. *Probabilités, analyse de données et Statistiques*. Technip, Paris, 1990.

[85] J.W. Scannell, C. Blakemore, and M.P. Young. Analysis of connectivity in the cat cerebral cortex. *Journal of Neuroscience*, 15(2) :1463–1483, 1995.

[86] S. Schiffman, M. Reynolds, and F. Young. *Introduction to Multidimensional Scaling : Theory, Methods and Applications*. Academic Press, New-York, 1981.

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

[88] C.P. Sison and J. Glaz. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90(429) :366–369, 1995.

[89] P. Smets. The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5) :447–458, 1990.

[90] P. Smets. Belief functions : the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9 :1–35, 1993.

[91] P. Smets. The axiomatic justification of the Transferable Belief Model. Technical Report TR/IRIDIA/95-8, IRIDIA, Bruxelles, 1995.

[92] P. Smets. Non standard probabilistic and non probabilistic representations of uncertainty. Technical Report TR/IRIDIA/95-2, IRIDIA, Bruxelles, 1995.

[93] P. Smets. Imperfect information : Imprecision - Uncertainty. In A. Motro and Ph. Smets, editors, *Uncertainty Management in Information Systems*, pages 225–254. Kluwer Academic Publishers, Dordrecht, 1997.

[94] P. Smets. The Transferable Belief Model for quantified belief representation. In D. M. Gabbay and P. Smets, editors, *Handbook of Defeasible reasoning and uncertainty management systems*, volume 1, pages 267–301. Kluwer Academic Publishers, Dordrecht, 1998.

[95] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66 :191–243, 1994.

[96] I. Spence. Proximity and preference : Problems in the multidimensional analysis of large data sets. In R.G. Golledge and J.N. Rayer, editors, *Incomplete experimental design for multidimensional scaling*, pages 29–46. MN : University of Minnesota Press, Minneapolis, 1982.

[97] L. Tsogo, M.H. Masson, and A. Bardot. Recovery of the metric structure of a pattern of points using minimal information. *IEEE Transactions on Systems, Man and Cybernetics A*, 31(1) :30–42, 2001.

[98] R. Viertl. *Statistical methods for non-precise data.* CRC Press, New-York, 1996.

[99] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[100] M.P. Windham. Numerical classification of proximity data with assignment measures. *Journal of classification*, 2 :157–172, 1985.

[101] R. R. Yager. On the normalization of fuzzy belief structure. *International Journal of Approximate Reasoning*, 14 :127–153, 1996.

[102] C. Yu. Correlation of fuzzy numbers. *Fuzzy sets and systems*, 55 :303–307, 1993.

[103] L. A. Zadeh. Fuzzy sets. *Inform. Control*, 8 :338–353, 1965.

[104] L. A. Zadeh. Similarity relations and fuzzy orderings. *Information Science*, 3 :177–200, 1971.

[105] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning (Part 1). *Information Sciences*, 8 :199–249, 1975.

[106] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1 :3–28, 1978.

[107] L. M. Zouhal and T. Denoeux. An evidence-theoretic k-nn rule with parameter optimization. *IEEE Transactions on Systems, Man and Cybernetics - Part C*, 28(2) :263–271, 1998.