Application de la théorie des croyances et des systèmes flous à l'estimation fonctionnelle en présence d'informations incertaines ou imprécises.

Simon Petit-Renaud

Thèse soutenue le 1er décembre 1999 devant le jury composé de :

Mme	BOUCHON-MEUNIER Bernadette	Rapporteur
MM.	DELECROIX Michel	Examinateur
	DENOEUX Thierry	Directeur de thèse
	DEVEUGHELE Stéphane	Examinateur
	GOVAERT Gérard	Président
	GRANDVALET Yves	Examinateur
	SMETS Philippe	Rapporteur

Université de Technologie de Compiègne
Laboratoire HeuDiaSyc-UMR CNRS 6599

A la mémoire d'Olivier,

à Amel, Fabienne et mes parents,

Résumé

L'estimation des relations de dépendance entre variables est généralement déterminée à partir de modèles probabilistes. Cependant, ces modèles sont souvent inadaptés aux données définies de façon imprécise ou lors de la prise en compte d'informations non numériques, comme le jugement d'un expert. La théorie des ensembles flous et la théorie des croyances permettent au contraire de tenir compte de ces imperfections.

Nous avons d'abord proposé un système neuro-flou pour la reconstruction de données manquantes. Le principe est d'utiliser une base de règles floues construites à partir des relations entre les composantes des vecteurs d'un ensemble d'apprentissage. Notre méthode permet d'estimer toutes les variables manquantes d'un vecteur dans un seul modèle, quel que soit le nombre de variables disponibles. Nous l'avons appliquée à des données environnementales, dans le cadre du projet européen EM^2S. Une comparaison avec certaines approches probabilistes a été étudiée.

Ensuite, nous avons proposé une méthode de régression généralisée basée sur la théorie des croyances. L'information apportée par chaque élément de l'ensemble d'apprentissage est représentée par une structure de croyance définie par la sortie associée au vecteur d'apprentissage et par la distance au vecteur étudié. Cette approche permet une caractérisation de différents types d'incertitudes sur la sortie. Pour optimiser les performances du modèle, un critère d'erreur entre deux structures de croyance a été défini, généralisant une distance classique entre intervalles réels. Afin de diminuer le temps de calcul pour l'obtention de la structure finale, deux types de méthodes ont été développés. L'un d'eux consiste simplement à résumer l'information par classification de l'ensemble d'apprentissage. L'autre repose sur l'approximation des structures de croyance par classification hiérarchique des éléments focaux ou par optimisation de critères d'information.

Mots-clés :

données manquantes, données imprécises, données incertaines, théorie des croyances, théorie de Dempster-Shafer, fonctions de croyance, approximation, systèmes flous, systèmes neuro-flous, régression non-paramétrique, mesures d'incertitude, fusion de données, surveillance de l'environnement.

Abstract

The estimation of dependence relationships between variables is generally performed using probabilistic models. However, these models are not adapted to imprecise data and they cannot easily take into account symbolic information such as experts opinions. On the contrary, fuzzy set theory and evidence theory allow to integrate these kinds of uncertainties.

First, we have proposed a neuro-fuzzy system for missing data reconstruction. The principle of this system is to use a collection of fuzzy rules defined from relationships between the components of training set vectors. In this method, all the missing variables of a vector are estimated using a single model, whatever the number of available variables. We have applied it to environmental data, in the framework of the European project EM^2S. A comparison with some probabilistic approaches has been studied.

Then, a generalized regression method based on evidence theory had been proposed. The information provided by each element of the training set is represented by a belief structure defined from the output associated to the training vector and the distance to the studied vector. This approach allows to characterize different kinds of uncertainties related to the output. In order to optimize the performances of the model, an error criterion between two belief structures has been defined. This criterion generalizes a classical distance between real numbers, intervals, and fuzzy numbers. Finally, we have focused on the problem of decreasing computation time to obtain the final structure. Two ways have been explored. The first one consists in summarizing the available information by making a classification of the training set. The principle of the second one is to approximate the belief structures by a hierarchical classification of the focal elements or by optimization of information criteria.

Keywords:

missing data, imprecise data, uncertain data, evidence theory, Dempster-Shafer theory, belief functions, approximation, fuzzy systems, neuro-fuzzy systems, non-parametric regression, uncertainty, data fusion, environmental monitoring.

Remerciements

Je tiens en premier lieu à exprimer ma profonde gratitude à mon directeur de thèse Thierry Denœux, Professeur à l'Université de Technologie de Compiègne (UTC). Etant de formation statistique, j'ai découvert grâce à lui et avec émerveillement une grande partie des théories et concepts que j'ai utilisés dans cette thèse, comme la théorie des croyances, la théorie des ensembles flous ou les réseaux de neurones. Durant ces trois années, il m'a prodigué des conseils toujours pertinents, aussi bien sur des aspects théoriques que pratiques. Sa grande rigueur m'a également permis de tempérer mon côté naturellement désorganisé.

Je suis très reconnaissant à Madame Bernadette Bouchon-Meunier, Directeur de Recherche CNRS au laboratoire d'Informatique de Paris 6, d'avoir accepté de rapporter ma thèse. Je remercie vivement Philippe Smets, Professeur à l'Université Libre de Belgique. Ses nombreux travaux sur la théorie des croyances ayant servi de base au développement d'une partie du mémoire, je suis d'autant plus heureux qu'il ait bien voulu être rapporteur. Les discussions fructueuses au cœur de la forêt de Compiègne, en plein air, resteront un souvenir fort agréable.

Je remercie beaucoup Gérard Govaert, Professeur à l'UTC, de m'avoir fait l'honneur de présider le jury. Je remercie également Michel Delecroix, Professeur à l'ENSAI de Rennes, d'avoir bien voulu participer à ce jury, bien que le sujet soit un peu éloigné de ses préoccupations. Les conseils précis d'Yves Grandvalet, chargé de recherche CNRS à l'UTC, en particulier en apprentissage statistique, m'ont éclairé sur bien des points de ma thèse. Sa disponibilité, l'intérêt qu'il porte à mon travail, et sa lecture très détaillée de mon mémoire ont été très stimulants.

Un grand merci également à Stéphane Deveughèle, Docteur-Ingénieur de recherche au Centre International des Techniques Informatiques (Lyonnaise-des-Eaux-Suez), pour ses précieux conseils et ses encouragements, en particulier pendant toute la première partie de ma thèse, qui s'est déroulée dans le cadre du projet européen EM^2S.

L'expérience du projet européen était très enrichissante, et je remercie tous ceux qui m'ont épaulé pendant les différentes étapes, en particulier Stéphane Canu, actuellement Professeur à l'INSA de Rouen, dont la bonne humeur communicative m'a par ailleurs souvent détendu, ainsi que Mylène Masson, Maître de Conférence à l'UTC.

Je remercie Nathalie Alexandre, Jacqueline Beusnel, Nathalie Laboureur et Dominique Porras, ainsi que Corinne Boscolo, Harry Claisse, Paul Crubillé et Jean-Claude Escande pour leur soutien logistique et technique.

Je remercie les collègues et amis, dont certains sont actuellement au bout du monde : Victor, Graciela, Alejandro, Valentina, Francisco (ah, les soirées mexicaines!), Skander, Mô, Manoocheer, Dimitrios, Léo, Emmanuel, Emmanuel, Isabelle, Fatima, Cyril, Dominique, Anbulagan, Nicolas, Pascale, Christophe, Nassim ... Toutes mes excuses à ceux ou celles que j'aurais malencontreusement oubliés.

Enfin, je remercie bien sûr mes parents, Fabienne et Amel pour m'avoir supporté pendant ces trois années.

Accessoirement, je salue au passage mes chers vélocypèdes et la région compiégnoise qui ont contribué à un rythme équilibré dans mon travail.

Table des matières

Introduction

L'augmentation continue de la capacité de stockage informatique des données a permis d'acquérir une quantité d'information de plus en plus importante. Les outils de collecte se sont spécialisés, les instruments de mesure se sont affinés, les données recueillies sont devenues plus complexes et porteuses d'une information plus riche. Le traitement de cette information a donné naissance à de nombreux modèles, en vue d'applications diverses. L'estimation des relations de dépendance [159] entre deux groupes de variables $\boldsymbol{x}$ et $\boldsymbol{y}$ d'un objet[1] constitue un des aspects essentiels de ces applications et couvre des disciplines diverses telles que l'estimation fonctionnelle [13], la théorie de l'apprentissage [15] ou l'identification de systèmes [143]. L'objectif commun est d'estimer les variables $\boldsymbol{y}$, sorties du système, par une « quantité raisonnable », en tenant compte des entrées $\boldsymbol{x}$ correspondantes, des *connaissances disponibles* sur le système et d'une base de vecteurs d'apprentissage $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

Estimation dans un environnement incertain

Dans une application réelle, les connaissances dont on dispose sont en général imparfaites. Ces imperfections peuvent prendre des formes très variées. Les informations sur le système peuvent être *peu nombreuses, parcellaires, contradictoires* et certaines variables pertinentes sont souvent ponctuellement *manquantes* ou *peu fiables*. Les causes de ces imperfections sont multiples. L'obtention même des informations constitue une des sources d'incertitude. L'acquisition des connaissances, au moyen de capteurs ou d'experts humains, est généralement sujette à des erreurs ou des imprécisions. La représentation, le codage des phénomènes observés, qu'ils soient numériques, linguistiques ou logiques entraînent nécessairement une perte d'information.

La terminologie adoptée par Dubois et Prade [40] permet de distinguer clairement les notions voisines d'incertitude et d'imprécision. Une information quelconque sur la caractéristique d'un objet, d'un individu ou d'un phénomène donné peut être représentée par deux éléments : une valeur ou un ensemble de valeurs associé à l'objet et la confiance dans l'information délivrée.

On peut ainsi considérer deux types fondamentaux d'imperfections (cf.[16, 83, 40]) :

- l'imprécision sur la valeur d'une donnée,
- l'incertitude de l'information.

1. On notera en gras les vecteurs multidimensionnels.

L'imprécision est donc relative au contenu de l'information, tandis que l'incertitude correspond au jugement, à la croyance en la véracité de cette information. Considérons la proposition suivante : « Il est certain qu'il fera chaud demain. ». Ici, la variable est la température, la valeur est représentée de manière imprécise par l'expression linguistique : « chaud ». En revanche, il n'y a aucune mise en doute de la proposition. Au contraire, la proposition « Il fera probablement 25^oC demain. » est incertaine, mais la valeur fournie est précise.

Imprécision des données

La connaissance que l'on possède des données $\boldsymbol{x}$ et $\boldsymbol{y}$ peut être très imparfaite et difficilement représentée directement sous forme de valeurs réelles. Si les informations sont délivrées par des experts, ils ne fourniront sans doute pas une valeur numérique précise, mais plutôt une quantité *approximative* (« environ 25^oC »), un intervalle de valeurs « plus de 20^oC » ou une expression appartenant au langage naturel (« chaud »). La théorie des ensembles flous [176], associée à la théorie des possibilités [179], offre un formalisme adapté à ce type d'imprécision.

Le cas particulier important de l'imprécision totale est celui de la donnée manquante. Le vecteur d'entrée $\boldsymbol{x}$ peut être incomplet. Un expert est sans opinion sur les valeurs de certaines de ses composantes. Si les valeurs sont acquises par des capteurs, ceux-ci sont en panne. Si les données sont collectées par sondage, il s'agit de « non-réponse » à certaines questions. Dans ces conditions, on peut envisager de reconstruire ces valeurs manquantes à l'aide de la base d'apprentissage avant d'estimer $\boldsymbol{y}$.

Incertitude d'une proposition

L'incertitude peut provenir du manque de fiabilité de la source d'information. Elle peut également être générée par le conflit entre plusieurs sources. Supposons que l'on dispose de deux capteurs de la variable « température ». Chacun des capteurs fournit une valeur numérique pour cette variable : 24^oC et 26^oC. La variable est donc entachée d'une incertitude entre deux valeurs conflictuelles.

Le concept d'incertitude, définie en tant qu'information déficiente, a été longtemps intimement lié, depuis le XVII^e siècle, à la théorie des probabilités. Depuis les années 1960, plusieurs théories mathématiques, dont la théorie des croyances [28, 130] et la théorie des probabilités imprécises [163], ont apporté un cadre plus souple et un formalisme plus général de représentation de l'incertitude. En particulier, l'ignorance, cas extrême de l'incertitude totale, est mieux prise en compte par ces deux théories.

Modélisation probabiliste

Les techniques statistiques d'estimation fonctionnelle supposent que l'ensemble d'apprentissage est composé de données $(\boldsymbol{x}, \boldsymbol{y})$ fiables, complètes, bien connues. Les modèles statistiques définissent des relations fonctionnelles *précises* entre les variables d'entrée et de sortie. L'erreur de mesure, la déviation entre la prédiction définie par le modèle et la véritable valeur est caractérisée par une variable aléatoire. En général, le principe consiste à sélectionner parmi une classe de fonctions de $\boldsymbol{x}$ du système considéré, celle qui « s'adapte le mieux » à $\boldsymbol{y}$ selon

un critère déterminé. Le choix du critère et de la classe conduisent à différentes méthodes d'approximation fonctionnelle.

Les modèles probabilistes sont capables de faire face à certaines perturbations ou erreurs qui entachent ponctuellement certaines caractéristiques. Par exemple, dans le cas de données qualifiées d' « aberrantes », il existe des estimateurs robustes [91], tenant compte de ces données qui s'écartent de la fraction dominante de l'échantillon.

Certaines méthodes, comme les arbres de régression [17] s'adaptent à l'existence de composantes manquantes d'un vecteur, mais elles sont plutôt rares.

Approche floue

La théorie des possibilités, jointe aux concepts de la théorie des ensembles flous, permet la manipulation d'objets et quantités aux contours imprécis, comme les intervalles ou les quantités floues. Deux types principaux de méthodes sont capables d'intégrer ces quantités dans des modèles de régression : il s'agit des systèmes flous [150, 78] et de la régression floue ou par intervalles [151, 39].

Le principe des systèmes flous est d'exploiter une base de règles du type « $Si \ldots Alors$ », définissant des relations imprécises entre les variables. Deux types principaux de modèles peuvent être utilisés. Dans le modèle de Mamdani-Zadeh [103], les entrées $\boldsymbol{x}$ et les sorties $\boldsymbol{y}$ peuvent être définies de façon imprécise. Dans le modèle de Takagi-Sugeno [150], seules les entrées sont imprécises.

Les méthodes de régression floue visent à chercher une relation fonctionnelle de type floue entre des variables, qui elles-mêmes peuvent être précises ou floues. Un premier type de méthodes, les techniques possibilistes, définit de nouveaux estimateurs en utilisant un formalisme emprunté à la théorie des possibilités. La plupart de ces méthodes [151] peuvent se ramener à un problème de programmation linéaire. L'avantage essentiel de ces méthodes est de fournir des informations pertinentes même pour un petit nombre de données. Une deuxième catégorie de méthodes est basée sur l'extension aux données imprécises de type intervalle ou flou, d'un critère d'erreur entre nombres réels [38, 39]. Le principal avantage de ces techniques par rapport aux précédentes est qu'il est possible d'estimer la précision des modèles.

Modélisation par la théorie des croyances

La théorie des croyances [130, 142] permet de combiner les connaissances diverses, ou la *croyance*, que l'on possède d'un phénomène à partir de différentes sources, de manière plus souple que le formalisme probabiliste. En particulier, certains types d'imperfections, comme l'incertitude totale, sont mal représentés par la théorie des probabilités. Dans le cas de données à la fois imprécises et multi-valuées, un formalisme très général, combinant la théorie de croyances et la théorie des ensembles flous, permet de représenter ces deux types d'incertitude.

La théorie des croyances a été appliquée par Denœux [31, 30] en discrimination. Mais peu de travaux ont été menés dans le cas où l'ensemble de référence est continu.

Contributions de la thèse

Dans ce mémoire, nous nous sommes particulièrement attachés à l'estimation de variables explicatives dans un contexte où les informations globales disponibles peuvent être imprécises, disparates ou incertaines. Cette thèse comporte deux parties distinctes.

La première, orientée sur un problème particulier d'estimation, le traitement de données manquantes, est motivée par la participation au projet européen EM^2S (*Environmental Monitoring and Management System*)[2]. Dans le cas où, ponctuellement, les entrées du système ne sont elles-mêmes pas disponibles ou mal spécifiées, nous avons proposé de reconstruire ces données manquantes. Nous avons proposé une méthode originale, basée sur un système neuro-flou. Le principe est d'utiliser une base de règles construites à partir des relations entre les composantes des vecteurs de l'ensemble d'apprentissage. Ce système peut se voir comme un réseau de neurones auto-associatif, c'est-à-dire dont les entrées et les sorties sont les mêmes vecteurs. L'intérêt essentiel de cette méthode est de pouvoir estimer toutes les variables manquantes d'un vecteur dans un même modèle, quel que soit le nombre de variables disponibles. Cependant, les systèmes flous ne permettent pas de traiter tous les types d'incertitude de façon satisfaisante.

Dans la deuxième partie de la thèse, nous avons proposé une nouvelle méthode de régression généralisée basée sur la théorie des croyances. Les fonctions de croyance permettent de définir des degrés de confiance sur les éléments de l'ensemble d'apprentissage. L'information apportée par chaque élément $\boldsymbol{x}_i$ est représentée sous la forme d'une structure de croyance qui repose sur deux composantes, la proximité de deux vecteurs observés $\boldsymbol{x}$ et $\boldsymbol{x}_i$, et l'information *a priori* fournie par le vecteur $\boldsymbol{x}_i$. Cette information *a priori* peut être un nombre réel, comme dans le cas probabiliste, un intervalle ou un ensemble flou, comme dans le cas des systèmes flous, ou une structure de croyance, éventuellement floue. La définition très générale de structures de croyance floues m_i permet de généraliser les résultats obtenus dans la première partie. Ainsi, cette approche répond à deux objectifs principaux :

- la gestion des données imprécises (les éléments focaux peuvent être des intervalles, voire des ensembles flous).
- la représentation claire des différents types d'imperfection dans un objectif ultérieur d'aide à la décision ou de fusion d'informations.

Un aspect essentiel de la méthode concerne l'identification du modèle. La sortie estimée correspondant au vecteur $\boldsymbol{x}$ est définie sous la forme d'une structure de croyance. Afin de permettre l'apprentissage du modèle, nous avons défini un critère d'erreur entre deux structures de croyance, généralisant le critère d'erreur quadratique classique, dans le cas où les données de l'ensemble d'apprentissage sont des nombres réels. Ce critère est construit comme l'extension d'une distance entre intervalles réels [183].

Un deuxième aspect largement développé concerne la diminution du temps de calcul pour l'obtention de la structure estimée. Différentes approches sont proposées. L'une d'elle, spécifique à notre méthode, repose sur l'approximation des structures de croyance par classification de leurs éléments focaux.

2. Projet Esprit 22442 EM^2S : « Environmental Monitoring and Management System ». Partenaires : Computas (Norvège), CNRS (France), Danfoss (Danemark), Hitec (Norvège), Suez-Lyonnaise-des-eaux (France), VKI (Danemark).

Enfin, pour chacun de nos deux modèles, nous nous sommes efforcés de comparer nos résultats avec ceux de méthodes de régression classiques.

Plan du mémoire

Ce mémoire est donc divisé en deux parties distinctes, l'une consacrée à la modélisation par les systèmes flous (chapitre 1 et 2), l'autre, à l'application de la théorie des croyances en régression (chapitre 3, 4 et 5).

Dans le chapitre 1, nous étudions en détail le mécanisme des systèmes flous. Nous tentons ensuite de clarifier les différentes catégories de systèmes neuro-flous. Le chapitre 2 est consacré à l'estimation de données manquantes à l'aide de notre méthode neuro-floue. La reconstruction des données réelles environnementales du projet EM^2S est proposée comme exemple d'application.

Le chapitre 3 est une introduction à la théorie des croyances. Nous développons particulièrement l'extension de la théorie aux ensembles de référence continus et à l'existence d'éléments focaux flous. Dans les chapitres 4 et 5, nous développons notre méthode de régression. Le chapitre 4, plus général, est consacré aux fondements de la méthode et à l'identification du modèle. Nous comparons notre méthode à certaines techniques de régression classiques ainsi qu'aux systèmes flous. Le chapitre 5 est plus particulièrement dédié aux aspects pratiques. Nous détaillons les diverses techniques de simplification envisagées et nous proposons quelques simulations et exemples réels afin d'illustrer les différents points de notre méthode.

En outre, dans les annexes A et B, nous avons proposé deux études bibliographiques portant sur les principales méthodes d'estimation fonctionnelle et de traitement de données manquantes.

Chapitre 1

Systèmes flous et neuro-flous

1.1 Introduction

Le manque de souplesse des techniques quantitatives traditionnelles dans la description de phénomènes a amené à définir une manière plus flexible de modéliser un système. Afin de modéliser le raisonnement humain à l'aide de valeurs linguistiques floues plutôt que réelles, Zadeh a proposé de développer une nouvelle classe de systèmes appelés *systèmes flous* [177]. Les systèmes flous étendent la définition des systèmes classiques en fournissant une représentation des éléments essentiels d'un système à l'aide de la théorie des ensembles flous. L'utilisation d'ensembles flous permet de représenter des informations imprécises, comme des valeurs approximatives (« environ 2 heures »), des limites mal définies (« l'appartement est grand ») ou des situations intermédiaires entre deux états (« il fait presque jour »). L'utilisation conjointe des méthodes connexionistes et floues dans des systèmes hybrides permet de tirer avantage des capacités des unes et des autres : faculté d'apprentissage des techniques neuronales, lisibilité et facilité de manipulation des objets dans les techniques floues. Dans ce chapitre, nous présentons successivement les systèmes flous et neuro-flous.

1.2 Systèmes d'inférence flous

1.2.1 Théorie des ensembles flous

Dans cette section, nous rappelons brièvement des notions et définitions que nous utiliserons par la suite.

Définitions de base

Soit $\mathcal{X}$ un ensemble de référence quelconque. Un sous-ensemble flou F de $\mathcal{X}$ est représenté par sa *fonction d'appartenance* μ_F, définie de $\mathcal{X}$ dans $[0,1]$, où $\mu_F(x)$ indique le degré d'appartenance de x à F. La notion d'ensemble flou généralise la définition de l'ensemble classique, dont la fonction d'appartenance est une fonction binaire, à valeurs dans $\{0,1\}$, dite fonction caractéristique ou indicatrice (cf. figure 1.1). On notera $\mathcal{F}(\mathcal{X})$ l'ensemble des sous-ensembles flous de $\mathcal{X}$. Dans la suite, on emploiera souvent la même notation pour l'ensemble et sa

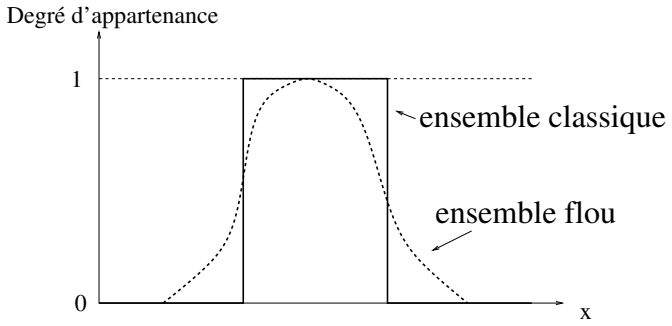


FIG. 1.1 – *Ensemble classique et ensemble flou.*

fonction d'appartenance et l'on notera $F(x) \triangleq \mu_F(x)$.

Un ensemble flou F est *normalisé* s'il existe au moins un élément $x \in \mathcal{X}$ tel que $F(x) = 1$. Le noyau de F, noté *Noyau(F)*, est l'ensemble des éléments dont le degré d'appartenance est 1 :

$$Noyau(F) \triangleq \{x \in \mathcal{X} | F(x) = 1\}. \tag{1.1}$$

La *hauteur* d'un ensemble flou F, notée $h(F)$, est la borne supérieure des degrés d'appartenance parmi les éléments de F. Ainsi,

$$h(F) = \sup_{x \in \mathcal{X}} F(x). \tag{1.2}$$

Le *support* de $F \in \mathcal{F}(X)$, noté $supp(F)$, est l'ensemble classique de $\mathcal{X}$ dont les éléments ont un degré d'appartenance à F non nul.

$$supp(F) = \{x \in \mathcal{X} | F(x) > 0\}. \tag{1.3}$$

L'α-coupe de F, notée F_α, $\alpha \in [0, 1]$, est le sous-ensemble classique de $\mathcal{X}$ constitué de tous les éléments de $\mathcal{X}$ pour lesquels $F(x) \geq \alpha$:

$$F_\alpha \triangleq \{x \in \mathcal{X} | F(x) \geq \alpha\}. \tag{1.4}$$

On définit de même l'α-coupe stricte $F_{\alpha+}$:

$$F_{\alpha+} \triangleq \{x \in \mathcal{X} | F(x) > \alpha\}. \tag{1.5}$$

Le noyau et le support sont donc des coupes particulières d'un ensemble flou : $Noyau(F) = F_1$ et $Supp(F) = F_{0+}$. Ces notions sont illustrées dans la figure 1.2.

Opérations sur les ensembles flous

La plupart des concepts définis sur les ensembles classiques peuvent s'étendre aux ensembles flous [176]. Les notions d'inclusion et d'égalité ainsi que les principales opérations binaires (intersection, union) et la complémentation se définissent de la manière suivante :

Inclusion. *Soient A et $B \in \mathcal{F}(\mathcal{X})$. L'ensemble flou A est inclu dans B ($A \subseteq B$) si :*

$$A(x) \leq B(x) \ \forall x \in \mathcal{X}.$$

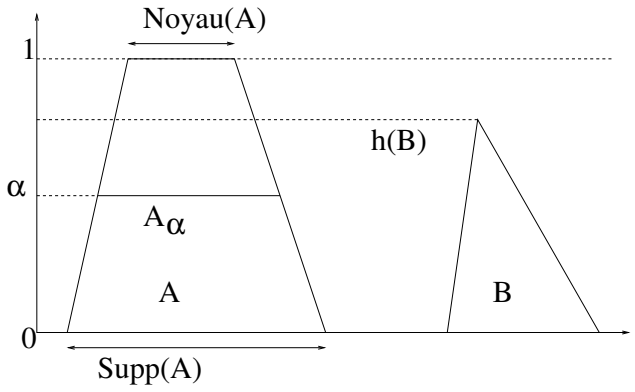


FIG. 1.2 – *Caractéristiques d'ensembles flous*

Egalité. *A et B sont égaux si et seulement si $A \subseteq B$* et *$B \subseteq A$.*

Complémentation. *L'ensemble flou $\overline{F}$, complément de F dans $\mathcal{X}$, est tel que :*

$$\forall x \in X, \quad \overline{F}(x) = 1 - F(x).$$

Intersection. *L'intersection $C = A \cap B$ de A et $B \in \mathcal{F}(\mathcal{X})$ est définie par :*

$$\forall x \in \mathcal{X}, \quad C(x) = \min(A(x), B(x)). \tag{1.6}$$

Union. *L'union $D = A \cup B$ de A et $B \in \mathcal{F}(\mathcal{X})$ est définie par :*

$$\forall x \in \mathcal{X}, \quad D(x) = \max(A(x), B(x)). \tag{1.7}$$

Toutes ces définitions généralisent les opérations sur les ensembles ordinaires, mais elles sont arbitraires et ne sont pas uniques. En particulier, on définit deux classes plus larges d'opérateurs, appelées *normes triangulaires* (ou *t-normes*, en abrégé), et *co-normes triangulaires* (ou *t-conormes*), généralisant respectivement l'intersection et l'union.

Une t-norme T est un opérateur binaire de $[0,1]\times[0,1]$ dans $[0,1]$ vérifiant quatre conditions : la commutativité, l'associativité, la monotonie et la neutralité de l'élément 1.

Une t-conorme S est un opérateur de $[0,1]\times[0,1]$ dans $[0,1]$ commutatif, associatif, monotone et possédant un élément neutre : 0.

D'autres propriétés peuvent être requises pour les t-normes et co-normes, parmi lesquelles l'*idempotence*, qui satisfait la condition: $T(a,a) = a$ pour tout $a \in [0,1]$. Les opérateurs *min* et *max* sont les seuls possédant ces cinq propriétés. On montre qu'il est toujours possible de construire une t-conorme S à partir d'une t-norme quelconque T de la façon suivante: $S(a,b) = 1 - T(1-a, 1-b)$. Pour cette même paire (S,T), on obtient la relation réciproque $T(a,b) = 1 - S(1-a, 1-b)$. S et T sont alors dites duales. Le tableau 1.1 présente les t-normes et conormes duales usuelles.

De nombreuses autres classes d'opérateurs binaires d'agrégation ont été définies dans la littérature, en particulier des opérateurs de type moyenne [168, 172].

Quantités floues

Dans la plupart des applications, les ensembles flous représentent des propriétés de variables à valeurs dans $\mathbb{R}$. L'univers de référence $\mathcal{X}$ est alors un sous-ensemble de $\mathbb{R}$ et les ensembles

t - norme	t - conorme
$\min(a,b)$	$\max(a,b)$
produit: $a * b$	somme probabiliste: $a+b-a\,b$
$\max(0,a+b-1)$	somme bornée: $\min(1,a+b)$

TAB. 1.1 – *t-normes et t-conormes duales usuelles.*

flous sont appelés *quantités floues*. Ils correspondent à l'idée de voisinage d'une valeur précise ou d'intervalles de valeurs aux bornes mal spécifiées.

Quantité floue. *Une quantité floue A est un sous-ensemble flou normalisé de* $\mathbb{R}$. *Une* valeur modale *de A est un élément x de* $\mathbb{R}$ *tel que* $A(x)=1$.

Convexité. *Une quantité floue A est* convexe si

$$\forall x,y\in\mathcal{X}\quad\forall\lambda\in[0,1],\quad A\left(\lambda x+(1-\lambda)y\right)\geq\min(A(x),A(y)).$$

Si A est convexe, pour tout $\alpha\in[0,1]$, l'α-coupe A_α est un intervalle fermé de $\mathbb{R}$ (μ_A est semi-continue supérieurement).

Intervalle flou *Une quantité floue A est un* intervalle flou *si A est convexe.*

Nombre flou. *Un intervalle flou A est un* nombre flou *s'il admet une* unique *valeur modale et si μ_A est bornée.*

Les intervalles flous les plus naturels sont définis par une fonction d'appartenance trapézoïdale, paramétrée par 4 constantes $\{a,b,c,d\}$, telles que $a<b\leq c<d$:

$$\mathrm{Trap}(x;a,b,c,d)=\max\left(\min\left(\frac{x-a}{b-a},1,\frac{d-x}{d-c}\right),0\right)\quad\forall x\in\mathbb{R}.$$

Les nombres flous peuvent par exemple être définis par une fonction d'appartenance triangulaire,

$$\mathrm{Tri}(x;a,b,c)=\max\left(\min\left(\frac{x-a}{b-a},\frac{c-x}{c-b}\right),0\right)\quad\forall x\in\mathbb{R},$$

avec $a<b<c$, ou gaussienne :

$$\mathrm{Gauss}(x;c,\sigma)=\exp\left\{-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2\right\}\quad\forall x\in\mathbb{R}.$$

La figure 1.3 illustre ces trois types de représentation.

Le concept de nombre flou $L-R$, introduit par Dubois et Prade [41], fournit une représentation simplifiée pour la manipulation d'intervalles flous. Soient L et R deux fonctions de $[0,\infty[$ dans $[0,1]$, dites de référence, semi-continues supérieurement, telles que $L(0)=R(0)=1$.

Un nombre flou A, noté $(c,a,b)_{LR}$ peut être défini par :

$$A(x)=\begin{cases}L\left(\dfrac{c-x}{a}\right), & \text{si } x\leq c\\ R\left(\dfrac{x-c}{b}\right), & \text{si } x\geq c\end{cases}\tag{1.8}$$

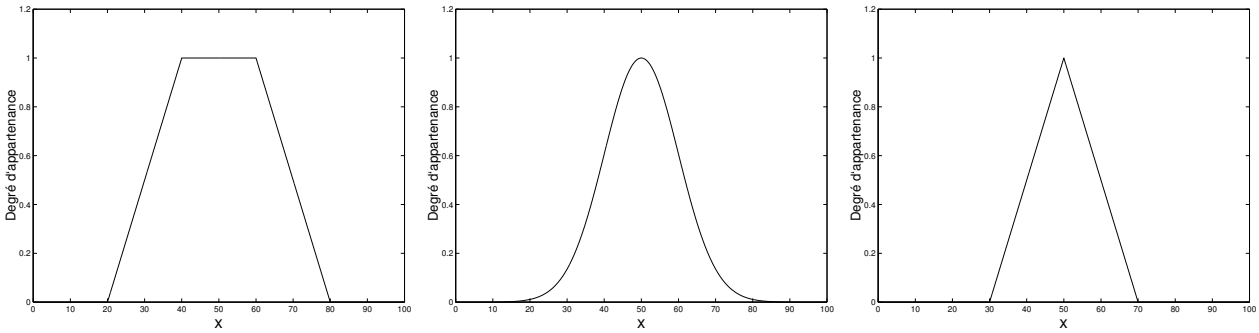


FIG. 1.3 – *Exemple de trois classes paramétrées de fonctions d'appartenance représentant l'expression linguistique « environ 50 ». A gauche, la fonction d'appartenance trapézoïdale Trap(x;20,40,60,80). Au milieu et à droite, respectivement, les fonctions d'appartenance gaussienne Gauss(x;50,10) et triangulaire Tri(x;30,50,70).*

où les paramètres c, a et b représentent respectivement le centre et l'imprécision à gauche et à droite de ce centre et où L et R sont des fonctions de référence. Les fonctions de référence les plus fréquemment employées, $x \longmapsto \exp -(|x|^p), p > 0$, et $x \longmapsto \max(0, 1 - |x|^p), p > 0$ donnent lieu respectivement aux nombres flous gaussiens et triangulaires.

Produit Cartésien et relations floues

La description d'un système fait généralement intervenir plusieurs ensembles de référence, relatifs à des variables différentes. Lorsque ces variables ne sont liées par aucune relation, les ensembles flous définis sur leurs ensembles de définition sont dits *non interactifs* [16].

Produit Cartésien *Soient $A_1, A_2, \ldots, A_r$, des ensembles flous respectifs des ensembles de référence $\mathcal{X}_1, \mathcal{X}_2, \ldots \mathcal{X}_r$. Leur produit Cartésien $A = A_1 \times A_2 \times \ldots \times A_r$, est un sous-ensemble flou de l'espace produit $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_r$, de fonction d'appartenance :*

$$A(x_1, x_2, \ldots, x_r) = A_1(x_1) \wedge A_2(x_2) \wedge \ldots \wedge A_r(x_r) \quad \forall (x_1, x_2, \ldots, x_r) \in \mathcal{X},$$

où $\wedge$ est une t-norme.

On peut ainsi définir des fonctions d'appartenance multidimensionnelles.

L'*extension cylindrique* est un cas particulier de produit Cartésien. Elle permet d'induire une connaissance sur toutes les composantes d'un espace à partir d'une information sur certaines d'entre elles seulement.

Extension cylindrique. *L'extension cylindrique de $A \in \mathcal{F}(\mathcal{X})$ à $\mathcal{X} \times \mathcal{Y}$ est l'ensemble flou $A' = A \times \mathcal{Y}$. Ainsi, $A'(x, y) = A(x)$ pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$.*

La figure 1.4 représente le produit Cartésien de deux ensembles flous gaussiens ainsi que l'extension cylindrique d'un nombre flou gaussien. L'opération réciproque est la *projection*.

Projection. *La projection de $A \in \mathcal{X} \times \mathcal{Y}$ sur $\mathcal{X}$, noté $Proj_{\mathcal{X}}(A)$ est l'ensemble flou $A_X \in \mathcal{X}$ défini par : $A_X(x) = \vee_{y \in \mathcal{Y}} A(x, y)$, où $\vee$ est une t-conorme.*

Relation floue. *Une relation floue R entre deux ensembles $\mathcal{X}$ et $\mathcal{Y}$ est définie comme un sous-ensemble flou du produit Cartésien $\mathcal{X} \times \mathcal{Y}$: $R \in \mathcal{F}(\mathcal{X} \times \mathcal{Y})$.*

Le produit Cartésien est donc une relation floue particulière. Divers types de relations floues

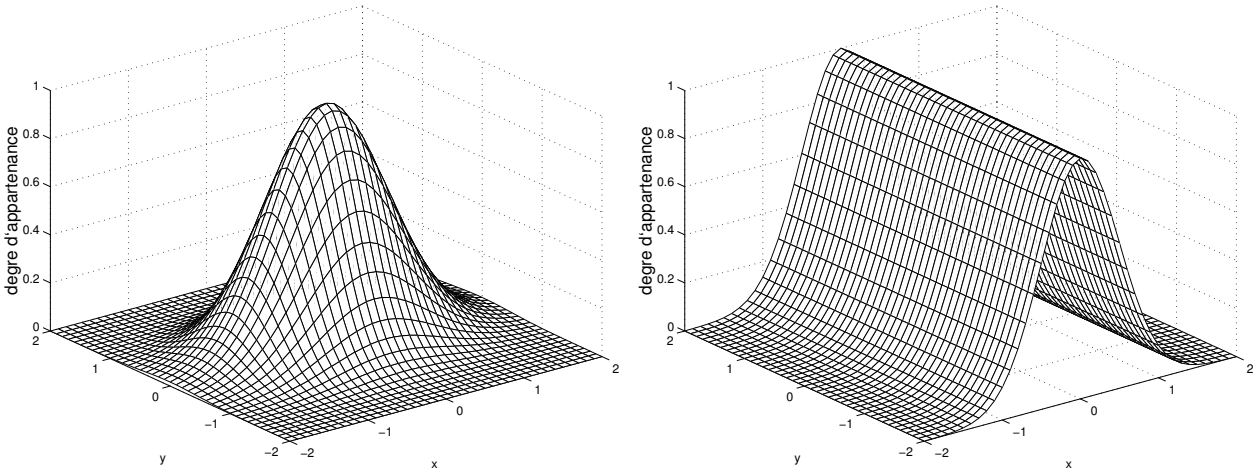


FIG. 1.4 – **A gauche** *: produit Cartésien de deux nombres flous gaussiens.* **A droite** *: extension cylindrique d'un nombre flou gaussien dans* $\mathbb{R}^2$*.*

peuvent être définis, entre deux univers $\mathcal{X}$ et $\mathcal{Y}$, comme par exemple :

- « x est proche de y » : $R(x,y) = \exp(-(x-y)^2), \quad \forall (x,y) \in \mathcal{X} = \mathcal{Y} = \mathbb{R}$,
- « si x est grand, y est petit »,

la deuxième expression étant à la base des systèmes flous.

Les relations floues étant des cas particuliers d'ensembles flous, toutes les propriétés et définitions concernant les ensembles flous leur sont donc applicables. Ainsi, on peut définir la hauteur, le noyau ou le support d'une relation floue.

Principe d'extension et arithmétique floue

Le principe d'extension définit des relations fonctionnelles entre les ensembles flous. Soit $f : \mathcal{X} \longmapsto \mathcal{Y}$ une application quelconque. Le principe d'extension élargit la définition de f aux ensembles flous. On définit ainsi la fonction $\phi : \mathcal{F}(\mathcal{X}) \longmapsto \mathcal{F}(\mathcal{Y})$, telle que, pour tout $A = A_1 \times \ldots \times A_r \in \mathcal{X}, B = \phi(A)$, avec

$$\forall y \in \mathcal{Y}, \quad B(y) = \begin{cases} \sup_{\{x \in \mathcal{X} | f(x) = y\}} A(x) & \text{si } f^{-1}(y) \neq \emptyset \\ 0 & \text{sinon.} \end{cases} \tag{1.9}$$

où f^{-1} est la fonction réciproque de f. Si f est bijective, la détermination de B devient simplement : $B(y) = A(f^{-1}(y)) \quad \forall y \in \mathcal{Y}$.

En particulier, si ∇ est une opération arithmétique binaire, $C = A \nabla B, A, B \in \mathcal{X}$, est défini par :

$$C(z) = \sup_{(x,y) | x \nabla y = z} A(x) \wedge B(y).$$

L'addition et la soustraction de nombres flous de même type LL sont également des nombres flous de type LL [41]. La multiplication et l'addition de deux nombres triangulaires est illustrée en figure 1.5.

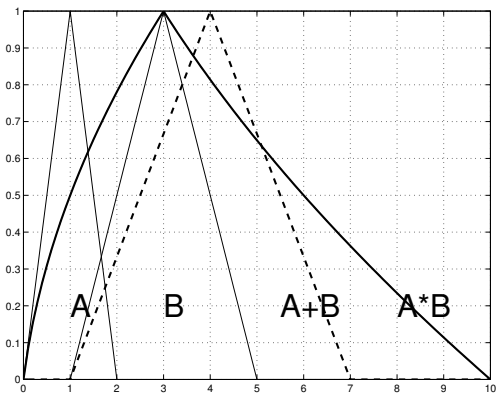


FIG. 1.5 – *Addition (–) et multiplication (-) de 2 nombres flous triangulaires : A = Tri(x;0,1,2) et B=Tri(x;1,3,5).*

1.2.2 Raisonnement approximatif

La théorie des ensembles flous permet la manipulation de concepts vagues ou imprécis à travers la théorie du *raisonnement approximatif* ou *raisonnement flou.*

Variable linguistique

L'un des concepts fondamentaux du raisonnement flou est celui de la *variable linguistique*, qui est une variable dont les valeurs sont des termes ou expressions appartenant au langage naturel, par essence imprécis, et sont représentés par des ensembles flous. Formellement, selon Bouchon-Meunier [16], une variable linguistique peut être décrite par un triplet $(x, \mathcal{X}, \mathcal{A}_x)$, où x est le nom de la variable, $\mathcal{X}$ son domaine de référence et $\mathcal{A}_x$, un ensemble de *valeurs* linguistiques pouvant être affectées à x, chacune d'entre elles étant caractérisée par un ensemble flou de $\mathcal{X}$.

La figure 1.6 présente un exemple de variable linguistique. La variable x est la température de l'eau d'un lac.

Proposition floue

L'affectation d'une valeur particulière $F \in \mathcal{A}_x$ à une variable linguistique x est réalisée à l'aide d'une *proposition floue élémentaire*, en utilisant par convention la syntaxe « x est A ».

On peut combiner plusieurs propositions floues élémentaires à l'aide d'opérateurs binaires logiques de conjonction, disjonction ou négation, et former ainsi des propositions plus complexes, comme

$$(x_1 \text{ est } A) \text{ ou } \big((x_2 \text{ est } B) \text{ et } (x_3 \text{ est } C)\big),$$

où A, B, C sont trois ensembles flous définis sur les espaces de référence respectifs de trois variables x_1, x_2, x_3.

La traduction ensembliste d'une proposition floue peut être réalisée en utilisant les opérateurs flous définis précédemment en section 1.2.1. La conjonction de deux propositions « x est A » et « y est B » définit la relation floue $C = A \times B$ entre les deux variables x et y. De façon similaire, la disjonction de ces deux propositions produit l'ensemble flou D tel que

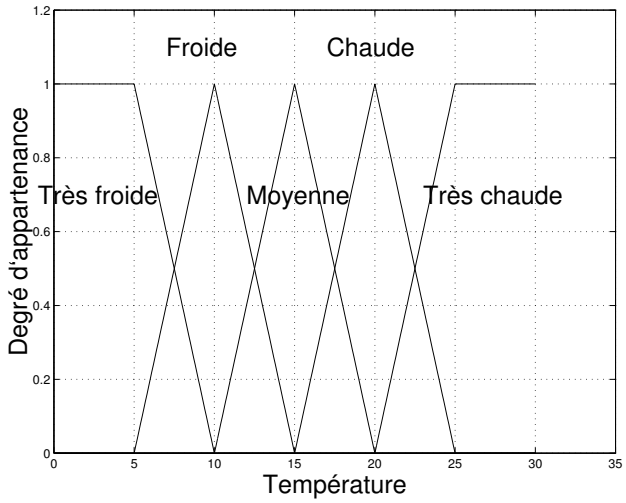


FIG. 1.6 – *Exemple de variable linguistique. La température de l'eau d'un lac est décrite (subjectivement) par 5 valeurs linguistiques {très froide, froide, moyenne, chaude, très chaude} dont on a représenté les fonctions d'appartenance.*

$D(x,y) = A(x) \vee B(y)$. On obtient donc une relation floue globale entre les différentes variables intervenant dans la proposition.

Dans l'exemple précédent, l'ensemble flou résultant R est défini par :

$$R(x_1, x_2, x_3) = A(x_1) \vee (B(x_2) \wedge C(x_3)).$$

Raisonnement flou

Le raisonnement flou [178] est une procédure d'inférence qui permet de déduire certaines conclusions à partir de relations fonctionnelles imprécises ou floues. Ce mécanisme constitue la base de la modélisation par les systèmes flous.

Le raisonnement flou généralise le raisonnement élémentaire suivant, issu de la définition habituelle d'une fonction. Soit une fonction f définissant une relation précise « $y = f(x)$ » entre deux variables classiques $x \in \mathcal{X}$ et $y \in \mathcal{Y}$. Pour une valeur particulière de x égale à a, la relation f nous permet de conclure que $y = b = f(a)$. Si la connaissance de f et de a est imprécise, le raisonnement n'est plus immédiat. Supposons que la valeur soit définie par un intervalle A et que la fonction devienne une courbe multivaluée par intervalles F (cf. figure 1.7). On construit alors l'extension cylindrique A' de A sur $\mathcal{X} \times \mathcal{Y}$, puis on cherche l'intersection G de A' avec la courbe F. En projetant G sur $\mathcal{Y}$, on obtient un intervalle B.

On peut étendre cette démarche aux ensembles flous. Soit R une relation floue définie sur $\mathcal{X} \times \mathcal{Y}$ et A un sous-ensemble flou quelconque de $\mathcal{X}$. L'objectif est de déterminer le sous-ensemble flou B de $\mathcal{Y}$, connaissant l'ensemble A, à travers la relation floue R. Pour cela, on procède aux mêmes étapes que précédemment. On construit l'extension cylindrique de A, $A' = A \times \mathcal{Y}$. Par définition :

$$A'(x,y) = A(x).$$

On définit ensuite l'intersection G de A' et de R :

$$G(x,y) = A'(x,y) \wedge R(x,y).$$

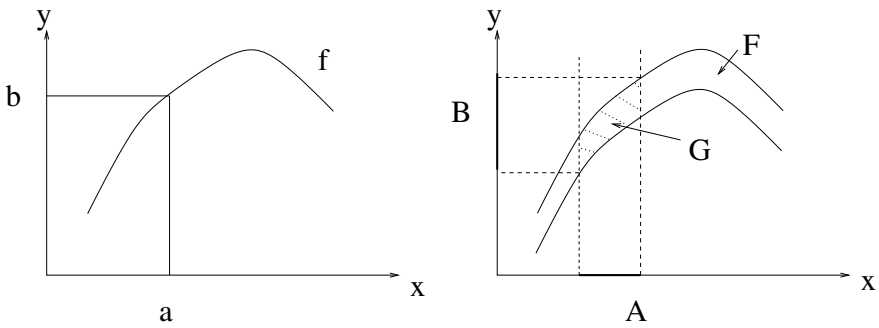


FIG. 1.7 – *Déduction de la valeur de sortie d'une fonction pour une entrée selon deux cas différents. A gauche, l'entrée a et la fonction f sont précises. La sortie induite est donc précise : $b = f(a)$. A droite, L'entrée est un intervalle A et la fonction F est à valeur dans un intervalle. La sortie B est un intervalle, déterminé par la projection de G sur l'axe des ordonnées, où G est défini comme l'intersection entre l'extension cylindrique de A au plan des deux variables et la courbe F.*

Enfin, on projette G sur $\mathcal{Y}$ pour obtenir la conclusion B. Ainsi,

$$B(y) = \vee_x (A(x) \wedge R(x, y)). \tag{1.10}$$

Cette formule est connue sous le nom de *règle d'inférence compositionnelle* et on note

$$B \triangleq A \circ R. \tag{1.11}$$

La figure 1.8 présente un exemple bidimensionnel de cette règle d'inférence.

On peut remarquer que si R est défini comme le produit Cartésien $A \times C$, alors $A \circ (A \times C) = C$.

L'application du raisonnement flou aux règles floues est l'élément clé des systèmes d'inférence flous que nous allons maintenant aborder.

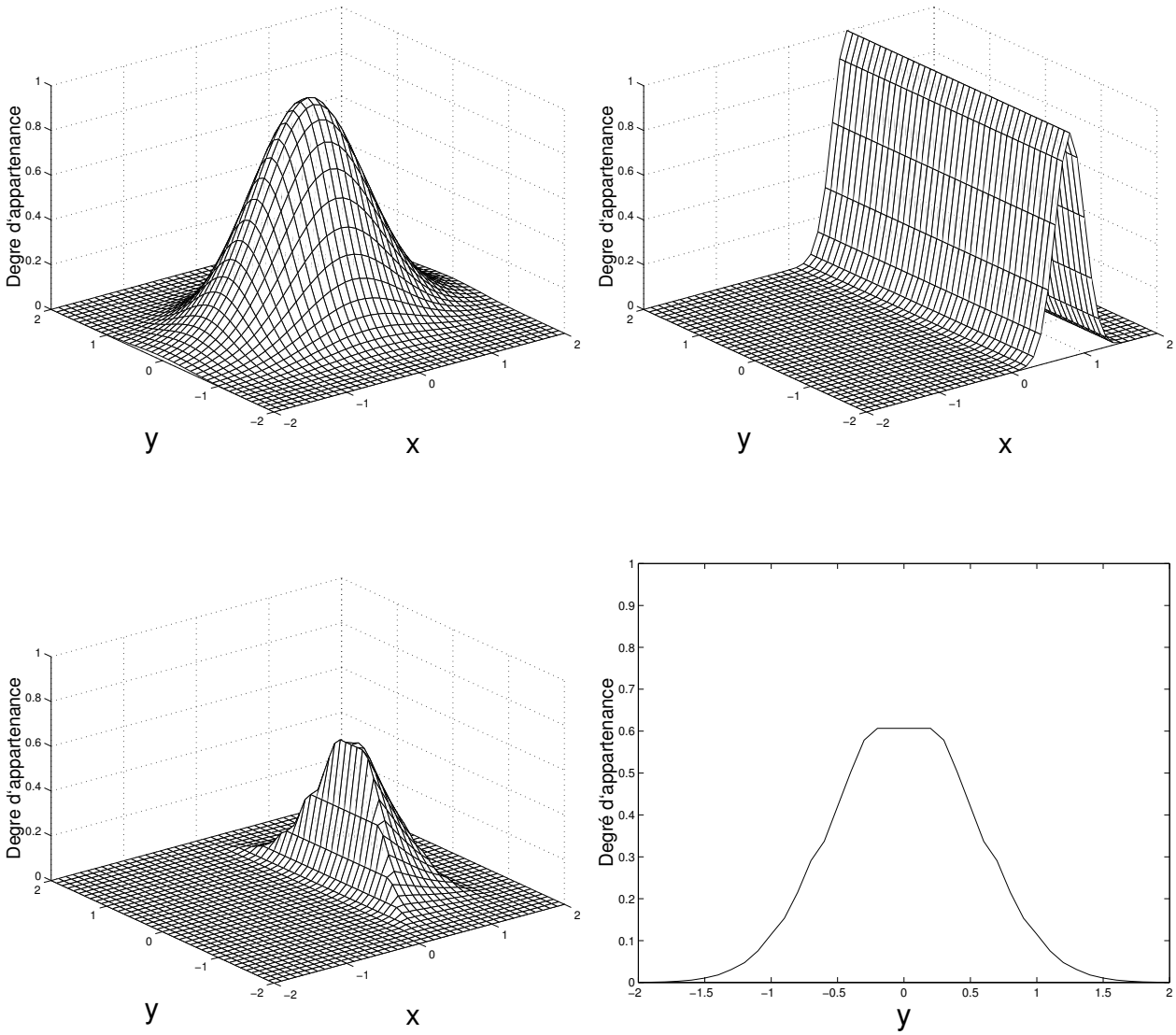


FIG. 1.8 – *Les étapes de la règle d'inférence compositionnelle. En haut, à gauche : la relation floue R, définie par le produit de deux fonctions d'appartenance gaussiennes Gauss(x;0,0.5) et Gauss*$(y;0,0.7)$*. A droite : l'extension cylindrique A′ de l'ensemble flou A, défini par la fonction Gauss(x;0.8,0.2). En bas, à gauche, l'intersection G de A′ et R. A droite, l'ensemble résultant B.*

1.2.3 Système d'inférence flou

L'une des directions principales dans la théorie des systèmes flous est l'approche linguistique. Celle-ci a été initialement introduite par Zadeh [177], puis développée par de nombreux auteurs [149, 76, 155, 114]. Un *modèle linguistique* ou *système d'inférence flou*, est un système décrit à l'aide de règles du type $SI - ALORS$ utilisant des propositions floues. C'est un système à base de connaissance qui contient des informations vagues, provenant de l'observation de phénomènes réels. Les systèmes flous ont été appliqués à des domaines très variés, comme le contrôle flou, la classification, la robotique, la reconnaissance des formes ou les séries temporelles, sous des appellations différentes : les *mémoires associatives floues* [87], les modèles flous [150], les *systèmes experts flous* [79] ou les *systèmes à base de règles floues.*

Dans l'optique de l'estimation fonctionnelle, nous considérerons pour l'instant que les vecteurs d'entrée du système sont multidimensionnels et la sortie monodimensionnelle (système MISO), sans perte de généralité, puisqu'un système à plusieurs sorties peut toujours se décomposer en un ensemble de systèmes à sortie unique. Soient $\boldsymbol{x} = \{x_1, \ldots, x_r\}$ les variables d'entrée du système appartenant aux ensembles de référence $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_r$ et y, la variable de sortie appartenant à l'espace $\mathcal{Y}$.

Un système d'inférence flou est essentiellement composé de deux blocs fonctionnels différents :

- une base de connaissance constituée d'une base de règles floues et d'une base de données définissant les fonctions d'appartenance des ensembles flous ;
- un mécanisme d'inférence flou [92] qui détermine la sortie du système sous la forme d'un ensemble flou.

On peut y ajouter deux blocs optionnels (cf. figure 1.9) :

- une interface de fuzzification, qui transforme des entrées ponctuelles en ensembles flous, le cas échéant ;
- une interface de défuzzification, qui transforme la sortie floue en sortie ponctuelle.

Les deux premiers blocs définissent des relations entre les entrées et sorties floues du système.

Les entrées du système $\{x_1, \ldots, x_r\}$ peuvent être fixes ou floues. La fuzzification n'est qu'une étape technique permettant d'introduire des entrées observées fixes dans le mécanisme de raisonnement proprement dit en convertissant artificiellement ces entrées en ensembles flous. Si ces données sont déjà définies par des ensembles flous, cette étape est donc inutile. Dans le cas contraire, la fuzzification dite « singleton » consiste simplement à transformer un *élément* fixe x_k de $\mathcal{X}_k$ en un *sous-ensemble classique* $\{x_k\}$ de $\mathcal{X}_k$, noté $\widetilde{x}_k$. C'est l'équivalent d'une loi de Dirac en théorie des probabilités.

Ce type de fuzzification ne fait pas intervenir l'imprécision éventuelle des données. Une autre approche, la fuzzification « non singleton », intègre l'imprécision ou le bruit dû à l'observation, en représentant l'entrée x_k sous la forme d'un ensemble flou quelconque $\widetilde{x_k}$, comme par exemple un nombre flou gaussien centré en x_k [1].

Nous allons maintenant détailler les autres composantes du système.

Base de règles floues

Une *règle linguistique* ou *règle floue* est une proposition floue de la forme « si p alors q », utilisant une implication entre deux propositions floues quelconques p et q.

Un système flou est basé sur un ensemble de règles floues r_i qui définissent des relations floues entre les variables :

$$r_i : \quad \text{SI} \quad (x_1 \text{ est } B_{i1}) \ldots \text{et} \ldots (x_r \text{ est } B_{ir}) \quad \text{ALORS} \quad (y \text{ est } D_i), \tag{1.12}$$

1. Cette procédure présente des analogies intéressantes avec des méthodes bayésiennes où l'on définit une distribution de probabilité *a priori*. Ici, cela revient à définir une distribution de possibilité *a priori*. Ce qui peut être contestable, comme dans le cas bayésien, c'est de définir *arbitrairement* ces distributions *a priori*.

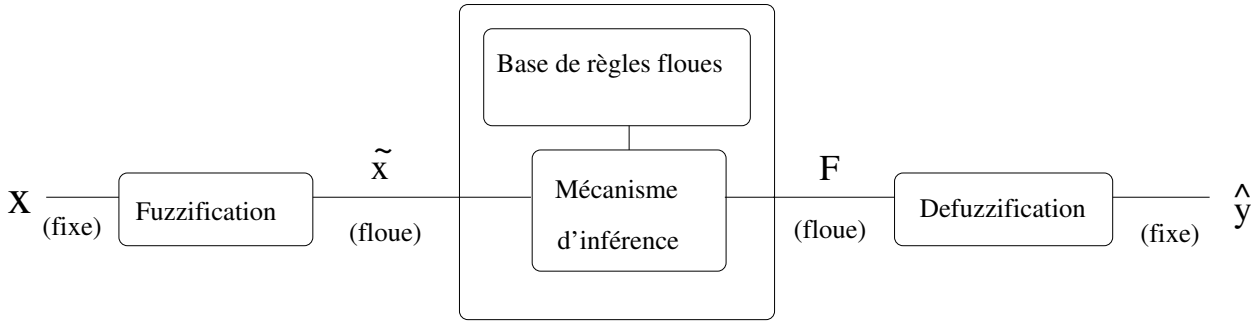


FIG. 1.9 – *Architecture d'un système flou. Les éléments essentiels sont la base de règles et le mécanisme d'inférence, qui permettent de déterminer la sortie sous forme d'un ensemble flou. L'interface de fuzzification n'est nécessaire que si les entrées sont fixes. L'interface de défuzzification permet de définir une sortie fixe.*

où $B_{i1}, \ldots, B_{ir}$ et D_i sont des valeurs linguistiques de $x_1, \ldots, x_r$ et y respectivement. Ce sont des ensembles flous dont les fonctions d'appartenance définissent les paramètres du modèle. Pour chaque x_k, B_{ik} est l'une quelconque des n_k valeurs linguistiques différentes caractérisant la variable et D_i est une des J valeurs linguistiques possibles de y. Dans le cas général, les n_k ne sont pas forcément égaux. Le nombre de combinaison total de règles est donc au maximum de $I = J \prod_{k=1}^{r} n_k$. On peut réécrire le système précédent de manière condensée :

$$r_i : \quad \text{SI} \quad \boldsymbol{x} \text{ est } B_i \text{ ALORS} \quad y \text{ est } D_i, \tag{1.13}$$

avec $B_i = B_{i1} \times \ldots B_{ir}$, le produit Cartésien des ensembles flous monodimensionnels. La proposition « $\boldsymbol{x}$ est B_i » est appelée prémisse ou partie antécédente et la proposition « y est D_i » est la partie conséquente de la règle floue r_i. La règle floue étant elle-même une proposition, on peut lui associer une relation floue R_i.

A chaque règle r_i, il est possible d'associer un niveau de confiance $p_i \in [0, 1]$, qui peut être défini par un expert. Si $p_i = 0$, la règle n'est pas pertinente et peut être supprimée. Afin d'alléger les notations, nous ne tiendrons pas compte des niveaux de confiance dans le reste de ce chapitre.

Généralement, le modèle défini par l'équation (1.13) peut se simplifier car bon nombre de règles sont impossibles ou peu probables et l'abondance de règles rend le système difficilement exploitable. Dans la littérature, le nombre de règles est souvent restreint. Le cas particulier le plus courant est celui des *règles binaires* : toutes les variables x_k et y ont le même nombre I de valeurs linguistiques et chaque valeur linguistique B_i ou D_i est employée pour une seule règle. On obtient donc uniquement I règles. La figure 1.10 illustre le cas général et le cas des règles binaires.

Les exemples de propositions modélisées par des règles floues abondent dans le langage courant :

- Si la pression est élevée, alors le volume est faible.
- Si la pente est raide, alors la vitesse est faible.

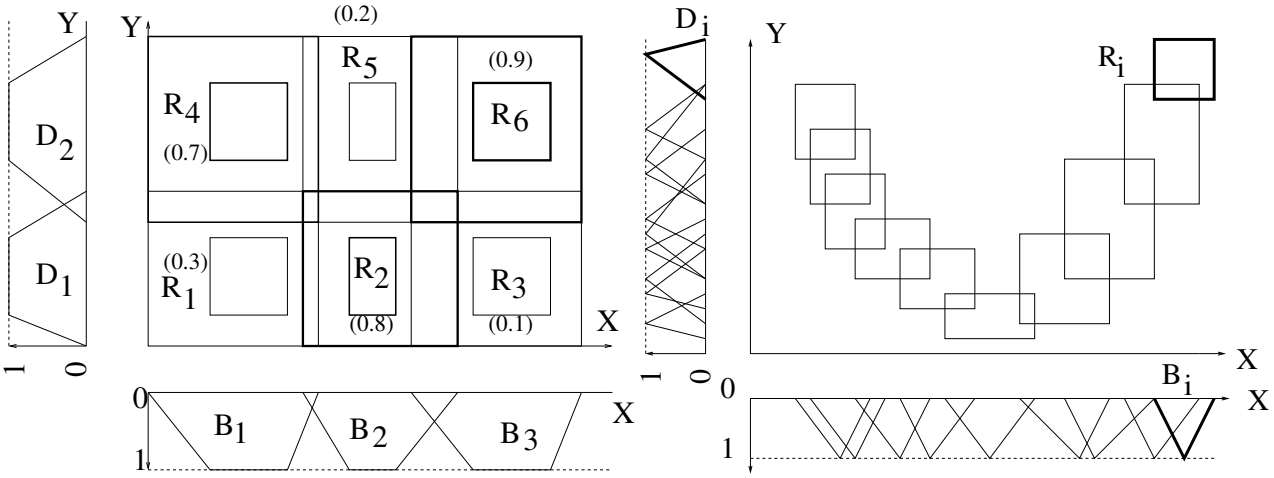


FIG. 1.10 – *Règles floues. A gauche, cas général. Les variables x et y sont décrites respectivement par 3 et 2 valeurs linguistiques, définissant 6 règles et donc 6 relations floues. Les facteurs de confiance p_i, entre parenthèses, déterminent l'influence de R_i. A droite, cas de 9 règles binaires. Les règles floues définissent des relations locales imprécises entre les variables.*

Fonctions d'appartenance.

Dans la base de connaissance, le choix des fonctions d'appartenance des ensembles flous dépend de l'objectif désiré. Les deux types les plus populaires sont les B-splines et les gaussiennes. Les B-splines [14, 19] sont une classe particulière de polynômes locaux par morceaux définis par des nœuds, reliés entre eux par des relations de récurrence et présentant un certain nombre de propriétés intéressantes. Les fonctions d'appartenance gaussiennes sont souvent utilisées pour l'expression simple de la fonction multivariée :

$$B_i(\boldsymbol{x}) = \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{c}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{c}_i)\right\}, \tag{1.14}$$

où $\boldsymbol{\Sigma}_i$ est une matrice définie positive diagonale (r, r) et $\boldsymbol{c}_i \in \mathcal{X}$. La fonction B_i est une fonction de base radiale et peut être interprétée comme une mesure de similarité entre $\boldsymbol{x}$ et $\boldsymbol{c}_i$. Un avantage également appréciable est celui de la couverture totale du domaine $\mathcal{X}$, car $supp(B_i) = \mathcal{X}$, ce qui n'est pas le cas des fonctions triangulaires, par exemple. En effet, dans le cas contraire, pour un $\boldsymbol{x}$ particulier, atypique, on peut avoir $B_i(\boldsymbol{x}) = 0$ pour tout i, ce qui pose des problèmes lors de la défuzzification notamment (cf. section 1.2.3).

Mécanisme d'inférence flou

Le mécanisme d'inférence flou est un mécanisme de décision, généralisant le *modus ponens* de la logique classique, qui utilise les règles de la base afin de déterminer les sorties floues F correspondant aux entrées floues ou fuzzifiées $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_r)$.

A chaque règle r_i on peut associer une relation floue R_i définie sur le produit Cartésien $\mathcal{X} \times \mathcal{Y}$. Il existe environ une quarantaine de façons différentes d'interpréter l'opération d'*implication* [92]. Si nous utilisons les plus fréquemment utilisées, celle de Mamdani [103], où l'implication s'interprète comme une opération de conjonction, nous obtenons :

$$R_i(\boldsymbol{x}, y) = B_i(\boldsymbol{x}) \wedge D_i(y), \tag{1.15}$$

avec $B_i(\boldsymbol{x}) = \bigwedge_k B_{ik}(x_k)$. Les opérations floues d'intersection (« et ») et d'implication (« si...alors ») sont représentées par la t-norme $\wedge$.

Les relations floues sont agrégées par un opérateur d'union $\vee$, produisant la relation globale R du système :

$$R(\boldsymbol{x}, y) = \bigvee_{i=1}^{I} R_i(\boldsymbol{x}, y). \tag{1.16}$$

Le problème d'inférence que l'on se pose peut se résumer de la façon suivante :

Fait observé :	$\boldsymbol{x}$ est $\widetilde{\boldsymbol{x}}$
Règles floues r_i :	si $\boldsymbol{x}'$ est B_i alors y' est D_i
Conclusion :	y est F

Connaissant $\boldsymbol{x}$ et les règles de la base, comment peut-on en déduire la sortie F ? Selon la théorie du raisonnement approximatif (cf. 1.2.2), la sortie F correspondant à l'entrée fuzzifiée $\widetilde{\boldsymbol{x}}$ est obtenue par la règle d'inférence compositionnelle (cf. équation 1.10 et 1.11). Soit

$$F = \widetilde{\boldsymbol{x}} \circ R.$$

Nous rappelons qu'il s'agit de la projection sur $\mathcal{Y}$ de l'ensemble flou G défini par l'intersection entre R et l'extension cylindrique de $\widetilde{\boldsymbol{x}}$ à $\mathcal{X} \times \mathcal{Y}$.

A partir des équations (1.10) et (1.16), nous obtenons :

$$\begin{aligned} \forall y \in \mathcal{Y} \quad F(y) &= \bigvee_{u \in \mathcal{X}} \widetilde{\boldsymbol{x}}(\boldsymbol{u}) \wedge R(\boldsymbol{u}, y) \\ &= \bigvee_{u \in \mathcal{X}} \left[\widetilde{\boldsymbol{x}}(\boldsymbol{u}) \wedge \bigvee_{i=1}^{I} R_i(\boldsymbol{u}, y) \right]. \end{aligned} \tag{1.17}$$

L'opérateur $\circ$ étant distributif par rapport à $\vee$, nous en déduisons, successivement, d'après l'équation (1.15) :

$$\begin{aligned} \forall y \in \mathcal{Y} \quad F(y) &= \bigvee_{i=1}^{I} \bigvee_{u} [\widetilde{\boldsymbol{x}}(\boldsymbol{u}) \wedge R_i(\boldsymbol{u}, y)] \\ &= \bigvee_{i=1}^{I} \bigvee_{u_1, \ldots, u_r} \left[\bigwedge_{k=1}^{r} \widetilde{\boldsymbol{x}}_k(u_k) \right] \wedge \left[\bigwedge_{k=1}^{r} B_{ik}(u_k) \wedge D_i(y) \right] \end{aligned} \tag{1.18}$$

L'opérateur $\wedge$ étant commutatif, nous avons :

$$\forall y \in \mathcal{Y} \quad F(y) = \bigvee_{i=1}^{I} \left\{ \bigwedge_{k=1}^{r} \left[\bigvee_{u_k} (B_{ik}(u_k) \wedge \tilde{x}_k(u_k)) \right] \wedge D_i(y) \right\} \tag{1.19}$$

$$\text{Ainsi,} \quad F(y) = \bigvee_{i=1}^{I} \tau_i \wedge D_i(y), \tag{1.20}$$

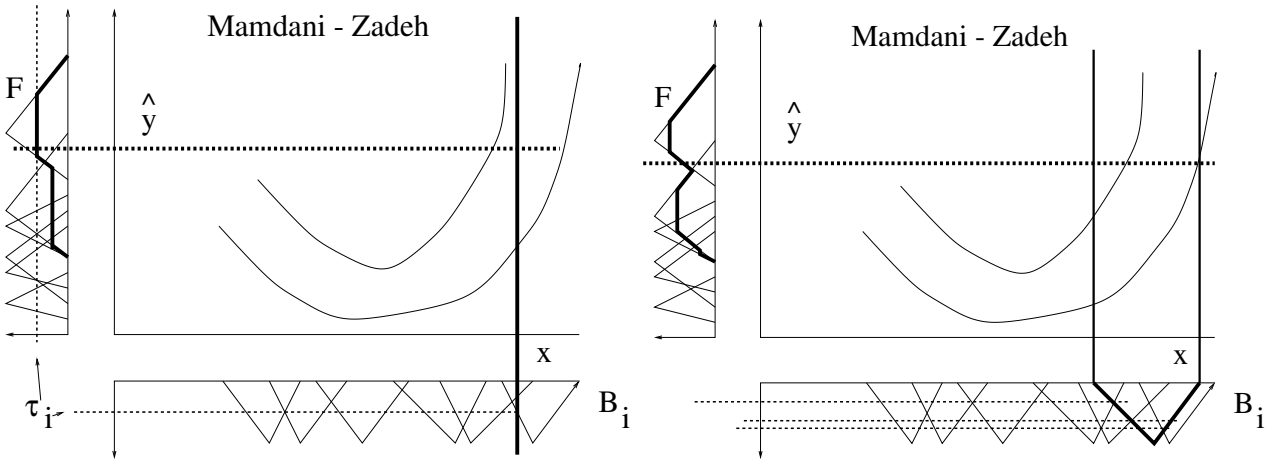


FIG. 1.11 – *Mécanisme d'inférence dit de Mamdani (règle d'inférence max-min).* **A gauche**, *les entrées sont fixes.* **A droite**, *les entrées sont floues.*

où τ_i, appelé *degré de déclenchement* de la règle r_i, est défini à partir des degrés de possibilité[2] de B_{ik} sachant $\tilde{x}_k$:

$$\tau_i = \bigwedge_{k=1}^{r} \mathrm{Poss}(B_{ik}|\tilde{x}_k) = \bigwedge_{k} \left[\bigvee_{u_k} (B_{ik}(u_k) \wedge \tilde{x}_k(u_k)) \right] . \tag{1.21}$$

Dans le cas de la fuzzification «singleton», les entrées sont des valeurs réelles, l'ensemble G est égal à R et F est donc la projection de R sur $\mathcal{X}$. Le seuil de déclenchement τ_i est alors égal à $B_i(\boldsymbol{x})$.

Le cas particulier où la t-norme et la t-conorme utilisées sont les opérateurs *min* et *max* correspond à la règle proposée initialement par Mamdani. En résumé, la mécanisme d'inférence peut être divisé en trois étapes.

1. Pour chaque règle r_i, calcul du degré de déclenchement τ_i.
2. Pour chaque règle r_i, calcul de la sortie floue correspondante, $F_i = \tau_i \wedge D_i$.
3. Agrégation des règles : $F = \vee_i F_i$.

La figure 1.11 illustre ce mécanisme dans le cas des entrées floues et fixes.

Défuzzification

Dans la plupart des applications, on est intéressé par une sortie ponctuelle $\widehat{y}$ du système. Pour cela, on peut procéder à une étape de « défuzzification » de la sortie floue F. La défuzzification est une application de $\mathcal{F}(\mathcal{Y})$ dans $\mathcal{Y}$, qui permet de sélectionner la valeur la plus représentative de F dans un certain sens. Il existe là encore de nombreuses techniques [19, 92, 78]. La méthode la plus courante et qui offre de notre point de vue les propriétés les

2. Soient deux ensembles flous A et $B \in \mathcal{F}(\mathcal{X})$. Par définition, la possibilité de A sachant B est la quantité $\mathrm{Poss}(A|B) = \vee_{x \in \mathcal{X}} A(x) \wedge B(\boldsymbol{x})$. Elle mesure le degré maximal avec lequel un élément de $\mathcal{X}$ peut appartenir à A et B.

plus intéressantes (cf. [92]) est celle du Centre de Gravité, qui définit la valeur défuzzifiée y_F^* de F de la façon suivante :

$$y_F^\star \triangleq \frac{\int_{\mathcal{Y}} uF(u)du}{\int_{\mathcal{Y}} F(u)du}, \tag{1.22}$$

si l'ensemble de référence $\mathcal{Y}$ est continu. On choisit alors $\widehat{y} = y_F^*$.

Si l'on s'intéresse à l'estimation de la sortie y par $\widehat{y}$, l'utilisation des opérateurs *max* et *min* n'est pas toujours conseillée, car ces opérateurs ne sont pas dérivables en tout point et rendent la tâche d'optimisation des paramètres du système délicate. Il est préférable d'employer par exemple le *produit* comme opérateur d'implication et d'intersection et la *somme* comme opérateur disjonctif, bien que ce dernier opérateur ne soit pas une t-conorme[3].

On suppose ici que les entrées sont *fixes*. Les expressions de τ_i (équation 1.21), de F (équation 1.20) et de $\widehat{y}$ deviennent alors, en fonction de l'entrée $\boldsymbol{x}$:

$$\tau_i(\boldsymbol{x}) = \prod_{k=1}^{r} B_{ik}(x_k), \tag{1.23}$$

$$F(y) = \sum_{i=1}^{I} \tau_i(\boldsymbol{x}) D_i(y), \tag{1.24}$$

$$\widehat{y}(\boldsymbol{x}) = \frac{\sum_{i=1}^{I} \tau_i(\boldsymbol{x}) y_{D_i}^* \int_{\mathcal{Y}} D_i(u)du}{\sum_i \tau_i(\boldsymbol{x}) \int_{\mathcal{Y}} D_i(u)du}, \tag{1.25}$$

où $y_{D_i}^*$ est le centre de gravité de D_i. Ainsi la valeur $\hat{y}(\boldsymbol{x})$ est la moyenne pondérée du centre de gravité des ensembles flous de la partie conséquente, dont les poids dépendent de l'influence des règles sur $\boldsymbol{x}$. La fonction $\widehat{y}$ appartient à la classe des fonctions de base généralisées (cf. section 1.2.5).

Cette méthode est appelée « méthode de raisonnement flou simplifié » [172], car elle permet de formuler analytiquement la relation entre les entrées et la sortie du système. Le modèle peut ainsi être vu comme un *réseau de neurones* (cf. figure 1.12). Cette méthode de raisonnement simplifié offre la possibilité d'un apprentissage du système flou (cf. section 1.3).

1.2.4 Modèles flous de Takagi-Sugeno

Takagi et al. [150] ont développé un modèle hybride dont les règles sont constituées de la façon suivante :

$$r_i : \quad \text{SI} \quad \boldsymbol{x} \text{ est } B_i \quad \text{ALORS} \quad y = f_i(\boldsymbol{x}), \tag{1.26}$$

où les f_i sont des fonctions de $\mathcal{X}$ dans $\mathcal{Y}$. La partie antécédente de la règle r_i est une proposition floue comme précédemment mais la partie conséquente est un modèle local conventionnel

3. L'inconvénient est qu'on peut obtenir un ensemble résultant F d'une hauteur $h(F) > 1$ difficilement interprétable. D'autres opérateurs sont envisageables, comme la t-conorme duale du produit, la *somme bornée*, mais elle est également non dérivable en tout point, ou des opérateurs de type moyenne, mais ils ne sont pas associatifs.

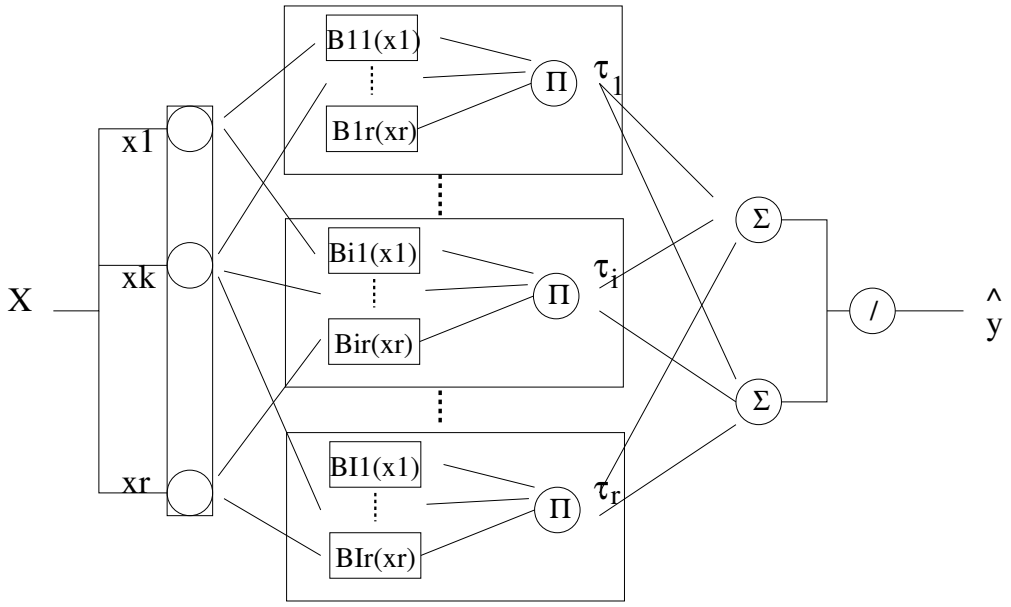


FIG. 1.12 – *Représentation neuronale d'un système flou. Les premières couches calculent les degrés de déclenchement des règles τ_i. La couche suivante combine les règles et normalise les τ_i. La dernière couche calcule la sortie finale $\widehat{y}$.*

utilisé pour l'approximation de la sortie du système dans la région de l'espace représentée par B_i.

Si l'on utilise la méthode d'inférence *somme - produit* comme précédemment, la sortie ponctuelle du système est alors :

$$\widehat{y}(\boldsymbol{x}) = \frac{\sum_i B_i(\boldsymbol{x}) f_i(\boldsymbol{x})}{\sum_i B_i(\boldsymbol{x})}. \tag{1.27}$$

La sortie finale est la moyenne de la sortie correspondant à la règle r_i, pondérée par le degré de déclenchement normalisé de la règle. Des méthodes d'estimation locales basées sur cette équation ont été proposées par plusieurs auteurs [77, 108, 109] selon la forme de f_i. Puisqu'il s'agit d'approximations locales, les fonctions f_i sont souvent très simples, linéaires ou polynomiales, voire constantes. Cette approche, fréquente en modélisation de systèmes, suppose que la dynamique est différente mais polynomiale selon les régions de l'espace. Le modèle présente certaines analogies avec la régression polynomiale par morceaux (cf. annexe A).

1.2.5 Equivalence fonctionnelle entre systèmes flous et réseaux de neurones

L'expression fonctionnelle obtenue par les deux principaux types de systèmes flous (de Mamdani et de Takagi-Sugeno) permet de les comparer à des techniques conventionnelles comme les fonctions de base généralisées. Cette vaste classe de modèles contient une grande partie des régresseurs décrits dans l'annexe A, comme les estimateurs à noyaux, les arbres de régression, les modèles polynomiaux et certains types de réseaux de neurones [4]. Des théorèmes,

4. En annexe A, une description des principales méthodes d'estimation fonctionnelle est proposée.

analyses ou traitements établis pour ces modèles peuvent alors être directement appliqués aux systèmes flous.

La forme générale de ces régresseurs est la suivante [132] :

$$f(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{I} a_i h_i(\boldsymbol{x}; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i), \tag{1.28}$$

où $\boldsymbol{w}$ est le vecteur des paramètres, $\boldsymbol{w} = [\boldsymbol{a}\ \boldsymbol{C}\ \boldsymbol{\Sigma}]$, $\boldsymbol{a} = (a_1, \ldots, a_I)^T$, $\boldsymbol{C} = (\boldsymbol{c}_1 \ldots \boldsymbol{c}_I)^T$ et $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_I)$. Chaque paramètre a une signification particulière : a_i représente le poids ou l'amplitude de la fonction de base h_i, le vecteur $\boldsymbol{c}_i$, la position ou la translation et la matrice $\boldsymbol{\Sigma}_i$, l'échelle ou la direction.

Or, l'expression de $\widehat{y}(\boldsymbol{x})$ dans les équations (1.25) et (1.27) peut se réécrire sous la forme précédente (1.28) et cette fonction fait donc partie de la classe des fonctions de base généralisées. Elle est alors appelée *fonction de base floue* [164].

Une forme très générale de (1.28), appelée *produit tensoriel* par certains auteurs [19], est déterminée par la composition de deux classes de fonctions de base plus simples [132]. Nous allons voir maintenant les liens entre certains systèmes flous et ces deux autres classes de fonctions de base dont les représentants sont les réseaux de neurones les plus fréquemment utilisés : les fonctions de base radiale et les perceptrons multi-couches.

Fonction de base radiale généralisée

On peut montrer que les fonctions de base radiale sont des modèles locaux d'approximation fonctionnelle (cf. annexe A, section A.3.3). Pour ces modèles, l'expression de h_i peut se mettre sous la forme :

$$h_i(\boldsymbol{x}; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i) = G_i\big(\|\boldsymbol{x} - \boldsymbol{c}_i\|_{\Sigma_i}\big)$$

où $\|.\|_{\Sigma_i}$ est une norme sur $\mathcal{X}$ et G_i est une fonction de $\mathbb{R}^+$ dans $\mathbb{R}$. En général, $\boldsymbol{\Sigma}_i$ est une matrice semi-définie positive de dimension (r, r). L'exemple le plus fréquent est celui des fonctions de base gaussiennes B_i définies par l'équation (1.14).

Ces modèles, qui peuvent se mettre sous la forme d'un réseau de neurones artificiel à 2 couches, la couche intermédiaire contenant les unités cachées $\boldsymbol{c}_i$, sont définis comme des *Réseaux à Fonction de Base Radiale* [105]. Les paramètres a_i représentent les poids de connexion entre les unités $\boldsymbol{c}_i$ et la sortie finale. La quantité $g_i(\boldsymbol{x}) = G_i\big(\|\boldsymbol{x} - \boldsymbol{c}_i\|_{\Sigma_i}\big)$ représente le degré d'activation de l'unité $\boldsymbol{c}_i$. La sortie finale représente alors la *somme* pondérée des a_i, sorties associées aux unités $\boldsymbol{c}_i$, par ces degrés d'activation (cf. figure 1.13). Afin de pallier le faible recouvrement des fonctions d'appartenance dans certaines régions, on peut normaliser ces degrés d'activation, ce qui revient à utiliser la fonction de base suivante :

$$h_i(\boldsymbol{x}; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i) = \frac{g_i(\boldsymbol{x})}{\sum_i g_i(\boldsymbol{x})}.$$

La sortie devient alors la *moyenne* pondérée des a_i par les degrés d'activation.

Jang [77] et Hunt et al. [72] ont montré l'équivalence fonctionnelle entre les réseaux à fonctions de base gaussiennes normalisées et certains systèmes flous de type Takagi-Sugeno sous certaines conditions. Soit le réseau caractérisé par I unités cachées $\boldsymbol{c}_i$ de paramètres de

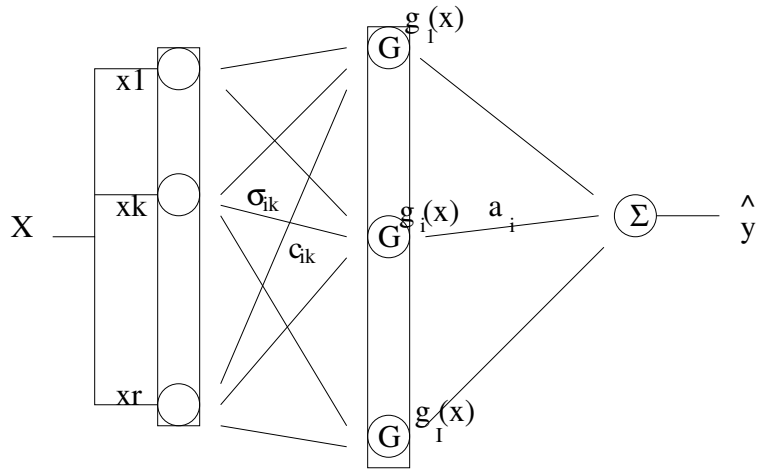


FIG. 1.13 – *Représentation d'un réseau à fonction de base radiale. La première couche calcule le degré d'activation* $g_i(\boldsymbol{x}) = G_i\big(\|\boldsymbol{x} - \boldsymbol{c}_i\|_{\Sigma_i}\big)$ *de l'unité* $\boldsymbol{c}_i$*. La deuxième couche calcule la sortie finale* $\widehat{y}$*.*

dispersion $\boldsymbol{\Sigma}_i$ et les poids a_i. Alors, le système flou vérifiant les hypothèses suivantes est équivalent à ce réseau :

1. Le nombre de règles du système est le même que le nombre d'unités I du réseau.
2. Les opérateurs du mécanisme d'inférence sont les mêmes que ceux du réseau (le produit et la somme).
3. Les fonctions f_i du système flou sont les fonctions constantes égales à a_i (cf. équation (1.27)).
4. Les fonctions d'appartenance multi-dimensionnelles B_i des parties antécédentes des règles sont les fonctions de base (gaussiennes normalisées) h_i, de mêmes paramètres $\boldsymbol{c}_i, \boldsymbol{\Sigma}_i$. C'est en particulier le cas si $\boldsymbol{\Sigma}_i$ est diagonale et si l'opérateur de conjonction est le produit. Le cas fréquent où $\boldsymbol{\Sigma}_i = \sigma_i \boldsymbol{I}_r$ permet de limiter le nombre de paramètres à estimer.

On peut étendre la définition de ces réseaux en remplaçant la constante a_i par une fonction quelconque, par exemple polynomiale. On obtient alors des réseaux équivalents aux systèmes flous quelconques, de type Takagi-Sugeno.

Perceptrons multi-couches et systèmes flous

Les perceptrons à une couche cachée font partie des fonctions de base dont la forme est la suivante :

$$h_i(\boldsymbol{x}; \boldsymbol{c}_i, b_i) = G\big(\boldsymbol{c}_i^t \boldsymbol{x} - b_i\big).$$

où b_i est ici un scalaire. On peut également montrer l'équivalence fonctionnelle de certains types de systèmes flous et certains perceptrons multi-couches. Etant donné un perceptron dont les fonctions d'activation sont continues, il existe un système flou capable de l'*approcher*,

à un degré de précision quelconque [22]. Réciproquement, on peut toujours construire un perceptron capable d'approcher un système flou à un niveau de précision donné.

Les perceptrons multi-couches étant des approximateurs universels (cf. annexe A), c'est-à-dire capables d'approcher indéfiniment toute fonction continue d'un compact de $\mathcal{X} \subset \mathbb{R}^r$ dans $\mathcal{Y} \subset \mathbb{R}$, les systèmes flous sont de ce fait également des approximateurs universels.

Ces résultats fondamentaux établissant la dualité des systèmes flous et des réseaux de neurones ont naturellement amené la définition d'une nouvelle classe de systèmes, combinant les avantages des systèmes flous et des réseaux de neurones : les systèmes neuro-flous.

1.3 Systèmes neuro-flous

1.3.1 Introduction

Les systèmes neuro-flous permettent de combiner les avantages de deux techniques complémentaires. Les systèmes flous fournissent une bonne représentation des connaissances. L'intégration de réseaux de neurones au sein de ces systèmes améliore leurs performances grâce à la capacité d'apprentissage des réseaux de neurones. Inversement, l'injection de règles floues dans les réseaux de neurones, souvent critiqués pour leur manque de lisibilité, clarifie la signification des paramètres du réseau et facilite leur initialisation, ce qui représente un gain de temps de calcul considérable pour leur *identification*.

De nombreux types de réseaux ou systèmes neuro-flous ont été développés ces dernières années et leur définition, loin d'être uniformisée, est parfois ambiguë et confuse. Afin de clarifier les définitions, nous proposons de répartir cette classe de systèmes en trois catégories : les systèmes neuro-flous classiques [77, 78], qui peuvent se voir à la fois comme des réseaux de neurones et comme des systèmes flous conventionnels. Il s'agit de systèmes flous auxquels on incorpore des concepts issus des réseaux de neurones : apprentissage des paramètres, ou des règles. Les réseaux de neurones auxquels on incorpore des concepts issus de la théorie des ensembles flous (fuzzification des opérations, des poids, des entrées, des sorties) constituent une deuxième catégorie. Afin de ne pas les confondre avec les précédents, nous les nommerons *réseaux de neurones fuzzifiés*. Enfin, les systèmes complexes se décomposent souvent en plusieurs tâches et requièrent différentes méthodes pour deux sous-problèmes différents. Les réseaux de neurones et les systèmes flous peuvent alors être utilisés *séparément* pour résoudre deux tâches différentes en parallèle ou successivement. Ces systèmes constituent une troisième catégorie. Dans la suite, nous ne décrirons que les deux premières catégories : les systèmes neuro-flous classiques et les réseaux de neurones fuzzifiés.

1.3.2 Systèmes neuro-flous classiques

Apprentissage des systèmes flous

Il n'est pas toujours possible de construire un système flou uniquement en traduisant la connaissance transmise par des experts en termes de règles et de fonctions d'appartenance. En effet, ce type d'information est souvent incomplet, épisodique et difficile à collecter de manière efficace et systématique. Pour remédier à ces problèmes d'acquisition et de traduction de connaissances, quand des données numériques sont disponibles, il est possible d'automatiser

le système, qui devient alors un système neuro-flou, à l'aide de méthodes d'identification.

L'identification d'un système consiste à déterminer, parmi une classe de modèles, celui qui semble le plus adapté, selon un critère donné, aux relations entre les variables d'entrée et de sortie du système [143]. Dans le cas des systèmes neuro-flous définis par l'équation (1.25), les techniques classiques sont envisageables mais des techniques plus spécifiques, utilisant les particularités de ces systèmes, comme la capacité d'intégration de connaissances *a priori*, peuvent être plus adaptées. On peut faire en particulier la distinction entre les modèles de type « boîte grise », où la base de règles est fournie par un expert et les modèles de type « boîte noire » pure, où la base de règles elle-même est estimée.

Dans le cas d'un modèle de type boîte noire, la connaissance du système se résume à l'existence d'un ensemble d'apprentissage $\mathcal{T} = (\boldsymbol{x}_k, y_k), k = 1, \ldots, N \in \mathcal{X} \times \mathcal{Y}$. L'identification du système se décompose alors essentiellement en deux étapes, même si elles ne sont pas indépendantes : l'identification de la structure du modèle et l'estimation des paramètres. Le modèle de type « boîte grise » requiert uniquement la deuxième étape.

Identification de la structure du modèle

La structure du modèle inclut la détermination des régresseurs $\boldsymbol{x}$ et les relations entre les variables $\boldsymbol{x}$ et y, le nombre de règles et la forme des fonctions d'appartenance. A ce stade, on suppose que les régresseurs ont été convenablement choisis, par exemple à l'aide de tests statistiques ou de techniques d'analyse des données. La forme des fonctions d'appartenance n'est pas forcément cruciale (cf. section 1.2.3). On s'intéresse alors essentiellement à la détermination du nombre optimal de règles. Plusieurs types d'approches sont envisageables [65], parmi lesquels la classification automatique de l'ensemble d'apprentissage [7] ou les techniques basées sur des critères d'erreur de prédiction, comme les techniques de rééchantillonnage, que l'on a évoquées dans l'annexe A.

Classification automatique

Les méthodes les plus fréquentes reposent sur la classification de l'espace représentant les données à partir de l'ensemble d'apprentissage. Soit $P = \{P_1, \ldots P_I\}$ un ensemble de I classes partitionnant l'espace $\mathcal{X} \times \mathcal{Y}$. Ces classes peuvent avoir été obtenues à l'aide d'un algorithme quelconque de partitionnement, comme par exemple les centres mobiles [99], ou les cartes auto-organisatrices [86] (cf. annexe D). Afin de tenir compte du passage progressif d'une classe à une autre, elles peuvent également être déterminées par une méthode de partitionnement *flou*, comme les centres mobiles flous [11]. Chaque classe P_i va permettre de construire une règle floue r_i en tenant compte des relations locales entre variables. Chaque règle correspond ainsi à une région de l'espace. La détermination du nombre de règles revient ainsi à choisir le nombre optimal de classes. La sélection du nombre de classes correspond à un arbitrage entre la précision et la complexité du modèle. Un grand nombre de règles permet une grande précision dans l'estimation de la sortie du système mais est coûteuse en temps de calcul. Inversement, si le nombre de règles est trop faible, le calcul sera rapide, mais l'estimation sera de mauvaise qualité.

Différentes variantes sont possibles pour construire les classes. Il est courant de partitionner séparément les espaces $\mathcal{X}$ et $\mathcal{Y}$ ou de partitionner uniquement l'espace $\mathcal{X}$ en classes $\{C_1, \ldots C_I\}$ et d'induire les classes P_i sur $\mathcal{X} \times \mathcal{Y}$ en utilisant par exemple une méthode de

régression [7].

Quant à la sélection du nombre de classes, de nombreux critères ont été proposés dans le contexte de la reconnaissance des formes [112]. Plaçons-nous dans le cas le plus fréquent consistant à partitionner uniquement $\mathcal{X}$. Soit $\boldsymbol{c}_i$ le centre de la classe C_i. On désigne par u_{ik} le degré d'appartenance de l'élément $\boldsymbol{x}_k$ de l'ensemble d'apprentissage à la classe C_i. Ces critères sont basés sur les matrices de covariance (floues) :

$$\boldsymbol{\Sigma}'_i = \sum_{k=1}^{N} \frac{u_{ik}}{\sum_k u_{ik}} (\boldsymbol{x}_k - \boldsymbol{c}_i)(\boldsymbol{x}_k - \boldsymbol{c}_i)^T,$$

ou des mesures de dispersion intra-classes :

$$SSW = \sum_{i=1}^{m} \sum_{k=1}^{N} u_{ik} \|\boldsymbol{x}_k - \boldsymbol{c}_i\|^2. \tag{1.29}$$

En particulier, une approche simple mais efficace consiste à considérer le pourcentage d'« inertie expliquée » par la partition, SSW/SST, où SST est la dispersion totale : $SST = \sum_k \|\boldsymbol{x}_k - \bar{\boldsymbol{x}}\|^2$, avec $\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{x}_k$.

Ces méthodes déterminent une bonne répartition des prototypes dans l'espace de représentation des données. Cette répartition ne garantit pas une bonne estimation des sorties du système mais elle permet de résumer l'information contenue dans l'ensemble d'apprentissage. La perte d'information est compensée par une grande lisibilité des règles.

Rééchantillonnage

L'identification de la structure peut se formuler comme un problème d'estimation fonctionnelle par un réseau de neurones (cf. annexe A). Elle représente la première étape qui consiste à contrôler la complexité du modèle afin de définir sa taille, c'est-à-dire, si on se limite à un réseau à une couche cachée, le nombre I d'unités cachées.

On peut alors utiliser des techniques de rééchantillonnage, telles que la validation croisée, le jackknife ou le bootstrap [48], basées sur la division de l'échantillon en un ensemble de généralisation et d'apprentissage (cf. annexe A, section A.4).

Nous présentons uniquement une des variantes de la validation croisée, connue sous le nom de «leave one out». Pour un certain nombre de valeurs de I, on effectue la procédure suivante. Le point $(\boldsymbol{x}_k, y_k)$ est retiré de l'échantillon et on estime la variable y en $\boldsymbol{x}_k$ à l'aide des $N-1$ exemples restants. L'estimateur de y_k obtenu étant noté $\widehat{f}_I^{(-k)}(\boldsymbol{x}_k)$, on construit alors le critère de validation croisée suivant :

$$CV(I) = \frac{1}{N} \sum_{k=1}^{N} (y_k - \widehat{f}_I^{(-k)}(\boldsymbol{x}_k))^2. \tag{1.30}$$

La valeur choisie, $\hat{I}$, est celle qui minimise ce critère.

Estimation des paramètres

Une fois la première phase achevée, la taille du modèle est déterminée et il ne reste plus qu'à sélectionner les paramètres, représentés par un vecteur $\boldsymbol{w}$, de l'équation (1.25) ou de

l'équation (1.27) selon qu'il s'agisse d'un modèle de type Mamdani ou Takagi-Sugeno. Ce vecteur contient les paramètres des fonctions d'appartenance des ensembles flous B_i et D_i dans le cas d'un modèle de type Mamdani. Dans le cas d'un modèle de type Takagi-Sugeno, il contient les paramètres de B_i ainsi que les coefficients des polynômes f_i.

Cette deuxième phase correspond au problème classique de l'optimisation des paramètres $\boldsymbol{w}$ d'un réseau de neurones, c'est-à-dire d'une fonction non linéaire $\boldsymbol{w} \longmapsto f(\boldsymbol{x}, \boldsymbol{w})$. Etant donné l'ensemble d'apprentissage, le problème consiste à *ajuster* les poids $\boldsymbol{w}$ en minimisant par rapport à $\boldsymbol{w}$, l'un des critères suivants, définis dans l'annexe A : le critère quadratique empirique (cf. équation A.2),

$$J_N(\boldsymbol{w}) = \frac{1}{N} \sum_k (y_k - f(\boldsymbol{x}_k, \boldsymbol{w}))^2,$$

ou le critère $J_{pen}(\boldsymbol{w})$ (cf. équation A.11) si on ajoute un terme de pénalisation. Ceci est un problème très classique de minimisation d'une fonction non linéaire.

Le minimum $\widehat{\boldsymbol{w}}$ de la fonction de coût J ne peut être calculé analytiquement mais à l'aide de méthodes *itératives*, dont la plupart peuvent être vues comme un cas particulier de la méthode de Newton et pour lesquelles la convergence vers un minimum local est assurée.

On initialise les paramètres à $\boldsymbol{w}(0)$ en utilisant les informations *a priori* sur le modèle. A l'étape t+1, dans la méthode de Newton, l'estimation des paramètres devient :

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - [\nabla^2 J(\boldsymbol{w}(t))]^{-1}.\nabla J(\boldsymbol{w}(t)),$$

où $\nabla J(.) = [\frac{\partial J}{\partial w_i}(.)]$ et $\nabla^2 J = [\frac{\partial^2 J}{\partial w_i \partial w_j}(.)]$ sont respectivement le vecteur gradient et la matrice hessienne de J. Les calculs étant souvent très lourds, des méthodes simplifiées, ne faisant pas intervenir la matrice hessienne, sont souvent utilisées en pratique, comme les méthodes de descente de gradient, de gradient conjugué ou de Levenberg-Marquart. L'algorithme le plus utilisé est celui de la descente de gradient pour lequel on calcule

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta(t) \nabla J(\boldsymbol{w}(t)), \quad \eta(t) \in]0, 1[.$$

L'algorithme s'arrête dès que J est inférieur à un seuil fixé, à l'itération T. Alors $\boldsymbol{w}(T)$ est l'estimation finale de $\boldsymbol{w}$. Le système neuro-flou est alors complètement spécifié[5].

1.3.3 Les réseaux de neurones fuzzifiés

Dans cette section, nous présentons brièvement les réseaux de neurones fuzzifiés. On peut les diviser en deux sous-groupes : les *réseaux neuro-flous hybrides*, pour lesquels les opérateurs neuronaux sont des *opérateurs flous*, et les *réseaux de neurones fuzzifiés réguliers* ou *standards*, pour lesquels les entrées, les poids et/ou les sorties ne sont plus des réels mais des *intervalles*, des *nombres flous* ou des *intervalles flous*. Par souci de clarté, nous nous limitons ici à des réseaux à une seule couche, d'entrées $\boldsymbol{x} = (x_1, \ldots, x_r)$, de poids $\boldsymbol{w} = (w_1, \ldots, w_r)$, de fonction de transfert g et de fonction cible ou sortie désirée y, sachant qu'une généralisation à plusieurs couches est immédiate (cf figure 1.14).

5. Selon la forme particulière des réseaux, d'autres méthodes d'optimisation sont envisageables. Par exemple, dans le cas des fonctions de base radiale, nous avons vu en annexe A que l'on pouvait déterminer analytiquement les paramètres de sortie des unités cachées.

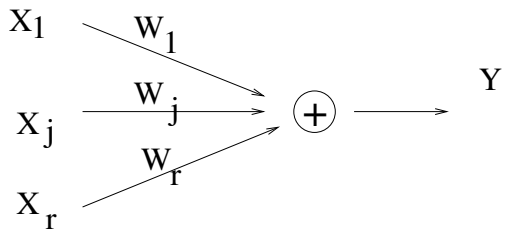


FIG. 1.14 – *Réseau neuro-flou : les entrées x_k et les poids w_k peuvent être réels ou flous. La sortie $\tilde{y} = g\big(\bigoplus_k(\tilde{w}_k \otimes \tilde{x}_k)\big)$ est calculée par l'arithmétique floue.*

Les réseaux hybrides

Dans le cas d'un réseau de neurones classique, connu sous le nom de perceptron, de fonction de transfert g (sigmoïdale, tangente hyperbolique, ou autre), la sortie $\hat{y}$ du réseau est définie par :

$$\hat{y} = f(\boldsymbol{x}, \boldsymbol{w}) = g(\boldsymbol{x}^t \boldsymbol{w}) = g\left(\sum_{k=1}^{r} w_k x_k\right).$$

Les poids, les entrées et les sorties sont des *valeurs réelles* et les opérateurs de combinaison des poids et des entrées sont le *produit* et la *somme*.

Si d'autres opérateurs de combinaison, de type flou, comme des t-normes ou des t-conormes, sont utilisés, nous obtenons un réseau de neurones dit *hybride* [66, 68]. Dans ce type de réseau, les entrées, les poids et les sorties restent des valeurs réelles, mais doivent appartenir à l'intervalle $[0, 1]$ pour pouvoir être combinés à l'aide d'opérateurs flous. Pour la même raison, la fonction g doit être définie de $[0, 1]$ dans $[0, 1]$.

Divers exemples de réseaux hybrides ont été proposés [66]. Les deux exemples les plus simples, où g est la fonction identité, sont les suivants [68]:

- le neurone flou « ET », où $y = \min_{k=1}^{r}(\max(w_k, x_k))$ représente la loi de composition $min - max$.
- le Neurone flou « OU », où $y = \max_{k=1}^{r}(\min(w_k, x_k))$ représente la loi de composition $max - min$.

Réseau de neurones fuzzifié standard

La fuzzification directe d'un réseau de neurones classique est obtenue lorsque les entrées $\boldsymbol{x}$, les poids $\boldsymbol{w}$ ou les sorties désirées y sont des nombres flous. Les opérations fonctionnelles (addition, multiplication, fonction de transfert) sont définies par le principe d'extension et l'arithmétique floue (cf. section 1.2.1). On peut distinguer trois types de réseaux fuzzifiés utilisés dans l'implémentation de règles linguistiques [21].

- les réseaux fuzzifiés de type 1, dont les entrées sont des nombres réels mais les poids flous,
- les réseaux fuzzifiés de type 2, dont les entrées sont floues et les poids, réels,
- les réseaux fuzzifiés de type 3, dont les entrées et les poids sont flous.

Dans tous les cas, les sorties du réseau sont floues. Il existe un quatrième type, où les entrées sont floues, les poids et les sorties sont des nombres réels, utilisé pour la classification d'entrées imprécises [67, 73].

Si on reprend l'exemple du réseau précédent, où les entrées $\widetilde{x}_k$ et les poids $\widetilde{w}_k$ sont flous[6], on obtient un réseau fuzzifié de type 3 dont la sortie floue $\widetilde{y}$ est définie par :

$$\tilde{y} = g\left(\bigoplus_{k=1}^{r}(\widetilde{w}_k \otimes \widetilde{x}_k)\right),$$

où $\oplus$ et $\otimes$ sont les opérateurs flous généralisant l'addition et la multiplication (cf. section 1.2.1). Si la fonction de transfert g est bijective, la sortie $\widetilde{y}$ est alors définie par

$$\tilde{y}(u) = \left[\bigoplus_{k=1}^{r}(\widetilde{w}_k \otimes \widetilde{x}_k)\right]\left(g^{-1}(u)\right) \quad \forall\, u \in [0,1].$$

Il est possible de combiner les deux types de fuzzification des réseaux de neurones en permettant à la fois l'utilisation d'opérateurs flous et de poids ou d'entrées flous [21].

Nous allons maintenant nous intéresser à l'identification de ces réseaux fuzzifiés. Certains auteurs ont généralisé les mécanismes d'apprentissage des réseaux de neurones classiques, comme l'algorithme de rétropropagation du gradient, aux réseaux neuro-flous. Cela suppose de généraliser la définition de la fonction de coût aux nombres flous :

$$C\left(\widetilde{y}, f(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{w}})\right)$$

c'est-à-dire, de définir la notion de *distance* ou de *dissimilarité* entre deux nombres flous.

« Distance » entre ensembles flous

De nombreuses mesures de « distance » entre ensembles flous ont été définies dans la littérature [183, 129, 39]. On peut définir deux types de distances entre deux ensembles flous F et F' définis sur un ensemble de référence $\mathcal{Y} \subset \mathbb{R}$:

- les opérateurs $\tilde{d}$ de type classique, $\tilde{d}$: $\mathcal{F}(\mathcal{Y}) \times \mathcal{F}(\mathcal{Y}) \to \mathbb{R}^+$
- les opérateurs de type flou, $\tilde{d} : \mathcal{F}(\mathcal{Y}) \times \mathcal{F}(\mathcal{Y}) \to \mathcal{F}(\mathbb{R}^+)$

Parmi les mesures classiques, l'une d'elles est définie à partir de la distance de Minkowski :

$$\tilde{d}_1(F, F') = \left(\int_{\mathcal{Y}} [F(y) - F'(y)]^p dy\right)^{1/p}, \quad p \in \mathbb{N}. \tag{1.31}$$

Cependant, cette quantité ne généralise pas la distance de Minkowski entre nombres réels. En fait, la plupart des mesures de dissimilarité ou de distance entre ensembles flous ne tiennent pas véritablement compte du caractère ordonné de l'ensemble de référence $\mathcal{Y}$.

La généralisation de la distance de Hausdorff entre ensembles classiques paraît plus judicieuse. Soient deux intervalles $F = [f_1, f_2]$ et $F' = [f'_1, f'_2]$. La distance de Hausdorff entre F

6. Dans la suite, la quantités floues sont signalés par la présence du symbole « ˜ ».

et F' s'écrit : $h(F,F') = \max\{|f_1 - f_1'|, |f_2 - f_2'|\}$. On peut définir différentes extensions à la classe des ensembles flous *convexes* et *normalisés*, utilisant les $\alpha-$coupes des ensembles flous [183] :

$$\tilde{d}_2(F,F') = \int_0^1 h(F_\alpha, F'_\alpha) d\alpha \tag{1.32}$$

$$\tilde{d}'_2(F,F') = \sup_\alpha h(F_\alpha, F'_\alpha) \tag{1.33}$$

Dans le cas de nombres réels y et y', $h(y,y') = |y - y'|$ et donc

$$\tilde{d}_2(F,F') = \tilde{d}'_2(F,F') = \frac{|y-y'|}{2},$$

ce qui est tout à fait satisfaisant.

On peut définir une distance $\tilde{d}$ de type flou d'après le principe d'extension. Au lieu de définir une distance comme un nombre réel, on définit une distribution de valeurs possibles. Soit $\Delta = \tilde{d}(F,F') \in \mathcal{F}(\mathbb{R}^+)$. Alors

$$\Delta(u) = \sup_{\{(y,y') \in F \times F') | d(y,y') = u\}} \min[F(y), F(y')],$$

où d est une distance classique de $\mathcal{Y} \times \mathcal{Y}$ dans $\mathbb{R}^+$.

Identification des réseaux fuzzifiés de types 2 et 3

Supposons que l'on possède un ensemble d'apprentissage constitué d'entrées et sorties *floues* : $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)$. Soit $F = f(\tilde{\boldsymbol{x}}, \boldsymbol{w})$ la sortie floue d'un réseau fuzzifié associée à l'entrée floue $\tilde{\boldsymbol{x}}$. Nous traitons pour l'instant le cas des réseaux de type 2 où les poids sont des nombres réels.

Pour l'apprentissage du réseau, il est possible de choisir l'une des mesures de distance introduites dans la section précédente. Ishibushi et al. [73] ont proposé d'utiliser une variante de la distance (1.33), basée sur la décomposition d'un nombre flou en α-coupes :

$$C(\tilde{y}, F) = \int_0^1 e(\alpha)\alpha d\alpha, \text{ avec } e(\alpha) = \frac{1}{2}\{(F_\alpha^g - \tilde{y}_\alpha^g)^2 + (F_\alpha^d - \tilde{y}_\alpha^d)^2\}, \tag{1.34}$$

où F_α^g et F_α^d sont les bornes inférieure et supérieure de l'intervalle F_α et $\tilde{y}_\alpha^g$ et $\tilde{y}_\alpha^d$ sont les bornes inférieure et supérieure de $\tilde{y}_\alpha$ (cf. figure 1.15). Pour des raisons pratiques, les auteurs calculent en fait une somme discrète sur un nombre fini K d'α-coupes : $\sum_{\alpha \in K} e(\alpha)\alpha$. Rappelons que, les nombres flous étant convexes, leurs $\alpha-$coupes sont des intervalles. Le calcul de F_α se détermine donc par l'arithmétique des intervalles, sachant que, pour deux intervalles A et B,

- $(A \oplus B)_\alpha = [A_\alpha^g + B_\alpha^g, A_\alpha^d + B_\alpha^d]$
- $(A \otimes B)_\alpha = [A_\alpha^g B_\alpha^g, A_\alpha^d B_\alpha^d]$, si $A_\alpha^g, B_\alpha^g, A_\alpha^d, B_\alpha^d \in \mathbb{R}^+$
- si f est croissante, $f(A_\alpha) = [f(A_\alpha^g), f(A_\alpha^d)]$.

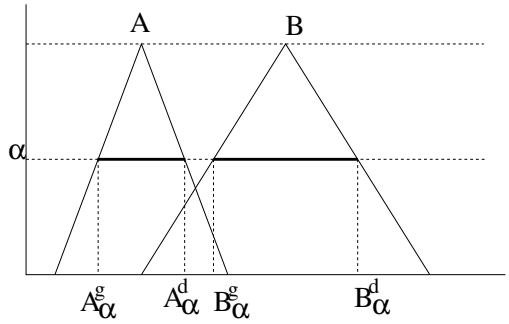


FIG. 1.15 – *Distance entre deux nombres flous A et B : utilisation des α-coupes. Pour chaque $\alpha-$coupe, on calcule la moyenne de la distance entre les extrémités inférieure et supérieure des intervalles A_α et B_α. On intègre ensuite en pondérant par la valeur de α.*

On obtient ainsi, dans le cas d'un réseau de type 2, $F_\alpha = [f(B^g_\alpha), f(B^d_\alpha)]$ avec $B_\alpha = [\sum_k (A_k)^g_\alpha, \sum_k (A_k)^d_\alpha]$ et $(A_k)_\alpha = [w_k(x_k)^g_\alpha, [w_k(x_k)^d_\alpha]$. Le choix de ce coût se justifie par le fait que

$$C(\tilde{y}, F) \to 0 \quad \Leftrightarrow F(u) \to \tilde{y}(u) \quad \forall u \in \mathcal{Y}.$$

La fonction de coût totale de l'ensemble d'apprentissage $(\tilde{\boldsymbol{x}}_k, \tilde{y}_k)_{k=1}^N$ est alors

$$E = \frac{1}{N} \sum_k C(\tilde{y}_k, F_k).$$

Pour l'apprentissage, puisque les poids $\boldsymbol{w}$ sont réels, les algorithmes de la section 1.3.2 comme celui du gradient peuvent alors s'appliquer. A chaque itération t, on calcule :

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) - \eta(t) \nabla E(\boldsymbol{w}(t)).$$

Le détail des calculs est effectué dans [73]. Pour les réseaux à plusieurs couches, on peut généraliser l'algorithme de rétroprogation du gradient.

Dans le cas des réseaux de type 1 et 3 où interviennent des poids flous, on peut également définir des règles d'apprentissage en définissant un *paramétrage* de $\boldsymbol{w}$ [74].

De nombreuses applications ont été réalisées à partir des réseaux fuzzifiés, comme l'approximation fonctionnelle ou la résolution d'équations matricielles floues.

1.4 Conclusion

Dans ce chapitre, nous avons présenté les mécanismes d'inférence des systèmes flous. La représentation des connaissances fondée sur la théorie des ensembles flous conduit à un traitement souple de l'information dont les systèmes flous constituent l'un des outils majeurs. La structure d'un système flou est constituée de deux principaux éléments : une base de règles, qui contient un ensemble de règles floues et un mécanisme de raisonnement, en général le raisonnement flou, qui définit la procédure d'inférence. Les règles floues sont un outil efficace permettant de modéliser les expressions provenant du langage naturel. Le raisonnement flou est une procédure d'inférence qui permet de déduire certaines conclusions à partir de relations fonctionnelles imprécises ou floues entre variables. Les systèmes d'inférence sont un

outil bien établi, qui connaît un vaste champ d'application, comme le contrôle automatique, les systèmes experts, la prédiction des séries temporelles, la robotique ou la reconnaissance des formes. L'intérêt essentiel des systèmes flous est de permettre de modéliser des relations imprécises entre différentes variables. Quand les entrées et les sorties sont précises, les systèmes définissent une relation fonctionnelle non linéaire et peuvent être représentés par des réseaux de neurones classiques. L'utilisation des systèmes neuro-flous permet de tirer avantage de la capacité d'identification des réseaux de neurones et de la lisibilité des ensembles flous.

Afin de clarifier les définitions des systèmes neuro-flous, nous avons proposé une classification personnelle de ces systèmes en trois catégories : les systèmes neuro-flous classiques, qui sont des systèmes flous adaptatifs, dont les entrées et les sorties sont des valeurs réelles ; les réseaux de neurones fuzzifiés, qui sont des réseaux de neurones auxquels on a incorporé des concepts issus de la théorie des ensembles flous (fuzzification des opérations, des poids, des entrées, des sorties) ; les systèmes complexes divisés en sous-tâches dont le traitement est effectué soit par un réseau de neurones, soit par un système flou.

Dans le chapitre suivant, nous nous intéressons à une application des systèmes neuro-flous classiques à un problème d'estimation particulier, le traitement de données manquantes.

Chapitre 2

Application au traitement de données manquantes

Dans ce chapitre, nous proposons d'appliquer le principe des systèmes flous au problème de l'estimation de données manquantes. Dans un tableau de données, l'existence de valeurs manquantes peut être due à des causes très diverses. La donnée peut être indisponible à cause d'un dysfonctionnement de l'appareil de mesure qui la délivre. Dans la collecte de données par sondages, il peut s'agir d'absence de réponses ou de réponses contradictoires invalidées par l'analyste. Une absence de réponse peut elle-même refléter deux types de comportement : la donnée est complètement inconnue par le sondé ou au contraire elle provient d'un refus de répondre. Face à un problème de données manquantes, il y a deux attitudes possibles. On peut éluder la question en ne retenant dans la base de données que les vecteurs complets. Cette méthode d'élimination, d'une simplicité évidente, présente l'avantage de permettre l'application de techniques classiques d'analyse de données sans modification pour des traitements ultérieurs. Mais elle ne donne des résultats satisfaisants que lorsque les vecteurs incomplets sont peu nombreux, ce qui s'avère difficile dans les cas de vecteurs de grande dimension. De plus, elle ne fournit aucune estimation, même imprécise de ces données manquantes.

L'autre attitude consiste donc à remplacer les valeurs manquantes par une quantité « raisonnable ». Les méthodes d'estimation de données manquantes font appel à des techniques très variées. Elles supposent en général un cadre probabiliste. Une description détaillée des principales approches envisagées dans la littérature [97, 127, 26], est proposée en annexe B. Les méthodes heuristiques, souvent utilisées par le praticien, comme le remplacement par la moyenne, la médiane ou une valeur quelconque de référence, permettent d'éluder le problème rapidement à l'aide de solutions peu coûteuses. L'estimation des données incomplètes n'est souvent pas un but en soit, mais un prétraitement de données. Néanmoins les algorithmes ultérieurs de classification, d'estimation de sortie d'un système ou d'apprentissage peuvent être perturbés par une mauvaise reconstruction de données. Si le nombre de caractéristiques et le nombre de données à reconstruire sont élevés, les méthodes basées sur la régression manquent de souplesse, car elles exigent un grand nombre de modèles. Les méthodes paramétriques basées sur la maximisation de la vraisemblance, comme l'algorithme EM («Expectation-Maximisation») [29], ou la simulation de probabilités, comme les méthodes de Monte-Carlo, ont prouvé leur efficacité et sont largement utilisées, mais elles requièrent la connaissance ou l'estimation des lois des variables et sont de ce fait généralement

coûteuses en temps de calcul.

La plupart de ces techniques supposent un cadre probabiliste. Nous présentons dans ce chapitre une nouvelle méthode, basée sur la construction d'un système neuro-flou, définissant des relations floues entre variables, qui offre des similitudes avec certaines méthodes de régression, mais qui présente l'avantage de traiter un nombre quelconque de données manquantes vues comme les sorties du système. De plus, cette méthode tolère également la présence de données imprécises, représentées sous forme de nombres flous. L'imprécision sur la valeur de substitution est caractérisée par une distribution de possibilité.

Notre méthode a été appliquée à des données réelles environnementales, dans le cadre du projet européen EM^2S. Les résultats sont confrontés à ceux d'autres techniques existantes.

2.1 Principe de la méthode

2.1.1 Formulation du système flou

La connaissance relative au système se résume à un ensemble d'apprentissage de N vecteurs de r caractéristiques réelles : $\mathcal{T} = \{\boldsymbol{x}_i \in \mathcal{X}, i \in \{1, \ldots, N\}\}$, avec $\boldsymbol{x}_i = \{x_{i1}, \ldots, x_{ir}\}$, que l'on suppose pour l'instant complets. Une généralisation à un ensemble plus réaliste $\mathcal{T}$ contenant des données manquantes, imprécises ou « aberrantes », sera abordée plus loin.

Soit $\boldsymbol{x}$ un nouveau vecteur, composé de deux parties $\boldsymbol{x}^o$ et $\boldsymbol{x}^m$, contenant respectivement les variables observées et manquantes[1]. L'objectif est de reconstruire $\boldsymbol{x}^m$ à l'aide des informations fournies par l'ensemble d'apprentissage. Intuitivement, la méthode que nous proposons repose sur les similitudes entre ce nouveau vecteur $\boldsymbol{x} = (\boldsymbol{x}^o, \boldsymbol{x}^m)$ et les vecteurs $\boldsymbol{x}_i$. Soient $\mathcal{X}^o$ et $\mathcal{X}^m$ les sous-espaces de $\mathcal{X}$ restreints aux variables observées et manquantes de $\boldsymbol{x}$ et $\boldsymbol{x}_i^o$ et $\boldsymbol{x}_i^m$ les projections de $\boldsymbol{x}_i$ sur ces sous-espaces. Plus $\boldsymbol{x}^o$ est « proche » de $\boldsymbol{x}_i^o$ au sens d'une certaine distance, plus on aura de chances que $\boldsymbol{x}^m$ soit « proche » de $\boldsymbol{x}_i^m$, à condition que les variables soient bien corrélées entre elles. On désigne par $O(\boldsymbol{x}) \subset \{1, \ldots, r\}$ et $M(\boldsymbol{x})$, les ensembles d'indices des composantes observées et manquantes de $\boldsymbol{x}$.

La proximité étant une notion floue, nous pouvons traduire cette idée sous la forme d'un système flou S à base de règles binaires r_i,

$$r_i : \text{SI } \boldsymbol{x}^o \text{ est } Z_i^o \text{ ALORS } \boldsymbol{x}^m \text{ est } Z_i^m, \tag{2.1}$$

où $Z_i^o \in \mathcal{F}(\mathcal{X}^o)$ et $Z_i^m \in \mathcal{F}(\mathcal{X}^m)$ sont des ensembles flous multidimensionnels représentant respectivement $\boldsymbol{x}_i^o$ et $\boldsymbol{x}_i^m$. On note $Z_i = Z_i^m \times Z_i^o$ l'ensemble flou r-dimensionnel représentant $\boldsymbol{x}_i$ et Z_{ij} l'ensemble flou représentant la j$^{\text{ème}}$ variable de $\boldsymbol{x}_i$, qui est aussi la projection de Z_i sur $\mathcal{X}_j$.

La proposition floue « $\boldsymbol{x}^o$ est Z_i^o» est bien la traduction de l'expression linguistique : $\boldsymbol{x}^o$ est « proche» de $\boldsymbol{x}_i^o$.

Il est possible d'étudier séparément les variables manquantes, en remarquant que le système flou S peut se décomposer en systèmes flous $S_l, l \in M(\boldsymbol{x})$, possédant chacun N règles r_{il} :

$$r_{il} : \quad \text{SI } \boldsymbol{x}^o \text{ est } Z_i^o \text{ ALORS } \boldsymbol{x}_l \text{ est } Z_{il} \tag{2.2}$$

Dans la suite, nous étudions séparément chacun des systèmes flous S_l.

1. Dans toute la suite, les indices m et o correspondent respectivement aux variables manquantes et observées.

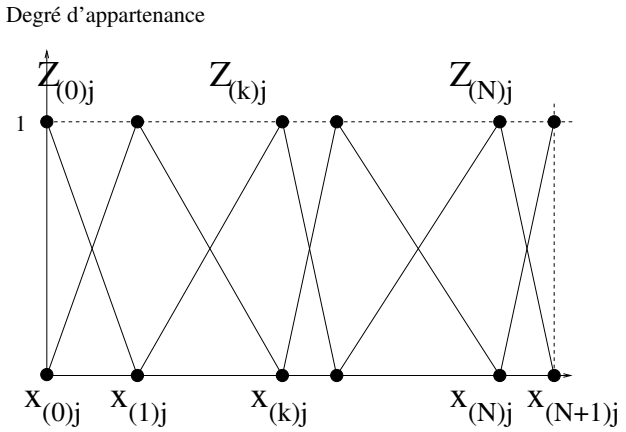


FIG. 2.1 – *Construction des fonctions d'appartenance Z_{ij} par interpolation de B-splines d'ordre 1. On obtient des fonctions d'appartenance triangulaires.*

2.1.2 Fonctions d'appartenance du système

Les x_{ij} peuvent être fuzzifiés de nombreuses façons différentes. L'utilisation de fonctions d'appartenance du type des B-splines, qui sont des polynômes par morceaux, semble *a priori* judicieuse [14]. En effet, celles-ci se définissent de façon *unique* par interpolation des nœuds, une fois l'ordre des polynômes d'interpolation choisi, sans identification préalable de paramètres. Si k est l'ordre des polynômes d'interpolation, les ensembles flous Z_{ij} décrivant la variable j sont définis comme les B-splines d'ordre k, les nœuds $x_{(i)j}$ étant réordonnés par ordre croissant selon i. De plus, elles permettent un bon recouvrement du domaine, car elles possèdent la propriété de normalisation suivante :

$$\forall j \in \{1, \ldots, r\} \quad \forall u \in \mathcal{X}_j \quad \sum_{i=1}^{N} Z_{ij}(u) = 1.$$

Il existe des relations de récurrence permettant de les déterminer très facilement [14]. Si on se limite à l'ordre 1, on obtient des fonctions d'appartenance triangulaires (cf. figure 2.1).

L'utilisation d'autres fonctions d'appartenance nécessite l'identification de certains paramètres. Par exemple, dans le cas des fonctions d'appartenance gaussiennes, centrées en x_{ij} et d'écart-type σ_{ij}, le paramètre σ_{ij} doit être choisi *a priori* ou estimé.

2.1.3 Estimation des données manquantes

Les systèmes flous S_l étant du type de Mamdani (équation 1.13), étudié dans le chapitre précédent, nous pouvons donc appliquer les résultats décrits en section 1.2.3.

Le déclenchement de chaque règle r_{il} est défini par $\tau_i = Z_i^o(\boldsymbol{x}^o)$ et mesure la proximité de $\boldsymbol{x}$ et $\boldsymbol{x}_i$. On peut noter que, la partie antécédente des règles r_{il} étant la même quel que soit l, le degré de déclenchement τ_i de r_{il} est indépendant de l.

D'après l'équation (1.20), pour chaque $l \in M(\boldsymbol{x})$, l'ensemble flou associé F_l déduit du système S_l s'écrit :

$$F_l(x_l|\boldsymbol{x}^o) = \bigvee_{i=1}^{N} Z_i^o(\boldsymbol{x}^o) \wedge Z_{il}(x_l) \quad \forall l \in M(\boldsymbol{x}). \tag{2.3}$$

Un estimateur $\widehat{x}_l$ de x_l peut alors être défini par défuzzification de F_l, la t-norme utilisée étant le *produit*:

$$\widehat{x}_l = \frac{\bigvee_{i=1}^{N} Z_i^o(\boldsymbol{x}^o) z_{il} \int_{\mathcal{X}_l} Z_{il}(u_l) du_l}{\bigvee_i Z_i^o(\boldsymbol{x}^o) \int_{\mathcal{X}_l} Z_{il}(u_l) du_l}, \tag{2.4}$$

où z_{il} est le centre de gravité de Z_{il}, qui peut être différent de x_{il}, selon le choix de Z_{il}, comme dans le cas des B-splines. De ce point de vue, il peut être conseillé d'utiliser par exemple des fonctions d'appartenance gaussiennes ou triangulaires symétriques.

On obtient donc une relation fonctionnelle entre x_l et $\boldsymbol{x}^o$, qui peut s'écrire:

$$\widehat{x}_l = \widehat{f}_l(\boldsymbol{x}^o), \tag{2.5}$$

pour $l \in M(\boldsymbol{x})$. En section 2.3, nous verrons que l'expression (2.4) généralise celle de l'estimateur classique de Nadaraya -Watson (cf. annexe A), sous certaines hypothèses.

2.1.4 Données imprécises ou floues

L'un des intérêts de cette méthode est de pouvoir utiliser des données de type flou. Le résultat est immédiat. Si $\boldsymbol{x}$ est flou, ce que l'on note $\widetilde{\boldsymbol{x}} = (\widetilde{\boldsymbol{x}}^o, \widetilde{\boldsymbol{x}}^m)$, le degré de déclenchement de r_i ou r_{il} s'écrit:

$$\tau_i = \bigwedge_{j \in O(x)} \mathrm{Poss}(Z_{ij} | \widetilde{x}_j),$$

où $\mathrm{Poss}(Z_{ij}|\widetilde{x}_j) \triangleq \bigvee_{u_j \in \mathcal{X}_j} Z_{ij}(u_j) \wedge \widetilde{x}_j(u_j)$ est le degré de possibilité de Z_{ij} sachant $\widetilde{x}_j$. On peut exprimer τ_i de manière synthétique:

$$\tau_i = \mathrm{Poss}(Z_i^o | \tilde{\boldsymbol{x}}^o).$$

Il va de soi que si l'ensemble $\mathcal{T}$ est lui-même contitué de données floues, l'étape de fuzzification devient inutile.

2.1.5 Base d'apprentissage incomplète

Si l'ensemble d'apprentissage est lui-même incomplet, chaque $\boldsymbol{x}_i$ se décompose en deux parties, observée et manquante: $\boldsymbol{x}_i = (\boldsymbol{x}_i^{o_i}, \boldsymbol{x}_i^{m_i})$. On ne peut donc plus calculer τ_i, car certaines composantes Z_{ij} de Z_i^o peuvent ne pas être définies. Plusieurs stratégies sont possibles.

Dans une première approche, on ne remplace pas les données manquantes $\boldsymbol{x}_i^m$ et on modifie la règle r_{il} en conséquence. Pour cela, on se base uniquement sur les variables observées communes de $\boldsymbol{x}$ et $\boldsymbol{x}_i$. Par exemple, soit un vecteur de trois variables, $\boldsymbol{x} = \{x_1, x_2, x_3\}$ dont la variable x_3 est manquante. La variable x_{i1} du vecteur $\boldsymbol{x}_i$ est également manquante. On ne s'appuie donc que sur la deuxième variable pour reconstruire x_3. Soient $\mathcal{X}^{o,o_i} = \mathcal{X}^o \cap \mathcal{X}^{o_i}$, l'espace des variables observées simulanément pour $\boldsymbol{x}$ et $\boldsymbol{x}_i$, et $\boldsymbol{x}^{o,o_i}$, la restriction de $\boldsymbol{x}^o$ à cet espace. On obtient donc une nouvelle règle r'_{il}:

$$r'_{il}: \quad \text{SI } \boldsymbol{x}^{o,o_i} \text{ est } Z_i^{o,o_i} \text{ ALORS } \boldsymbol{x}_l \text{ est } Z_{il} \tag{2.6}$$

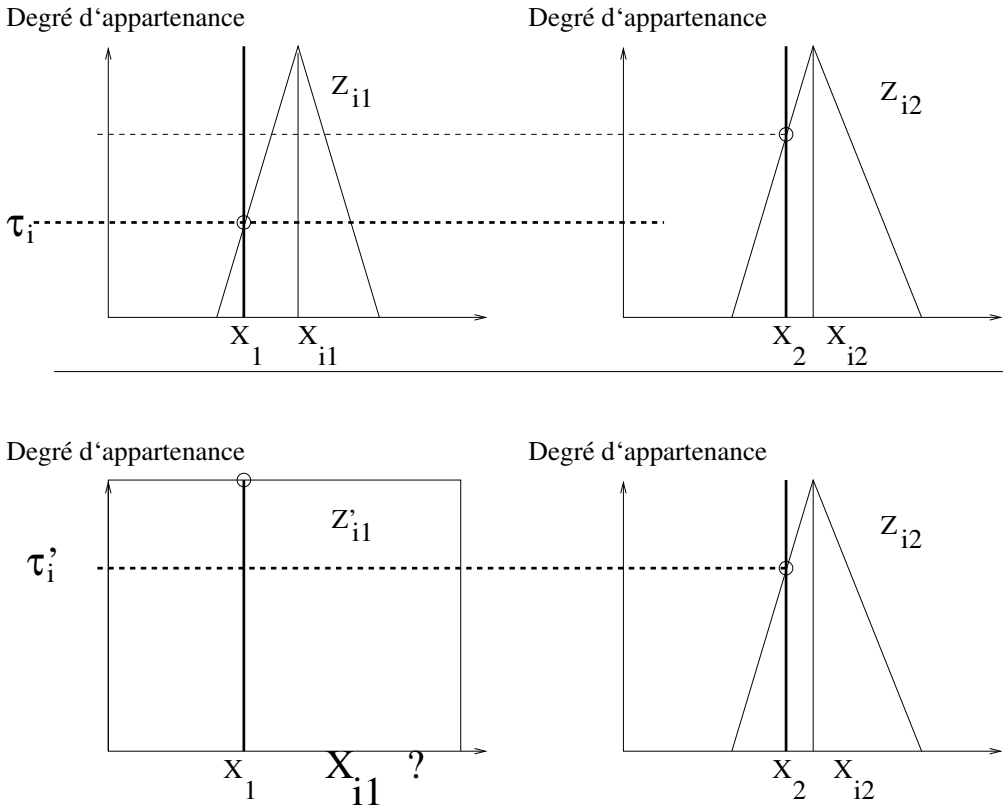


FIG. 2.2 – *Existence de données manquantes dans l'ensemble d'apprentissage. Dans cet exemple,* $\boldsymbol{x}^o = (x_1, x_2)$. **En haut**, *les variables* x_{i1} *et* x_{i2} *sont connues. Elles sont représentées par deux ensembles flous triangulaires* Z_{i1} *et* Z_{i2}. *On obtient* $\tau_i = min(Z_{i1}(x_1), Z_{i2}(x_2))$. **En bas**, x_{i1}, *maintenant indisponible, est représentée par l'ensemble de référence* $Z'_{i1} = \mathcal{X}_1$. *Alors,* $\tau'_i = Z_{i2}(x_2) \geq \tau_i$.

où Z_i^{o,o_i} est la restriction de Z_i^o à $\mathcal{X}^{o,o_i}$. On note τ'_i le degré de déclenchement de r'_{il}. L'inconvénient est que ce degré de déclenchement est plus élevé que le degré τ_i que l'on a obtenu lorsque toutes les composantes de $\boldsymbol{x}_i$ étaient connues! En effet,

$$\tau'_i = Z_i^{o,o_i}(\boldsymbol{x}^{o,o_i}) \geq Z_i^o(\boldsymbol{x}^o) = \tau_i.$$

Par conséquent, on obtient le résultat aberrant suivant : plus le nombre de coordonnées manquantes de $\boldsymbol{x}_i$ est grand, plus son influence sera importante![2]

On peut remarquer que cette transformation revient à conserver la règle initiale r_{il} en définissant l'ensemble flou Z_{ij} associé à la valeur manquante x_{ij} comme un ensemble particulier, l'ensemble de référence : $Z_{ij} = \mathcal{X}_j$. Cette représentation correspond en effet à une ignorance totale sur la valeur de x_{ij}, c'est-à-dire à une imprécision « maximale » (cf figure 2.2).

Le même problème se pose dans certaines méthodes de régression classique, comme les noyaux ou les plus proches voisins, où l'on utilise la distance aux éléments de l'ensemble d'apprentissage. Si le vecteur $\boldsymbol{x}_i$ est incomplet, doit-on définir la distance à $\boldsymbol{x}_i$ sur un sous-espace de $\mathcal{X}$?

2. Il est possible de contrôler ce résultat en accordant moins de poids à la règle, c'est-à-dire en faisant intervenir des facteurs de confiance (cf. chapitre 1)

Une autre approche consiste à remplacer d'abord les données incomplètes à l'aide d'une heuristique simple telles que celles définies en annexe B (remplacement par la moyenne, la médiane ou le plus proche voisin). Dans ce cas, on introduit une information qui peut être erronée, selon la qualité de la méthode de reconstruction.

Enfin, si la taille de l'échantillon est assez grande, on peut résumer l'information fournie par $\mathcal{T}$ à l'aide d'un algorithme de classification automatique capable de traiter les données manquantes, comme les centres-mobiles ou les cartes de Kohonen. On obtient alors un ensemble de vecteurs de référence complets qui définissent de nouvelles règles. Nous allons développer cette approche dans la section suivante.

2.2 Utilisation de prototypes

2.2.1 Justification

L'utilisation de prototypes se justifie pour de nombreuses raisons. Tout d'abord, nous l'avons déjà vu, la classification non supervisée est une méthode très classique d'identification d'un système flou quelconque [65]. La présence de données manquantes nécessite cependant quelques adaptations.

Jusqu'à présent, à chaque élément de l'ensemble d'apprentissage $\boldsymbol{x}_i$, nous avons fait correspondre une règle. Lorsque la taille de l'échantillon devient grande, les règles deviennent redondantes, trop nombreuses pour être facilement interprétables et le temps de calcul augmente. Il est donc nécessaire de résumer l'information globale délivrée par $\mathcal{T}$ en recourant par exemple à des méthodes de classification automatique, que $\mathcal{T}$ possède des données manquantes ou non. Rappelons que les algorithmes classiques s'adaptent très bien à l'existence éventuelle de données manquantes, en utilisant l'une des méthodes heuristiques suggérées plus haut (cf. annexe B).

Ces prototypes peuvent être également obtenus indépendamment de $\mathcal{T}$ à l'aide d'experts, dans le cas d'un modèle de type « boîte grise ».

Enfin, nous venons de voir que la construction d'un ensemble de vecteurs de référence complets est une réponse *a priori* pertinente au problème de données manquantes de la base d'apprentissage.

2.2.2 Description de la base de règles

Supposons donc que l'on possède un ensemble de n prototypes $C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$. Chaque prototype $\boldsymbol{c}_k$ est représentatif d'une classe C_k de vecteurs de $\mathcal{X}$. L'espace de référence $\mathcal{X}$ est ainsi partitionné en n classes $(C_k)_{k=1}^n$.

Si les prototypes fournissent une « bonne représentation » des données, nous pouvons espérer reconstruire les valeurs manquantes d'un vecteur proche d'un prototype donné à l'aide des autres coordonnées, de façon analogue au système décrit par l'équation (2.2).

La base de règles définie précédemment devient donc :

$$r_k : \text{SI } \boldsymbol{x}^o \text{ est } B_k^o \text{ ALORS } \boldsymbol{x}^m \text{ est } B_k^m \quad k = 1, \ldots, n \leq N, \tag{2.7}$$

où les $B_k = B_k^o \times B_k^m$ sont des ensembles flous représentant la proximité aux vecteurs $\boldsymbol{c}_k$

ou, de manière équivalente, le degré d'appartenance à la classe C_k. Chaque coordonnée c_{kj} de $\boldsymbol{c}_k$ sera donc représentée par la projection B_{kj} de B_k sur $\mathcal{X}_j$. Chaque règle traduit les relations entre les composantes d'un vecteur, pour une classe donnée. Les paramètres des B_{kj} peuvent être déterminés à partir des caractéristiques de la classe correspondante. On définit également les règles pour chacune des coordonnées x_l à reconstruire :

$$r_{kl} : \quad \text{SI } \boldsymbol{x}^o \text{ est } B_k^o \text{ ALORS } \boldsymbol{x}_l \text{ est } B_{kl}. \tag{2.8}$$

2.2.3 Construction des classes et des ensembles flous

A chaque vecteur éventuellement incomplet $\boldsymbol{x}_i = (\boldsymbol{x}_i^{o_i}, \boldsymbol{x}_i^{m_i}) \in \mathcal{T}$, on associe son plus proche prototype $\boldsymbol{c}(\boldsymbol{x}_i)$, selon une certaine distance. On supposera ici pour simplifier que les données sont réelles, non floues. Alors

$$\boldsymbol{c}(\boldsymbol{x}_i) = \arg\min_{c \in C} \|\boldsymbol{x}_i^{o_i} - \text{Proj}_{\mathcal{X}^{o_i}} \boldsymbol{c}\|.$$

On notera $\boldsymbol{c}_k^o$, la projection de $\boldsymbol{c}_k$ sur $\mathcal{X}^o$. Pour chaque $k \in \{1, \ldots, n\}$, la classe C_k est définie par l'ensemble des vecteurs de $\mathcal{T}$ associés à $\boldsymbol{c}_k$: $C_k = \{\boldsymbol{x} \in \mathcal{T}, \boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{c}_k\}$.

Pour chaque $k \in \{1, \ldots, n\}$, l'ensemble flou B_{kj} représentant la variable j de la classe C_k peut être défini par deux paramètres représentant ses caractéristiques :

- c_{kj}, ou $\mu_{kj} = \frac{1}{|C_{kj}|} \sum_{x_{ij} \in C_{kj}} x_{ij}$, la moyenne de la j^{eme} composante de C_k où $C_{kj} \subset C_k$ représente l'ensemble des vecteurs de C_k pour lesquels la variable x_j est connue. Selon l'algorithme de classification, c_{kj} et μ_{kj} peuvent ne pas être égaux. C'est le cas par exemple quand on remplace les coordonnées manquantes des x_{ij} par la moyenne globale $\overline{x}_i$ de la variable.

- σ_{kj}, l'écart-type de la j^{eme} composante de la classe C_k,

$$\sigma_{kj} = \left(\frac{1}{|C_{kj}|} \sum_{x_{ij} \in C_{ij}} (x_{ij} - c_{kj})^2 \right)^{1/2}.$$

  Là encore, on peut remplacer c_{kj} par μ_{kj}.

Si l'estimation est correcte, ces paramètres ont une signification évidente. Plus σ_{kj} est grand, plus l'incertitude sur la composante j d'un vecteur proche du prototype $\boldsymbol{c}_k$ est grande. La forme des fonctions d'appartenance est plus ou moins arbitraire, par exemple :

- gaussienne, avec $B_{kj}(u_k) = \text{Gauss}(u_k, c_{kj}, \sigma_{kj})$,
- triangulaire symétrique, avec $B_{kj}(u_k) = \text{Tri}\left(u_k, c_{kj} - \sigma_{kj}\sqrt{2\pi}, c_{kj}, c_{kj} + \sigma_{kj}\sqrt{2\pi}\right)$.

On peut aussi définir des B-splines par interpolation sur les $c_{kj}, k = 1, \ldots, n$. Dans ce cas, on ne tient pas compte de la dispersion à l'intérieur des classes, mais uniquement de la position des prototypes $\boldsymbol{c}_k$ par rapport à leurs voisins.

Pour des raisons déjà évoquées, nous choisissons de préférence les fonctions d'appartenance gaussiennes normalisées. En effet, outre l'interprétation concrète des paramètres, les gaussiennes possèdent les propriétés suivantes : le domaine est totalement recouvert par la fonction

($supp(B_{kj}) = \mathcal{X}_j$). Si tel n'est pas le cas, pour les points $\boldsymbol{x}$ aberrants ou simplement éloignés des prototypes, la situation indésirable suivante peut se produire : pour tout $k, \tau_k(\boldsymbol{x}) = 0$. Dans ce cas, l'ensemble flou F_l déterminé par le système est l'ensemble vide (cf. paragraphe suivant). Une valeur ponctuelle ne peut alors être définie que par convention.

2.2.4 Estimation des données manquantes

Si nous utilisons la règle d'inférence *somme-produit* les expressions de τ_k et de la sortie floue F_l, correspondant à $\boldsymbol{x}^o$ s'écrivent comme dans les équations (1.21 et 1.24) du chapitre précédent :

$$\tau_k(\boldsymbol{x}^o) = \exp\left\{-\frac{1}{2}\sum_{j\in O(x)}\left(\frac{x_j - c_{kj}}{\sigma_{kj}}\right)^2\right\}, \tag{2.9}$$

$$F_l(x_l|\boldsymbol{x}^o) = \sum_{k=1}^{n}\tau_k(\boldsymbol{x}^o)B_{kl}(x_l) \quad \forall l \in M(\boldsymbol{x}), \tag{2.10}$$

et l'estimation ponctuelle $\widehat{x}_l$ est définie par :

$$\forall l \in M(\boldsymbol{x}) \quad \widehat{x}_l(\boldsymbol{x}^o) = \frac{\sum_{k=1}^{n}\tau_k(\boldsymbol{x}^o)c_{kl}\sigma_{kl}}{\sum_{k=1}^{n}\tau_k(\boldsymbol{x}^o)\sigma_{kl}}, \tag{2.11}$$

le centre de gravité et l'aire de B_{kl} étant respectivement c_{kl} et $\sqrt{2\pi}\sigma_{kl}$. Nous rappelons que d'autres mécanismes d'inférence peuvent être proposés, en fonction des choix des fonctions d'appartenance et des opérateurs d'inférence. La figure 2.3 en donne un exemple.

Si on note

$$v_{kl}(\boldsymbol{x}^o) = \frac{\tau_k(\boldsymbol{x}^o)\sigma_{kl}}{\sum_{q=1}^{n}\tau_q(\boldsymbol{x}^o)\sigma_{ql}},$$

$$\forall l \in M(\boldsymbol{x}^o), \quad \widehat{x}_l = \sum_{k=1}^{n}v_{kl}(\boldsymbol{x}^o)c_{kl} \quad \text{avec} \sum_{k=1}^{n}v_{kl}(\boldsymbol{x}^o) = 1. \tag{2.12}$$

Cette sortie peut s'interpréter de la façon suivante. La sortie estimée est combinaison linéaire des coordonnées correspondantes des prototypes. Le poids v_{kl} reflète l'influence de la règle r_k en termes de distance entre le vecteur $\boldsymbol{x}$ et le prototype considéré sur l'espace des données connues de $\boldsymbol{x}$. Si on se place dans un contexte probabiliste et non flou, on peut voir le calcul de cette sortie ponctuelle comme une légère modification de la méthode des fonctions de base radiale à noyau gaussien (cf. section 2.3).

En effet, soit G_{kl} la fonction de base définie par :

$$G_{kl}\left(\|\boldsymbol{x}^o - \boldsymbol{c}_k^o\|\right) = \frac{\tau_k(\boldsymbol{x}^o)\sigma_{kl}}{\sum_{q=1}^{n}\tau_q(\boldsymbol{x}^o)\sigma_{kq}}.$$

Alors

$$\widehat{x}_l(\boldsymbol{x}^o) = \widehat{f}_l(\boldsymbol{x}^o) = \sum_{k=1}^{n}G_{kl}\left(\|\boldsymbol{x}^o - \boldsymbol{c}_k^o\|\right)c_{kl}. \tag{2.13}$$

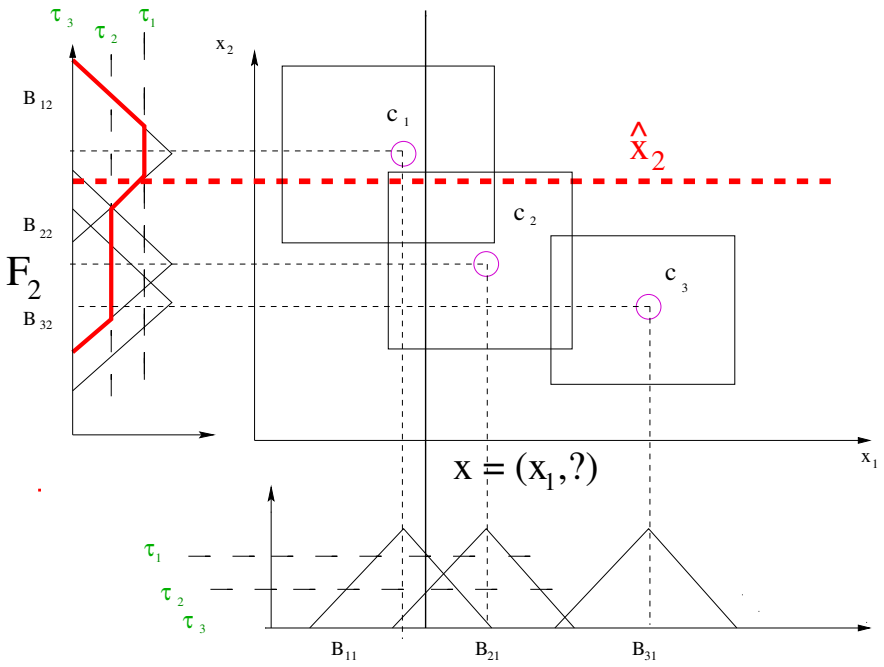


FIG. 2.3 – *Illustration du mécanisme d'inférence en deux dimensions, avec des fonctions d'appartenance B_{kj} triangulaires, les opérateurs d'inférence $\vee = max$ et $\wedge = min$. La variable $x^m = x_2$, inconnue, est estimée par la sortie floue F_2 (en gras) et la sortie ponctuelle $\widehat{x}_2$ (–, en gras).*

La fonction précédente $\widehat{f}_l$, pour un $\boldsymbol{x}$ donné, combinaison linéaire de fonctions de base radiale, peut donc être considérée comme une fonction de base radiale généralisée car les poids c_{kl} sont indépendants des fonctions G_{kl}.

2.2.5 Représentation neuronale

Cependant, les prédicteurs des fonctions $\widehat{f}_l$ étant les variables observées $x_j, j \in O(\boldsymbol{x})$, ils peuvent donc changer en fonction du vecteur en présence $\boldsymbol{x}$. De nombreux cas de figure différents peuvent se présenter selon l'ensemble $O(\boldsymbol{x})$ des valeurs connues. Plus précisément, chacune des r variables peut être estimée à l'aide d'une combinaison quelconque des $r-1$ autres variables, selon leur disponibilité éventuelle. On obtient donc un total de

$$r(2^{r-1} - 1)$$

configurations possibles.

Pour chacune de ces configurations, d'après les expressions (2.11) ou (2.13) et les résultats du chapitre précédent, on sait que l'on peut représenter notre modèle comme un réseau de neurones. Mais il est également possible de synthétiser l'ensemble des configurations possibles en *un seul* réseau de neurones, auto-associatif, où la première couche sépare les cellules dites actives $x_j, j \in O$ des autres, dites cellules inactives (cf. figure 2.4). Ce réseau n'est pas tout à fait à fonction de base radiale, car les poids c_{kl} de la troisième couche *ne sont pas indépendants* de ceux de la première couche. Ils ne peuvent donc pas être optimisés indépendamment les uns des autres. Afin de le distinguer des réseaux classiques, nous définissons ce modèle comme un « réseau à fonction de base quasi-radiale auto-associatif ». Notre modèle fait donc partie

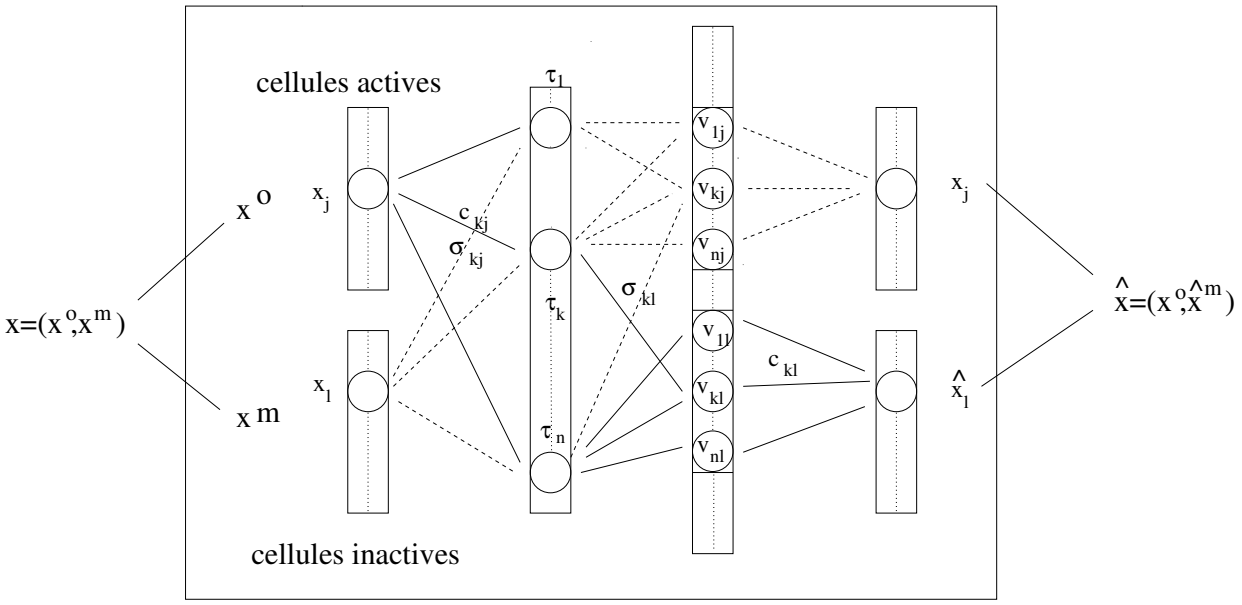


FIG. 2.4 – *Représentation du système flou par un réseau de neurones à 3 couches à fonction de base « quasi »-radiale. Les deux premières couches calculent les degrés de déclenchement des règles τ_k et leur normalisation v_{kl}. La dernière couche calcule les sorties finales $\widehat{x}_l$, $l \in O(\boldsymbol{x})$. Les poids c_{kj} interviennent dans la première et la troisième couches.*

de la classe des *systèmes neuro-flous standards*. Nous pourrons ainsi exploiter les avantages classiques des réseaux de neurones, en particulier l'apprentissage des paramètres (cf. section 2.4). Notre méthode se rapproche donc des méthodes d'estimation de données incomplètes par régression, mais la présentation de notre méthode sous forme d'un système neuro-flou permet d'intégrer dans un même modèle l'estimation de *toutes* les données manquantes d'un vecteur. L'utilisation d'une technique de régression standard nécessiterait la construction de $r(2^{r-1}-1)$ modèles différents afin de tenir compte des différentes situations possibles.

2.3 Comparaison avec la régression classique

2.3.1 Caractéristiques de la sortie floue

L'information fournie par la sortie $F_l(.|\boldsymbol{x})$ est multiple, plus riche que celle de la simple valeur ponctuelle $\widehat{x}_l$. Ses caractéristiques principales nous renseignent sur la distribution des valeurs possibles dont $\widehat{x}_l$ n'est que l'exemple le plus significatif, selon un certain sens.

Degré de confiance global

La hauteur de F_l définit la confiance *globale* que l'on a dans la sortie. Elle dépend essentiellement de la distance entre le vecteur $\boldsymbol{x}$ et les prototypes $\boldsymbol{c}_k$ (ou les vecteurs d'apprentissage $\boldsymbol{x}_i$), c'est-à-dire, de la valeur des τ_k. On supposera ici que $\vee$ est l'opérateur max. Dans ce cas, $h(F_l)$ est comprise entre 0 et 1 et F_l est en général sous-normalisé. Quand $\boldsymbol{x}$ s'éloigne

de l'ensemble d'apprentissage, la valeur des τ_k tend vers 0,

$$\forall x_l \in \mathcal{X}_l, \quad F(x_l) = \bigvee_{k=1}^{n} \tau_k B_{kl}(x_l) \to 0,$$

et donc la hauteur $h(F_l)$ tend vers 0. Dans ce cas, une valeur défuzzifiée $\widehat{x}_l$ n'a que peu d'intérêt. Au contraire, quand $\boldsymbol{x}$ se rapproche de l'un des éléments $\boldsymbol{c}_k$ (ou $\boldsymbol{x}_i$), $h(F_l)$ tend vers 1.

Nonspécificité

L'imprécision sur la sortie peut être quantifiée à l'aide de mesures d'incertitude généralisant la notion d'entropie en probabilités. Parmi les différents types de mesures d'incertitude qui peuvent être définis, la nonspécificité permet de caractériser l'imprécision d'un ensemble flou. La définition la plus répandue dans le cas flou est la suivante :

$$U(F_l) = \frac{1}{h(F_l)} \int_0^{h(F_l)} \log_2 |F_l^\alpha| d\alpha \tag{2.14}$$

où $|F_l^\alpha|$ désigne la mesure de Lebesgue de F_l^α. Cette mesure généralise une mesure définie en théorie des ensembles (classiques) caractérisant l'imprécision, la mesure de Hartley [63].

Normalisations de la sortie floue

A partir de F_l, on peut définir deux types de normalisations : la normalisation « probabiliste » et la normalisation « possibiliste »:

- La normalisation « possibiliste » est la normalisation classique d'un ensemble flou, définissant une mesure de possibilité K_l :

$$K_l(x_l) = \frac{F_l(x_l)}{h(F_l)}.$$

- La normalisation probabiliste transforme F_l en une densité de probabilité H_l :

$$H_l(x_l) = \frac{F_l(x_l)}{\int_{\mathcal{X}_l} F(x_l) dx_l}.$$

Le passage de F_l à K_l ou H_l se traduit par la perte de l'information contenue dans la hauteur de F_l. La normalisation peut cependant être rendue nécessaire par certaines opérations ultérieures sur les ensembles flous.

2.3.2 Analogie avec les modèles de mélange

On s'intéresse uniquement ici à la version de notre méthode utilisant une classification de l'ensemble d'apprentissage. On note $\text{Gauss}^*(.; \boldsymbol{c}, \boldsymbol{\Sigma})$ la densité de probabilité de la loi normale multivariée de moyenne $\boldsymbol{c}$ et de matrice de covariance $\boldsymbol{\Sigma}$ et $\text{Gauss}(.; \boldsymbol{c}, \boldsymbol{\Sigma})$ la fonction d'appartenance de mêmes paramètres, avec la restriction suivante : $\boldsymbol{\Sigma} = (\sigma_j^2)_{j=1}^r$ diagonale.

Puisque l'on suppose que l'espace de représentation des données est divisé en classes, il est possible de traiter ce problème d'estimation de x_l, connaissant $\boldsymbol{x}^o$, par une approche probabiliste basée sur les modèles de mélange. Le principe est le suivant. On suppose que les $(x_l)_{l=1}^r$ sont générés indépendamment, conditionnellement à C_k, par une densité mélange de n gaussiennes :

$$p(x_l) = \sum_{k=1}^{n} p(x_l|C_k)P(C_k),$$

où $P(C_k)$ peut se définir comme la probabilité *a priori* d'appartenance à C_k et $p(x_l|C_k)$ est la densité de probabilité univariée $\text{Gauss}^*(.; c_{kl}, \sigma_{kl})$.

La connaissance de $\boldsymbol{x}^o$ permet de définir la distribution de probabilité conditionnelle Q_l de x_l sachant $\boldsymbol{x}^o$:

$$Q_l(x_l) = p(x_l|\boldsymbol{x}^o) = \sum_{i=1}^{n} p(x_l|C_k)P(C_k|\boldsymbol{x}^o), \tag{2.15}$$

où $P(C_k|\boldsymbol{x}^o)$ est la probabilité *a posteriori* d'appartenance à C_k. On notera $\theta_k(\boldsymbol{x}^o) = P(C_k|\boldsymbol{x}^o)$. Puisque les $(C_k)_{k=1}^n$ forment une partition de $\mathcal{X}$, la somme des probabilités d'appartenance aux classes est égale à 1 : $\sum_{k=1}^n \theta_k(\boldsymbol{x}^o) = 1$. D'après le théorème de Bayes,

$$P(C_k|\boldsymbol{x}^o) = \frac{P(\boldsymbol{x}^o|C_k)P(C_k)}{\sum_{q=1}^K P(\boldsymbol{x}^o|C_q)P(C_q)}.$$

On suppose que les probabilités $P(C_k)$ sont toutes égales à $\frac{1}{K}$. Les densités *a priori* étant gaussiennes, les probabilités $\theta_k(\boldsymbol{x}^o)$ sont définies de façon suivante :

$$\theta_k(\boldsymbol{x}^o) = \frac{\text{Gauss}^*(\boldsymbol{x}^o; \boldsymbol{c}_k^o, \boldsymbol{\Sigma}_k^o)}{\sum_{q=1}^n \text{Gauss}^*(\boldsymbol{x}^o; \boldsymbol{c}_q^o, \boldsymbol{\Sigma}_q^o)}.$$

La quantité Q_l, prise comme une fonction de $\boldsymbol{x}^o$, peut alors être vue comme une fonction de base radiale normalisée (cf. [157, 53]).

Par analogie avec le cas probabiliste, on peut alors définir F_l comme la possibilité conditionnelle de x_l sachant $\boldsymbol{x}^o$. En effet, si on note $\tilde{\boldsymbol{x}}$ la fuzzification « singleton » de $\boldsymbol{x}$, le degré de déclenchement de r_k s'écrit : $\tau_k = \text{Poss}(B_k^o|\tilde{\boldsymbol{x}}^o)$. De même, $B_{kl}(x_l) = \text{Poss}(B_{kl}|\tilde{x}_l)$. Par symétrie, on a également : $B_{kl}(x_l) = \text{Poss}(\tilde{x}_l|B_{kl})$.

On peut alors écrire F_l sous la forme :

$$F_l(x_l) = \sum_{k=1}^{n} \text{Poss}(\tilde{x}_l|B_{kl})\text{Poss}(B_k^o|\tilde{\boldsymbol{x}}^o),$$

et on peut exprimer F_l de façon suivante :

$$F_l(x_l) = \text{Poss}(\tilde{x}_l|\tilde{\boldsymbol{x}}^o)$$

La quantité F_l est donc l'analogue possibiliste de Q_l.

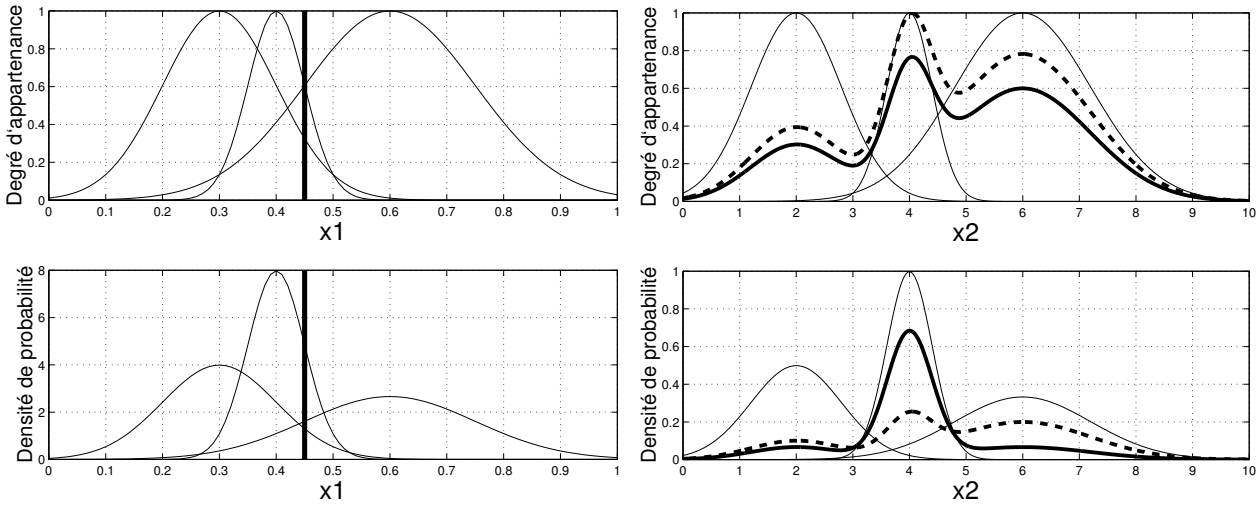


FIG. 2.5 – *Mélange de trois classes gaussiennes bidimensionnelles. Comparaison entre les points de vue probabiliste et possibiliste. Les sorties sont calculées pour* $x_1 = 0.45$ *en utilisant les modèles de mélanges et notre méthode.* **A gauche, en haut** *: fonctions d'appartenance* $Gauss(.; c_{k1}, \sigma_{k1})$ *sur* $\mathcal{X}_1$*. Les* τ_k *sont directement lisibles :* $\tau_1 = 0.35$*,* $\tau_2 = 0.6$*,* $\tau_3 = 0.6$*.* **En bas** *: densités de probabilité correspondantes* $Gauss^*(,; c_{k1}, \sigma_{k1})$*. Les* θ_k *sont déterminés par normalisation.* **A droite, en haut** *: fonctions d'appartenance sur* $\mathcal{X}_2$*, sortie floue* F_l *(-, en gras) et sa normalisation « possibiliste »* K_l *(- -).* **En bas***, densités a priori sur* $\mathcal{X}_2$*, normalisation « probabiliste »* H_l *(- -, en gras) de* F_l*, et la densité* Q_l *(-, en gras). Le pic de la deuxième classe dans* Q_l *n'apparaît pas dans* H_l*. Les deux courbes sont ici assez distantes car les variances sont très différentes.*

Exemple.

Dans la figure 2.5, nous avons représenté les sorties F_l, Q_l, H_l et K_l dans le cas de trois classes gaussiennes bi-dimensionnelles, de paramètres :

- $\boldsymbol{c}_1 = \begin{pmatrix} 0.3 \\ 2 \end{pmatrix}, \boldsymbol{\sigma}_1 = \begin{pmatrix} 0.1 \\ 0.8 \end{pmatrix}$;
- $\boldsymbol{c}_2 = \begin{pmatrix} 0.4 \\ 4 \end{pmatrix}, \boldsymbol{\sigma}_2 = \begin{pmatrix} 0.05 \\ 0.4 \end{pmatrix}$;
- $\boldsymbol{c}_3 = \begin{pmatrix} 0.6 \\ 6 \end{pmatrix}, \boldsymbol{\sigma}_3 = \begin{pmatrix} 0.15 \\ 1.2 \end{pmatrix}$.

Pour $x_1 = 0.45$, on obtient $\tau_1 = 0.35$, $\tau_2 = 0.6$, $\tau_3 = 0.6$ et $\theta_1 = 0.13$, $\theta_2 = 0.67$, $\theta_3 = 0.2$.

Les distributions H_l et Q_l peuvent être très différentes quand les variances des fonctions d'appartenance sont elles-mêmes très différentes. Il peut être intéressant de comparer de manière théorique les deux densités de probabilité Q_l et H_l. On peut montrer facilement la proposition suivante :

Proposition. *Les densités H_l et Q_l sont égales si et seulement la condition de proportionalité suivante est vérifiée* :

$$\forall i = 1, \ldots, n \quad \sigma_{il} \prod_{j \in M(x)} \sigma_{ij} = constante,$$

c'est-à-dire, si les classes ont *même volume*. En particulier, le cas où les variances des classes sont identiques pour chaque variable ($\sigma_{ij} = \sigma_{qj}$ pour tout i, j, q) vérifie cette condition. Ce cas particulier important intervient notamment quand on utilise la totalité de l'ensemble d'apprentissage, et non pas un ensemble de prototypes, et que l'on « fuzzifie » arbitrairement les x_{ij} en les remplaçant par des nombres flous gaussiens $\text{Gauss}(.; x_{ij}, \sigma_{ij})$, où σ_{ij} est donc à déterminer.

2.3.3 Etude de la sortie ponctuelle

Modèle complet

Si l'on utilise la totalité de l'ensemble d'apprentissage $\mathcal{T}$ et que les $\boldsymbol{x}_i$ sont fuzzifiés à l'aide de fonctions d'appartenance gaussiennes de variances *identiques*, la sortie défuzzifiée correspond *exactement* au régresseur de Nadaraya-Watson à noyau gaussien (cf. équation (A.6)). En effet, d'après l'équation 2.4, on obtient par un calcul direct, en utilisant la règle d'inférence *somme-produit* :

$$\widehat{x}_l = \sum_{i=1}^{N} w_{il}(\boldsymbol{x}^o) x_{il} \text{ avec } w_{il}(\boldsymbol{x}^o) = \frac{\text{Gauss}^*(\boldsymbol{x}^o; \boldsymbol{x}_i^o, \boldsymbol{\Sigma}_i^o)}{\sum_{q=1}^{N} \text{Gauss}^*(\boldsymbol{x}^o; \boldsymbol{x}_q^o, \boldsymbol{\Sigma}_q^o)}. \tag{2.16}$$

La sortie est donc une combinaison linéaire directe des x_{il}.

Modèle avec prototypes

Si l'on utilise les prototypes c_k, nous avons déjà vu que l'expression de $\widehat{x}_l$ (équations (2.12) et (2.13)) était celle d'une *fonction de base radiale*. Elle peut être comparée à l'expression analogue, obtenue à partir de Q_l :

$$\widehat{x}_l{}' = \sum_{i=1}^{n} \theta_i(\boldsymbol{x}^o) c_{kl},$$

qui est également une fonction de base radiale. Si la condition de proportionalité des écarts-types que nous venons de voir est vérifiée, les quantités $\theta_{kl}(\boldsymbol{x}^o)$ et $v_{kl}(\boldsymbol{x}^o)$ sont égales et les sorties ponctuelles $\widehat{x}_l$ et $\widehat{x}'_l$ aussi.

Analogie avec la fonction de régression

La sortie défuzzifiée de F_l est également définie comme l'espérance de x_l selon la densité de probabilité H_l. De façon générale, la technique de défuzzification par centre de gravité revient en effet à convertir un ensemble flou en probabilité, c'est-à-dire, à le normaliser suivant un certain sens, et à calculer la valeur moyenne selon cette probabilité :

$$\widehat{x}_l = \mathbb{E}_{H_l}(X_l | X^o = \boldsymbol{x}^o),$$

où X_l est vue comme une variable aléatoire de loi H_l. Soit $P_{X_l|X^o}$ la probabilité conditionnelle de x_l sachant $\boldsymbol{x}^o$. La plupart des méthodes de régression classique cherchent à estimer la fonction dite de régression (cf. annexe A) :

$$f(x_l) = \mathbb{E}_{P_{X_l|X^o}}(X_l | X^o = \boldsymbol{x}^o),$$

le problème majeur étant que la probabilité $P_{X_l|X^o}$ n'est en général pas connue. La densité H_l peut se voir comme une estimation de la vraie probabilité conditionnelle $P_{X_l|X^o}$.

2.3.4 Conclusion partielle

On peut faire quelques remarques concernant les différentes quantités introduites dans cette section.

La transformation de F_l en une densité de probabilité H_l, même si sa construction peut paraître artificielle, puisque, dans notre modèle, on ne manipule pas de variable aléatoire, permet de définir un certain nombre de caractéristiques comme des « fractiles », des « intervalles de confiance » sur la valeur estimée. Ces quantités sont cependant à utiliser avec prudence.

Les quantités F_l, H_l, K_l tout comme Q_l, n'étant pas nécessairement convexes, on pourra, si on le désire, remplacer arbitrairement F_l par l'ensemble flou N_l le plus « proche » parmi une famille particulière convexe, normalisée (cf [118]) ou conservant une mesure de nonspécificité. Cependant, cette démarche conduit à une perte d'information dans le cas où la sortie est ambiguë ou multivaluée. Ce type de cas est représenté dans la figure 2.6 où la variable $\boldsymbol{x}^m$ à reconstruire est bimodale pour le vecteur observé $\boldsymbol{x}^o$. Nous étudierons ce type de cas dans les chapitres 4 et 5. Il est simplement utile de remarquer que l'information ponctuelle n'a

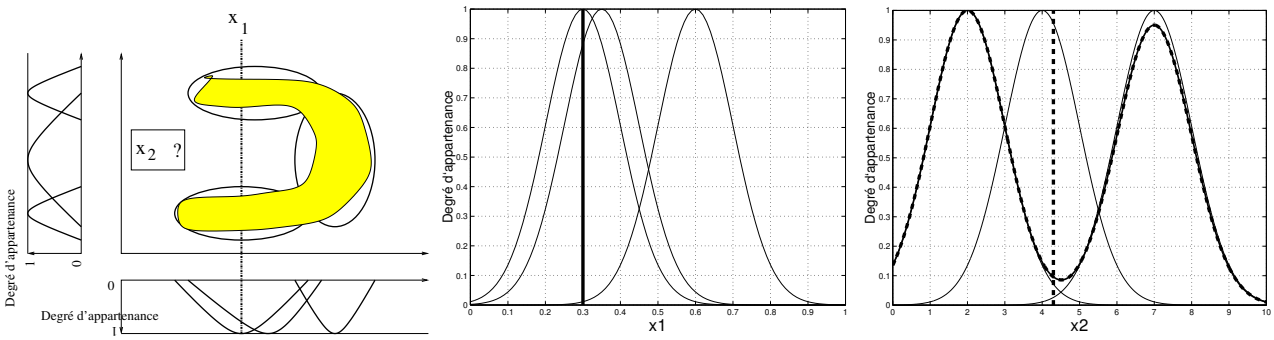


FIG. 2.6 – *Estimation de données multivaluées. Dans cet exemple, la connaissance de la variable x_1 ne suffit pas à déterminer x_2 de façon unique. A gauche, on a représenté trois classes gaussiennes bidimensionnelles et leur projection sur $\mathcal{X}_1$ et $\mathcal{X}_2$. Au milieu, les projections sur $\mathcal{X}_1$. A droite, les projections sur $\mathcal{X}_2$ et l'estimation en $x_1 = 0.3$. L'information fournie par la sortie ponctuelle (- -) n'est pas satisfaisante. La sortie F_l (-, en gras) est bimodale.*

pas de sens ici. La donnée complète de F_l semble nécessaire. L'utilisation des modèles de mélange donne le même type de résultats.

Dans cette section, nous avons considéré que les variables étaient, soit parfaitement connues, soit manquantes. Le cas des valeurs imprécises ou floues n'a pas été traité car le but est ici de comparer les modèles flous et probabilistes. Les modèles probabilistes ne sont en effet pas adaptés pour ce type de données. Le gros avantage des systèmes flous est bien entendu de pouvoir tenir compte également de ce type de valeurs.

Selon l'information que l'on possède sur le vecteur $\boldsymbol{x}$, sur la densité de prototypes au voisinage de $\boldsymbol{x}$ ou le nombre de valeurs manquantes, on aura plus ou moins confiance dans la sortie définie par le système. Le choix des fonctions d'appartenance, d'une part, et l'introduction de poids dans les règles, d'autre part, permettent de mettre en valeur cette information. Nous avons vu que la hauteur $h(F_l)$ donnait une indication de la confiance que l'on a dans la sortie, ce qui représente un avantage par rapport à la modélisation probabiliste.

De ce point de vue, les fonctions d'appartenance du type des B-splines présentent l'inconvénient de ne pas tenir compte de la distance *globale* aux prototypes. On aura toujours : $\sum_k \tau_k = 1$. Or, si un vecteur $\boldsymbol{x}$ est éloigné de l'ensemble d'apprentissage, on aimerait que cette quantité soit très faible : $\sum_i \tau_k(\boldsymbol{x}) << 1$. Cela n'a pas d'importance pour la sortie ponctuelle, mais cela en a pour la sortie floue F_l.

2.4 Identification du modèle

2.4.1 Identification de la structure

Nous rappelons que l'identification d'un système neuro-flou peut se diviser en deux phases : l'identification de la structure du modèle et l'optimisation des paramètres de la structure choisie. Dans la version de notre méthode utilisant la totalité de l'ensemble d'apprentissage, seule la deuxième phase peut s'avérer nécessaire. Dans cette section, nous nous intéressons au modèle défini par classification de l'ensemble d'apprentissage, pour lequel la phase d'identifi-

cation de la structure se résume alors au choix de la taille du modèle, c'est-à-dire du nombre de règles ou de prototypes n. Le problème du choix des variables intervenant dans la structure est un problème d'extraction de caractéristiques, en amont de la chaîne de traitement, que l'on suppose résolu ici.

Comme nous l'avons vu dans le chapitre précédent, différents types de méthodes d'identification peuvent être envisagées : en particulier, les méthodes basées sur la minimisation du coût empirique, éventuellement pénalisé, et les méthodes basées sur des critères de classification. Ces deux types de méthodes étant parfois lourds à mettre en œuvre, nous avons envisagé également plusieurs heuristiques afin de limiter le temps de calcul, comme des méthodes basées sur la suppression ou au contraire la création de prototypes [120, 93]. Ainsi, dans une approche constructive [120], pendant la phase d'apprentissage, un vecteur que l'on considère trop loin des prototypes existants, selon une certaine distance, constitue une nouvelle unité. Dans [117], nous avons au contraire utilisé une approche destructive, très simple, basée sur l'élimination des prototypes les moins représentatifs. Si la distance entre deux prototypes $\boldsymbol{c}_k$ et $\boldsymbol{c}_j$ est inférieure à une constante δ donnée, ils sont alors remplacés par leur centre de gravité $\boldsymbol{c}_0$:

$$\boldsymbol{c}_0 = \frac{|C_k|\boldsymbol{c}_k + |C_j|\boldsymbol{c}_j}{|C_k| + |C_j|}.$$

Cette approche peut se voir comme une méthode de classification ascendante hiérarchique [165].

2.4.2 Estimation des paramètres

Une fois le nombre de règles choisies, et les paramètres σ_{kj} et c_{kj} du modèle initialisés par une technique de classification non supervisée, la représentation neuronale de notre système (cf. figure 2.4) permet d'ajuster plus finement ces paramètres.

Le problème devient donc le suivant : pour un vecteur $(\boldsymbol{x}^o, \boldsymbol{x}^m)$, nous cherchons à minimiser le coût quadratique moyen par variable $x_j, j \in O(\boldsymbol{x})$, pouvant être reconstruite :

$$J = \frac{1}{2|O(\boldsymbol{x})|} \sum_{l \in O(x)} (\widehat{x}_l - x_l)^2 \tag{2.17}$$

par rapport à σ_{kj} and c_{kj}, chaque x_l étant estimé à l'aide des autres variables connues. Il s'agit d'un apprentissage partiel ou semi-supervisé, puisque les $\boldsymbol{x}^m$ ne sont pas connues. Parmi les différents algorithmes d'optimisation présentés dans le chapitre précédent, en section 1.3.2, nous avons choisi la méthode classique de descente du gradient du coût, à pas adaptatif.

Les paramètres initiaux $c_{kj}(0)$ et $\sigma_{kj}(0)$ sont les paramètres de la partition initiale.

A chaque itération t, nous calculons, pour tout $i \in \{1, \ldots, n\}$ et tout $j \in \{1 \ldots, r\}$:

$$c_{kj}(t+1) = c_{kj}(t) - \eta(t)\frac{\partial J(t)}{\partial c_{kj}},$$
$$\sigma_{kj}(t+1) = \sigma_{kj}(t) - \eta(t)\frac{\partial J(t)}{\partial \sigma_{kj}},$$

le pas $\eta(t)$ dépendant de l'itération t. Par un calcul direct, dont le détail est donné en annexe C, nous obtenons les formules suivantes :

$$\frac{\partial J}{\partial c_{kj}} = \begin{cases} \dfrac{1}{|O(\boldsymbol{x})|} \left\{ \dfrac{x_j - c_{kj}}{(\sigma_{kj})^2} \displaystyle\sum_{l \in O(x)\setminus\{j\}} (\widehat{x}_l - x_l)(c_{kl} - \widehat{x}_l) v_{kl} + v_{kj}(\widehat{x}_j - x_j) \right\} \\ \text{si } j \in 0(\boldsymbol{x}) \\ \dfrac{\partial J}{\partial c_{kj}} = 0 \text{ sinon} \end{cases} \tag{2.18}$$

$$\frac{\partial J}{\partial \sigma_{kj}} = \begin{cases} \dfrac{1}{|O(\boldsymbol{x})|} \left\{ \dfrac{(x_j - c_{kj})^2}{(\sigma_{kj})^3} \displaystyle\sum_{l \in O(x)\setminus\{j\}} (\widehat{x}_l - x_l)(c_{kl} - \widehat{x}_l) v_{kl} + \frac{v_{kj}}{\sigma_{kj}} (\widehat{x}_j - x_j)(c_{ij} - \widehat{x}_j) \right\} \\ \text{si } j \in 0(\boldsymbol{x}) \\ \dfrac{\partial J}{\partial \sigma_{kj}} = 0 \text{ sinon.} \end{cases} \tag{2.19}$$

L'algorithme s'arrête à l'itération T, quand la diminution de la fonction d'erreur $J(t)$ devient inférieure à un seuil fixé. L'algorithme de rétropropagation du gradient de l'erreur optimise localement les paramètres. Le problème de l'initialisation des paramètres est donc crucial. Il est nécessaire d'avoir une information *a priori* raisonnable sur les ensembles flous intervenant dans le système. Dans notre méthode comme dans de nombreux cas, cette information est fournie par la classification non supervisée de l'ensemble d'apprentissage.

Si les entrées ou les poids sont flous, notre modèle fait alors partie de la classe des réseaux de neurones fuzzifiés présentés dans le chapitre précédent. Il est alors possible d'adapter l'algorithme de rétropropagation à ce nouveau problème en entreprenant une démarche équivalente à celle proposée dans [73].

2.4.3 Détection de données aberrantes

Comme nous l'avons déjà souligné, la notion de donnée « aberrante » est intimement liée à celle de donnée manquante [8, 26], bien que ces deux problèmes soit rarement traités simultanément. En effet, une méthode de détection de données aberrantes consiste souvent à éliminer la donnée suspecte, c'est-à-dire créer une donnée manquante et à comparer cette donnée à l'estimation obtenue. Nous proposons d'appliquer ce principe à notre méthode d'estimation.

Il est facile d'adapter notre algorithme au problème de validation multi-capteur. Nous proposons ici de reconstruire les données *connues* $x_l, l \in O(\boldsymbol{x})$ d'un vecteur $\boldsymbol{x}$ en utilisant les autres données connues $\boldsymbol{x}^o_{-l}$. Si les relations entre les variables sont de bonne qualité, on pourra par exemple décider d'invalider la donnée réelle si elle est « très différente » de son estimation. Les règles deviennent donc :

$$\forall l \in O(\boldsymbol{x}) \quad \text{Si } \boldsymbol{x}^o_{-l} \text{ est } B^o_{k,-l} \text{ alors } x_l \text{ est } B_{kl}$$

où $B^o_{k,-l}$ est le produit Cartésien des $B_{kj}, j \in O(\boldsymbol{x})\setminus l$. La procédure est exactement la même que pour le problème d'estimation de données manquantes. Voici l'expression des différentes

quantités introduites dans les sections précédentes :

1. le degré de déclenchement :

$$\tau_{kl} = \exp\left\{ -\frac{1}{2} \sum_{j \in O(x)\backslash\{l\}} \left(\frac{x_j - c_{kj}}{\sigma_{kj}} \right)^2 \right\},$$

2. la normalisation des τ_{kl} :

$$v_{kl} = \frac{\tau_{kl}\sigma_{kl}}{\sum_{q=1}^{m} \tau_{ql}\sigma_{ql}}$$

3. la sortie floue :

$$F_l(x_l) = \sum_{k=1}^{n} \tau_{kl} B_{kl}(x_l),$$

4. la sortie ponctuelle :

$$\forall l \in O(x)\backslash l \quad \widehat{x_l} = \sum_{k=1}^{n} v_{kl} c_{kl}.$$

Si a_l est une valeur fournie par un capteur, on peut par exemple décider de l'invalider si $F_l(a_l)$ est inférieur à un certain seuil défini par l'utilisateur.

2.5 Application à des données environnementales

2.5.1 Description et prétraitement des données

Description des données

Notre modèle a été appliqué à des données réelles environnementales, dans le cadre du projet européen EM^2S(*Environmental Monitoring and Management System*) [36, 117]. L'un des objectifs du projet était de développer un système de surveillance de la qualité de l'eau des rivières. Le site de Dijon, dont le réseau d'assainissement est géré par Suez-Lyonnaise-des-Eaux, a été choisi comme exemple d'application. Sur ce site, trois stations d'observations fournissent en continu des mesures de paramètres physico-chimiques dont la température de l'eau, le pH, la conductivité, l'oxygène dissout (O_2). Des indicateurs de qualité de l'eau peuvent être construits à partir de ces mesures. Celles-ci sont fournies par des capteurs et doivent subir une phase de prétraitement et de validation avant leur exploitation. Des techniques classiques d'analyse des données et la connaissance *a priori* apportée par des experts concernant la fiabilité et l'intérêt des paramètres nous ont conduit à ne retenir que les quatre variables précitées. Nous disposons d'un échantillon continu de taille $N = 1400$, dont le pas d'échantillonnage, initialement de 6 minutes, a été ramené à 30 minutes à l'aide d'un filtre moyenne mobile classique. Les données ont ensuite subi une phase de validation afin de détecter les valeurs aberrantes.

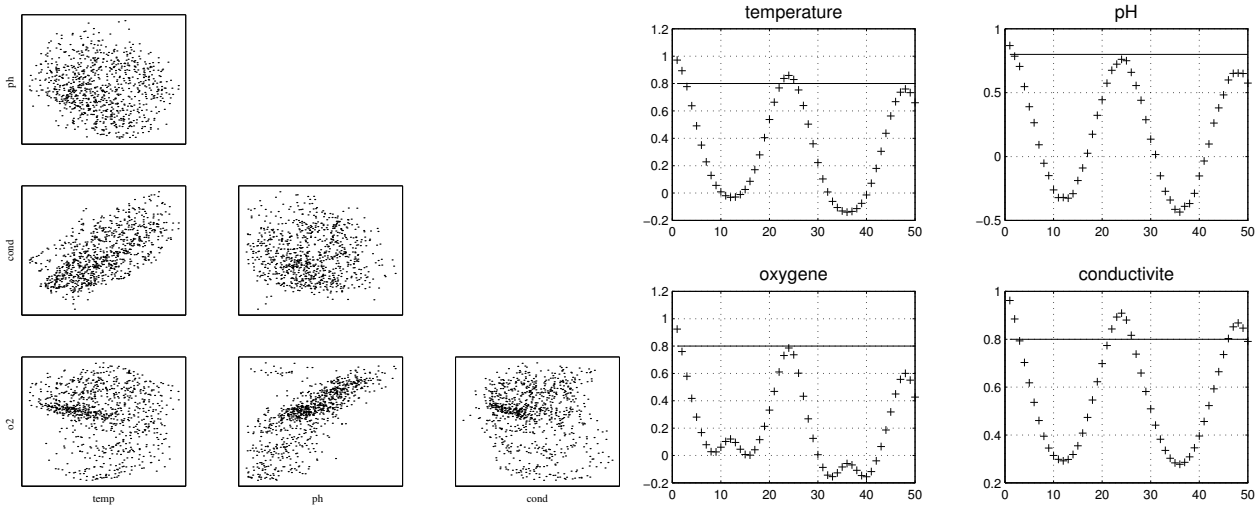


FIG. 2.7 – **A gauche :** *relations entre les variables, prises deux à deux.* **A droite :** *corrélogramme de chacune des variables. Exemple pour la variable pH : en ordonnée, le coefficient de corrélation linéaire $r(t, t-h)$ entre les séries temporelles pH(t) et pH(t-1), selon différentes valeurs du pas de temps h. A titre indicatif, on peut choisir un seuil minimal de « bonne » corrélation, ici égal à 0.8. Les variables sont périodiques, la période (h=24) correspond à une journée.*

Sélection des variables

L'objectif est donc ici d'estimer, à une date t, les valeurs manquantes d'un sous-ensemble quelconque de ces 4 paramètres. La première étape consiste à définir un ou plusieurs modèles adaptés aux données. Notre méthode ne suppose pas nécessairement l'existence de « bonnes » corrélations entre les variables. Néanmoins, la précision de l'estimation en dépend. La figure 2.7, à gauche, qui représente les relations entre les variables prises deux à deux, montre clairement que ce n'est globalement pas le cas. Il serait donc illusoire d'estimer un paramètre manquant uniquement à l'aide des trois autres, bien que l'on distingue des liens entre, d'une part, le pH et O_2, d'autre part, la conductivité et la température. Nous proposons donc d'utiliser également l'historique de la série temporelle de chacune des variables. On note pH(t), $O_2(t)$, cond(t) et temp(t) la valeur des variables à la date t. La figure 2.7, à droite, qui représente le corrélogramme des variables, c'est-à-dire le coefficient d'autocorrélation linéaire selon différents pas de temps h, montre (sans surprise) que l'on peut raisonnablement sélectionner les valeurs d'une variable dans un passé proche, aux temps t-1, t-2 pour la prédire à une date t.

Compte tenu de ce qui précède, plusieurs possibilités s'offrent à nous quant au choix du ou des systèmes neuro-flous adaptés au problème. La première consiste à utiliser un seul système contenant les 4 variables pH(t), $O_2(t)$, cond(t) et temp(t). La deuxième consiste à construire un système flou auto-régressif pour chacune des 4 variables, contenant les variables à des dates différentes, par exemple $O_2(t)$, $O_2(t-1)$,..., $O_2(t-k)$. Notre modèle présente alors des analogies avec des modélisations classiques du type AR(k) (*AutoRégressifs d'ordre k*) [18]. Cependant, lorsqu'une variable est manquante à une date t, les valeurs précédentes, à $t-1, \ldots t-k$ sont bien souvent également indisponibles et rendent ce type de modèles inopérants. La troisième possibilité, que nous avons retenue, combine les aspects multivarié et temporel

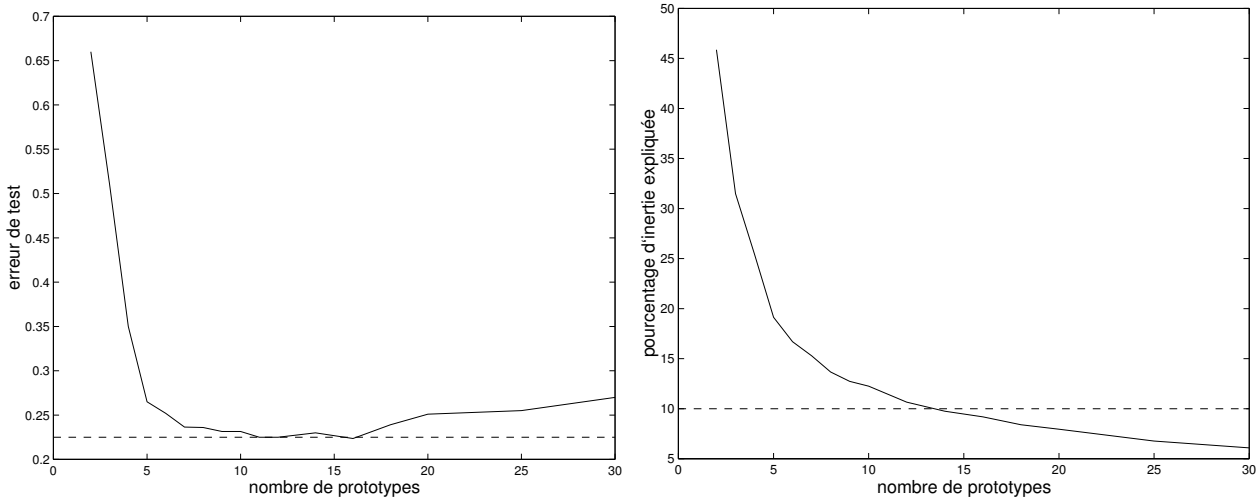


FIG. 2.8 – *Choix du nombre de prototypes de deux façons différentes.* **A gauche***: critère de validation croisée (-), valeur minimale (- -).* **A droite***: critère de partitionnement (pourcentage d'inertie expliquée SSW/SST). Exemple de seuil, SSW/SST=10% (- -)*

du problème. On définit deux systèmes neuro-flous contenant les variables suivantes :

1. $O_2(t)$, $O_2(t-1)$, pH(t), pH(t-1),
2. temp(t), temp(t-1), cond(t), cond(t-1).

Dans la suite, nous présentons essentiellement les résultats relatifs aux variables du système : $O_2(t)$, $O_2(t-1)$, pH(t), pH(t-1). L'intérêt de notre approche est qu'elle s'adapte à chaque pas de temps aux données manquantes du vecteur considéré. A une date t, si O_2 est inconnue, elle sera estimée par les autres paramètres. Mais si à $t+1$, O_2 est de nouveau connu, on pourra estimer une valeur manquante du pH à l'aide du même modèle.

2.5.2 Résultats

Identification du modèle

Le seul paramètre à régler est la taille du modèle, c'est-à-dire le nombre de prototypes n. Cette phase étant délicate, nous avons confronté les trois types de méthodes d'identification cités précédemment : une méthode de rééchantillonnage, la validation croisée ; une méthode de partitionnement, basée sur la dispersion intra-classes et une heuristique basée sur la suppression de prototypes.

Afin d'atténuer les problèmes de minima locaux, les résultats ont été calculés sur une moyenne de 10 essais (cf. figure 2.8). Pour la méthode de validation croisée, le nombre optimal n^* de prototypes, pour lequel l'estimation de l'erreur quadratique moyenne est la plus faible, est de 11. C'est cette valeur que nous avons choisie. Nous pouvons cependant remarquer que les valeurs de 12 à 16 sont également acceptables. Cependant, en vertu d'un principe de simplicité, il est inutile de multiplier le nombre de paramètres du modèle si le gain n'est pas substantiel. D'autre part, en terme de coût de calculs, l'optimisation ultérieure des paramètres sera plus rapide. Les résultats de la méthode basée sur la dispersion intra-classe

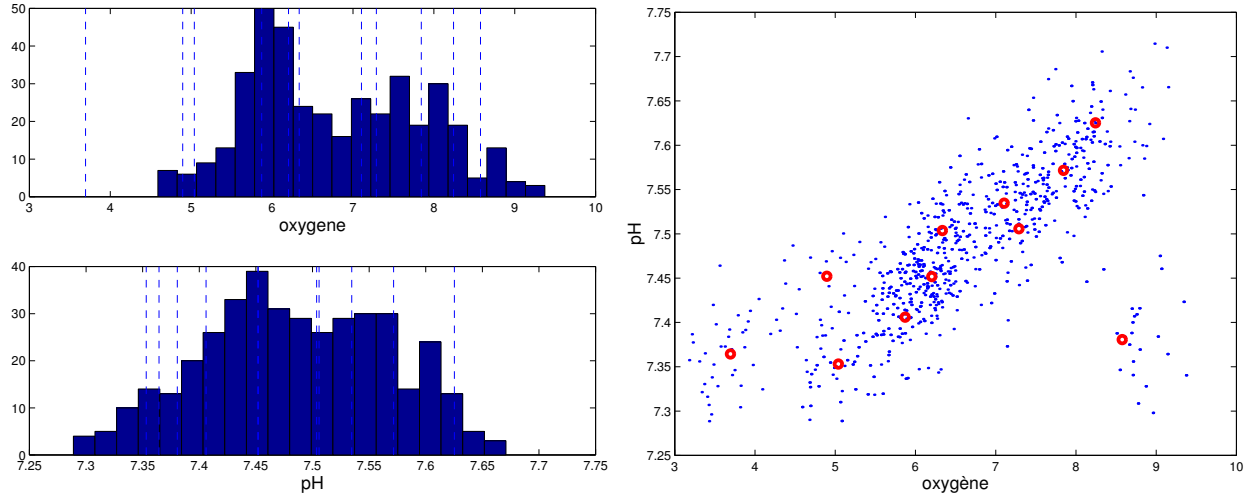


Fig. 2.9 – **A gauche** *: histogramme des variables O_2 et pH. Coordonnées initiales des prototypes (- -).* **A droite** *: ensemble d'apprentissage (.), prototypes des variables pH(t) et $O_2(t)$*

ne semblent pas contrevenir à ce choix (cf. figure 2.8, à droite). L'heuristique présentée dans [118] nécessite de définir la distance minimale δ acceptable entre les prototypes, ce qui dépend du choix de l'utilisateur. Son avantage essentiel est sa simplicité mais elle présente des inconvénients préjudiciables pour la qualité du modèle. En particulier, si le nombre de prototypes initial est très grand, les prototypes résultant d'agrégations successives sont de moins en moins représentatifs des données et la reconstruction se dégrade.

Pour initialiser les paramètres du modèle choisi, à savoir les centres $\boldsymbol{c}_k$ et les écarts-type $\boldsymbol{\sigma}_k$ des classes C_k, il faut d'abord s'assurer de la robustesse de leur estimation. Les points déclarés aberrants de chaque classe ont été éliminés, à l'aide d'un test statistique basé sur un rapport de vraisemblances [8]. Nous nous sommes également assurés que l'effectif de chaque classe était « suffisant » pour une « bonne » approximation de la variance. Cependant, il est inutile de recourir à des méthodes coûteuses dans la mesure où ces paramètres sont par la suite optimisés.

Optimisation des paramètres

La figure 2.9, représente l'histogramme des variables pH et O_2 ainsi que les coordonnées des 11 prototypes initiaux.

Afin d'améliorer la précision des résultats, nous avons optimisé les paramètres à l'aide de l'algorithme de rétropropagation du gradient (cf. équations (2.17) (2.18) et (2.19)). L'ensemble des données a été divisé classiquement en deux parties, l'une pour l'apprentissage des paramètres, l'autre pour le test. La figure 2.10 compare l'évolution de l'erreur quadratique moyenne relative à l'ensemble de test sur une moyenne de 10 essais pour le modèle choisi et le modèle auto-régressif d'ordre 2 selon le nombre de prototypes et pour la variable O_2. Les résultats montrent que l'erreur de généralisation diminue de 20 à 35% après 100 itérations. Pour le modèle choisi, on obtient presque la valeur optimale après seulement une dizaine d'itérations. Si l'on choisit un modèle auto-régressif, l'erreur continue de diminuer sensiblement après 100 itérations. Cela s'explique par le fait que les prédicteurs de ce modèle, O_2(t-1) et O_2(t-2), sont plus fortement corrélées à O_2 que dans l'autre cas. L'apprentissage

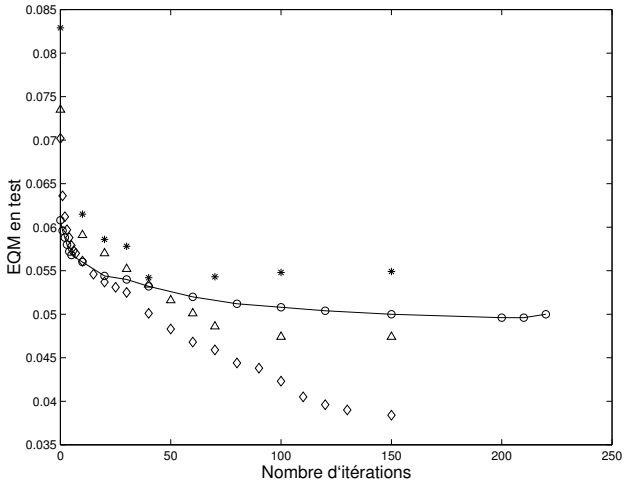


FIG. 2.10 – *Erreur quadratique des paramètres optimaux pour différents modèles : autorégressif d'ordre 2, avec 6 prototypes (△), 11 prototypes (◇) ; modèle mixte $O_2(t)$, $O_2(t-1)$, pH(t), pH(t-1) avec 6 prototypes (*) et 11 prototypes (o-).*

peut ainsi se faire presque « par cœur». Ainsi, le modèle autorégressif semble meilleur, mais suppose la connaissance de la variable aux pas de temps précédents.

Estimation des données manquantes

Estimation ponctuelle

La précision de l'estimation des variables dépend bien entendu des prédicteurs. Comme nous l'avons souligné plus haut, chacune des 4 variables peut-être estimée de $2^3 - 1 = 7$ façons différentes, selon les disponibilités des variables restantes. Nous considérons dans un premier temps le cas *a priori* le plus favorable où chaque variable est estimée à l'aide de toutes les autres. La figure 2.11 présente les résultats de l'estimation des variables O_2 et pH à l'aide de 3 prédicteurs.

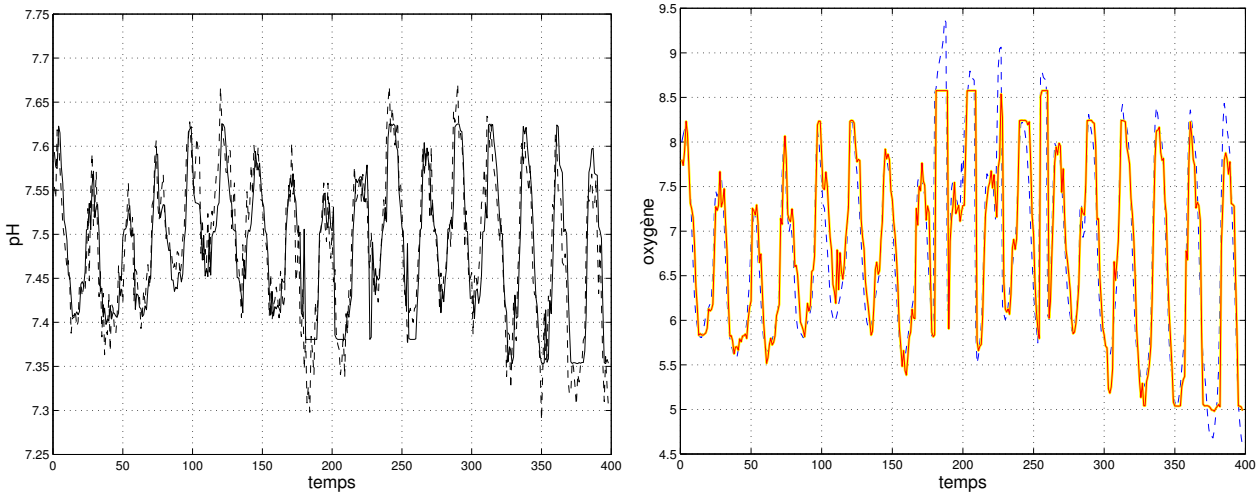


Fig. 2.11 – *Estimation à l'aide de 3 prédicteurs de pH(t) et $O_2(t)$ par notre système neuro-flou (11 prototypes). Prédicteurs de pH(t) : pH(t-1), $O_2(t)$, $O_2(t-1)$; prédicteurs de $O_2(t)$ pH(t), pH(t-1), $O_2(t-1)$. Valeurs estimées (-), vraies valeurs (- -).*

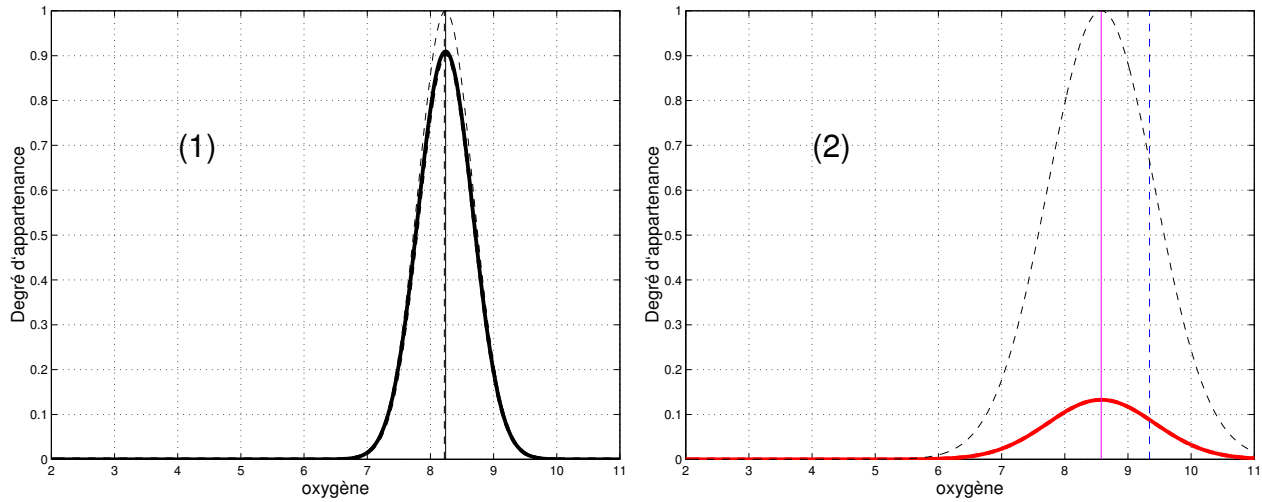


Fig. 2.12 – *Exemples de sortie F_1 (-, en gras) et ses normalisations H_1 (-.-), K_l (–), par la méthode d'inférence somme-produit, sortie floue obtenue par l'inférence max-produit (- -, en gras). Estimation ponctuelle (-) et valeur réelle (- -).* **A gauche***: cas (1), sortie « précise »;* **à droite***: cas (2), sortie « imprécise ».*

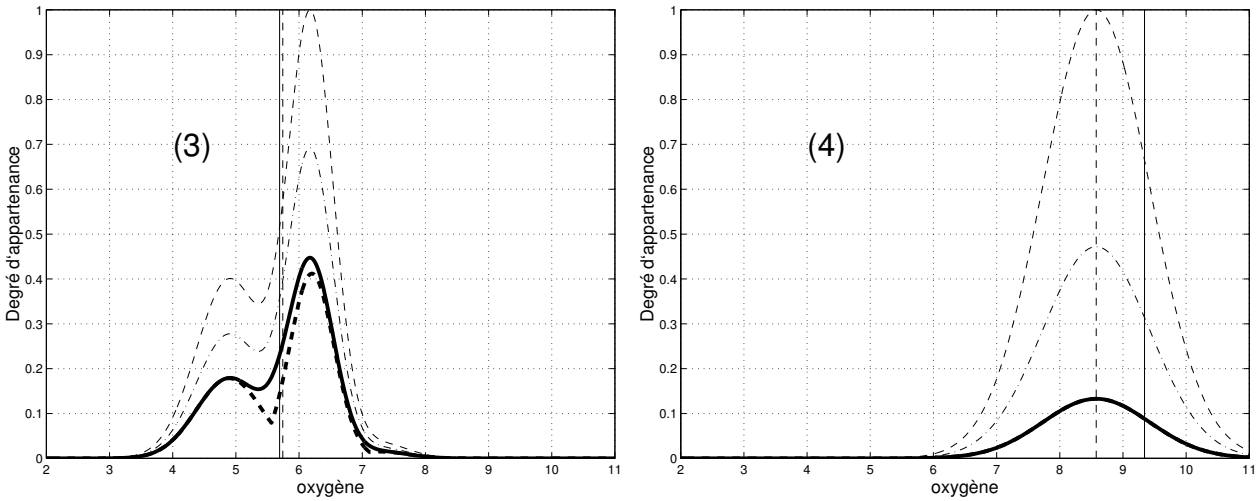


FIG. 2.13 – *Même légende.* **A gauche***: cas (3), sortie ambigüe, bimodale, provenant de prédicteurs conflictuels ;* **à droite***: cas (4), sortie peu fiable, les prédicteurs sont éloignés des prototypes.*

Afin de juger de la précision de notre méthode, nous l'avons comparée à une méthode très simple, basée sur le plus proche prototype (1-ppv), ainsi que trois méthodes de régression : la régression linéaire, la régression spline gaussienne, la méthode des noyaux (cf. équation (A.6)). La méthode du plus proche prototype (cf. annexe B) consiste simplement à sélectionner le plus proche prototype $\boldsymbol{c}_k$ sur le sous-espace des valeurs connues et remplacer les valeurs manquantes par les coordonnées correspondantes du prototype choisi. Pour les deux dernières méthodes de régression, afin de faciliter les comparaisons, nous avons utilisé les prototypes $\boldsymbol{c}_k$ intervenant dans notre modèle. Pour la méthode linéaire, les paramètres ont été déterminés à l'aide de la totalité de l'ensemble d'apprentissage. Pour toutes ces méthodes de régression, il est nécessaire de *déterminer les* 28 *modèles possibles (7 modèles pour chacune des* 4 *variables*), c'est-à-dire, d'identifier la structure et d'estimer les paramètres pour *chaque* modèle. Le tableau 2.1 montre les résultats des différents modèles pour l'estimation de la variable O_2. Les résultats ponctuels de notre méthode sont sensiblement équivalents à ceux de la régression spline et de la méthode des noyaux, dans tous les cas. La régression linéaire ne fournit évidemment de bons résultats que lorsque la seule variable linéairement corrélée à O_2, O_2(t-1), est présente.

Estimation par ensembles flous

L'estimation ponctuelle ne représente qu'une partie réduite de l'information fournie par notre méthode. Elle est issue de la sortie floue du système, dont les caractéristiques rendent compte de la précision et de la confiance. Nous décrivons la sortie floue F_1 associée à la première variable O_2(t), dans le cas de trois prédicteurs O_2(t-1), pH(t) et pH(t-1). Nous pouvons distinguer quatre situations reflétant le type d'information délivré par ces prédicteurs selon leur position par rapport aux prototypes, illustrés par les figures 2.12 et 2.13. Pour chacune de ces situations, nous avons représenté la sortie F_1 et ses normalisations H_1, K_1, calculées par la méthode d'inférence *somme-produit*, ainsi que la sortie floue obtenue par l'inférence *max-produit*, que nous noterons F_1'. Dans le premier cas, le point $\boldsymbol{x}^o$ représenté par les 3 variables connues est proche d'un prototype donné $\boldsymbol{c}_k$ et la densité de l'ensemble d'apprentissage

prédicteurs	Système flou	Splines	Noyaux	Linéaire	1-ppv
$O_2(t-1)$, pH(t), pH(t-1)	0.060	0.056	0.066	0.073	0.071
$O_2(t-1)$	0.058	0.058	0.069	0.113	0.064
$O_2(t-1)$, pH(t)	0.059	0.059	0.063	0.062	0.072
$O_2(t-1)$, pH(t-1)	0.061	0.059	0.070	0.110	0.079
pH(t),pH(t-1)	0.264	0.240	0.239	0.915	0.441
pH(t)	0.236	0.250	0.262	1.109	.486
pH-1(t-1)	0.250	0.250	0.262	1.466	0.468

TAB. 2.1 – *Erreur quadratique moyenne pour la variable $O_2(t)$ selon différents prédicteurs et différentes méthodes de régression*

autour de $\boldsymbol{x}^o$ est élevée. On obtient des sorties floues F_1 et F_1' précises, de hauteur proche de 1. Dans le deuxième cas, il existe toujours un prototype proche de $\boldsymbol{x}^o$, mais la sortie est moins précise, car la densité autour de ce prototype, représentée par $\boldsymbol{\sigma}_k$, est plus faible. Le troisième cas correspond à une ambiguïté sur la valeur de la sortie. La valeur des prédicteurs ne suffit pas à déterminer celle de 0_2(t). Dans l'exemple de la figure 2.13, à gauche, la valeur de pH(t) est voisine celle de deux prototypes (7.45), pour lesquelles la valeur de 0_2(t) est très différente (4.8 et 6.2). Les sorties F_1 et surtout, F_1', qui utilise l'opérateur discontinu max, sont bimodales, les deux « pics » correspondant approximativement à ces deux valeurs. Les résultats sont encore plus marqués quand on n'utilise pas le prédicteur $0_2(t-1)$, mais uniquement les deux valeurs pH(t) et pH(t-1). Nous n'avons pas présenté ces résultats car la qualité de l'estimation ponctuelle n'était pas satisfaisante (cf. tableau 2.1). Le quatrième cas correspond à une densité très faible autour de $\boldsymbol{x}^o$. Les valeurs des prédicteurs ne correspondent à aucune des situations de référence représentées par les prototypes. Les degrés de déclenchement des règles sont tous peu élevés. La hauteur de la sortie est donc faible et traduit le peu de confiance allouée à cette sortie.

2.6 Conclusion

Nous avons proposé un système neuro-flou pour la reconstruction de variables manquantes ou la validation multi-capteurs. Cette méthode est capable d'exprimer la connaissance acquise à partir de données brutes sous forme de règles d'inférence floues. Le principe est basé sur les relations entre les variables en différentes régions de l'espace de représentation des données. La méthode propose d'estimer toutes les valeurs manquantes d'un vecteur dans un même modèle, quel que soit le nombre de variables disponibles. L'imprécision du résultat est donnée sous forme d'une distribution de possibilité définissant un degré de confiance dans les valeurs de l'espace des variables de sortie.

Comme tout système neuro-flou, notre modèle présente des analogies avec certaines méthodes de régression. Sur le plan purement fonctionnel, l'expression de la valeur ponctuelle de chaque variable reconstruite, prise individuellement, est voisine de celles d'une méthode de noyau, si la totalité de l'ensemble d'apprentissage est utilisée. Si l'on procède à une classification préalable de l'ensemble d'apprentissage, son expression est celle d'une fonction de base radiale généralisée. Une analogie avec les modèle de mélange de densités de probabilités

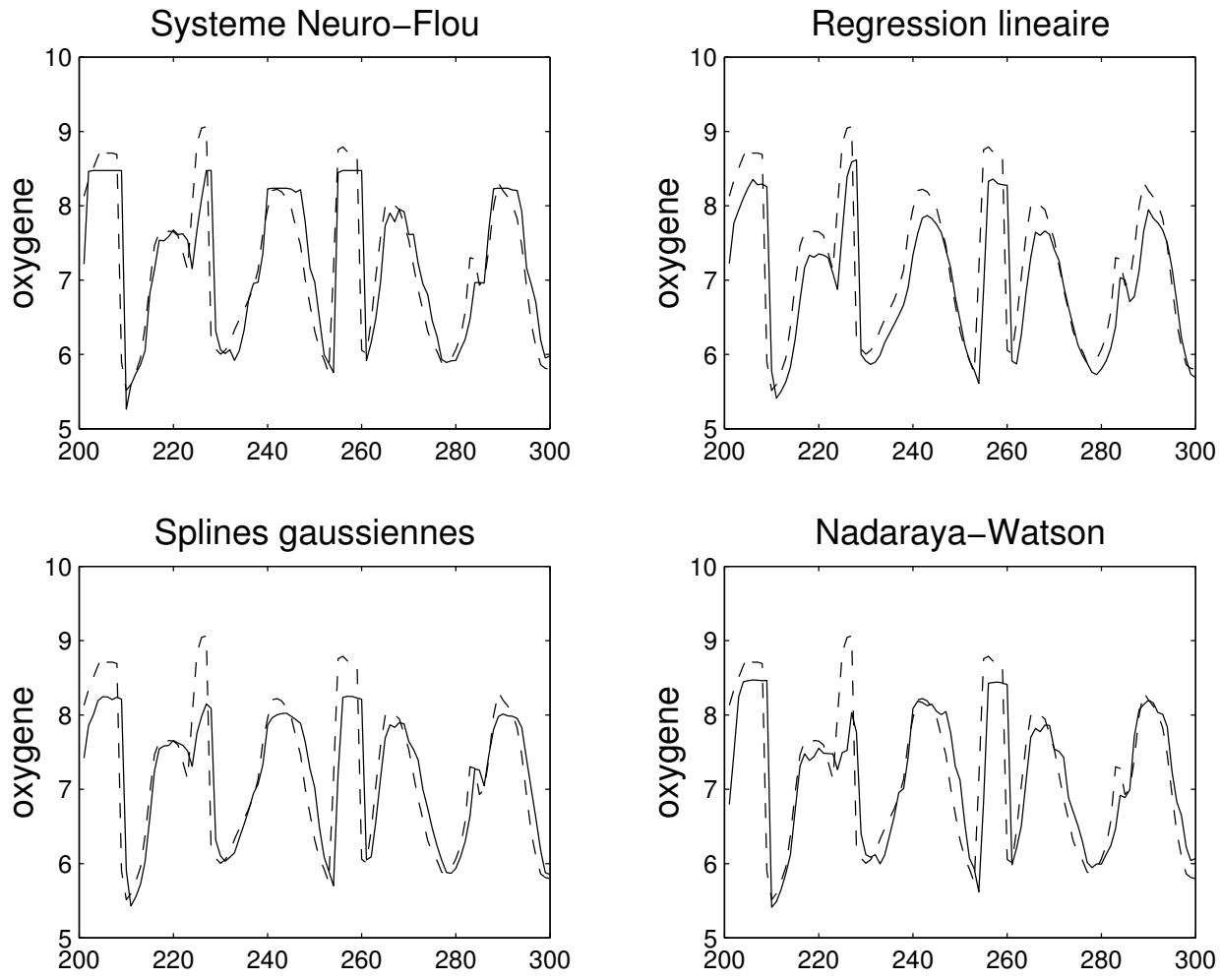


FIG. 2.14 – *Estimation ponctuelle : comparaison avec différentes méthodes de régression*

a été également développée dans la version utilisant le partitionnement.

L'étude des systèmes flous ne permet pas de traiter tous les types d'incertitude de façon satisfaisante. Dans les chapitres suivants, nous allons aborder le problème de l'estimation fonctionnelle à l'aide d'autres outils que les systèmes flous, les fonctions de croyance, qui permettent de définir des degrés de confiance *a priori* sur les éléments de l'ensemble d'apprentissage.

Chapitre 3

Théorie des croyances - Extension au flou

3.1 Introduction

La théorie des croyances est issue des travaux de Dempster [28] sur les bornes inférieure et supérieure d'une famille de distributions de probabilités induites par une fonction multivaluée, et a été développée par Shafer [130]. Elle permet de combiner les connaissances diverses que l'on possède relativement à un phénomène à partir de différentes sources, de manière plus souple que le formalisme probabiliste. En particulier, certains types d'imperfection de l'information, comme l'incertitude totale, sont mal représentés par la théorie des probabilités. Devant les critiques de certains probabilistes, des modifications et généralisations successives de la théorie initiale de Shafer ont été proposées dans différentes directions [171, 85, 59], occasionnant une certaine confusion dans la définition de la théorie dite de *Dempster et Shafer* [140]. Cette appellation recouvre en fait trois types de modèles distincts qui sont des extensions ou des variantes du modèle de Shafer : la théorie des probabilités imprécises ou des probabilités inférieures [163], le modèle initial de Dempster, développé par Kohlas et al. [28, 85] et le modèle des croyances transférables [142].

La théorie des probabilités imprécises suppose l'*existence* d'une mesure de probabilité précise P sur le référentiel d'étude Ω, mais celle-ci *n'est pas parfaitement connue*. Soit $\mathcal{P}$ l'ensemble des probabilités compatibles avec l'information disponible. Le modèle est défini, soit directement par $\mathcal{P}$, soit par ses enveloppes inférieure P_* et supérieure P^*, définies respectivement, pour tout $A \subseteq \Omega$, par $P_*(A) = \{\min P(A),\ P \in \mathcal{P}\}$ et $P^*(A) = \max\{P(A),\ P \in \mathcal{P}\}$. Sous certaines contraintes, P_* est une fonction de croyance. Ce modèle est une généralisation du modèle probabiliste. Le modèle de Dempster est un cas particulier des probabilités imprécises. On suppose qu'il existe une relation multivaluée entre Ω et un espace sous-jacent Θ sur lequel une mesure de probabilité est supposée exister. On peut là encore définir des probabilités inférieure et supérieure compatibles avec ces informations. Le modèle des croyances transférables proposé par Smets [142, 137] est très différent des modèles précédents, car il définit des fonctions de croyance *indépendamment de tout modèle probabiliste*. La transformation des fonctions de croyance en une mesure de probabilité n'intervient qu'à un niveau décisionnel. Dans ce modèle, Smets a complété la théorie des croyances en proposant une justification axiomatique [141]. Dans ce chapitre, nous adopterons généralement ce point de vue non probabiliste.

Dans la perspective de l'application ultérieure de la théorie des croyances en régression, nous insistons particulièrement sur les aspects suivants : l'existence d'un référentiel continu, la définition de l'espérance et les mesures d'incertitudes. Enfin, une généralisation de la théorie des croyances aux ensembles flous, proposée par divers auteurs [178, 134, 167, 175], permet d'intégrer les avantages des deux théories dans un même modèle. Nous présentons ici une synthèse de ces travaux.

3.2 Théorie des croyances

3.2.1 Structures de croyance

Soit Ω un ensemble, appelé *cadre de discernement*, que nous supposerons pour l'instant fini, et $S(\Omega)$ l'ensemble des sous-ensembles de Ω. Le concept fondamental de la théorie des croyances est celui de la structure de croyance, définie comme une fonction m de $S(\Omega)$ dans $[0,1]$ et vérifiant :

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{3.1}$$

La quantité $m(A)$ représente la croyance en la proposition A *uniquement*, c'est-à-dire en aucune proposition plus restrictive, en ne considérant que l'information disponible. La quantité $m(\emptyset)$ représente la masse qui ne peut être dédiée à aucune des propositions de Ω. Dans l'hypothèse d'un monde *ouvert*, le référentiel Ω n'est pas exhaustif et on suppose l'existence sous-jacente d'un ensemble Ψ contenant toutes les propositions inconnues. La masse $m(\emptyset)$ représente donc la masse allouée à Ψ. Une structure de croyance m telle que $m(\emptyset) = 0$ est dite *normalisée*. Dans le modèle initial de Shafer [130], cette condition est imposée. C'est l'hypothèse d'un monde *fermé*. Les sous-ensembles de Ω tels que $m(A) > 0$ sont appelés *éléments focaux* de m. Nous noterons $F(m)$ l'ensemble des éléments focaux de m. L'union de tous les éléments focaux est appelée *noyau*. La procédure de normalisation de Dempster convertit une structure de croyance non normalisée m en une structure normalisée m^* en répartissant proportionnellement la masse de l'ensemble vide sur les autres éléments focaux :

$$m^*(A) = \begin{cases} \dfrac{m(A)}{1 - m(\emptyset)} & \text{si } A \neq \emptyset \\ 0 & \text{si } A = \emptyset. \end{cases} \tag{3.2}$$

L'information fournie par une structure de croyance peut être également représentée par une *fonction de croyance* ou une *fonction de plausibilité* définies, respectivement, par :

$$\text{bel}(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \text{ et } \text{pl}(A) = \sum_{B \cap A \neq \emptyset} m(B) = \text{bel}(\Omega) - \text{bel}(\bar{A}), \tag{3.3}$$

où $\bar{A} = \Omega \backslash A$ est le complémentaire de A dans Ω. Notons que $\text{bel}(\emptyset) = 0$ et que $\text{bel}(\Omega) = 1 - m(\emptyset)$. Le référentiel n'est un événement certain que dans le cas d'un monde fermé. On peut retrouver l'expression de m si l'on connaît la fonction bel à partir de la transformation de Möbius :

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} \text{bel}(B) \quad \text{si } A \neq \emptyset, \text{ et } m(\emptyset) = 1 - \text{bel}(\Omega), \tag{3.4}$$

où $|A\backslash B|$ est le cardinal de $A \cap \bar{B}$.

Une fonction de crédibilité ou de croyance bel de $S(\Omega)$ dans $[0,1]$ peut être également définie, indépendamment d'une structure de croyance associée, comme une fonction *monotone complète, ou d'ordre infini*, c'est-à-dire :

$$\begin{cases} \forall (A_1 \ldots A_n) \subseteq \Omega \quad \text{bel}(A_1 \cup \ldots \cup A_n) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,n\}} (-1)^{|I+1|} \text{bel}(\cap_{i \in I} A_i) \\ \text{bel}(\Omega) \in]0,1]. \end{cases} \tag{3.5}$$

Elle vérifie par conséquent la propriété de sur-additivité (ou de monotonie d'ordre 2) :

$$\text{bel}(A \cup B) \geq \text{bel}(A) + \text{bel}(B) - \text{bel}(A \cap B) \quad \forall A,\ B \subseteq \Omega. \tag{3.6}$$

On peut définir de manière analogue une fonction de plausibilité par :

$$\begin{cases} \forall (A_1 \ldots A_n) \subseteq \Omega \quad \text{pl}(A_1 \cap \ldots \cap A_n) \leq \sum_{\emptyset \neq I \subseteq \{1,\ldots,n\}} (-1)^{|I+1|} \text{pl}(\cup_{i \in I} A_i) \\ \text{pl}(\Omega) \in]0,1]. \end{cases} \tag{3.7}$$

Cette fonction est donc sous-additive :

$$\text{pl}(A \cup B) \leq \text{pl}(A) + \text{pl}(B) - \text{pl}(A \cap B) \quad \forall A,\ B \subseteq \Omega. \tag{3.8}$$

La quantité $\text{bel}(A)$ s'interprète comme la croyance *totale* en A et $\text{pl}(A)$ comme la croyance qui pourrait être allouée à A, compte tenu des éléments qui ne contredisent pas cette proposition.

Par analogie avec la théorie des probabilités, on peut définir l'espace crédibilisable $(\Omega, S(\Omega))$ et l'espace crédibilisé : $(\Omega, S(\Omega), \text{bel})$, $(S(\Omega), \cap, \cup)$ étant une σ-algèbre sur Ω.

Nous noterons $\mathcal{M}(\Omega)$ l'ensemble des structures de croyances définies sur Ω. Parmi les diverses notions utilisées dans la théorie des croyances, signalons la fonction de communalité, qui présente en particulier un intérêt pratique dans le calcul de la combinaison de structures :

$$q(A) = \sum_{B \supseteq A} m(B) \tag{3.9}$$

La fonction bel peut s'exprimer en fonction de q. L'une quelconque des quatre fonctions bel, pl, m et q détermine ainsi de manière unique les trois autres. Par la suite, si le contexte n'est pas ambigu, quand nous définirons une structure par l'une de ces fonctions, les autres fonctions seront définies implicitement.

La théorie des croyances peut facilement se généraliser à un espace Ω continu si le nombre des éléments focaux $|F(m)|$ est fini (cf. section 3.2.6).

3.2.2 Transformation pignistique

Les quantités m ou bel décrivent l'état d'une croyance. Dans le modèle des croyances transférables, la prise de décision se fait par l'intermédiaire d'une mesure de probabilité induite par m ou bel, appelée *probabilité pignistique*, définie par[1] :

$$p_{bet}(\omega) = \sum_{A \subset \Omega} \frac{m^*(A)}{|A|} \delta_A(\omega) \quad \forall \omega \in \Omega, \tag{3.10}$$

1. Dans toute la suite, nous noterons par une lettre capitale P la mesure de probabilité définie sur un *ensemble* $\{\omega\}$, et par son équivalent minuscule p la probabilité de l'*élément* ω.

où m^* est la structure normalisée associée à m et δ_A est la fonction indicatrice de l'ensemble A. Si l'on impose un certain nombre d'axiomes raisonnables, cette solution est unique [142, 136]. En l'absence d'information supplémentaire, la masse de croyance de chaque élément focal est distribuée uniformément entre les éléments qui la composent.

Dans le modèle de Dempster ou des probabilités imprécises, le point de vue est différent. Les fonctions bel et pl peuvent également être considérées respectivement comme des bornes inférieure et supérieure d'un ensemble de distributions de probabilité $\mathcal{P}$ compatibles avec la structure de croyance : $\mathcal{P} = \{P |\ \forall A \subseteq \Omega,\ \text{bel}(A) \leq P(A) \leq \text{pl}(A)\}$. L'intervalle $[\text{bel}(A), \text{pl}(A)]$ peut alors être vu comme reflétant l'ignorance relative à l'événement A.

3.2.3 Liens avec d'autres mesures floues

Dans cette section, les structures de croyances sont supposées normalisées. Les fonctions de plausibilité et de crédibilité normalisées sont des cas particuliers d'un ensemble de mesures floues définies de la façon suivante par Sugeno [148]:

Mesure floue. *Une mesure floue μ est une fonction de $S(\Omega)$ dans $[0,1]$, telle que :*

1. μ est bornée : $\mu(\emptyset)$ et $\mu(\Omega) = 1$.
2. μ est monotone : $\forall A, B \in S(\Omega) \quad A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$.
3. μ est continue : pour toute suite d'ensembles emboîtés $(A_n)_n$, tels que $A_1 \subseteq \ldots \subseteq A_n \subseteq \ldots$ ou $A_1 \supseteq \ldots \supseteq A_n \supseteq \ldots$, on a: $\lim_{n\to\infty} \mu(A_n) = \mu(\lim_{n\to\infty} A_n)$.

Les mesures de possibilité, de nécessité et les mesures de probabilité sont elles aussi des mesures floues. En effet, l'axiome de monotonie est équivalent à :

$$\mu(A \cup B) \geq \max(\mu(A), \mu(B)) \quad \text{ou} \quad \mu(A \cap B) \leq \min(\mu(A), \mu(B)). \tag{3.11}$$

Les égalités dans les deux dernières expressions définissent alors respectivement une mesure de *possibilité* et une mesure de *nécessité*.

Si μ est une mesure floue et vérifie l'axiome d'additivité :

$$\forall A, B \subseteq \Omega, \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B), \tag{3.12}$$

alors μ est une mesure de probabilité. D'après les équations (3.8), (3.6) et l'expression précédente, les mesures de probabilité sont donc à la fois des mesures de plausibilité et de crédibilité.

Les relations entre toutes ces mesures floues sont illustrées dans la figure 3.1. Nous allons préciser le lien entre certaines structures particulières et ces différentes mesures floues.

Crédibilité consonante. *Une structure consonante bel est une structure dont les éléments focaux sont emboîtés, au sens de l'inclusion*:

$$F(m) = (A_k)_{k=1}^n \text{ avec } A_1 \subseteq \ldots \subseteq \ldots \subseteq A_n.$$

Les trois propositions suivantes sont équivalentes [130] :

1. bel est une fonction de crédibilité consonante.

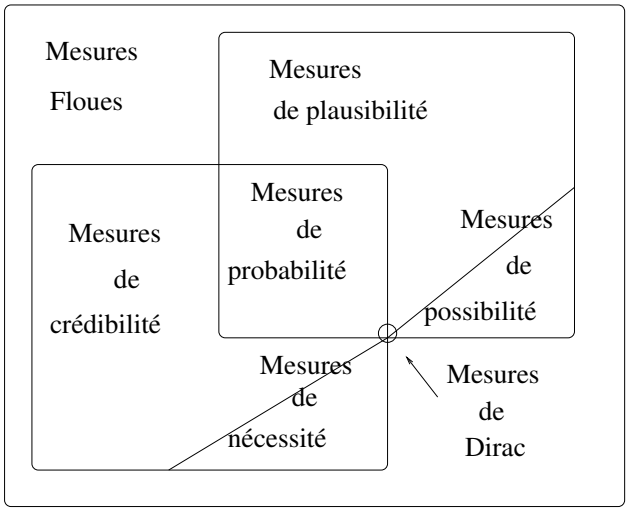


FIG. 3.1 – *Liens entre différentes mesures floues.*

2. $\forall A, B \subseteq \Omega, \quad \text{bel}(A \cap B) = \min(\text{bel}(A), \text{bel}(B))$.

3. $\forall A, B \subseteq \Omega, \quad \text{pl}(A \cup B) = \max(\text{pl}(A), \text{pl}(B))$.

Les mesures de crédibilité et de plausibilité consonantes sont donc formellement équivalentes respectivement aux mesures de nécessité et de possibilité. Remarquons toutefois que l'interprétation donnée à ces fonctions diffère selon que l'on se place du point de vue de la théorie des possibilités ou de celui de la théorie des croyances [138].

Crédibilité bayésienne. *La fonction de crédibilité bel est bayésienne si ses éléments focaux sont des singletons, dans le cas où Ω est discret.*

Les fonctions de crédibilité et plausibilité bayésiennes sont confondues avec une mesure de probabilité P :

$$\forall A \subseteq \Omega \quad \text{bel}(A) = \text{pl}(A) = P(A).$$

L'ensemble $\mathcal{P}$ des probabilités compatibles avec la structure m se réduit alors à P.

Structure à support simple. *Une structure de croyance est à support simple si elle est focalisée au maximum sur Ω et un seul sous-ensemble A : $F(m) = \{\Omega, A\}$ ou $F(m) = \{A\}$. Elle est alors définie par* :

$$\begin{cases} m(A) = s, \quad s \in [0,1] \\ m(\Omega) = 1 - s \\ m(B) = 0 \quad \forall B \notin \{A, \Omega\} \end{cases} \tag{3.13}$$

Le cas particulier où $s = 1$ correspond aux sous-ensembles classiques de Ω.

Structure monofocale. *On appellera structure monofocale ou ensembliste, une structure focalisée sur un unique ensemble A.* On la notera m_A (ou bel_A). Ces structures peuvent donc être assimilées à des ensembles classiques. Il existe une bijection entre l'ensemble de ces structures et Ω. Parmi ces structures, on peut définir,

- les **structures à support certain**. Si A est un singleton $\{a\}$ de Ω, la structure $m_{\{a\}}$, est dite *à support certain*. Elle exprime une certitude sur une valeur particulière de Ω.

Elle correspond à la mesure de probabilité de Dirac. C'est la seule structure à la fois consonante et bayésienne (cf. figure 3.1).

- la **structure vide (ou triviale)**. Dans ce cas, $F(m) = \{\Omega\}$, c'est-à-dire, $m(\Omega) = 1$ et $\forall A \neq \Omega,\ m(A) = 0$. Elle représente l'élément neutre dans la règle de combinaison de structures (cf. section 3.2.4). Elle exprime une absence totale d'information sur le cadre de discernement. C'est l'une des caractéristiques essentielles qui distinguent la théorie des croyances de l'inférence probabiliste. En l'absence totale d'information sur un phénomène, la théorie bayésienne définit en général une distribution uniforme, notée $\mathcal{U}_\Omega$, sur l'espace considéré, fini ou continu. Cette distribution correspond à la probabilité pignistique p_{bet} associée à la structure vide.

3.2.4 Inférence et combinaison de structures

Règle de combinaison conjonctive

En présence d'informations incertaines, la fusion multi-sources se présente comme une solution permettant d'accéder à une information plus fiable. De façon générale, elle offre de nombreux avantages parmi lesquels la complémentarité et la redondance de l'information fournie par chaque source. La théorie des croyances possède un outil d'agrégation, la règle de combinaison de Dempster [28], qui permet de combiner des structures de croyances supposées indépendantes $m_1, \ldots, m_k$ définies sur un même espace crédibilisable $(\Omega, S(\Omega))$.

On définit l'opérateur binaire conjonctif $\cap$ sur $\mathcal{M}(\Omega)$ de façon suivante [142] :

$$\begin{array}{ll} \bigcap : & \mathcal{M}(\Omega) \times \mathcal{M}(\Omega) \to \mathcal{M}(\Omega) \\ & (m_1, m_2) \longmapsto m \\ & \text{tel que } \forall A \subseteq \Omega, \quad m(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \end{array} \tag{3.14}$$

Propriétés

La *règle de combinaison conjonctive* (3.14) vérifie un certain nombre de propriétés primordiales. Elle est associative, commutative, possède un élément neutre, la structure de croyance triviale. Cette règle de combinaison est la seule vérifiant un certain nombre d'axiomes raisonnables [137, 44, 81] (cf. section 3.2.4). La combinaison d'une structure bayésienne avec une structure quelconque produit une nouvelle structure bayésienne. Par contre, on peut montrer que cette règle n'est pas idempotente.

On peut alors définir la combinaison de n structures $m_1, \ldots, m_n$ de $\mathcal{M}(\Omega)$ par :

$$m = m_1 \cap \ldots \cap m_n,$$

avec

$$m(A) = \sum_{A_1 \cap \ldots \cap A_n = A} \prod_{i=1}^{n} m_i(A_i) \quad \forall A \subseteq \Omega.$$

Normalisation

Cette opération produit en général une structure de croyance non normalisée, c'est-à-dire, telle que $m(\emptyset) > 0$. La masse $m(\emptyset)$ représente alors le *conflit* entre les sources. Dans l'hypothèse d'un monde fermé, il est alors nécessaire de recourir à la procédure de normalisation

de Dempster (cf. équation (3.2)), en multipliant les masses des éléments de m par le facteur de normalisation $K = \frac{1}{1-m(\emptyset)}$, si $m(\emptyset) \neq 0$. On définit alors une structure normalisée m^*. Cette loi de combinaison conjonctive normalisée définie par m^*, connue sous le nom de *règle de combinaison de Dempster* [28] et que l'on notera $\oplus$ pour la différencier de sa version non normalisée $\cap$, peut donner lieu à des résultats criticables en cas de fort conflit. En particulier, si le conflit est maximal ($m(\emptyset) = 1$), la structure m^* n'est pas définie. Les structures $m_1, \ldots, m_n$ sont dites non compatibles. Cet aspect inconsistant de la règle a été largement critiqué par divers auteurs [156, 137] et justifie l'usage de la règle non normalisée.

Le résultat de la combinaison des fonctions de communalité par la règle non normalisée s'exprime de manière simple :

$$q = \prod_{i=1}^{n} q_i.$$

Dans le cas de la règle normalisée, on obtient : $q^* = q/K$.

Généralisation à d'autres opérateurs

Différentes généralisations de la règle conjonctive ont été proposées [45, 173, 139, 141]. La règle a d'abord été étendue à l'opérateur disjonctif $\cup$ [45], puis à toute opération binaire ∇. La fusion de structures de croyances m_1 et m_2, notée $m = m_1 \nabla m_2$, est définie ainsi [173]:

$$m(A) = \sum_{B \nabla C = A} m_1(B) m_2(C), \quad \forall A \in \Omega. \tag{3.15}$$

Certains de ces opérateurs, dont l'opérateur disjonctif, ne génèrent pas de conflit, au sens où la combinaison de deux structures normales est normale :

$$m_1(\emptyset) = 0 \text{ et } m_2(\emptyset) = 0 \Rightarrow (m_1 \cup m_2)(\emptyset) = 0.$$

Du point de vue de l'interprétation, l'opérateur disjonctif est utilisé quand l'une au moins des sources est valide, tandis que l'opérateur conjonctif suppose que toutes les sources le sont.

Une autre règle de normalisation, consistant à transférer la part de croyance conflictuelle de l'ensemble vide vers Ω, a été proposée par Yager [173] et Kohlas [85] et conduit à une structure m^o définie par :

$$m^o(A) = \begin{cases} m(A) \quad \text{si} \quad A \in F(m) \backslash \{\emptyset, \Omega\} \\ m(\Omega) + m(\emptyset) \quad \text{si} \quad A = \Omega \\ 0 \quad \text{si} \quad A = \emptyset. \end{cases} \tag{3.16}$$

Cependant, l'utilisation de cette normalisation avec la règle de combinaison conjonctive définit un opérateur qui n'est plus associatif.

D'autres règles de combinaison ont été proposées. Celle de Tonn [156] est un compromis entre les méthodes conjonctive et disjonctive. Elle permet de résoudre le problème de conflit entre deux ensembles A et B en répartissant les masses sur A, B et $A \cup B$ si $A \cap B = \emptyset$ (et sur A, B et $A \cap B$ sinon).

Règle de conditionnement

Le conditionnement consiste à réviser une croyance initiale quelconque m lorsqu'une proposition $B \subseteq \Omega$ est devenue vraie. Dans le modèle des croyances transférables, la structure de croyance conditionnelle de m à B, notée $m(.|B)$ est définie de la façon suivante [137] :

$$m(A|B) = \begin{cases} \sum_{C \subseteq \bar{B}} m(A \cup C) & \text{si } A \subseteq B \\ 0 & \text{sinon,} \end{cases} \tag{3.17}$$

avec $\bar{B} = \Omega \backslash B$. On en déduit les fonctions de crédibilité et de plausibilité conditionnelles :

$$\text{bel}(A|B) = \text{bel}(A \cup \bar{B}) - \text{bel}(\bar{B}) \text{ et } \text{pl}(A|B) = \text{pl}(A \cap B).$$

Soit m_B la structure focalisée sur B, c'est-à-dire telle que $m_B(B) = 1$. La structure $m(.|B)$ peut également être définie par :

$$m(.|B) = m \cap m_B. \tag{3.18}$$

Dans le modèle des croyances transférables, contrairement à Dempster et Shafer, Smets définit d'abord cette règle de conditionnement puis en déduit axiomatiquement l'unicité de la règle de combinaison conjonctive [137].

Si les structures m_B et m sont compatibles, au sens de la règle normalisée, c'est-à-dire si $\text{bel}(\bar{B}) < 1$, on peut également définir la règle de conditionnement normalisée, ou règle de combinaison de Dempster, $m^*(.|B)$:

$$m^*(.|B) = m \oplus m_B. \tag{3.19}$$

Les expressions de la crédibilité et de la plausibilité correspondantes sont données par les équations suivantes [130] :

$$\text{bel}^*(A|B) = \frac{\text{bel}^*(A \cup \bar{B}) - \text{bel}^*(\bar{B})}{1 - \text{bel}^*(\bar{B})} \text{ et } \text{pl}^*(A|B) = \frac{\text{pl}^*(A \cap B)}{\text{pl}^*(B)}.$$

Cette dernière équation rappelle la règle de conditionnement de Bayes. La règle de Dempster généralise la règle de Bayes aux structures de croyances. Si les structures de croyances sont bayésiennes, les deux règles sont équivalentes.

Produit cartésien - croyances marginales

On peut généraliser la combinaison à des espaces différents Ω_1 et Ω_2. Soit $\Omega = \Omega_1 \times \Omega_2$ le produit cartésien de Ω_1 et Ω_2. L'opérateur $\otimes$ définit la *croyance jointe* m sur Ω :

$$\begin{array}{rl} \otimes : & \mathcal{M}(\Omega_1) \times \mathcal{M}(\Omega_2) \to \mathcal{M}(\Omega) \\ & (m_1, m_2) \longmapsto m = m_1 \otimes m_2 \\ & \text{avec } m(A) = \sum_{B \times C = A} m_1(B) m_2(C). \end{array} \tag{3.20}$$

Les structures m_1 et m_2 sont alors définies comme les *croyances marginales* de m sur Ω_1 et Ω_2. Le concept similaire à l'indépendance en théorie de l'évidence est celui de la noninteractivité.

Noninteractivité. *Les structures m_1 et m_2 sont dites* noninteractives *si et seulement si* :

$$m(A \times B) = m_1(A) m_2(B) \quad \forall A, B \subset \Omega_1 \times \Omega_2.$$

Raffinement et grossissement

Ces deux opérations se rapportent à la gestion des cadres de discernement.

Raffinement. *Une application multivaluée ρ d'un référentiel Ω vers un référentiel Θ est un* raffinement *si les ensembles $\{\rho(\omega), \omega \in \Omega\}$ constituent une partition de Θ.*

La relation inverse ρ^{-1} de $S(\Theta)$ vers Ω est un *grossissement.*

Soit une structure $m \in \mathcal{M}(\Omega)$. On définit l'extension minimale de m sur Θ par la structure $m' \in \mathcal{M}(\Theta)$ telle que :

$$m'(\rho(A)) = m(A) \quad \forall A \in F(m).$$

Soient deux structures m_1 et m_2 définies respectivement sur des espaces Ω_1 et Ω_2. Pour combiner m_1 et m_2, il suffit de déterminer deux raffinements ρ_1 et ρ_2 de Ω_1 et Ω_2 vers un référentiel Θ *commun* et de combiner les extensions minimales m'_1 et m'_2 de m_1 et m_2 sur Θ.

Affaiblissement

L'affaiblissement d'une fonction de crédibilité bel de taux $a \in [0,1]$ consiste à réduire le degré de certitude des éléments focaux non triviaux, c'est-à-dire différents de Ω, en fonction de la confiance qu'on accorde à la source ayant permis de définir la structure. On obtient ainsi une structure bel^a définie par:

$$\text{bel}(\Omega) = 1 \text{ et } \text{bel}^a(A) = (1-a)\text{bel}(A). \tag{3.21}$$

L'intérêt de l'opération d'affaiblissement est de maîtriser l'influence des sources selon leur fiabilité avant de les combiner.

3.2.5 Mesures d'incertitudes

Nous rappelons que l'incertitude relative à un phénomène caractérise une information qui peut être imprécise, fragmentaire, peu fiable. Les premières mesures d'incertitude ont été développées par Hartley en 1928 [63] en théorie des ensembles (classiques), puis en théorie des probabilités, avec l'entropie de Shannon en 1948 [131]. Depuis les années 80, le concept d'incertitude a été étendu à la théorie des ensembles flous, la théorie des possibilités et la théorie des croyances. Une harmonisation entre ces cinq théories a permis de définir deux types principaux d'incertitude :

- la nonspécificité, qui représente l'imprécision des ensembles ;
- la discordance, qui quantifie le conflit entre différentes possibilités.

En théorie des croyances, la recherche de mesures d'incertitude a fait l'objet de nombreux travaux, dus à Klir notamment [83], mais il n'existe pas actuellement de mesure « idéale ». Une présentation exhaustive de ces mesures étant illusoire, nous allons nous restreindre à certains types de mesures vérifiant des propriétés jugées intéressantes ou nécessaires.

Nonspécificité

En théorie classique, l'imprécision d'un ensemble A peut se mesurer par la fonction de Hartley: $U(A) = \log_2(|A|)$. Celle-ci se généralise en théorie des croyances par la mesure de nonspécificité [43] :

$$N(m) = \sum_{A \in F(m)} m(A) \log |A|, \tag{3.22}$$

où m est une structure de croyance. Ramer [123] a prouvé que, sous certaines conditions, la fonction N était l'unique mesure de nonspécificité en théorie des croyances. Elle représente la moyenne pondérée par leur masse de l'imprécision des éléments focaux. Les valeurs possibles de N appartiennent à $[0, \log_2 |\Omega|]$. Le maximum est atteint pour la structure de croyance triviale, c'est-à-dire en cas d'ignorance totale: $m(\Omega) = 1$. Le minimum, $N(m) = 0$, est atteint pour une structure bayésienne quelconque, pour laquelle tous les éléments focaux sont des singletons : il n'y a pas d'imprécision. Cela signifie que les mesures de probabilité sont totalement spécifiques et donc pareillement informatives sur le plan de la nonspécificité.

Exemple. Soit $\Omega = \{a_1, \ldots, a_n\}$, m_1 la structure certaine, focalisée sur $\{a_1\}$ et m_2 est la loi uniforme $\mathcal{U}_\Omega$, une structure bayésienne, focalisée uniformément sur Ω. Alors $N(m_1) = N(m_2) = 0$.

On peut donc se demander quel type d'incertitude représente l'entropie de Shannon [131], bien connue en théorie des probabilités. Nous allons voir qu'il s'agit en fait d'une mesure de conflit.

Mesures de conflit

Entropie de Shannon

L'entropie de Shannon, définie uniquement pour les mesures de probabilité, n'est donc applicable que pour une structure de croyance bayésienne m et s'écrit dans ce cas :

$$H(m) = -\sum_{a \in \Omega} m(\{a\}) \log_2 m(\{a\}), \tag{3.23}$$

Si on reprend l'exemple précédent, on obtient $H(m_1) = 0$ et $H(m_2) = \log_2 |\Omega|$.

On peut exprimer l'entropie sous la forme suivante :

$$H(m) = -\sum_{a \in \Omega} m(\{a\}) \log_2 (1 - \mathrm{Confl}_m(\{a\})) \text{ où } \mathrm{Confl}_m(\{a\}) = \sum_{b \in \Omega \setminus \{a\}} m(\{b\}). \tag{3.24}$$

Il est clair que $\mathrm{Confl}_m(\{a\})$ représente l'information en conflit avec la proposition soutenant l'événement $\{a\}$. L'entropie de Shannon représente le conflit total pour l'ensemble de la structure.

Généralisations de l'entropie de Shannon

Différentes mesures ont été proposées afin de généraliser l'entropie de Shannon à la théorie des croyances. Les extensions les plus naturelles sont la mesure de dissonance $E(m)$ [174], la

mesure de confusion $C(m)$ [69] et la mesure de discordance $D(m)$ [82] :

$$E(m) = -\sum_{A \in F(m)} m(A) \log_2 \mathrm{pl}(A), \tag{3.25}$$

$$C(m) = -\sum_{A \in F(m)} m(A) \log_2 \mathrm{bel}(A), \tag{3.26}$$

$$D(m) = -\sum_{A \in F(m)} m(A) \log_2 P_{bet}(A). \tag{3.27}$$

On montre facilement que $E(m) \leq D(m) \leq C(m)$ pour tout m de $\mathcal{M}(\Omega)$.
Alternativement à $D(m)$, Klir [83] a proposé une autre mesure appelée « strife » :

$$S(m) = -\sum_{A \in F(m)} m(A) \log_2 \sum_{B \in F(m)} m(B) \frac{|A \cap B|}{|A|}. \tag{3.28}$$

Klir a analysé ce que mesurent réellement ces quantités en tenant compte du degré de conflit entre les éléments focaux de m. Ces quatre mesures peuvent se réécrire sous la forme générale d'une mesure de conflit MC de manière analogue à (3.24),

$$MC(m) = -\sum_{A \in F(m)} m(A) \log_2(1 - \mathrm{Confl}_m(A)). \tag{3.29}$$

où $\mathrm{Confl}_m(A)$ représente le conflit entre A et m, c'est-à-dire le conflit entre les 2 structures m_A et m. Pour chacune des 4 mesures, le conflit $\mathrm{Confl}_m(A)$ s'exprime de la façon suivante :

- Dissonance : $\mathrm{Confl}_m(A) = \sum_{A \cap B = \emptyset} m(B)$.
  On retrouve le conflit entre les deux structures m_A et m, défini dans la règle de Dempster : $\mathrm{Confl}_m(A) = (m_A \oplus m)(\emptyset)$.
- Confusion : $\mathrm{Confl}_m(A) = \sum_{B \not\subset A} m(B)$.
- Discordance : $\mathrm{Confl}_m(A) = \sum_{B \in F(m)} m(B) \frac{|B \setminus A|}{|B|}$: on ne retient que la part de B réellement en conflit avec A.
- Strife : $\mathrm{Confl}_m(A) = \sum_{B \in F(m)} m(B) \frac{|A \setminus B|}{|A|}$.

George et Pal [52] ont défini une axiomatique raisonnable déterminant une nouvelle mesure de conflit de manière unique. Ils se proposent de définir une *distance* D entre deux éléments focaux :

$$D(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

La mesure de conflit, exprimée sous la forme de l'équation (3.29) devient :

$$\Delta(m) = -\sum_{A \in F(m)} m(A) \log_2 \left(1 - \sum_{B \in F(m)} m(B) D(A, B) \right) \tag{3.30}$$

Ici, $\mathrm{Confl}_m(A) = \sum_{B \in F(m)} m(B) D(A, B)$.

Propriétés

- Ces cinq mesures se réduisent à l'entropie de Shannon si m est une structure bayésienne.
- $\forall m \in \mathcal{M}(\Omega), \quad MC(m) \in [0, \log_2 |\Omega|]$. Le minimum est atteint pour une structure certaine, le maximum, pour une structure bayésienne.

On peut vérifier facilement les inégalités suivantes: $E(m) \leq D(m) \leq \Delta(m)$, $E(m) \leq S(m) \leq \Delta(m)$.

Conclusion

Il est difficile de dire qu'une mesure est meilleure qu'une autre. L'axiomatique définie par George et Pal peut amener à choisir la mesure $\Delta(m)$. Cependant, du point de vue de l'interprétation, l'emploi des mesures $S(m)$ et $D(m)$ semble parfaitement justifiable.

En conclusion, une bonne mesure de conflit doit :

- généraliser l'entropie de Shannon,
- mesurer correctement le conflit selon le degré d'inclusion entre les éléments focaux de la structure.

Incertitude « totale »

Jusqu'à présent, nous avons défini deux types d'incertitude distincts en théorie de l'évidence : la nonspécificité et le conflit. Dans certaines applications, il peut être intéressant de connaître la quantité totale d'information manquante.

La première mesure d'incertitude totale a été proposée par Lamata et Moral [89]:

$$NE(m) = N(m) + E(m).$$

On peut définir ainsi 5 mesures d'incertitude totale par la somme de $N(m)$ et d'une quelconque des 5 mesures de conflit définies précédemment.

Plusieurs auteurs [62, 101] ont défini une autre mesure $AU(m)$ (« *amount of uncertainty* »), qui satisfait une série d'axiomes souhaitables [83], parmi lesquels, la généralisation de la fonction de Hartley et la généralisation de l'entropie de Shannon :

$$AU(m) = \max_{P \in \mathcal{P}} H(P).$$

Cette mesure représente le maximum de l'entropie de Shannon parmi toutes les mesures de probabilité compatibles avec la structure de croyance m.

En dehors de ce cadre axiomatique, signalons encore une mesure d'incertitude $I(m)$, proposée par Smets [135], basée sur la fonction de communalité :

$$I(m) = -\sum_{A \in F(m)} \log_2 q(A).$$

Cette mesure présente l'avantage d'être additive dans la combinaison conjonctive:

$$m = m_1 \cap m_2 \Rightarrow I(m) = I(m_1) + I(m_2).$$

3.2.6 Extension à un référentiel continu

La littérature s'est presque essentiellement consacrée au cas d'un cadre de discernement Ω fini, bien que de nombreuses applications utilisent des variables réelles continues. Certains auteurs ont cependant cherché à généraliser la théorie aux espaces continus. En vue d'une application à l'estimation fonctionnelle, nous allons maintenant nous placer dans ce cadre. On distingue deux approches principales. Dans la première approche [133, 146, 23], Ω est un sous-ensemble de $\mathbb{R}$ et les éléments focaux sont des intervalles de Ω. Le nombre d'éléments focaux n'étant pas nécessairement limité, la distribution de masse m devient analogue à une distribution de probabilité continue. L'autre approche, développée par Guan et Bell [60], permet de généraliser la théorie aux algèbres de Boole si le nombre d'éléments focaux est fini.

Espaces continus à éléments focaux contigus

Discrétisation

On se restreint ici au cas où Ω est un intervalle réel. Soit donc $\Omega = [a, b]$, avec $a, b \in \mathbb{R}$. La première étape consiste à se ramener au cas d'un cadre de discernement fini en *discrétisant* Ω en un espace Θ_n de n éléments ordonné: $\{a_1, \ldots, a_n\}$, chaque a_i pouvant par exemple être défini de façon suivante: $a_i = a + \frac{i-1}{n-1}(b-a)$. On peut alors définir une structure de croyance m_n sur Θ_n de manière habituelle. Puisqu'il existe une relation d'*ordre* entre les éléments de Θ_n, Strat [146] propose de se limiter aux éléments focaux définis comme réunion d'éléments contigus, adjacents de Θ_n, ce qui correspond dans le cas continu à des intervalles de Ω. Ainsi, l'ensemble $\{\{a_1\}, \{a_3\}\}$ ne peut être un élément focal. Sur le plan des applications, cette restriction semble raisonnable. L'exemple précédent suppose que l'on possède une information permettant d'exclure explicitement la valeur intermédiaire a_2. Ce postulat présente l'avantage de réduire considérablement le nombre d'éléments focaux: $F(m_n) \leq \frac{n(n+1)}{2}$. Il permet également de représenter graphiquement en deux dimensions la structure de croyance (cf. [146]). On peut alors construire une structure $m_n^\Omega \in \mathcal{M}(\Omega)$ dont les éléments focaux sont des intervalles du type $[a_i, a_j]$. Formellement, on a :

$$\forall i, j \in \{1, \ldots, n\} \quad m_n^\Omega([a_i, a_j]) = m_n\left(\bigcup_{k=i}^{j} \{a_k\}\right). \tag{3.31}$$

La fonction de communalité s'exprime alors simplement en fonction de la structure m_n :

$$q_n([a_i, a_j]) = \sum_{p=1}^{i} \sum_{q=j}^{n} m_n([p, q]) \quad \forall a_i < a_j. \tag{3.32}$$

Structure de croyance limite

On peut augmenter indéfiniment le nombre de points de discrétisation et définir ainsi une suite m_n de structures. L'ensemble limite devient l'ensemble non dénombrable $\Omega = [a, b]$, puisque $\lim_{n\to\infty} \Theta_n = \Omega$. La structure limite m est alors continue et devient analogue à une densité de probabilité. Le nombre d'éléments focaux peut donc être infini.

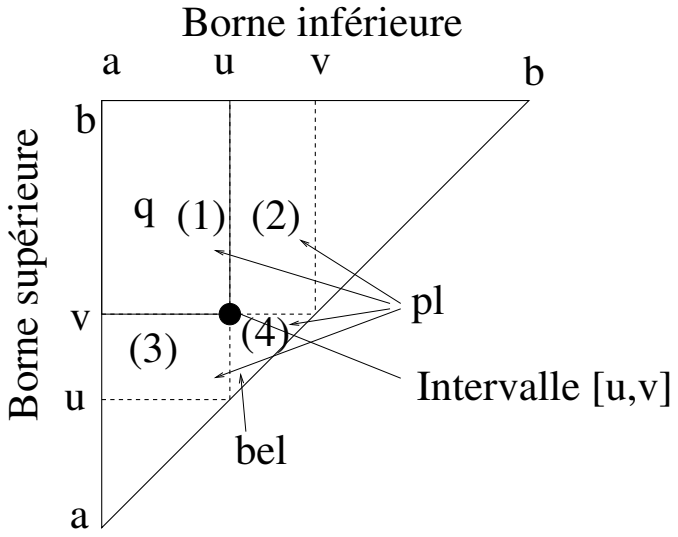


FIG. 3.2 – *Représentation d'une structure continue m définie sur un intervalle $[a,b]$. Les éléments focaux sont des intervalles de $[a,b]$. Chaque point ($\bullet$) du triangle représente un intervalle $[u,v]$ dont les extrémités sont définies par la projection sur les deux côtés droits du triangle. La zone (1) représente les intervalles contenant $[u,v]$. Par définition, l'intégrale de m sur (1) vaut $q([u,v])$. La zone (4) représente les intervalles inclus dans $[u,v]$. L'intégrale de m sur cette zone vaut $bel([u,v])$. Enfin, $pl([u,v])$ est représentée par l'intégrale sur les 4 zones (1, 2, 3 et 4).*

On peut calculer la fonction de communalité limite à partir de l'équation (3.32) :

$$q([u,v]) = \int_a^u \int_v^b m([x,y])dxdy \quad \forall u < v. \tag{3.33}$$

On obtient de la même façon les expressions de la crédibilité et de la plausibilité limites :

$$\mathrm{bel}([u,v]) = \int_u^v \int_x^v m([x,y])dxdy \quad \forall u < v. \tag{3.34}$$

$$\mathrm{pl}([u,v]) = \int_a^v \int_{max(u,x)}^b m([x,y])dxdy \quad \forall u < v. \tag{3.35}$$

La densité de probabilité pignistique s'écrit, en généralisant l'équation (4.18),

$$p_{bet}(u) = \int_a^b \int_a^b \frac{m([x,y])}{|y-x|}\delta_{[x,y]}(u)dxdy \quad \forall u \in [a,b]. \tag{3.36}$$

Toutes ces quantités peuvent être représentées facilement dans un graphique en deux dimensions (cf. figure 3.2).

Généralisation aux algèbres de Boole

La théorie des croyances peut facilement se généraliser au cas d'une algèbre de Boole si le nombre d'éléments focaux est *fini* [60].

Soit $(\mathcal{A}, \cap, \cup, -, \emptyset, \Omega, \subseteq)$ une algèbre de Boole définie sur un ensemble *quelconque* Ω, où $-$ est la complémentation, l'ensemble $\emptyset$, l'élément absorbant. L'ensemble Ω est l'élément neutre de $\cap$, le plus grand élément au sens de $\subseteq$, une relation d'ordre partiel.

La fonction m de $\mathcal{A}$ dans $[0,1]$ est alors une structure de croyance sur $\mathcal{M}(\Omega)$ s'il existe un ensemble fini $F(m)$ tel que

$$\sum_{A \in \mathcal{A}} m(A) = \sum_{F \in F(m)} m(F) = 1.$$

Ainsi, $\forall A \notin F(m), m(A) = 0$. L'ensemble $F(m)$ reste l'ensemble des éléments focaux. Ici encore, la structure est normalisée si $m(\emptyset) = 0$. On suppose à partir de maintenant que $\mathcal{A}$ est l'ensemble des boréliens de $\mathbb{R}$. Sous la contrainte de finitude de $F(m)$, toutes les notions vues dans les sections précédentes se généralisent au cas continu. La seule modification majeure consiste à remplacer la cardinalité d'un ensemble A par sa mesure de Lebesgue: $|A| = \int_\Omega \delta_A(\omega) d\omega$.

La définition de la nonspécificité définie dans le chapitre 2 n'est plus valable, car $\log_2 |A|$ peut être négatif. On utilisera dans ce cas la définition suivante:

$$U(A) = \log_2(1 + |A|) \tag{3.37}$$

Si Ω est continu, p_{bet} devient la densité de probabilité d'une variable aléatoire continue. En particulier, si Ω est un intervalle de $\mathbb{R}$ et $\mathcal{A}$, l'ensemble des intervalles de Ω, p_{bet} devient une loi de probabilité réelle:

$$\forall \omega \in \Omega \quad p_{bet}(\omega) = \sum_{A \in F(m)} \frac{m(A)}{\int_\Omega \delta_A(y) dy} \delta_A(\omega).$$

Plus précisément, dans ce cas, la fonction $\omega \to p_{bet}(\omega)$ est une fonction constante par morceaux, discontinue aux extrémités des éléments focaux.

Exemples. $\Omega = \mathbb{R}, \mathcal{A} = \mathcal{B}(\mathbb{R})$, ensemble des boréliens de $\mathbb{R}$.

- $a < c < d < b$ et $m([a,d]) = 0.6$ et $m([c,b]) = 0.4$.

$$p_{bet}(\omega) = \begin{cases} 0 & \text{si } \omega \notin [a,b] \\ \frac{0.6}{d-a} & \text{si } \omega \in [a,c] \\ \frac{0.6}{d-a} + \frac{0.4}{b-c} & \text{si } \omega \in [c,d] \\ \frac{0.4}{b-c} & \text{si } \omega \in [d,b]. \end{cases}$$

- $m([a,b]) = 1$: structure triviale sur $[a,b]$ et p_{bet} est la loi uniforme continue sur $[a,b]$.

$$p_{bet}(\omega) = \begin{cases} 0 & \text{si } \omega \notin [a,b] \\ \frac{1}{b-a} & \text{si } \omega \in [a,b]. \end{cases}$$

Parmi les deux types d'extension à un référentiel continu, nous adopterons cette approche pour sa facilité de mise en œuvre.

3.2.7 Espérance en théorie des croyances

Dans cette sous-section, nous allons voir comment il est possible de généraliser la notion d'espérance mathématique classique aux structures de croyance. Cette notion est en effet essentielle pour différentes applications comme l'estimation fonctionnelle ou la prise de décision, dont un cas particulier est la discrimination.

Rappel probabiliste

Soit Y une variable aléatoire réelle d'un espace probabilisé $(\Omega, S(\Omega), P)$, Ω étant un ensemble discret ou continu et f une fonction de Ω dans $\mathbb{R}$. L'espérance mathématique de la variable aléatoire $f(Y)$ relativement à P, si elle existe, est définie par :

$$\mathbb{E}_P[f(Y)] = \int_\Omega f(y)dP. \tag{3.38}$$

Si Ω est discret, l'espérance de $f(Y)$ s'écrit :

$$\mathbb{E}_P[f(Y)] = \sum_{y\in\Omega} f(y)P(\{y\}). \tag{3.39}$$

Si Ω est continue et P est une mesure absolument continue, soit p la densité correspondante[2]. On a alors :

$$\mathbb{E}_P[f(Y)] = \int_{y\in\Omega} f(y)p(y)dy. \tag{3.40}$$

Application en estimation fonctionnelle

Dans le cas de l'estimation d'une fonction g inconnue, définie d'un espace d'observation $\mathcal{X}$ vers Ω, ces deux espaces étant ordonnés et continus, connaissant $\boldsymbol{x} \in \mathcal{X}$, on peut chercher à estimer $y = g(\boldsymbol{x}) \in \Omega$ par l'intermédiaire de la fonction de régression, c'est-à-dire l'espérance conditionnelle :

$$\mathbb{E}_{P_{Y|X}}[Y|X = \boldsymbol{x}] = \int_\Omega y p_{Y|X}(y|\boldsymbol{x})dy. \tag{3.41}$$

où $p_{Y|X}$ est la distribution conditionnelle de la variable Y sachant $\boldsymbol{x}$ (cf. annexe A), et X est une variable aléatoire sous-jacente de l'espace $\mathcal{X}$. De manière plus générale, pour une fonction f quelconque, l'équation (3.40) devient :

$$\mathbb{E}_{P_{X|Y}}[f(Y)|X = \boldsymbol{x}] = \int_\Omega f(y)p_{Y|X}(y|\boldsymbol{x})dy, \tag{3.42}$$

si cette quantité est définie.

2. Dans toute la suite, dans le cas de variable continue, on notera en majuscule les mesures de probabilité et en minuscule les densités correspondantes.

Application à la discrimination

En théorie de la décision (ou de l'utilité), Ω représente l'ensemble souvent fini des hypothèses possibles, qui peuvent par exemple être des classes, dans le cas de la discrimination. On définit un ensemble d'actions possibles $\mathcal{D}$, auxquelles on associe une fonction de coût (ou fonction d'utilité) $f_d, \Omega \longmapsto \mathbb{R}$, pour tout $d \in \mathcal{D}$. La théorie bayésienne de la décision préconise de choisir l'action d pour laquelle l'espérance du coût $\mathbb{E}_P(f_d(Y))$ est la plus faible.

Dans le cas de la discrimination, les actions correspondent en général à l'affectation aux différentes classes possibles. Ainsi les espaces $\mathcal{D}$ et Ω sont confondus. Soit $\boldsymbol{x}$ un vecteur d'un espace d'observation $\mathcal{X}$. L'objectif est de déterminer la classe de $\boldsymbol{x}$ sur Ω. On suppose connues les lois de probabilité conditionnelles sachant $\boldsymbol{x}$, $P_{Y|X}$. On utilise en général des fonctions de coûts binaires, selon que la classification est correcte ou non. Ainsi,

$$f_d(y) = \begin{cases} 0 & \text{si} \quad y = d \\ 1 & \text{si} \quad y \neq d \end{cases} \tag{3.43}$$

D'après l'équation (3.39), pour chacune des actions ou classes $d \in \Omega$, l'espérance de coût correspondant est donc

$$\mathbb{E}_{P_{Y|X}}\left[f_d(Y)|X = \boldsymbol{x}\right] = \sum_{\{y \in \Omega | y \neq d\}} P_{Y|X}(y|\boldsymbol{x}). \tag{3.44}$$

Espérance en théorie des probabilités imprécises

La première étape dans la généralisation de la définition de l'espérance aux structures de croyance consiste à étendre la notion d'espérance mathématique aux probabilités inférieure et supérieure [28, 163]. Soit $\mathcal{P}$ la famille supposée finie de probabilités compatibles avec l'information disponible, bornée par les probabilités inférieure et supérieure P_* et P^*. Soit Y une variable aléatoire définie sur $\Omega \subseteq \mathbb{R}$, mais de loi de probabilité *imprécise* $P \in \mathcal{P}$. On s'intéressera ici plus spécialement au cas de variables continues. On peut définir les fonctions de répartition inférieure et supérieure F_* et F^* par les formules :

$$\forall y \in \mathbb{R} \quad F_*(y) = P_*(Y \leq y) \text{ et } F^*(y) = P^*(Y \leq y).$$

On peut alors définir les espérances inférieure et supérieure de la variable aléatoire $f(Y)$, où f est une fonction de Ω dans $\mathbb{R}$, par les intégrales de Lebesgue-Stieltjes respectives :

$$\mathbb{E}_*(f(Y)) = \int_{\mathbb{R}} f(y) dF_*(y), \tag{3.45}$$

$$\mathbb{E}^*(f(Y)) = \int_{\mathbb{R}} f(y) dF^*(y), \tag{3.46}$$

si ces intégrales sont définies. La famille $\mathcal{P}$ étant bornée, ces quantités peuvent être également définies par les expressions :

$$\mathbb{E}_*(f(Y)) = \min_{P \in \mathcal{P}} \mathbb{E}_P(f(Y)), \tag{3.47}$$

$$\mathbb{E}^*(f(Y)) = \max_{P \in \mathcal{P}} \mathbb{E}_P(f(Y)). \tag{3.48}$$

Espérance en théorie des croyances

Lorsque l'incertitude est caractérisée non plus par une mesure de probabilité, mais plus généralement par une structure de croyance m, plusieurs approches sont envisageables. La première consiste à s'appuyer sur la notion d'espérance mathématique dans le cadre des probabilités inférieure et supérieure, définies par les équations (3.45) et (3.46), où P_* devient la fonction de crédibilité bel et P^*, la fonction de plausibilité pl. Soit Y une variable de l'espace crédibilisé $(\Omega, S(\Omega), m)$. Par analogie avec les probabilités imprécises, les espérances inférieure et supérieure de $f(Y)$ deviennent:

$$\mathbb{E}_*(f(Y)) = \int_\Omega f(y) d\mathrm{bel}(\{y_i \in \Omega | y_i \leq y\}), \tag{3.49}$$

$$\mathbb{E}^*(f(Y)) = \int_\Omega f(y) d\mathrm{pl}(\{y_i \in \Omega | y_i \leq y\}). \tag{3.50}$$

Ces équations peuvent s'exprimer de la façon suivante [134] :

$$\mathbb{E}_*(f(Y)) = \sum_{A \in F(m)} m(A) \inf_{y \in A} f(y) \text{ et } \mathbb{E}^*(f(Y)) = \sum_{A \in F(m)} m(A) \sup_{y \in A} f(y) \tag{3.51}$$

Dans le cas de l'estimation fonctionnelle, si l'on se place dans le cadre des probabilités imprécises, ces deux quantités représentent les valeurs extrêmes d'un intervalle dont la largeur représente l'incertitude sur l'estimation de la fonction. En revanche, en théorie des croyances, cet intervalle n'a pas de signification particulière, puisque la structure ne suppose pas de probabilité sous-jacente. On définit simplement deux valeurs ponctuelles, correspondant à une estimation « pessimiste » et « optimiste ». Dans le cas de la décision, deux stratégies, basées sur la minimisation des coûts inférieur et supérieur, peuvent mener à des conclusions *différentes*. Si l'on est intéressé par une valeur ponctuelle unique, comme c'est souvent le cas pour les systèmes de décision ou l'estimation, il faut donc recourir à des approches intermédiaires. Là encore, on peut avoir recours aux probabilités imprécises et choisir une probabilité particulière dans $\mathcal{P}$.

Dans le modèle des croyances transférables, la solution consiste à utiliser la probabilité pignistique p_{bet} associée à m. On obtient alors

$$\mathbb{E}_{p_{bet}}(f(Y)) = \int_\Omega f(y) p_{bet}(y) dy = \sum_{A \in F(m)} m(A) \overline{f_A}, \tag{3.52}$$

où

$$\overline{f_A} = \frac{\int_\Omega f(y) \delta_A(y) dy}{\int_\Omega \delta_A(y) dy}$$

représente la moyenne des valeurs de f sur A.

Si f est la fonction identité, on obtient l'espérance de Y, qui s'écrit donc :

$$\mathbb{E}_{p_{bet}}(Y) = \sum_{A \in F(m)} m(A) y_A^* \tag{3.53}$$

où y_A^* est le centre de gravité de A (si A est un intervalle $[a, b]$, $y_A^* = \frac{a+b}{2}$).

Une autre approche, proposée par Strat [147] et Jaffray [?], consiste à introduire un paramètre permettant de choisir de façon plus souple une mesure de probabilité dans $\mathcal{P}$. Pour $\rho \in [0,1]$, l'espérance de $f(Y)$ peut alors se définir par :

$$\mathbb{E}_\rho(f(Y)) = \sum_{A \in F(m)} m(A)[\rho \min_{y \in A} f(y) + (1-\rho) \max_{y \in A} f(y)]. \tag{3.54}$$

Cette expression correspond à une interpolation linéaire des espérances inférieure et supérieure:

$$\mathbb{E}_\rho(f(Y)) = \mathbb{E}^*(f(Y)) + \rho(\mathbb{E}_*(f(Y)) - \mathbb{E}^*(f(Y))).$$

Yager [169, 170] propose de généraliser les expressions des équations (3.51), (3.53) et (3.54) à l'écriture suivante, où $\boldsymbol{A}$ est maintenant un ensemble aléatoire de Ω et ϕ est une fonction de $S(\Omega)$ dans $\mathbb{R}$:

$$\mathbb{E}_m(\phi(\boldsymbol{A})) = \sum_{A \in F(m)} m(A)\phi(A). \tag{3.55}$$

En fait, Yager définit simplement l'expression de l'espérance d'une fonction ϕ de $S(\Omega)$ dans $\mathbb{R}$ sans faire intervenir la notion d'ensemble aléatoire. Par abus de notation et afin d'alléger les notations, on peut écrire :

$$\mathbb{E}_m(\phi) = \sum_{A \in F(m)} m(A)\phi(A). \tag{3.56}$$

Une méthode particulière concernant le *choix* de la fonction ϕ, basée sur la maximisation d'une fonction d'entropie, a été présentée par Nguyen et Walker [111].

3.3 Extension de la théorie aux ensembles flous

La théorie de la mesure floue, qui généralise la théorie de la mesure classique, doit être clairement différenciée de la théorie des ensembles flous, qui généralise la théorie des ensembles. Ces deux théories sont cependant complémentaires. Dans la théorie des ensembles flous, les objets manipulés sont précis mais leur représentation ensembliste est imprécise. Inversement, dans la théorie de la mesure floue, les ensembles considérés sont classiques, mais l'appartenance d'un objet à ces ensembles est de nature incertaine (ambiguë ou imprécise). On peut combiner les avantages des deux types de théories en définissant la *mesure* floue d'un *ensemble* flou. Alors que les mesures floues classiques μ (probabilités, possibilités, crédibilités) sont définies de $S(\Omega) \to \mathbb{R}$, on peut étendre la notion aux ensembles flous. La mesure μ est alors une fonction de $\mathcal{F}(\Omega) \to \mathbb{R}$.

3.3.1 Probabilités et flou

La théorie des ensembles flous est souvent vue comme une alternative à la théorie des probabilités. La littérature est riche en travaux visant, soit à la discréditer par rapport aux probabilités comme représentation efficace de l'incertain [95, 90], soit au contraire à mettre en avant ses avantages.

Plutôt que d'opposer les deux théories, il peut sembler intéressant de combiner leurs avantages, en étendant la théorie des probabilités aux ensembles flous. Soit un espace probabilisé classique $(\Omega, S(\Omega), P)$. Un sous-ensemble flou de Ω est alors appelé *événement flou*.

Probabilité d'un événement flou

La probabilité d'un événement flou F est définie comme l'espérance de sa fonction d'appartenance [178]:

$$P(F) = \int_{\Omega} F(\omega) dP \tag{3.57}$$

si Ω est continu. Cette expression généralise le cas où F est un ensemble classique.

Un certain nombre de propriétés, comme la propriété d'additivité sont conservées :

$$\forall A, B \in \mathcal{F}(\Omega), \quad P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

où les opérateurs d'intersection et d'union sont remplacés par leur équivalent flou (t-normes et t-conormes, cf. chapitre 1).

On peut également définir la probabilité conditionnelle d'un événement flou $P(A|B)$ par :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{si } P(B) > 0$$

avec $P(A \cap B) = \int_{\Omega} A(\omega) B(\omega) dP$. La t-norme utilisée est en général le produit. L'indépendance de deux événements A et B se définit par $P(A \cap B) = P(A)P(B)$.

Variable aléatoire floue

Le concept de variable aléatoire floue a été introduit afin de décrire des données floues lors d'une expérience aléatoire. On peut définir une *variable aléatoire floue* comme une fonction $X : \Omega \to \mathcal{F}(\mathbb{R})$, vérifiant certaines conditions de mesurabilité, basées sur des α-coupes de X [122]. Ce concept généralise à la fois celui des variables aléatoires classiques et celui des ensembles aléatoires, où les variables sont des ensembles classiques, et est adapté à des situations où des informations de nature probabiliste et possibiliste interviennent. Des théorèmes classiques comme la Loi des Grands Nombres ou le Théorème Central Limite ont été établis pour les variables aléatoires floues. La théorie a été appliquée en particulier à l'inférence statistique en présence de données imprécises.

3.3.2 Théorie des croyances floues

Une généralisation de la théorie des croyances a été proposée par différents auteurs [178, 134, 167, 175]. L'idée générale est de permettre aux éléments focaux A d'être flous : $A \in \mathcal{F}(\Omega)$. Selon la définition de Yager [167], une fonction m de $\mathcal{F}(\Omega)$ vers $[0, 1]$ est une structure de croyance floue sur Ω, s'il existe une collection d'ensembles flous de Ω, $F(m) \subseteq \mathcal{F}(\Omega)$, telle que :

$$\begin{cases} m(A) = 0 \quad \text{si} \quad A \notin F(m) \\ \sum_{A \in F(m)} m(A) = 1. \end{cases} \tag{3.58}$$

Ici encore, A est un élément focal de m si $m(A) \neq 0$.

Structure normalisée. *Si tous les éléments focaux flous de m sont normalisés, m est dite normalisée.*

On peut distinguer deux types de généralisations de définitions de fonctions de crédibilité. La première est basée sur l'extension directe d'opérations ensemblistes au flou, la deuxième, sur la décomposition des éléments focaux en $\alpha-$coupes.

Extension directe des fonctions de croyance

Interprétation possibiliste

Zadeh [178] a été le premier à généraliser la théorie des croyances aux ensembles flous, en se basant sur les concepts de la théorie des possibilités et de la « granularité » de l'information.

Soit Π une mesure de possibilité sur Ω, π la distribution de possibilité associée et F un ensemble flou sur Ω. Nous rappelons que la mesure de possibilité de F est définie par [40]:

$$\Pi(F) = \sup_{\omega \in \Omega} [\min(\pi(\omega), F(\omega))] \tag{3.59}$$

La possibilité de F peut s'interpréter comme une mesure d'intersection entre F et l'ensemble flou défini par π.

Soit m une structure de croyance floue de Ω et $B \in F(m)$. Pour un ensemble flou A de Ω, soit $\Pi(A|B)$ la mesure de possibilité conditionnelle de A sachant B, définie par

$$\Pi(A|B) = \sup_{\omega \in \Omega} \min(A(\omega), B(\omega)).$$

Zadeh [178] définit alors la plausibilité de la proposition floue A de la façon suivante :

$$\mathrm{pl}(A) = \sum_{B \in F(m)} m(B)\Pi(A|B). \tag{3.60}$$

Si A et B sont des ensembles classiques, $\Pi(A|B) = 1$ si $A \cap B \neq \emptyset$; cette expression généralise donc bien la fonction de plausibilité (équation (3.3)).

Pour tout ensemble flou A, on note Π_A, la fonction de $\mathcal{F}(\Omega)$ dans $[0,1]$ définie par $\Pi_A(B) = \Pi(A|B)$ pour tout B.

On peut étendre la notion d'espérance définie à la section précédente (équation (3.56)) à toute fonction ϕ de $\mathcal{F}(\Omega)$ dans $[0,1]$, par un nouvel abus de notation[3] :

$$\mathbb{E}_m(\phi) = \sum_{B \in F(m)} m(B)\phi(B), \quad \phi : \mathcal{F}(\Omega) \to [0,1]. \tag{3.61}$$

On obtient alors :

$$\mathbb{E}_m(\Pi_A) = \sum_{B \in F(m)} m(B)\Pi_A(B) = \mathrm{pl}(A).$$

La fonction de plausibilité se définit donc comme l'espérance de la fonction de possibilité conditionnelle.

3. On devrait plutôt définir l'espérance d'un ensemble flou aléatoire $\boldsymbol{B}$ de Ω, puis l'espérance de $\phi(\boldsymbol{B})$. Ceci est valable pour les équations suivantes.

Si les éléments focaux sont classiques (i.e. $m \in \mathcal{M}(\Omega)$) mais A est flou, on obtient l'expression suivante :

$$\mathrm{pl}(A) = \sum_{B \in F(m)} m(B) \sup_{\omega \in B} A(\omega) = \mathbb{E}^*(\mu_A), \tag{3.62}$$

définie par Smets [134] comme l'espérance supérieure de la fonction d'appartenance de A (cf. équation (3.46)). La plausibilité d'un événement flou peut donc aussi être vue comme une généralisation de la probabilité d'un événement flou proposée par Zadeh (équation 3.57).
De même, si $N(A|B)$ représente la nécessité de A sachant B, définie par

$$N_A(B) = N(A|B) = \inf_{\omega \in \Omega} \max(A(\omega), \bar{B}(\omega)),$$

la crédibilité de A se définit comme l'espérance conditionnelle de N_A :

$$\mathrm{bel}(A) = \sum_{B \in F(m)} m(B) N(A|B) \tag{3.63}$$

Si les ensembles B sont classiques et A est flou, d'après [134],

$$\mathrm{bel}(A) = \mathbb{E}_*(\mu_A). \tag{3.64}$$

Interprétation ensembliste

Les fonctions de plausibilité et de crédibilité classiques étant basées sur des opérations ensemblistes (inclusion, intersection), on peut généraliser les expressions (3.60) et (3.63) en définissant des mesures d'intersection et d'inclusion à l'aide des opérateurs flous correspondants [167]:

$$\mathrm{pl}(A) = \sum_{B \in F(m)} m(B) \mathrm{Int}(A, B), \tag{3.65}$$

$$\mathrm{bel}(A) = \sum_{B \in F(m)} m(B) \mathrm{Inc}_A(B), \tag{3.66}$$

où $\mathrm{Int}(A, B)$ et $\mathrm{Inc}_A(B)$ sont respectivement des mesures d'intersection de A et B et d'inclusion de B dans A. Les équations (3.60) et (3.63) deviennent alors des cas particuliers des expressions précédentes avec

$$\mathrm{Int}(A, B) = \sup_{\omega} \min(A(\omega), B(\omega)) \text{ et } \mathrm{Inc}_A(B) = \inf_{\omega} \max(A(\omega), \bar{B}(\omega)).$$

Yager [167] a proposé de remplacer les opérateurs min et max par d'autres t-normes $\wedge$ et t-conormes $\vee$. Ainsi, on obtient :

$$\mathrm{Int}(A, B) = \bigvee_{\omega} \wedge(A(\omega), B(\omega)) \text{ et } \mathrm{Inc}_A(B) = \bigwedge_{\omega} \vee(A(\omega), \bar{B}(\omega)).$$

Propriété. La relation $\mathrm{pl}(A) = \mathrm{bel}(\Omega) - \mathrm{bel}(A)$ est toujours vérifiée.

Décomposition en α-coupes

En partant du modèle initial de Dempster, Yen [175] propose une autre généralisation des concepts de crédibilité et de plausibilité, basée sur la décomposition de chaque élément focal (flou) A d'une structure de croyance m en un ensemble fini de n éléments focaux classiques, ses α_i-coupes $\{A_{\alpha_i}\}_{i=1}^n$ tels que :

$$\forall i \in \{1, \ldots, n\}\ m(A_{\alpha_i}) \geq 0 \quad \text{et} \quad \sum_{i=1}^{n} m(A_{\alpha_i}) = m(A) \quad \forall A \in F(m). \tag{3.67}$$

L'ensemble Ω est donc ici supposé fini. Yen s'est inspiré des travaux de Dubois et Prade [42] sur la relation entre une structure consonante et une distribution de possibilité, et donc la fonction d'appartenance d'un ensemble flou. Chaque élément focal A_{α_i} a pour masse $m(\alpha_i)(\alpha_i - \alpha_{i-1})$, $i = 1, \ldots, n$ avec $0 = \alpha_0 < \ldots < \alpha_n = 1$. Puisque m devient une structure à éléments focaux non flous, on peut utiliser la définition de plausibilité et de crédibilité de Smets (équations (3.64) et (3.62) d'un ensemble flou A :

$$\text{bel}(A) = \sum_{B \in F} m(B) \sum_{i=1}^{n} (\alpha_i - \alpha_{i-1}) \min_{\omega \in B_{\alpha_i}} A(\omega),$$

$$\text{pl}(A) = \sum_{B \in F} m(B) \sum_{i=1}^{n} (\alpha_i - \alpha_{i-1}) \max_{\omega \in B_{\alpha_i}} A(\omega).$$

La relation $\text{pl}(A) = \text{bel}(\Omega) - \text{bel}(A)$ est là encore vérifiée. Si la structure est définie sur des éléments focaux non flous, on retrouve les équations (3.64) et (3.62).

Dans la méthode de Yen, les mesures de plausibilité et de crédibilité sont plus sensibles à la modification des fonctions d'appartenance de leurs éléments focaux que les autres approches. Ce manque de robustesse, présenté pourtant comme un avantage par Yen, peut s'avérer préjudiciable si l'on considère que le choix de ces fonctions présente souvent un caractère relativement arbitraire.

Dans l'approche de Yen, si Ω est continu, il peut y avoir une infinité d'α-coupes d'un élément focal. En revanche, la première approche (de Zadeh- Smets-Yager), permet facilement l'extension à un référentiel Ω continu. Dans la suite, c'est la première approche que nous retiendrons.

Probabilité pignistique

Le concept de probabilité pignistique peut facilement s'étendre à la théorie des ensembles flous, en utilisant la définition usuelle de la cardinalité d'un ensemble flou A, $|A| = \sum_{\omega \in \Omega} A(\omega)$ si Ω est discret et $|A| = \int_\Omega A(y)dy$ si Ω est continu. L'expression de la probabilité pignistique est alors la suivante:

$$\forall \omega \in \Omega, \quad p_{bet}(\omega) = \sum_{A \in F(m)} \frac{m(A)}{|A|} A(\omega). \tag{3.68}$$

Combinaison de structures

L'étape suivante dans la généralisation de la théorie des croyances aux événements flous est la combinaison des structures. Comme l'a proposé Yager [173], quel que soit l'opérateur de combinaison ∇ utilisé, il peut être remplacé par une de ses versions floues.

Ainsi, soient m_1 et m_2 deux structures de croyance d'éléments focaux respectifs $(A_i)_{i=1}^n$ et $(B_j)_{j=1}^p$. L'opérateur de combinaison disjonctive $\cup$ se généralise facilement. Les éléments focaux C_k de $m_1 \cup m_2$ sont définis par: $C_k = A_i \cup B_j$ où $C_k(\omega) = A_i(\omega) \vee B_j(\omega)$ et $\vee$ est une t-conorme. La masse correspondante est

$$(m_1 \cup m_2)(C_k) = \sum_{A_i \cup B_j = C_k} m_1(A_i) m_2(B_j)$$

De même, les éléments focaux D_k de $m_1 \cap m_2$ sont définis par: $D_k = A_i \cap B_j$ où $D_k(\omega) = A_i(\omega) \wedge B_j(\omega)$ et $\wedge$ est une t-norme, et

$$(m_1 \cap m_2)(D_k) = \sum_{A_i \cap B_j = D_k} m_1(A_i) m_2(B_j).$$

Normalisation

La combinaison de deux structures de croyance normalisées peut donner lieu à une structure de croyance non normalisée m. C'est le cas de la combinaison conjonctive.

Si la condition de normalisation s'avère nécessaire, il est possible d'étendre la règle de normalisation de Dempster aux structures de croyance floues. Yager [173] a proposé la procédure suivante, appelée « *smooth normalization procedure* » (SNP) pour normaliser une structure de croyance floue.

Si m est une structure d'éléments focaux $F \in F(m)$, la SNP convertit m en une structure normalisée m^* d'éléments focaux $E \in F(m^*)$, tels que :

$$\begin{cases} E(\omega) = \dfrac{F(\omega)}{h(F)} \quad \forall \omega \in \Omega \quad \forall F \in F(m) \backslash \emptyset \\ m^*(E) = \dfrac{\sum_{F^*=E} m(F) h(F)}{\sum_{F \in F(m)} m(F) h(F)}, \end{cases} \tag{3.69}$$

où F^* représente la normalisation de l'ensemble flou F définie dans la première équation.

Cette méthode généralise à la fois la normalisation classique d'un ensemble flou et la normalisation des structures de croyances de Dempster. En effet,

1. si m a un unique élément focal flou A ($m = m_A$ est monofocale), elle peut être assimilée à un ensemble flou et la SNP produit l'ensemble normalisé correspondant A^*.

2. si m est une structure classique non normalisée, $F(m^*) = F(m) \backslash \emptyset$. De plus, $h(F) = 0$ si $F = \emptyset$ et $h(F) = 1$ sinon. Ainsi,

$$m^*(F^*) = \frac{m(F)}{\sum_{F \in F(m) \backslash \emptyset} m(F)}.$$

La distribution des masses entre les éléments focaux se justifie intuitivement de la manière suivante. Plus la hauteur d'un élément focal est faible, moins on lui accordera d'importance, et donc plus la masse relative de l'ensemble normalisé correspondant sera également faible. Et si la hauteur est la même pour tous les éléments focaux A de m, alors $m^*(A^*) = m(A)$.

Enfin, une généralisation de la procédure de normalisation de Yager (cf. équation 3.16) est définie ainsi :

$$m^o(A) = \sum_{B^o=A} m(B)$$

où B^o est un ensemble flou normalisé défini par : $B^o(\omega) = B(\omega) + 1 - h(B)$ pour tout ω de Ω.

Mesures d'incertitude

Nous avons vu précédemment que l'on pouvait distinguer deux types d'incertitude en théorie des croyances : la nonspécificité et le conflit. La nonspécificité, mesure d'imprécision, définie par la mesure de Hartley en théorie des ensembles classiques, a été généralisée également en théorie des ensembles flous et en théorie des possibilités. La mesure de conflit, issue de la théorie des probabilités, se généralise également à la théorie des possibilités. On dispose donc d'un cadre commun concernant la définition de l'incertitude pour ces cinq théories (cf. [83]).

Nonspécificité et conflit

La généralisation de la nonspécificité à la théorie des croyances floue est immédiate, si on utilise la définition de la cardinalité d'un ensemble flou :

$$N_1(m) = \sum_{A \in F(m)} m(A) \log_2 |A|, \tag{3.70}$$

avec $|A| = \sum_\Omega A(\omega)$ si Ω est discret. On peut également la définir comme la moyenne pondérée de la nonspécificité des éléments focaux flous :

$$N_2(m) = \sum_{A \in F(m)} m(A) U(A),$$

où U est la mesure de nonspécificité d'un ensemble flou, dont l'expression est définie par l'équation (2.14) si Ω est discret et par

$$U(A) = \frac{1}{h(A)} \int_0^{h(A)} \log_2(1 + |A_\alpha|) d\alpha \tag{3.71}$$

si Ω est continu.

Les différentes mesures de conflit Δ, S, D, C, E présentées en section 3.2.5 se généralisent elles aussi en utilisant la définition de la cardinalité d'un ensemble flou et en remplaçant le cas échéant les opérateurs d'intersection et d'union par leur équivalent en théorie des ensembles flous.

3.3.3 Généralisation des structures de croyances floues

Les structures de croyances floues fournissent un cadre de représentation des croyances en propositions vagues, comme celles qui peuvent être définies par le langage. Cependant, chaque élément focal F reste associé à un nombre réel, précis, sa masse $m(F)$. Or l'importance d'un élément focal, représentée par sa masse en théorie des croyances, peut être elle-même entachée d'incertitude. Denœux [34, 35] a proposé une extension supplémentaire de la théorie en permettant aux masses de croyances d'être des intervalles ou des nombres flous.

3.4 Conclusion

La théorie des croyances présente l'avantage d'offrir une certaine flexibilité dans la représentation et la modélisation des informations incertaines. Elle a été utilisée dans de nombreux domaines d'application [59], comme la fusion multi-sources [80, 75], les systèmes experts [121] ou la reconnaissance des formes [180, 32].

Elle fait partie de nombreuses théories de l'incertain qui ont émergé depuis une trentaine d'années comme alternative à la théorie des probabilités: la théorie des possibilités [179], la théorie des ensembles flous, la théorie des probabilités imprécises [163].

Dans ce chapitre, nous nous sommes particulièrement intéressés à trois aspects : la définition de l'espérance, l'extension à un ensemble de référence continu et la généralisation aux éléments focaux flous. L'association des structures de croyance aux ensembles flous permet de définir les propositions de manière plus souple. L'extension aux ensembles continus et la notion d'espérance sont cruciales dans l'application des croyances à l'estimation fonctionnelle, que nous allons voir dans les deux derniers chapitres.

Chapitre 4

Application de la théorie des croyances en régression

4.1 Introduction

Bien que les méthodes de régression statistiques soient les plus répandues pour modéliser les relations entre variables, elles ne sont pas toujours bien adaptées aux différents types d'incertitude que l'on peut rencontrer dans des applications réelles. Nous avons vu dans les chapitres précédents que la théorie des croyances, la théorie des ensembles flous, la théorie des possibilités, permettent la représentation de plusieurs types d'incertitude, comme le conflit, l'imprécision ou l'ignorance [84] et offrent un cadre plus général que la théorie des probabilités pour l'analyse des données. Plusieurs approches, comme la régression linéaire par intervalle ou floue [39], et les systèmes flous [173, 150] ou neuro-flous [78] intègrent des éléments flous, entrées ou paramètres, dans l'estimation fonctionnelle. Une méthode neuronale basée sur la théorie de l'évidence a également été proposée comme méthode d'approximation fonctionnelle [31].

Nous proposons dans ce chapitre une nouvelle méthode de régression basée sur la théorie des croyances floues, que l'on peut appliquer non seulement à des valeurs réelles, mais aussi à des données imprécises comme les intervalles ou les nombres flous. Le cas des sorties conflictuelles est également pris en compte : pour une même entrée, on dispose de plusieurs sorties qui peuvent être contradictoires. L'information fournie sur la sortie est délivrée sous la forme très générale d'une structure de croyance. Cette structure est déterminée à partir de l'ensemble d'apprentissage, selon un principe récemment développé par Denœux [31, 30] dans un contexte de discrimination. La généralisation de l'espérance en théorie des croyances permet de définir une sortie ponctuelle à partir d'une distribution de probabilité compatible avec la structure de croyance obtenue.

4.2 Présentation générale

4.2.1 Modèle complet

Principe de la méthode

Nous proposons d'utiliser la théorie des croyances dans le cadre de la régression. Le principe de notre approche a été introduit par Denœux [30, 31] dans le contexte de la classification supervisée. Nous présentons dans un premier temps une approche générale dont la régression et la discrimination sont deux cas particuliers.

Soit $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \ldots \mathcal{X}_r$ et $\mathcal{Y}$ des espaces de référence réels, domaines de variations respectifs de l'entrée de dimension r et de la sortie monodimensionnelle d'un système. Dans cette section, l'espace $\mathcal{X}$ est supposé continu mais $\mathcal{Y}$ peut être *discret ou continu, ordonné ou non*. Les entrées $\boldsymbol{x}$ sont réelles mais la forme des sorties y est très variée. Celles-ci peuvent être définies par des réels, des intervalles, des nombres flous ou des quantités floues. On suppose que la connaissance relative à la sortie peut être représentée de façon générale par une structure de croyance, éventuellement floue. On note $\mathcal{MF}(\mathcal{Y})$ l'ensemble des structures de croyance floues définies sur $\mathcal{Y}$.

Soit $(\boldsymbol{x}_i, m_i)_{i=1,\ldots,N}$ un ensemble de vecteurs d'apprentissage appartenant à $\mathcal{X} \times \mathcal{MF}(\mathcal{Y})$, où m_i est une structure de croyance floue d'éléments focaux réels ou flous

$$F(m_i) = \{\tilde{y}_{i1}, \ldots, \tilde{y}_{ij}, \ldots, \tilde{y}_{i,J(i)}\},$$

chaque $\tilde{y}_{ij}$ étant une quantité réelle ou floue et $m_i(\tilde{y}_{ij}) = p_{ij}$. On notera $\{x_{i1} \ldots, x_{ir}\}$ les coordonnées de $\boldsymbol{x}_i$.

Soit $\boldsymbol{x}$ un vecteur d'entrée arbitraire, et y la sortie correspondante, supposée inconnue. Chaque élément $(\boldsymbol{x}_i, m_i)$ de l'ensemble d'apprentissage apporte une information sur la valeur possible de y qui peut se représenter par une fonction de croyance. Si $\boldsymbol{x}$ est « proche » de $\boldsymbol{x}_i$ selon une certaine métrique $\|.\|$, nous pouvons raisonnablement supposer que y sera également « proche » de m_i. Ceci peut se traduire par l'affectation d'une certaine masse de croyance aux éléments focaux de m_i, dont la valeur dépend de la distance entre $\boldsymbol{x}$ et $\boldsymbol{x}_i$. En l'absence d'information supplémentaire, le complément à 1 de la masse peut être alloué à l'ensemble de référence $\mathcal{Y}$. Nous obtenons alors une structure de croyance normalisée $\widehat{m}_i(.|\boldsymbol{x})$ sur $\mathcal{Y}$, d'éléments focaux $F[\widehat{m}_i(.|\boldsymbol{x})] = F(m_i) \cup \mathcal{Y}$, définie de la façon suivante :

$$\left\{ \begin{array}{lcl} \widehat{m}_i(\tilde{y}_{ij}|\boldsymbol{x}) & = & p_{ij}\phi_i\left[\|\boldsymbol{x} - \boldsymbol{x}_i\|\right] \quad \forall j \in \{1, \ldots, J(i)\} \\ \widehat{m}_i(\mathcal{Y}|\boldsymbol{x}) & = & 1 - \phi_i\left[\|\boldsymbol{x} - \boldsymbol{x}_i\|\right] \\ \widehat{m}_i(A|\boldsymbol{x}) & = & 0 \quad \forall A \in \mathcal{F}(\mathcal{Y}) \setminus F(m_i(.|\boldsymbol{x})), \end{array} \right. \tag{4.1}$$

où ϕ_i est une fonction décroissante de $\mathbb{R}^+$ dans $[0, 1]$ telle que :

$$\phi_i(0) \in]0, 1[\text{ et } \lim_{d \to \infty} \phi_i(d) = 0. \tag{4.2}$$

On appellera ϕ_i la *fonction d'activation* de $\boldsymbol{x}_i$. Dans la suite, nous noterons $g_i(\boldsymbol{x}, \boldsymbol{\theta}_i) = \phi_i[\|\boldsymbol{x} - \boldsymbol{x}_i\|]$, où $\boldsymbol{\theta}_i$ est l'ensemble des paramètres intervenant dans la fonction ϕ_i.

L'opération définie par l'équation (4.1) correspond à un affaiblissement de la structure m_i dont le facteur est $\alpha_i = 1 - g_i(\boldsymbol{x}, \boldsymbol{\theta})$. Ainsi, la structure produite est définie de manière synthétique par :

$$\widehat{m}_i(.|\boldsymbol{x}) = m_i^{\alpha_i} \tag{4.3}$$

Ce facteur d'affaiblissement détermine l'influence de $\boldsymbol{x}_i$ sur $\boldsymbol{x}$. Si $\boldsymbol{x}$ est très proche de $\boldsymbol{x}_i$, α_i est proche de 0 et les structures $\widehat{m}_i$ et m_i sont presque identiques. Quand $\boldsymbol{x}$ s'éloigne indéfiniment de $\boldsymbol{x}_i$, α_i tend vers 0 et $\widehat{m}_i$ devient la structure triviale $\widehat{m}_{\mathcal{Y}}$.

Afin de combiner l'information fournie par chaque élément de l'ensemble d'apprentissage, on peut utiliser la généralisation aux structures de croyances floues de la règle de combinaison conjonctive non normalisée $\cap$, quoique d'autres opérateurs de combinaison soient envisageables *a priori*. La structure de croyance finale est alors définie par :

$$\widehat{m}(.|\boldsymbol{x}) = \bigcap_{i=1}^{N} \widehat{m}_i(.|\boldsymbol{x}) \tag{4.4}$$

On note $\widehat{m}^*(.|\boldsymbol{x})$ la structure de croyance floue obtenue en normalisant $\widehat{m}(.|\boldsymbol{x})$ par la procédure SNP (cf. équation 3.69).

Dans cette méthode, on est donc amené à faire deux types de choix *a priori* pour définir le modèle, avant même une éventuelle phase d'identification :

- les fonctions d'activation ϕ_i,
- et l'opérateur de combinaison $\cap$.

On notera SCF1 le modèle complet défini par l'équation (4.1). Le nombre d'éléments focaux de $\widehat{m}(.|\boldsymbol{x})$ peut donc atteindre :

$$\prod_{i=1}^{N} (J(i)+1).$$

Ceci peut poser des problèmes de temps de calculs. Nous verrons dans le chapitre suivant de façon détaillée différentes méthodes permettant de pallier ce type de problèmes. L'une des possibilités est de diminuer le nombre de structures à combiner. Ainsi, par exemple, la combinaison peut éventuellement se restreindre aux k plus proches voisins de $\boldsymbol{x}$, au sens de $\|.\|$, dans l'ensemble d'apprentissage : $\{\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(k)}\}$. La structure de croyance floue finale devient alors :

$$\widehat{m}(.|\boldsymbol{x}) = \bigcap_{i=1}^{k} \widehat{m}_{(i)}(.|\boldsymbol{x}) \tag{4.5}$$

Nous détaillerons dans le chapitre 5 d'autres techniques de simplifications de la méthode.

Choix de l'opérateur de combinaison

Le choix d'un opérateur de type conjonctif se justifie par la propriété de neutralité de la structure triviale $m_{\mathcal{Y}}$:

$$\forall m \in \mathcal{MF}(\mathcal{Y}), \quad m \cap m_{\mathcal{Y}} = m.$$

En effet, il est nécessaire de s'assurer que l'influence d'un vecteur $\boldsymbol{x}_i$ éloigné de $\boldsymbol{x}$ sur l'estimation de la sortie correspondante y est négligeable. L'utilisation d'un opérateur de type disjonctif est donc exclue. En revanche, le choix de la norme triangulaire représentant l'opérateur conjonctif n'a pas grande importance, comme des simulations l'ont confirmé.

Choix de la fonction d'activation

Une infinité de fonctions ϕ_i vérifient les conditions requises dans l'équation (4.2). Ces fonctions d'activation offrent des similitudes avec les fonctions de noyaux, mais la contrainte sur l'intégrale de la fonction est remplacée par une contrainte sur la valeur maximale.

Parmi les fonctions acceptables, on peut citer les deux suivantes :

- $\phi_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) = \exp[-(\boldsymbol{x} - \boldsymbol{x}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{x}_i)]$,
- $\phi_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) = \left(1 + (\boldsymbol{x} - \boldsymbol{x}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{x}_i)\right)^{-1}$,

où $\|.\|$ est la distance euclidienne induite par une matrice symétrique définie positive $\boldsymbol{\Sigma}_i$. L'usage de la fonction exponentielle a été justifiée théoriquement par Denœux [33] mais d'autres fonctions sont envisageables sur le plan pratique.

On peut remarquer que de nombreuses fonctions d'appartenance d'ensembles flous multidimensionnels conviennent à cette définition de fonction d'activation. Ainsi, par exemple, ϕ_i peut également être définie comme le produit de nombres flous triangulaires centrés sur x_{ij} :

$$\phi_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) = \prod_{j=1}^{r} \mathrm{Tri}\left(x_j; x_{ij} - a_{ij}, x_{ij}, x_{ij} + b_{ij}\right), \quad a_{ij} > 0,\ b_{ij} > 0.$$

Cette quantité peut donc se voir comme le degré d'appartenance de $\boldsymbol{x}$ à un ensemble flou représentant $\boldsymbol{x}_i$. Une analogie détaillée entre notre méthode et un certain type de systèmes flous fera l'objet de la section 4.6.

Informations fournies

Les informations fournies par la sortie $\widehat{m}(.|\boldsymbol{x})$ sont multiples. Différentes caractéristiques peuvent être déduites de $\widehat{m}(.|\boldsymbol{x})$. D'après le chapitre précédent, la probabilité pignistique est définie par :

$$\widehat{p}_{bet}(y|\boldsymbol{x}) = \sum_{A \in F(\widehat{m}^*)} \frac{\widehat{m}^*(A|\boldsymbol{x})}{|A|} A(y) \quad \forall y \in \mathcal{Y}.$$

On peut également définir la plausibilité $\widehat{\mathrm{pl}}(.|\boldsymbol{x})$ et la crédibilité $\widehat{\mathrm{bel}}(.|\boldsymbol{x})$ ainsi que différentes mesures d'incertitude.

Plusieurs types d'incertitude peuvent en effet être identifiés dans la structure finale :

1. L'imprécision ou la nonspécificité, relative à la cardinalité des $\tilde{y}_{ij}$. En particulier, l'ignorance est représentée par le poids affecté à l'ensemble $\mathcal{Y}$, l'ensemble vide et les éléments focaux de hauteur faible,

2. Le conflit, représenté par le choix entre les différents éléments focaux.

3. Le flou, si les $\tilde{y}_{ij}$ initiaux sont flous, caractérisé par les frontières des éléments focaux.

Exemple

La figure 4.1 donne un exemple de détermination de la sortie $\widehat{m}(.|\boldsymbol{x})$ à partir de deux éléments $(\boldsymbol{x}_1, m_1)$ et $(\boldsymbol{x}_2, m_2)$ d'un ensemble d'apprentissage. Le cadre de discernement $\mathcal{Y}$ est l'intervalle $[0, 10]$. Les structures m_1 et m_2 sont constituées de nombres flous triangulaires :

- $F(m_1) = \{A, B\}$, $A = \text{Tri}(y; 0, 2, 4)$, avec $m_1(A) = 0.4$, et $B = \text{Tri}(y; 2, 4, 6)$, avec $m_1(B) = 0.6$.
- $F(m_2) = \{C\}$, $C = \text{Tri}(y; 3, 6, 8)$.

La détermination de $\widehat{m}(.|\boldsymbol{x})$ se décompose en 3 étapes :

1. Calcul des facteurs d'affaiblissement :

   On suppose que $\alpha_1 = 0.8 > \alpha_2 = 0.4$. Le vecteur $\boldsymbol{x}$ est plus proche de $\boldsymbol{x}_1$ que de $\boldsymbol{x}_2$. L'influence de $\boldsymbol{x}_2$ sera faible par rapport à celle de $\boldsymbol{x}_1$.

2. Calcul des structures $\widehat{m}_i$.

   - $F(\widehat{m}_1(.|\boldsymbol{x})) = \{A, B, \mathcal{Y}\}$, avec $\widehat{m}_1(A|\boldsymbol{x}) = 0.32$, $\widehat{m}_1(B|\boldsymbol{x}) = 0.48$ et $\widehat{m}_1(\mathcal{Y}|\boldsymbol{x}) = 0.2$.
   - $F(\widehat{m}_2(.|\boldsymbol{x})) = \{C, \mathcal{Y}\}$, avec $\widehat{m}_2(C|\boldsymbol{x}) = 0.4$ et $\widehat{m}_2(\mathcal{Y}|\boldsymbol{x}) = 0.6$.

3. Calcul de la structure finale $\widehat{m}(.|\boldsymbol{x}) = \widehat{m}_1(.|\boldsymbol{x}) \cap \widehat{m}_2(.|\boldsymbol{x})$, constituée de 6 éléments focaux :

   $F(\widehat{m}(.|\boldsymbol{x})) = \{A, B, C, \mathcal{Y}, D, E\}$, où $E = A \cap C$ et $D = B \cap C$, avec $D = B \cap C$ et $E = A \cap C$, avec

   $\widehat{m}(A|\boldsymbol{x}) = 0.192$, $\widehat{m}(B|\boldsymbol{x}) = 0.288$, avec $\widehat{m}(C|\boldsymbol{x}) = 0.08$, $\widehat{m}(\mathcal{Y}|\boldsymbol{x}) = 0.12$, $\widehat{m}(A \cap C|\boldsymbol{x}) = 0.128$ et $\widehat{m}(B \cap C|\boldsymbol{x}) = 0.192$.

4.2.2 Cas particuliers

Plusieurs cas particuliers au modèle SCF1 peuvent être envisagés, selon la complexité de la sortie m_i (cf. figure 4.2) :

1. m_i est une structure de croyance classique. Les éléments focaux de m_i sont des ensembles classiques de $\mathcal{Y}$;
2. m_i est une structure bayésienne, c'est-à-dire, une distribution de probabilité sur $\mathcal{Y}$;
3. m_i est un ensemble flou ;
4. m_i est un ensemble classique ;
5. m_i est une valeur ponctuelle.

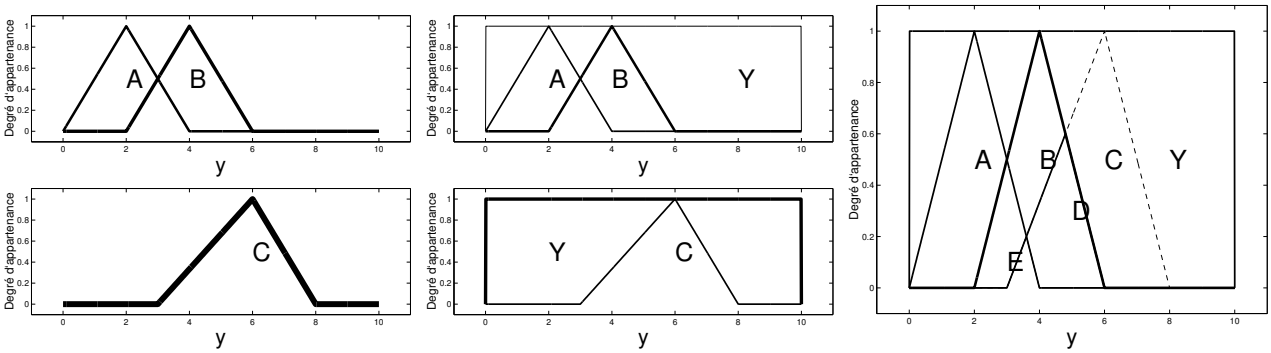


Fig. 4.1 – *Exemple de détermination de la sortie* $\widehat{m}(.|\boldsymbol{x})$ *avec un ensemble d'apprentissage comportant deux éléments* $\{(\boldsymbol{x}_1, m_1), (\boldsymbol{x}_2, m_2)\}$). **A gauche** m_1*, définie par deux éléments focaux flous* A *et* B*, et* m_2*, d'élément focal unique* C*.* **Au milieu.** $\widehat{m}_1(.|\boldsymbol{x})$ *et* $\widehat{m}_2(.|\boldsymbol{x})$*.* **A droite.**$\widehat{m}(.|\boldsymbol{x})$*.* $F(\widehat{m}(.|\boldsymbol{x}) = \{A, B, C, \mathcal{Y}, D, E\}$*, où* $E = A \cap C$ *et* $D = B \cap C$*. Le vecteur* $\boldsymbol{x}$ *étant plus proche de* $\boldsymbol{x}_1$*, que de* $\boldsymbol{x}_2$*, la sortie est également plus proche de* m_1 *que de* m_2*. L'influence de* $\boldsymbol{x}_2$ *est ici plus faible.*

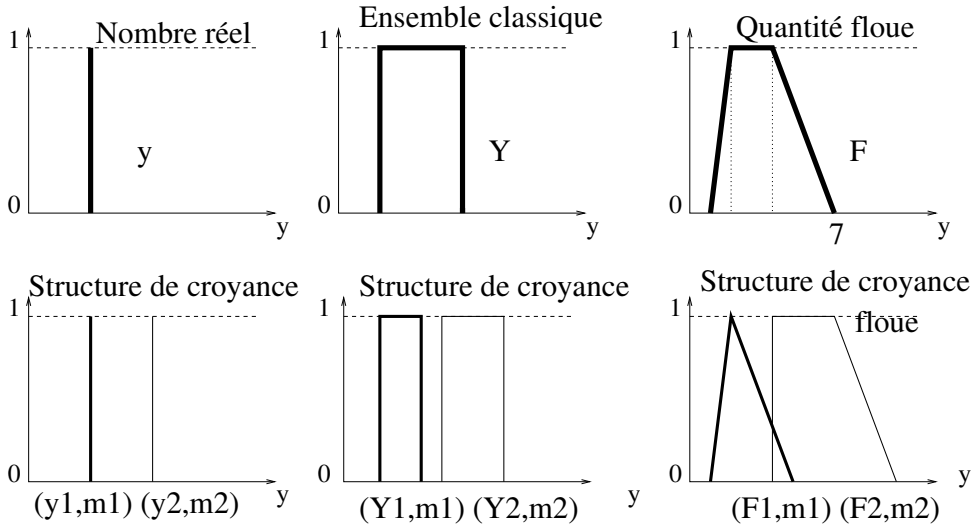


Fig. 4.2 – *Différents types de sorties* m_i *pris en compte par notre modèle.*

Dans les cas 1, 2, 4 et 5, on ne travaille qu'avec des éléments focaux classiques. La structure de croyance finale sera donc une SC classique.

Dans les cas 3 et 4, il n'y a pas d'ambiguïté entre différentes sorties possibles pour les $\boldsymbol{x}_i$, mais il subsiste une imprécision relativement à cette sortie. La structure de croyance m_i se résume à un unique élément focal (flou ou non) $\tilde{y}_i$ et peut donc être assimilée à cet ensemble. Pour le vecteur $\boldsymbol{x}$, la structure $\widehat{m}_i(.|\boldsymbol{x})$ induite par $(\boldsymbol{x}_i, \tilde{y}_i)$, ne possède alors que deux éléments focaux : $F[\widehat{m}_i(.|\boldsymbol{x})] = \{\tilde{y}_i, \mathcal{Y}\}$ et devient par conséquent :

$$\left\{ \begin{array}{lcl} \widehat{m}_i(\tilde{y}_i|\boldsymbol{x}) & = & \phi_i\left[\|\boldsymbol{x}-\boldsymbol{x}_i\|\right] \\ \widehat{m}_i(\mathcal{Y}|\boldsymbol{x}) & = & 1-\phi_i\left[\|\boldsymbol{x}-\boldsymbol{x}_i\|\right] \\ \widehat{m}_i(A|\boldsymbol{x}) & = & 0 \quad \forall A \in \mathcal{F}(\mathcal{Y})\backslash F(m_i(.|\boldsymbol{x})), \end{array} \right. \tag{4.6}$$

Par la suite, le modèle défini par l'équation (4.6) sera noté SCF2.

Comme cela a été souligné par Ph. Smets (communication personnelle), si les $\tilde{y}_i$ sont des nombres réels y_i (cas 5), $\widehat{m}_i(.|\boldsymbol{x})$ et $\widehat{m}(.|\boldsymbol{x})$ deviennent des structures de croyance classiques avec seulement $N+1$ éléments focaux et la structure finale normalisée peut s'exprimer facilement (modèle SCF3) :

$$\left\{ \begin{array}{lcl} \widehat{m}^*(\{y_i\}|\boldsymbol{x}) & = & \frac{1}{K}\phi_i\left[\|\boldsymbol{x}-\boldsymbol{x}_i\|\right]\prod_{j\neq i}[1-\phi_j\left[\|\boldsymbol{x}-\boldsymbol{x}_j\|\right] \\ \widehat{m}^*(\mathcal{Y}|\boldsymbol{x}) & = & \frac{1}{K}\prod_{i=1}^{N}[1-\phi_i\left[\|\boldsymbol{x}-\boldsymbol{x}_i\|\right] \\ \widehat{m}^*(A|\boldsymbol{x}) & = & 0 \quad \forall A \in \mathcal{F}(\mathcal{Y}) \setminus F[\widehat{m}(.|\boldsymbol{x})], \end{array} \right. \tag{4.7}$$

où

$$K = \prod_{i=1}^{N}\{1-\phi_i\left[\|\boldsymbol{x}-\boldsymbol{x}_i\|\right]\} + \sum_{i=1}^{N}\phi_i\left[\|\boldsymbol{x}-\boldsymbol{x}_i\|\right]\prod_{j\neq i}\{1-\phi_j\left[\|\boldsymbol{x}-\boldsymbol{x}_j\|\right]\}$$

est le facteur de normalisation.

4.3 Application en reconnaissance des formes

4.3.1 Traduction en terme de classification

Dans cette section, nous appliquons le principe vu précédemment à un problème de classification. Les résultats suivants ont été largement développés, sous différentes formes dans [30, 180, 33]. Nous nous contentons ici d'en retracer les grandes lignes. L'espace $\mathcal{Y}$ est défini ici comme un ensemble fini de K classes envisageables : $\mathcal{Y} = \{c_1, \ldots, c_K\}$. Dans le cas le plus simple (cas 5), l'ensemble d'apprentissage est constitué de N vecteurs étiquetés $(\boldsymbol{x}_i, y_i)$, où la variable d'étiquetage y_i est l'une quelconque des classes c_k. Dans cette situation, il n'y a ni imprécision, ni ambiguïté dans la description des classes des vecteurs d'apprentissage. Dans le cas d'un étiquetage imprécis (cas 3 et 4), la classe de $\boldsymbol{x}_i$ n'est plus parfaitement connue. L'information que l'on possède peut par exemple être du type : « $\boldsymbol{x}_1$ appartient à la classe c_1 *ou* à la classe c_2 ». Dans ces conditions, le vecteur $\boldsymbol{x}_i$ est étiqueté par l'ensemble classique $\{c_1, c_2\}$. De façon plus générale, on peut définir un degré d'appartenance $\mu_{ik} \in [0,1]$ de $\boldsymbol{x}_i$ à chacune des K classes. L'étiquetage $\tilde{y}_i$ devient alors l'ensemble flou de $\mathcal{Y}$ défini par les μ_{ik} [181, 182]. Enfin, l'information concernant la classe du vecteur d'apprentissage $\boldsymbol{x}_i$ peut être également définie sous forme d'une structure de croyance m_i (cas 1 ou 2), éventuellement

floue (cas général). Ce dernier cas peut par exemple correspondre à une agrégation d'avis d'experts.

Pour un nouveau vecteur $\boldsymbol{x}$, l'information fournie par le vecteur $(\boldsymbol{x}_i, m_i)$ se traduit par une structure de croyance $\widehat{m}_i(.|\boldsymbol{x})$ dont l'expression est définie par l'équation (4.7), (4.6) ou (4.1) selon le cas. Quelque soit le cas de figure, on obtient une information globale concernant la classe de $\boldsymbol{x}$, sous la forme d'une structure de croyance classique ou floue $\widehat{m}(.|\boldsymbol{x})$. Nous allons voir comment choisir la classe d'affectation de $\boldsymbol{x}$.

4.3.2 Prise de décision

Nous avons vu au chapitre précédent comment il était possible de généraliser la théorie bayésienne de la décision aux structures de croyances. Soit $\mathcal{D}$ l'ensemble des décisions possibles et f_d la fonction de coût ou *fonction d'utilité*, définie de $\mathcal{Y}$ dans $\mathbb{R}$, associée au choix de la décision particulière d. Soit Y une variable de l'espace crédibilisé $(\mathcal{Y}, S(\mathcal{Y}), m)$.

Cas des structures de croyance classiques

Dans le cas de structures de croyance classiques, parmi les différentes stratégies possibles (cf. chapitre 3, section 3.2.7), la plus simple consiste à choisir une probabilité particulière compatible avec $\widehat{m}(.|\boldsymbol{x})$, comme la probabilité pignistique, afin de se placer dans les conditions habituelles de la théorie bayésienne de la décision :

Ainsi, le risque (pignistique) conditionnel à $\boldsymbol{x}$ associé à la décision d s'écrit, d'après l'équation (3.53) :

$$R_{p_{bet}}(d|\boldsymbol{x}) = \mathbb{E}_{P_{bet}}[f_d(Y)|\boldsymbol{x}] = \sum_{k=1}^{K} f_d(c_k)\widehat{P}_{bet}(\{c_k\}|\boldsymbol{x}). \tag{4.8}$$

On peut également définir les risques conditionnels inférieur et supérieur, définis par :

$$R_*(d|\boldsymbol{x}) = \mathbb{E}_*[f_d(Y)|\boldsymbol{x}] = \sum_{A \in F[\widehat{m}(.|x)]} \widehat{m}(A|\boldsymbol{x}) \min_{c \in A} f_d(c), \tag{4.9}$$

$$R^*(d|\boldsymbol{x}) = \mathbb{E}^*[f_d(Y)|\boldsymbol{x}] = \sum_{A \in F(\widehat{m}(.|x))} \widehat{m}(A|\boldsymbol{x}) \max_{c \in A} f_d(c), \tag{4.10}$$

d'après l'équation (3.51), ou utiliser les généralisations de Yager (équation 3.56) ou Strat (équation 3.54).

Supposons que l'ensemble des décisions possibles $\mathcal{D}$ soit celui des classes $\mathcal{Y}$, et que les fonctions de coûts soient binaires, à valeur dans $\{0, 1\}$ (0 pour une bonne classification, 1 pour une mauvaise classification) : $f_{c_i}(c_j) = 0$ si $j = i$ et $f_{c_i}(c_j) = 1$ si $j \neq i$. Dans ces conditions, les expressions des risques pignistique, inférieur et supérieur deviennent :

$$R_{\widehat{p}_{bet}}(c_j|\boldsymbol{x}) = 1 - \widehat{P}_{bet}(\{c_j\}|\boldsymbol{x}),$$

$$R_*(c_j|\boldsymbol{x}) = 1 - \widehat{\mathrm{pl}}(\{c_j\}|\boldsymbol{x}),$$

$$R^*(c_j|\boldsymbol{x}) = 1 - \widehat{\mathrm{bel}}(\{c_j\}|\boldsymbol{x}).$$

Par analogie avec la théorie bayésienne de la décision, selon la stratégie utilisée, la minimisation du risque conduira à choisir la classe c^* de plus grande probabilité pignistique, de plus grande plausibilité ou de plus grande crédibilité.

Ces trois stratégies peuvent donner lieu à des décisions différentes. Dans le cas particulier de l'étiquetage parfait où les y_i sont réels (cas 5), les trois stratégies précédentes mènent à la même décision (dans le cas des coûts $\{0,1\}$).

D'autres stratégies de décision ont été étudiées dans [32], incluant la possibilité de rejets dits d'ambiguité et de distance [47] ainsi que l'existence de classes inconnues.

Structures de croyance floues

Dans le cas où la connaissance concernant les classes des vecteurs d'apprentissage est exprimée sous la forme de structures de croyance floues, on peut se baser sur une généralisation des critères définis précédemment dans le cadre des structures classiques. Le vecteur $\boldsymbol{x}$ sera là encore affecté à la classe de plus grande probabilité pignistique, de plus grande crédibilité ou de plus grande plausibilité. On utilise l'extension de la définition de ces quantités aux ensembles flous (cf. équations (3.68), (3.66) et (3.65)) [37].

4.3.3 Apprentissage

Critères existants

La qualité de la classification dépend du vecteur de paramètres $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_n]$, qui intervient dans les fonctions d'activation $g_i(.,\boldsymbol{\theta}_i)$ et donc dans la structure $\widehat{m}(.|\boldsymbol{x},\boldsymbol{\theta})$ et ses dérivés. La sortie m_i de chaque vecteur $\boldsymbol{x}_i$ peut être estimée comme précédemment par $\widehat{m}(.|\boldsymbol{x}_i)$ en utilisant une partie de l'ensemble d'apprentissage contenant d'autres vecteurs à l'aide d'une méthode classique de type rééchantillonnage. Plus m_i est « proche » de $\widehat{m}(.|\boldsymbol{x}_i,\boldsymbol{\theta})$, plus le vecteur de paramètres $\boldsymbol{\theta}$ semble adapté au modèle. Cette notion de proximité, de distance entre structures de croyance sera développée ultérieurement. Dans [182], les auteurs ont proposé de choisir ce paramètre en minimisant un critère d'erreur $J_{bet}(\boldsymbol{\theta})$ basé sur l'écart entre les probabilités pignistiques estimées et réelles des classes des vecteurs d'apprentissage, défini par :

$$J_{bet}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} C\left(\widehat{m}(.|\boldsymbol{x}_i,\boldsymbol{\theta}), m_i\right) \tag{4.11}$$

avec

$$C\left(\widehat{m}(.|\boldsymbol{x}_i,\boldsymbol{\theta}), m_i\right) = \sum_{k=1}^{K}[\widehat{P}_{bet}(\{c_k\}|\boldsymbol{x}_i,\boldsymbol{\theta}) - {P_{bet}}^i(\{c_k\})]^2 \tag{4.12}$$

où ${P_{bet}}^i$ est la probabilité pignistique associée à m_i. Dans le cas d'un étiquetage parfait, m_i correspond à une classe particulière c_k et ${P_{bet}}^i$ est la loi de Dirac δ_{c_k}. Si la classe de $\boldsymbol{x}_i$ n'est pas connue précisément, mais fait partie d'un sous-ensemble classique $A \subseteq \mathcal{Y}$, ${P_{bet}}^i$ est la loi uniforme sur A. Dans le cas général, m_i est une structure de croyance quelconque et ${P_{bet}}^i$ est la loi de probabilité correspondante.

Ce critère est évidemment un choix parmi d'autres. Il est satisfaisant si les m_i sont précises, c'est-à-dire, si l'étiquetage des classes est parfaitement connu. En revanche, si m_i est très incertaine, ce critère est un peu rigide : il est illusoire de chercher une sorte de « distance » précise entre deux structures, si la structure de référence est très incertaine. En particulier, dans le cas limite de la structure triviale où l'ignorance est totale, il n'y a plus d'apprentissage et toutes les solutions $\widehat{m}_i$ se valent ![1]. Ce n'est pas le cas si on utilise le critère précédent. Illustrons ce problème par un exemple.

Exemple. Soit $\mathcal{Y} = \{c_1, c_2\}$.

- Soit $\widehat{m}_i$ défini par : $\widehat{m}_i(\{c_1\}) = 0.7$; $\widehat{m}_i(\{c_2\}) = 0.1$ et $\widehat{m}_i(\mathcal{Y}) = 0.2$. Alors $\widehat{P}_{bet}(\{c_1\}) = 0.8$ et $\widehat{P}_{bet}(\{c_2\}) = 0.2$
- Soit $\widehat{m}'_i$ défini par : $\widehat{m}'_i(\{c_1\}) = 0.2$; $\widehat{m}'_i(\{c_2\}) = 0.2$ et $\widehat{m}'_i(\mathcal{Y}) = 0.6$. Alors $\widehat{P}'_{bet}(\{c_1\}) = 0.5$ et $\widehat{P}'_{bet}(\{c_2\}) = 0.5$

1. Si $m_i = \{c_1\}$, ${P_{bet}}^i(\{c_1\}) = 1$ et ${P_{bet}}^i(\{c_2\}) = 0$. Les coûts associés à $\widehat{m}_i$ et $\widehat{m}'_i$ sont respectivement de 0.08 et 0.5.

   Ce résultat est conforme à nos attentes. Plus $P_{bet}(\{c_1\})$ est proche de 1, plus l'estimation est correcte.

2. Si $m_i = m_{\mathcal{Y}}$, ${P_{bet}}^i(\{c_1\}) = 0.5$ et ${P_{bet}}^i(\{c_2\}) = 0.5$. Les coûts associés à $\widehat{m}_i$ et $\widehat{m}'_i$ sont respectivement de 0.18 et 0.

   Ce résultat n'est pas satisfaisant. Puisque l'on ne possède aucune information sur la classe de $\boldsymbol{x}_i$, il n'y a aucune raison de privilégier une estimation par rapport à une autre, comme c'est le cas ici.

Définition d'un nouveau critère

Afin d'éviter cette situation, nous proposons de remplacer le coût (4.12) par l'expression suivante :

$$C'\big(\widehat{m}(.|\boldsymbol{x}_i, \boldsymbol{\theta}), m_i\big) = \sum_{A \in F(m_i)} m_i(A)[\widehat{P}_{bet}(A|\boldsymbol{x}_i, \boldsymbol{\theta}) - 1]^2. \tag{4.13}$$

On obtient alors le coût moyen

$$J'_{bet}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} C'\big(\widehat{m}(.|\boldsymbol{x}_i, \boldsymbol{\theta}), m_i\big). \tag{4.14}$$

Ce nouveau critère a de bonnes propriétés. En effet, supposons que m_i soit une structure focalisée sur un unique élément A, qui peut être un nombre réel, un ensemble classique ou, de manière générale, un ensemble flou (cas 3, 4 et 5). La structure estimée $\widehat{m}_i$ est d'autant plus adaptée à m_i que la probabilité pignistique de A est proche de 1. Dans le cas de l'absence d'étiquetage, c'est-à-dire d'ignorance totale de la classe de $\boldsymbol{x}_i$, $m_i = m_{\mathcal{Y}}$, et

$$\widehat{{P_{bet}}^i}(\mathcal{Y}) = 1,$$

1. Ceci devrait d'ailleurs être érigé en axiome pour le choix d'un bon critère.

quelle que soit la structure $\widehat{m}_i$. Dans le cas d'un étiquetage parfait, m_i est focalisée sur l'une des classes c_k et dans ce cas, le critère devient

$$C(\widehat{m}_i, m_i) = (\widehat{P}_{bet}(\{c_k\}) - 1)^2.$$

Si la structure m_i a plusieurs éléments focaux, le critère global est simplement la moyenne pondérée par la masse de chaque élément A de l'écart quadratique entre $\widehat{P_{bet}}^i(A)$ et 1.

Si on reprend l'exemple précédent, on obtient :

1. Si $m_i = \{c_1\}$, $C(\widehat{m}_i, m_i) = 0.04$ et $C(\widehat{m}'_i, m_i) = 0.25$.
2. Si $m_i = m_{\mathcal{Y}}$, $C(\widehat{m}_i, m_i) = C(\widehat{m}'_i, m_i) = 0$, ce qui est conforme à nos attentes.

Des travaux visant à harmoniser les critères d'erreur dans le cas de la discrimination et de la régression, en théorie des croyances, sont actuellement en cours. Malheureusement, le critère défini par l'équation (4.13) n'est pas applicable en régression (cf. section 4.5).

4.4 Application en régression

4.4.1 Détermination de la sortie

Nous allons nous recentrer sur le problème de la régression. Cette fois, l'espace de sortie $\mathcal{Y}$ est *continu*. Comme dans les chapitres précédents, connaissant l'entrée $\boldsymbol{x}$, on cherche, avec toutes les informations que l'on possède, c'est-à-dire l'ensemble d'apprentissage $\{(\boldsymbol{x}_i, m_i)\}$, à estimer la sortie correspondante y.

Cette sortie est définie de façon très générale par la structure de croyance (floue) $\widehat{m}(.|\boldsymbol{x})$ des équations (4.1), (4.4), (4.6), ou (4.7).

Soit Y une variable aléatoire définie sur l'espace probabilisé $(\mathcal{Y}, S(\mathcal{Y}), P_{bet}(.|\boldsymbol{x}))$. Si on est intéressé par une estimation ponctuelle $\hat{y}$ de Y, celle-ci peut être obtenue en prenant l'espérance de Y (cf. chapitre précédent) :

$$\widehat{y}(\boldsymbol{x}) = \mathbb{E}_{\widehat{p}_{bet}(.|x)}(Y) = \sum_{A \in F[\widehat{m}^*(.|x)]} \frac{\widehat{m}^*(A|\boldsymbol{x})}{|A|} \int_{\mathcal{Y}} uA(u)du.$$

où $\widehat{m}^*(.|\boldsymbol{x})$ est la structure de croyance normalisée associée à $\widehat{m}(.|\boldsymbol{x})$, obtenue par la procédure SNP [2]. Soit y^*_A le centre de gravité de A :

$$y^*_A \triangleq \frac{\int_{\mathcal{Y}} yA(y)dy}{\int_{\mathcal{Y}} A(y)dy}$$

Alors

$$\widehat{y}(\boldsymbol{x}) = \sum_{A \in F(\widehat{m}^*(.|x))} \widehat{m}^*(A|\boldsymbol{x}) y^*_A. \tag{4.15}$$

L'utilisation de la probabilité pignistique permet de définir un certain nombre de caractéristiques, comme la médiane et un intervalle interfractile autour de cette valeur médiane, que l'on peut appeler abusivement un « intervalle de confiance ».

2. Rappelons que dans le cas de structures de croyance floues, l'étape de normalisation est nécessaire, car l'ensemble des éléments focaux non normalisés ne se limite pas à l'ensemble vide.

4.4.2 Cas particuliers - lien avec la régression classique.

On suppose dans cette section que les structures de croyance sont classiques (cas 2, 4 et 5). Plusieurs définitions de l'espérance sont possibles et donnent lieu à des valeurs ponctuelles différentes. On peut facilement calculer les espérances inférieure et supérieure de Y. D'après les équations (3.51), on obtient :

$$\hat{y}^*(\boldsymbol{x}) = \mathbb{E}^*(Y) = \sum_{A \in F(\widehat{m}(.|x))} \widehat{m}(A|\boldsymbol{x}) \sup_{y \in A} y, \tag{4.16}$$

$$\hat{y}_*(\boldsymbol{x}) = \mathbb{E}_*(Y) = \sum_{A \in F(\widehat{m}(.|x))} \widehat{m}(A|\boldsymbol{x}) \inf_{y \in A} y. \tag{4.17}$$

Si on se place dans le cadre des probabilités imprécises, on peut alors définir un intervalle de valeurs

$$[\hat{y}_*(\boldsymbol{x}), \hat{y}^*(\boldsymbol{x})]$$

dont la largeur représente l'incertitude sur la sortie[3]. La probabilité pignistique est une fonction constante par morceau, comme nous l'avons vu dans le chapitre 3.

Dans le cas particulier où les sorties sont des nombres réels (modèle SCF3), les éléments focaux sont les y_i et l'ensemble de référence et l'expression analytique de $\widehat{m}(.|\boldsymbol{x})$ est donnée par l'équation (4.7).

La probabilité pignistique devient alors :

$$\widehat{p}_{bet}(y|\boldsymbol{x}) = \sum_{i=1}^{N} \widehat{m}^*(\{y_i\}|\boldsymbol{x})\delta_{\{y_i\}}(y) + \frac{\widehat{m}(\mathcal{Y}|\boldsymbol{x})}{N}. \tag{4.18}$$

Il s'agit d'une loi de probabilité *mixte*: la probabilité $\widehat{p}_{bet}(.|\boldsymbol{x})$ est une somme pondérée de lois discrètes (lois de Dirac) et d'une loi uniforme continue. Elle est donc *discontinue* aux points y_i. La première partie de l'expression de p_{bet}, la somme des lois de Dirac, peut se voir comme une probabilité empirique sur $\mathcal{Y}$, pondérée par l'influence de $\boldsymbol{x}_i$ sur $\boldsymbol{x}$.

En prenant l'espérance sur l'ensemble des valeurs de y, on peut définir une sortie ponctuelle à partir de cette distribution :

$$\hat{y}(\boldsymbol{x}) = \sum_{i=1}^{N} \widehat{m}(\{y_i\}|\boldsymbol{x})y_i + \widehat{m}(\mathcal{Y}|\boldsymbol{x})\bar{y}, \tag{4.19}$$

où $\bar{y}$ est la valeur moyenne de y sur $\mathcal{Y}$.

La fonction de régression obtenue, qui est continue, est donc linéaire par rapport aux y_i et peut s'écrire :

$$\widehat{y}(\boldsymbol{x}) = \sum_{i=1}^{N} w_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|)y_i + b(\|\boldsymbol{x} - \boldsymbol{x}_i\|)\bar{y}, \tag{4.20}$$

3. Rappelons que dans le modèle des croyances transférables, cette interprétation est abusive, puisque le modèle suppose que les croyances m_i sont définies indépendamment de toute mesure de probabilité.

où w_i et b sont des fonctions de $\mathbb{R}^+$ dans $\mathbb{R}$. L'expression du noyau équivalent de $\widehat{y}$ en $\boldsymbol{x}_i$ (cf. équation A.7), est la fonction H, telle que : $\widehat{y}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{x}, \boldsymbol{x}_i) y_i$. On a donc

$$\forall \boldsymbol{x}, \boldsymbol{x}_i \in \mathcal{X} \quad H(\boldsymbol{x}, \boldsymbol{x}_i) = N w_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) + b(\|\boldsymbol{x} - \boldsymbol{x}_i\|).$$

Les espérances inférieure et supérieure deviennent, d'après les équations (4.17) et (4.16) :

$$\hat{y}^*(\boldsymbol{x}) = \sum_{i=1}^{N} w_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) y_i + b(\|\boldsymbol{x} - \boldsymbol{x}_i\|) \sup_{\mathcal{Y}} y, \tag{4.21}$$

$$\hat{y}_*(\boldsymbol{x}) = \sum_{i=1}^{N} w_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) y_i + b(\|\boldsymbol{x} - \boldsymbol{x}_i\|) \inf_{\mathcal{Y}} y. \tag{4.22}$$

Si on considère la définition de l'espérance plus générale définie par Strat (cf. section 3.54), on obtient :

$$\begin{array}{rcl} \hat{y}_\rho(\boldsymbol{x}) & = & \hat{y}_*(\boldsymbol{x}) + \rho\,(\hat{y}^*(\boldsymbol{x}) - \hat{y}_*(\boldsymbol{x})\,) \\ & = & \sum_{i=1}^{N} w_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) y_i + b(\|\boldsymbol{x} - \boldsymbol{x}_i\|)[\rho \inf_{\mathcal{Y}}(y) + (1-\rho)\sup_{\mathcal{Y}}(y)], \end{array} \tag{4.23}$$

pour $\rho \in [0, 1]$.

Si on compare ces quatre quantités, elles sont toutes composées de deux parties distinctes. La première, linéaire, définit l'influence des vecteurs d'apprentissage entre eux, traduite par les poids $w_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|)$, en fonction de la proximité de $\boldsymbol{x}$ et $\boldsymbol{x}_i$. C'est la composante *locale* du modèle. Cette quantité correspond au résultat obtenu par une méthode de régression classique, du type noyau. La deuxième représente la connaissance *globale* que l'ensemble d'apprentissage fournit selon la proximité de $\boldsymbol{x}$ à cet ensemble. Elle traduit l'ignorance de la sortie correspondant à $\boldsymbol{x}$. La quantité $m(\mathcal{Y}|\boldsymbol{x})$ tend vers 1 si $\boldsymbol{x}$ s'éloigne indéfiniment des $\boldsymbol{x}_i$. Son expression varie en fonction de la définition de l'espérance, contrairement à la partie classique, et n'est pas forcément linéaire. Ainsi, les trois dernières expressions ne sont plus linéaires en y_i.

Nous allons spécifier les résultats dans les deux cas extrêmes où $\widehat{m}(\mathcal{Y}|\boldsymbol{x})$ est égale à 1 (ignorance totale) ou tend vers 0.

Dans le cas limite où $\widehat{m}(\mathcal{Y}|\boldsymbol{x})$ tend vers 0, c'est-à-dire où $\boldsymbol{x}$ est l'un des vecteurs d'apprentissage $\boldsymbol{x}_i$, les quatre expressions de l'espérance sont confondues [4].

Notre méthode devient un régresseur à noyaux classique (cf. équation A.6) :

$$\hat{y}(\boldsymbol{x}) = \sum_{i=1}^{N} w_i(\|\boldsymbol{x} - \boldsymbol{x}_i) y_i \quad \text{avec} \sum_{i=1}^{N} w_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) = 1. \tag{4.24}$$

La loi de probabilité $\widehat{p}_{bet}(.|\boldsymbol{x})$ devient discrète. C'est une somme pondérée de lois de Dirac. On peut ainsi la voir comme la probabilité empirique de Parzen.

4. Remarquons que la masse affectée à $\mathcal{Y}$ ne peut pas être rigoureusement nulle, car $\phi_i(0) < 1$. Cette condition est nécessaire pour que les $\widehat{m}_i(.|\boldsymbol{x})$ soient combinables, car les éléments focaux $(y_i)_{i=1}^N$ sont disjoints! Il faut donc poser $\widehat{m}(\mathcal{Y}|\boldsymbol{x}) = \varepsilon$ et prendre la limite quand ε tend vers 0.

Dans l'autre cas extrême, où $\widehat{m}(\mathcal{Y}|\boldsymbol{x}) = 1$, c'est au contraire la partie classique qui disparaît de l'expression (4.18). La probabilité pignistique devient la loi de probabilité uniforme sur $\mathcal{Y}$:

$$\widehat{p}_{bet}(.|\boldsymbol{x}) = \mathcal{U}_{\mathcal{Y}}.$$

Une valeur ponctuelle peut toujours être définie mais ne présente aucun intérêt.

On peut cependant noter ici que le choix de $\mathcal{Y}$, si celui-ci n'est pas imposé, est à faire avec prudence lorsqu'on est intéressé par une sortie ponctuelle dans le cas général. En effet, l'influence de l'élément focal trivial $\mathcal{Y}$ ne dépend que de sa valeur moyenne. Cette valeur biaise les résultats si la masse affectée à $\mathcal{Y}$ est trop importante.

4.5 Identification du modèle de régression

4.5.1 Problèmes d'apprentissage

Afin de limiter les problèmes de biais dont il vient d'être question, on peut recourir à une méthode d'apprentissage permettant, d'une part, de contrôler l'influence globale de l'élément focal $\mathcal{Y}$, et d'autre part, d'ajuster l'influence relative des éléments de l'ensemble d'apprentissage.

Nous avons vu dans l'exemple de la discrimination que la sortie pouvait se mettre sous la forme $\widehat{m}(.|\boldsymbol{\theta}, \boldsymbol{x})$, où $\boldsymbol{\theta}$ est un vecteur à choisir selon un certain critère.

Par analogie avec le cas de la discrimination, on peut définir le critère suivant :

$$J_{bet1}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \int_{\mathcal{Y}} [\widehat{p}_{bet}(y|\boldsymbol{x}_i, \boldsymbol{\theta}) - {p_{bet}}^i(y)]^2 dy, \tag{4.25}$$

qui est l'équivalent de J_{bet} (équation 4.11) pour un espace continu. On peut également utiliser directement le critère J'_{bet} (équation 4.14). Cependant, ces critères ne semblent pas satisfaisants, car ils ne généralisent pas le coût quadratique moyen, dans le cas de la régression classique où les m_i sont des réels y_i:

$$J_N = \frac{1}{N} \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2.$$

En effet, dans ce cas, ${p_{bet}}^i$ est la distribution de Dirac δ_{y_i}, c'est-à-dire, ${p_{bet}}^i(y_i) = 1$ et

$$J_{bet1}(\boldsymbol{\theta}) = J'_{bet}(\boldsymbol{\theta}) = \sum_{i=1}^{N} [\widehat{p}_{bet}(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) - 1]^2 \neq J_N.$$

Nous allons voir comment définir un critère mieux approprié à l'identification de notre modèle.

4.5.2 Critères d'erreur entre structures de croyance

Objectifs

Nous proposons dans cette section de généraliser la notion de critère d'erreur aux structures de croyance. Un tel critère C peut être défini comme une fonction à valeurs dans $\mathbb{R}^+$:

$$C : \mathcal{MF}(\mathcal{Y}) \times \mathcal{MF}(\mathcal{Y}) \longrightarrow \mathbb{R}^+.$$

La recherche d'une comparaison entre deux structures m et m' répond à un objectif de l'identification de notre modèle où m et m' représentent une structure de croyance et son estimation obtenues indépendamment l'une de l'autre.

Nous commençons par discuter de trois propriétés *a priori* souhaitables pour ce type de fonction :

1. C généralise un critère d'erreur classique entre nombres réels,
2. C généralise un critère d'erreur classique entre ensembles flous (cf. chapitre 2),
3. C est une distance sur $\mathcal{MF}(\mathcal{Y})$.

Les axiomes 1 et 2 sont primordiaux. Si m et m' sont des réels y et y', $C(m, m')$ doit être une distance classique. Si les structures sont des ensembles flous, $C(m, m')$ doit généraliser une distance « raisonnable » entre ensembles flous, comme la distance de Hausdorff (cf. section 1.3.3). Mais, dans le cas général, cette fonction doit-elle être une distance ou une dissimilarité (axiome 3)? A notre avis, la notion de distance n'est pas adaptée aux structures de croyance. En effet, il n'est pas nécessairement judicieux de traiter une structure de croyance quelconque de la même façon qu'un nombre réel. La connaissance de m_i ne caractérise pas une valeur précise de y, mais définit plutôt un ensemble de *contraintes* sur les valeurs possibles. Si on prend l'exemple extrême de l'ignorance totale, où m est la structure triviale $m_{\mathcal{Y}}$, il n'y a aucune contrainte. Dans ces conditions, aucune structure m' de $\mathcal{MF}(\mathcal{Y})$ ne contredit l'information apportée par m. Il ne peut pas y avoir d'apprentissage, puisque, précisément, on ne possède aucune information. Nous avons déjà fait cette remarque dans le cas de la discrimination (cf. section 4.3.3).

Critères envisageables

Nous distinguons plusieurs types d'approches définissant $C(m, m')$:

1. les critères de type probabiliste : entropie ou information mutuelle ;
2. les critères basés sur la généralisation d'un critère d'erreur entre réels ou entre ensembles flous ;
3. les critères basés sur le principe d'extension.

Les critères de type probabiliste sont basés sur des distances entre diverses quantités issues de m et m' comme la probabilité pignistique, la crédibilité ou la plausibilité. Nous ne les détaillons pas dans cette section car ils ne généralisent pas les distances classiques (axiomes 1 et 2).

Généralisation de critères d'erreur classiques

De nombreuses mesures sont *a priori* envisageables. Nous proposons la mesure suivante, qui semble être la plus naturelle :

$$C(m,m') = \sum_{F\in F(m)} \sum_{F'\in F(m')} m(F)m(F')\tilde{d}(F,F'), \tag{4.26}$$

où $\tilde{d}$ est une distance classique entre ensembles flous (cf. chapitre 2), à valeurs dans $\mathbb{R}^+$:

$$\tilde{d} : \mathcal{F}(\mathcal{Y}) \times \mathcal{F}(\mathcal{Y}) \to \mathbb{R}^+$$

Par sa construction, cette mesure généralise le critère d'erreur $\tilde{d}$ entre ensembles flous (axiome 2). Il faut donc choisir maintenant une distance appropriée entre ensembles flous.

Parmi les différentes mesures définies dans le chapitre 2, les quantités $\tilde{d}_2$ (équation (1.32)) et $\tilde{d}'_2$ (équation (1.33)) semblent les plus judicieuses. Elles généralisent la distance de Hausdorff. Afin de généraliser le critère quadratique classique, on peut modifier légèrement la distance de Hausdorff de façon suivante. Soient deux intervalles $F = [f_1, f_2]$ et $F' = [f'_1, f'_2]$. La distance entre F et F' devient alors $h_2(F,F') = \max\{(f_1 - f'_1)^2, (f_2 - f'_2)^2\}$. Nous utiliserons dans la suite la mesure suivante :

$$\tilde{d}(F,F') = \int_0^1 h_2(F_\alpha, F'_\alpha)d\alpha. \tag{4.27}$$

Nous rappelons que ces distances nécessitent la normalisation des ensembles flous.

Critères basés sur le principe d'extension

On peut définir un deuxième type de critère $C(m,m')$, à valeur dans $\mathcal{F}(\mathbb{R}^+)$, en utilisant le principe d'extension, par analogie avec les « distances » entre ensembles flous.

On définit une distance

$$\Delta : \mathcal{MF}(\mathcal{Y}) \times \mathcal{MF}(\mathcal{Y}) \to \mathcal{MF}(\mathbb{R}^+)$$
$$(m,m') \longmapsto \Delta(m,m') = \tilde{m}_\Delta$$

d'éléments focaux :

$$F(\tilde{m}_\Delta) = \{d(F,F'), \forall F, F' \in \mathcal{F}\times\mathcal{F}'\}$$

avec $\tilde{m}_\Delta(\delta) = \sum_{d(F,F')=\delta} m(F)m'(F')$.

On définit la probabilité pignistique p_{bet_Δ} associée à m_Δ. Le critère d'erreur peut alors être défini par : $C(m,m') = \int_{\mathbb{R}^+} p_{bet_\Delta}(\delta)\delta d\delta$.

4.5.3 Identification du modèle

Choix *a priori*

La phase d'identification nécessite d'abord de définir certains choix dans notre modèle. Nous avons vu en particulier que l'opérateur de combinaison ainsi que le type des fonctions d'activation ϕ_i devaient être spécifiés. Nous avons choisi un opérateur conjonctif, et non disjonctif,

car notre méthode nécessite la *neutralité* de la structure de croyance triviale (focalisée sur $\mathcal{Y}$). En effet, si la sortie d'un élément $\boldsymbol{x}_i$ est totalement inconnue, c'est-à-dire, $m_i = \mathcal{Y}$, il ne doit avoir aucune influence sur les autres vecteurs. Quant au choix de la norme triangulaire dans le cas de structures de croyance floues, des simulations ont montré que différentes t-normes donnent des résultats sensiblement équivalents. Nous prendrons en général l'opérateur classique min. De même, on peut montrer de manière empirique que le type de la fonction ϕ_i n'a pas une influence primordiale.

Sélection du critère d'erreur

L'identification du système nécessite de définir un critère d'erreur entre $\widehat{m}(.|\boldsymbol{x}_i, \boldsymbol{\theta})$ et m_i, c'est-à-dire de définir une fonction de coût C entre deux structures quelconques m et m'. Parmi les critères définis dans la section précédente, le critère le plus adapté à cet objectif semble être celui qui généralise le critère d'erreur classique (cf. équation (4.26)) en utilisant la distance entre ensembles flous $\tilde{d}$ définie par l'équation (4.27). Cette distance étant définie pour des ensembles flous normalisés, le critère C doit se calculer sur les structures de croyance *normalisées*. Si nous récapitulons les résultats de la section précédente, le critère de sélection de la structure $\widehat{m}_i(.|\boldsymbol{x})$ est donc :

$$\begin{aligned} C(m_i, \widehat{m}_i(.|\boldsymbol{x}_i, \boldsymbol{\theta})) &= \sum_{F \in F(m_i^*)} \sum_{F' \in F(m_i^*(.|x))} m_i^*(F) \widehat{m}_i^*(F'|\boldsymbol{x}) \tilde{d}(F, F'), \\ \text{avec } \tilde{d}(F, F') &= \int_0^1 h_2(F_\alpha, F'_\alpha) d\alpha, \\ \text{et } h_2(F_\alpha, F'_\alpha) &= \max\{(F_\alpha^- - F_\alpha^-)^2, (F_\alpha^+ - F_\alpha^+)^2\}, \end{aligned} \tag{4.28}$$

où $F_\alpha^+ = \sup_{y \in \mathcal{Y}} F_\alpha(y)$ et $F_\alpha^- = \inf_{y \in \mathcal{Y}} F_\alpha(y)$. Le fait de calculer ce critère sur les structures normalisées n'est pas *a priori* très contraignant. Néanmoins, cette procédure entraîne nécessairement une perte d'information que nous allons analyser brièvement. Supposons que $\widehat{m}_i$ et m_i ne possèdent chacun qu'un seul élément focal, respectivement F et $\tilde{y}_i$. Le critère se ramène donc à la distance $\tilde{d}(F, \tilde{y}_i)$ entre les ensembles flous F et $\tilde{y}_i$. Supposons le cas de figure suivant, où $F = \tilde{y}_i / h(F)$. Alors, $\tilde{y}_i$ représente la normalisation F^* de F et par conséquent, $C(F, \tilde{y}_i) = 0$, quelle que soit la hauteur de F. Or $h(F)$ représente la confiance globale que l'on a dans la sortie F. En conclusion, l'information que l'on perd en normalisant est donc simplement une indication de la fiabilité de la structure $\widehat{m}_i$, ce qui n'est pas contraignant, car nous considérons que la normalisation par la méthode SNP définit une approximation cohérente de $\widehat{m}_i$.

Détermination de la fonction de coût globale

Un critère d'erreur entre deux structures ayant été défini, nous pouvons directement appliquer les méthodes classiques de validation et de sélection de modèle et, en particulier, les techniques de rééchantillonnage (cf. chapitre 1), comme la validation croisée, le jackknife et le bootstrap et leurs variantes [48]. Dans la suite, nous utilisons la méthode du *leave-one-out*.

Dans le cas du *leave-one-out*, la sortie de chaque vecteur $\boldsymbol{x}_i$ de l'ensemble d'apprentissage est estimée à l'aide des $N-1$ autres vecteurs $(\boldsymbol{x}_j, m_j), j \neq i$. Soit Θ l'ensemble des paramètres possibles $\boldsymbol{\theta}$ à sélectionner.

Pour chaque $\boldsymbol{\theta} \in \Theta$, on obtient une estimation de la sortie correspondant à $\boldsymbol{x}_i$, que l'on notera alors :

$$\widehat{m}_{-i}(.|\boldsymbol{x}_i, \boldsymbol{\theta}).$$

Le critère de validation est défini de façon classique par la formule suivante :

$$CV(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} C(m_i, \widehat{m}_i(.|\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{4.29}$$

Le vecteur de paramètres choisi $\widehat{\boldsymbol{\theta}}$ est celui qui minimise ce critère :

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} CV(\boldsymbol{\theta}). \tag{4.30}$$

Si on veut tenir compte de la qualité de l'information apportée par $(\boldsymbol{x}_i, m_i)$, on peut pondérer le critère d'erreur par un coefficient $\beta_i \in [0, 1]$, défini comme un facteur de confiance. En particulier, β_i devrait être nul en cas d'ignorance totale sur la valeur de y_i, c'est-à-dire, si $m_i = \mathcal{Y}$. Par contre, si la valeur de y_i est connue avec précision, β_i devrait être proche de 1. L'emploi d'une mesure de nonspécificité $N(m)$ semble être une bonne solution.

Le coefficient β_i peut alors être défini ainsi :

$$\beta_i = 1 - \frac{N(m_i^*)}{\log_2(Y)}$$

où m_i^* est la normalisation de la structure m_i.

Alors,

$$CV(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \beta_i C(m_i, \widehat{m}_i(\boldsymbol{\theta})).$$

Estimation des paramètres

Dans le modèle SCF1, les paramètres p_{ij} sont supposés connus, déterminés par un expert. Dans les modèles SCF1, SCF2 et SCF3, le vecteur de paramètres $\boldsymbol{\theta}$ est donc simplement constitué des paramètres des fonctions d'activation.

Par exemple, si $\phi_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|) = \exp[-(\boldsymbol{x} - \boldsymbol{x}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{x}_i)]$, où Σ_i est une matrice diagonale d'éléments diagonaux $(\sigma_{ij})_{j=1}^{r}$,

$$\boldsymbol{\theta} = [\sigma_{ij}], i = 1 \ldots N, j = 1 \ldots r$$

Comme il n'y a aucune raison de privilégier certaines entrées par rapport à d'autres, les $\boldsymbol{x}_i$ étant tous réels et bien définis, on considère que, pour tout i, les σ_{ij} sont égaux. Il reste donc r paramètres à identifier. La procédure de validation croisée étant relativement longue, si les données le permettent, on pourra prendre également les mêmes paramètres σ_{ij} dans toutes les dimensions, afin de limiter les calculs. Dans ce cas, il ne reste finalement qu'*un seul* paramètre θ à estimer.

4.6 Liens avec les systèmes flous

Dans la section 1, nous avons noté qu'il y avait quelques similitudes entre notre méthode d'inférence et celle définie par les systèmes flous. Dans cette section, nous proposons d'analyser les différences et les points communs entre certains types de systèmes flous et notre modèle. En particulier, nous étudions un modèle proche du nôtre qui a été proposé récemment par Yager [172, 173].

4.6.1 Le modèle de Yager

Définition du modèle

Soit le système flou classique de type Mamdani défini par l'ensemble des règles suivantes $(r_i)_{i=1}^N$ (cf. chapitre 1) :

$$r_{ij} : \quad \text{SI} \quad \boldsymbol{x} \text{ est } B_i \text{ ALORS} \quad y \text{ est } \tilde{y}_{ij} \quad (p_{ij}), \tag{4.31}$$

où les B_i et les $\tilde{y}_{ij}$ sont des ensembles flous respectifs de $\mathcal{X}$ et $\mathcal{Y}$, et p_{ij} est le facteur de confiance de la règle r_{ij}. La partie conséquente de chaque règle est une proposition de la forme « y est $\tilde{y}_{ij}$ » (cf. équations (1.13) et (4.31)). Pour chaque partie antécédente « $\boldsymbol{x}$ est B_i », la proposition « y est $\tilde{y}_{ij}$ » peut se voir comme l'une des solutions possibles de la sortie, à laquelle on accorde un certain crédit, représenté par le facteur p_{ij}. On peut distinguer ainsi deux types d'incertitudes relatives à la sortie y : l'imprécision, due à l'ensemble flou $\tilde{y}_{ij}$ et le conflit, représenté par le choix entre les différents ensembles $\tilde{y}_{ij}$. Si on normalise les p_{ij}, c'est-à-dire, si on impose la condition : $\sum_{j=1}^J p_{ij} = 1$ dans la règle r_{ij}, on peut rassembler les r règles $r_{ij}, j = 1 \ldots J$ à l'aide de la théorie des croyances. Si on définit la structure m_i, d'éléments focaux $\tilde{y}_{ij}$ tels que $m_i(\tilde{y}_{ij}) = p_{ij}$, l'ensemble des propositions « y est $\tilde{y}_{ij}$ », $j = 1, \ldots, J$ peut se synthétiser par l'écriture suivante :

$$y \text{ est } m_i.$$

Nous avons donc une équivalence entre ces deux groupes de propositions [172] :

$$y \text{ est } m_i \Leftrightarrow \begin{cases} y \text{ est } \tilde{y}_{i1} & (p_{i1}) \\ \ldots & \ldots \\ y \text{ est } \tilde{y}_{iJ} & (p_{iJ}) \end{cases}$$

On peut alors définir un nouveau système à base de règles r'_i, équivalent au système (4.31) :

$$r'_i : \quad \text{SI } \boldsymbol{x} \text{ est } B_i \text{ ALORS} \quad y \text{ est } m_i. \tag{4.32}$$

Méthode d'inférence

Nous allons maintenant montrer comment la méthode d'inférence de Mamdani définie dans le chapitre 1 se généralise à ce type de système. Nous rappelons que la méthode de Mamdani peut se décomposer en trois étapes :

- calcul du degré de déclenchement des règles τ_i,

– calcul de la sortie F_{ij} correspondant à chacune des règles r_{ij},

– agrégation des règles.

Nous allons reprendre ces trois étapes successivement.

La partie antécédente des règles est la même pour les deux systèmes (4.31) et (4.32). Le degré de déclenchement τ_i de r'_i est donc le même que celui de r_{ij} et est défini par :

$$\tau_i = B_i(\boldsymbol{x}).$$

Dans la deuxième étape, on détermine la sortie de chaque règle [172]. Celle-ci se définit comme une structure de croyance floue $\widetilde{m}_i$ définie sur $\mathcal{Y}$, d'éléments focaux flous $F(\widetilde{m}_i) = \{F_{ij}, j = 1, \ldots J, i = 1, \ldots N\}$ définis par :

$$F_{ij} = \tau_i \wedge \tilde{y}_{ij} \text{ avec } \widetilde{m}_i(F_{ij}) = m(\tilde{y}_{ij}) = p_{ij}. \tag{4.33}$$

La sortie finale du système est obtenue par agrégation des structures $\widetilde{m}_i$ selon un opérateur à définir. Par analogie avec la règle d'inférence de Mamdani, ces structures sont agrégées à l'aide de la règle de combinaison disjonctive $\cup$ généralisée aux ensembles flous [45], qui est définie par l'équation (3.15). On obtient ainsi la structure finale induite par l'ensemble du système, notée $\widetilde{m} \in \mathcal{MF}(\mathcal{Y})$, qui s'écrit :

$$\widetilde{m} = \bigcup \widetilde{m}_i. \tag{4.34}$$

et dont les éléments focaux $E \in F(\widetilde{m})$ sont de la forme

$$E = \bigcup_{i=1}^{N} F_{ij(i)}, \text{ où } F_{ij(i)} \in F(\widetilde{m}_i), \tag{4.35}$$

avec :

$$\widetilde{m}(E) = \prod_{i=1}^{N} \widetilde{m}_i(F_{ij(i)}).$$

Pour un vecteur $\boldsymbol{x} \in \mathcal{X}$ donné, la sortie du système est donc définie par une structure de croyance floue sur $\mathcal{Y}$, que l'on notera $\widetilde{m}(.|\boldsymbol{x})$.

De façon habituelle, à partir de $\widetilde{m}(.|\boldsymbol{x})$, on peut définir différentes caractéristiques, comme la crédibilité $\widetilde{\text{bel}}(.|\boldsymbol{x})$, la distribution pignistique $\widetilde{p}_{bet}(.|\boldsymbol{x})$. On peut en déduire une expression de la sortie ponctuelle comme la moyenne pondérée des centres de gravité y^*_E des éléments focaux de $\widetilde{m}$ (cf. équation (4.15)) :

$$\widetilde{y}(\boldsymbol{x}) = \sum_{E \in F(\widetilde{m}^*(.|x))} \widetilde{m}^*(E|\boldsymbol{x}) y^*_E,$$

où $\widetilde{m}^*(.|\boldsymbol{x})$ est la version normalisée de $\widetilde{m}(.|\boldsymbol{x})$. Rappelons que ce calcul est obtenu en calculant l'espérance de y selon la probabilité pignistique de $\widetilde{m}^*(.|\boldsymbol{x})$.

Les structures m_i sont normalisées, mais ce n'est pas le cas des $\widetilde{m}_i(.|\boldsymbol{x})$ ni par conséquent de leur combinaison $\widetilde{m}(.|\boldsymbol{x})$, quoique l'on utilise un opérateur disjonctif.

4.6.2 Analogie avec notre modèle

Notre méthode peut se voir comme une modification du système flou précédent [115]. Soit B_i, un ensemble flou représentatif du vecteur $\boldsymbol{x}_i$ (cf. chapitres 1 et 2). Par exemple, B_i est un nombre flou gaussien multidimensionnel centré sur $\boldsymbol{x}_i$. On suppose que B_i peut se mettre sous la forme suivante :

$$B_i(\boldsymbol{x}) = \phi_i(\|\boldsymbol{x} - \boldsymbol{x}_i\|),$$

où ϕ_i est une fonction d'activation, selon la définition de la section 1. C'est le cas de la plupart des fonctions d'appartenance, puisque $B_i(\boldsymbol{x}) \in [0,1]$. Le facteur d'affaiblissement est donc de $1 - B_i(\boldsymbol{x}) = 1 - \tau_i(\boldsymbol{x})$.

Dans notre méthode, nous transformons la structure m_i par affaiblissement de facteur $1 - \tau_i(\boldsymbol{x})$. Les structures $\widehat{m}_i(.|\boldsymbol{x})$ définies en section 1 peuvent donc s'exprimer en fonction de $\tau_i(\boldsymbol{x})$:

$$\widehat{m}_i(.|\boldsymbol{x}) = m_i^{1-\tau_i(\boldsymbol{x})}.$$

Notre méthode d'inférence permet de faire intervenir un facteur d'ignorance dans la partie conséquente : l'ensemble de référence $\mathcal{Y}$ devient un nouvel élément focal.

Enfin, l'agrégation des structures est réalisée à l'aide d'un opérateur conjonctif, et non disjonctif, comme dans l'approche de Yager, afin de contrôler la masse allouée à l'espace $\mathcal{Y}$. La structure finale est donc :

$$\widehat{m}(.|\boldsymbol{x}) = \bigcap_{i=1}^{N} \widehat{m}_i(.|\boldsymbol{x}).$$

Dans l'approche de Yager, si la norme triangulaire utilisée est le produit, les éléments focaux initiaux $\tilde{y}_{ij}$ sont transformés à l'aide d'un facteur multiplicatif et deviennent sous-normalisés, de hauteur $\tau_i(\boldsymbol{x})$. Mais leur masse est inchangée. Si nous re-normalisons ces éléments focaux à l'aide de la procédure SNP, nous retrouvons évidemment les $\tilde{y}_{ij}$, mais leur poids devient $\tau_i(\boldsymbol{x})p_{ij}$. La masse restante est affectée au domaine entier. Cette procédure correspond à notre approche.

Les avantages de notre approche par rapport à celle de Yager sont les suivants : d'abord, notre modèle est capable d'intégrer une mesure globale de confiance sur la sortie. De plus, le nombre maximal d'éléments focaux, même s'il est encore important, est malgré tout beaucoup plus faible. En effet, dans la méthode de Yager, le nombre d'éléments focaux de $\widehat{m}(.|\boldsymbol{x})$ peut atteindre J^N, contre 2^N dans notre méthode. Enfin, la structure de $F(\widehat{m}(.|\boldsymbol{x}))$ est stable par intersection et ne dépend pas du vecteur d'entrée : ce sont les masses qui sont distribuées différemment selon $\boldsymbol{x}$.

4.7 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle méthode de régression basée sur la théorie des croyances floues. Cette approche permet de prendre en compte différents types de données, comme les intervalles, les nombres flous ou plus généralement des structures de croyance floues. Ainsi, elle peut être considérée comme une généralisation de l'estimation fonctionnelle à des données incertaines. Deux types d'information sont proposés : une représentation de l'incertitude (imprécision, indétermination, ignorance) et une estimation ponctuelle.

Cependant, nous avons noté que, lorsque la taille N du modèle augmente, le nombre d'éléments focaux augmente de façon exponentielle et la combinaison des structures devient impossible. Dans le dernier chapitre, nous proposerons plusieurs méthodes de simplification de structures de croyance, afin de pallier l'augmentation indésirable du nombre d'éléments focaux pendant la phase de combinaison. Nous illustrerons notre méthode à travers différentes simulations.

Chapitre 5

Mise en oeuvre

5.1 Introduction

Dans le chapitre précédent, nous avons présenté une méthode de régression basée sur la théorie des croyances. Dans cette méthode, l'information fournie par chaque élément de l'ensemble d'apprentissage, de nature incertaine et imprécise, est représentée, non nécessairement par un nombre réel, mais par une structure de croyance floue dans le cas le plus général. Afin d'obtenir une information globale, ces informations individuelles doivent être fusionnées. Ces opérations, si la taille de l'ensemble d'apprentissage est grande, nécessitent un temps de calcul qui pénalise la méthode et la rend inopérante. Ce chapitre est en partie consacré à la résolution de ce problème.

Plusieurs techniques de simplification de notre méthode sont proposées. Deux groupes de méthodes peuvent être retenues :

- les méthodes réduisant le nombre de *structures* à combiner ;
- les méthodes réduisant le nombre d'*éléments focaux* des structures à combiner.

Le premier groupe de méthodes consiste essentiellement à résumer l'information délivrée par l'ensemble d'apprentissage, à l'aide de méthodes de classification notamment, où à n'utiliser qu'un voisinage de points significatifs, c'est-à-dire, par exemple, les k plus proches voisins dans l'ensemble d'apprentissage. Les méthodes de simplification des structures de croyance sont très spécifiques à notre méthode. Il s'agit d'éliminer les éléments focaux de peu de poids ou de regrouper les éléments focaux similaires. Les méthodes de classification (hiérarchique et partitionnement) peuvent ainsi être généralisées aux structures de croyances floues : l'approximation de structures en est une application.

Nous proposons d'appliquer la méthode à différents problèmes de régression. Nous étudions deux simulations mettant en valeur différents types d'incertitudes. Nous présentons également deux applications réelles faisant intervenir des données définies de façon imprécise.

5.2 Utilisation de prototypes

5.2.1 Motivation

Nous avons vu dans le chapitre précédent que notre méthode était difficilement applicable quand la taille de l'ensemble était trop importante. En effet, l'essentiel des calculs est effectué entre les éléments focaux, pendant la combinaison des structures. Il faut donc éviter que le nombre d'éléments focaux augmente trop rapidement. Si $J(i)$ représente le cardinal de $F(m_i)$, on peut avoir dans le cas le plus défavorable jusqu'à

$$\prod_{i=1}^{N}(J(i)+1)$$

éléments focaux dans le modèle général SCF1 (équation 4.1). Même dans le cas plus simple du modèle SCF2 (équation (4.6)) où il n'y a qu'un seul élément focal flou, le nombre d'éléments focaux est encore de 2^N. Si l'on considère uniquement les k plus proches voisins de $\boldsymbol{x}$, ce nombre peut se réduire sensiblement (2^k). Cette méthode permet d'extraire l'information *locale* au voisinage de $\boldsymbol{x}$. Une autre façon de diminuer le nombre d'éléments focaux consiste à réduire *globalement* la taille de l'échantillon par classification de l'ensemble d'apprentissage et la construction de prototypes. Il est évidemment avantageux de combiner ces deux techniques locale et globale. Nous supposerons dans cette section que les y_i sont réels. Si l'ensemble d'apprentissage est de « grande » taille, l'information qu'il contient peut être résumée par un ensemble de prototypes.

5.2.2 Partitionnement de l'ensemble d'apprentissage

Comme nous l'avons décrit en section 1.3.2, le partitionnement peut se faire de plusieurs façons différentes. Nous étudions successivement le cas d'un partitionnement indépendant de $\mathcal{X}$ et de $\mathcal{Y}$, et d'un partitionnement sur $\mathcal{X} \times \mathcal{Y}$. L'objectif n'est pas le même dans ces deux approches. Le premier cas permettra de mieux déterminer les problèmes où la variable y est multivaluée, pour un vecteur $\boldsymbol{x}$ donné. On privilégiera la deuxième approche dans les cas où y est seulement imprécise.

Partitionnement séparé des espaces d'entrée et de sortie

Soit

$$C = \{C_k, k \in \{1, \ldots, K\}\} \text{ et } E = \{E_j \in \mathcal{Y}, j \in \{1, \ldots, J\}\},$$

deux ensembles de K et J classes partitionnant respectivement les espaces $\mathcal{X}$ et $\mathcal{Y}$, obtenus indépendamment par classification de l'ensemble d'apprentissage à l'aide d'une méthode quelconque (cf. annexe D). Les classes $(C_k)_{k=1}^K$ sont caractérisées par les centres ou prototypes $\boldsymbol{c}_k$ et les matrices de dispersion $\boldsymbol{\Sigma}_k^x$. Nous noterons $\sigma_{kj}^x, j = 1 \ldots r$ les éléments diagonaux de $\boldsymbol{\Sigma}_k^x$, et $\boldsymbol{\sigma}_k^x$ le vecteur $(\sigma_{kj}^x)_{j=1}^r$. De même, les classes $(E_j)_{j=1}^J$ sont caractérisées par les centres e_j et les écarts-type σ_j^y.

Définition des fonctions d'activation ϕ_k

Soit $\boldsymbol{x}$ un vecteur de $\mathcal{X}$ dont on cherche à estimer la sortie correspondante. Au lieu d'utiliser directement les vecteurs $\boldsymbol{x}_i$ comme dans le chapitre précédent, on s'appuie maintenant uniquement sur l'information plus globale fournie par les classes C_k. Plus $\boldsymbol{x}$ est proche de C_k, ou du centre $\boldsymbol{c}_k$, plus cette classe aura une influence sur la sortie y. A chaque centre $\boldsymbol{c}_k$, on associe une fonction d'activation ϕ_k, définie par exemple par :

$$\phi_k(\|\boldsymbol{x} - \boldsymbol{c}_k\|) = \text{Gauss}(\boldsymbol{x}; \boldsymbol{c}_k, q_x \boldsymbol{\sigma}_k^x), \tag{5.1}$$

afin de tenir compte de la dispersion de la classe C_k. Le paramètre q_x est un paramètre de lissage réel positif qui sera utilisé ultérieurement pour l'identification des paramètres. Nous rappelons que la notation Gauss, utilisée dans le chapitre 2, définit la fonction d'appartenance d'un ensemble flou gaussien multidimensionnel. D'après le chapitre précédent (section 4.6), par analogie avec les systèmes flous, ϕ_k peut effectivement être vue comme la fonction d'appartenance d'un ensemble flou. D'autres types de fonctions ϕ_k tenant compte de manière équivalente de l'imprécision relative au degré d'appartenance à la classe considérée peuvent être bien entendu envisagés.

Construction des masses de croyance initiales m_k

De même, l'appartenance de y à la classe E_j peut être représentée par le nombre flou gaussien $\tilde{e}_j$ défini par :

$$\tilde{e}_j(u) = \text{Gauss}(u; e_j, q_y \sigma_j^y) \quad \forall u \in \mathcal{Y} \subset \mathbb{R},$$

faisant intervenir l'écart-type estimé σ_j^y de E_j. Là encore, q_y est un paramètre de lissage à identifier.

On suppose que $\boldsymbol{x}$ et y sont des réalisations de variables aléatoires sous-jacentes définies sur les espaces probabilisables $\mathcal{X}$ et $\mathcal{Y}$. On note p_{kj} la probabilité conditionnelle que y appartienne à E_j sachant que $\boldsymbol{x}$ appartient à C_k :

$$p_{kj} = \mathcal{P}_{\mathcal{Y}|\mathcal{X}}(E_j|C_k),$$

et P la matrice des $(p_{kj})_{k=1,\ldots,K, j=1,\ldots,J}$. On peut associer à chaque classe C_k, une structure de croyance m_k d'éléments focaux $F(m_k) = \{\tilde{e}_j, j = 1, \ldots, J\}$ tels que

$$m_k(\tilde{e}_j) = p_{kj}.$$

Les classes E_j formant une partition de $\mathcal{Y}$, on a donc $\sum_{j=1}^{J} p_{kj} = 1$. Puisque les $\tilde{e}_j$ sont normalisés, m_k est donc une structure normalisée.

Inférence

Une fois définis les m_k, on procède comme dans le chapitre 4 (équation 4.1). Compte tenu de ce qui précède, pour le vecteur $\boldsymbol{x}$, chaque classe C_k fournit une information concernant la sortie inconnue y, représentée par la structure de croyance floue

$$\widehat{m}_k(.|\boldsymbol{x}) = m_k^{\alpha_k},$$

où $\alpha_k = 1 - \phi_k(\|\boldsymbol{x} - \boldsymbol{c}_k\|)$. Plus précisément, la structure $\widehat{m}_k(.|\boldsymbol{x})$ est définie par :

$$\left\{ \begin{array}{lcl} \widehat{m}_k(\tilde{e}_j|\boldsymbol{x}) & = & p_{kj}\phi_k[\|\boldsymbol{x} - \boldsymbol{c}_k\|] \quad \forall j \in \{1, \ldots, J\} \\ \widehat{m}_k(\mathcal{Y}|\boldsymbol{x}) & = & 1 - \phi_k(\|\boldsymbol{x} - \boldsymbol{c}_k\|) \\ \widehat{m}_k(A|\boldsymbol{x}) & = & 0 \quad \forall A \in \mathcal{F}(\mathcal{Y}) \setminus \{\tilde{e}_1, \ldots \tilde{e}_J, \mathcal{Y}\}. \end{array} \right. \tag{5.2}$$

Si on veut limiter le nombre de paramètres, il est également possible de simplifier ce modèle en prenant des valeurs identiques *a priori* pour $\boldsymbol{\sigma}_k^x$ et σ_j^y pour toutes les classes.

Ici, $|F(\widehat{m}(.|\boldsymbol{x}))| \leq J^K$. Le nombre d'éléments focaux peut donc encore être assez important. On aura tout intérêt à n'utiliser que les s plus proches prototypes ou les s plus proches classes de $\boldsymbol{x}$. L'avantage essentiel de cette méthode est qu'elle fournit une information très riche, mais son principal inconvénient est finalement là encore son coût de calcul, même si, en règle pratique, de nombreux coefficients p_{kj} seront nuls. De plus, la différence avec le modèle SCF1 est que les p_{kj} ne font pas partie des données mais que ces quantités doivent être estimées par les probabilités empiriques correspondantes (cf. paragraphe 5.2.3). Par la suite, on notera ce modèle Proto1.

Partitionnement simultané de l'espace d'entrée et de sortie

Cette version, voisine de la précédente, diffère par la méthode d'obtention des classes et par son faible coût en temps de calcul. Au lieu de déterminer les classes C_k et E_j séparément, il est également possible de faire une classification directement sur l'espace $\mathcal{X} \times \mathcal{Y}$ ou, si la configuration des données le permet, à partir de $\boldsymbol{c}_k$, d'en déduire les centres e_k en utilisant une fonction de régression h simple. C'est cette solution que nous proposons ici. Afin d'alléger les calculs, on peut par exemple utiliser le régresseur de Nadaraya-Watson (cf. équation (A.6)), paramétré par un unique scalaire λ. On a alors

$$e_k = h(\boldsymbol{c}_k, \lambda). \tag{5.3}$$

Dans ce cas, $K = J$ et la matrice de probabilité condititionnelle P est simplement la matrice Identité d'ordre K: $p_{kj} = \delta_{kj}$ pour tout k, j. Chaque m_k ne possède donc qu'un seul élément focal, comme dans le modèle SCF2, ce qui réduit considérablement les calculs.

Le nombre d'éléments focaux est donc généralement beaucoup plus faible que dans le modèle complet SCF1 : $|F(\widehat{m}(.|\boldsymbol{x}))| \leq 2^K$.

Dans la suite, on notera ce modèle Proto2. De plus, comme nous avons une bonne représentation globale des données, il est possible de n'utiliser que les s plus proches prototypes. Par exemple, avec $s = 3$, on obtient au maximum 8 éléments focaux, ce qui est tout-à-fait raisonnable !

Le tableau 5.1 récapitule le nombre maximal d'éléments focaux pour les 5 variantes proposées. L'avantage semble être en faveur de SCF3. Mais ce modèle ne permet pas une bonne représentation de l'incertitude et de l'imprécision sur la sortie, contrairement au modèle Proto1, par exemple.

Enfin, on peut faire une remarque sur l'utilisation de données floues. Si les données de l'ensemble d'apprentissage sont floues, il est encore possible de réaliser une classification automatique. Nous proposons, dans le cadre de la simplification de structure (cf. section 5.3) une méthode de regroupement d'ensembles flous.

SCF1	SCF2	SCF3	Proto1	Proto2
$\prod_{i=1}^{N}(J(i)+1)$	2^N	$N+1$	J^K	2^K

TAB. 5.1 – *Nombre maximal d'éléments focaux selon la version de notre méthode : N : taille de l'ensemble d'apprentissage, K et J : nombre de prototypes ou de classes sur les espaces respectifs X et Y.*

5.2.3 Identification des modèles Proto1 et Proto2

Paramètres du modèle

Dans les modèles Proto1 et Proto2, l'ensemble des paramètres à estimer est plus complexe que dans les modèles décrits dans le chapitre précédent. La phase de classification impose de choisir les centres $\boldsymbol{c}_k$ et leur nombre, les paramètres de dispersion des classes $\boldsymbol{\sigma}_k$, ainsi que les « méta-paramètres » q_x et q_y. Une fois l'ensemble des classes choisi, dans la méthode Proto1, il faut ensuite estimer la matrice des probabilités conditionnelles. On peut découper l'apprentissage du modèle en deux temps :

- l'identification de la structure : choix de K et, le cas échéant, de J ou de λ ;
- l'estimation des paramètres q_x, q_y, $\boldsymbol{c}_k, e_j, \boldsymbol{\sigma}_k^x, \sigma_j^y, p_{kj}, \quad k = 1, \ldots K, j = 1 \ldots J$.

Identification de la structure

Dans cette approche se pose le problème de la détermination du nombre optimal de prototypes. Cet aspect dépasse le cadre spécifique de notre modèle et relève du problème général de l'identification de systèmes, dont il a déjà été question aux chapitres 1 et 2. On peut utiliser ici à nouveau des techniques de rééchantillonnage en introduisant par exemple un terme de pénalisation. Nous préférons utiliser d'autres techniques plus rapides, utilisées dans le contexte de l'identification de systèmes flous [65], puisque ceux-ci offrent des similitudes avec notre modèle. Ces méthodes, basées sur des critères de classification optimale, ont déjà été développées en section 1.3.2.

Dans le modèle Proto1, nous déterminons ainsi séparément les nombres optimaux K^* et J^* de prototypes $\boldsymbol{c}_k$ et e_j.

Dans la méthode Proto2, les prototypes e_j sont calculés comme évaluation d'une fonction de régression dépendant d'un paramètre de lissage λ : $e_j = h(\boldsymbol{c}_k, \lambda)$. La valeur optimale de λ peut être obtenue en minimisant le critère classique de validation croisée entre nombre *réels* (cf. section 1.3.2).

Estimation des paramètres du modèle

La phase d'identification précédente détermine en même temps les centres $\boldsymbol{c}_k$, e_j et les écarts-types σ_k^x, σ_j^y. Il reste donc à estimer les paramètres suivant :

- les méta-paramètres q_x et q_y ;
- les poids p_{ij} (dans la méthode Proto2).

Estimation des méta-paramètres

Ces paramètres des fonctions d'activation ϕ_k jouent le rôle de paramètres de lissage dans notre méthode. Pour estimer ces paramètres, nous utilisons la généralisation du critère de validation croisée aux structures de croyance défini en section 4.5.3. Le paramètre vectoriel $\boldsymbol{\theta} = (q_x, q_y)$ « idéal » minimise donc le coût défini par l'équation (4.29) :

$$CV(\boldsymbol{\theta}) = \frac{1}{K} \sum_{k=1}^{K} \beta_k C(m_i, \widehat{m}_i(\boldsymbol{\theta})),$$

où le coût C est défini par l'équation (4.28).

Estimation des probabilités conditionnelles (Proto2)

La matrice de probabilités conditionnelles d'appartenance aux différentes classes $P = (p_{kj})$ n'est pas connue. Nous proposons deux approches différentes pour déterminer de « bonnes » estimations $\hat{p}_{kj}$ de p_{kj}, dans un sens à définir : une approche fréquentiste classique et une approche « fréquentiste floue », basée sur les probabilités d'ensembles flous.

Approche fréquentiste. Cette approche est tout simplement basée sur l'estimation d'une probabilité théorique p par la probabilité empirique $\widehat{p}$ correspondante. Soit N_{kj} le nombre de couples $(\boldsymbol{x}_k, y_j)$ de l'ensemble d'apprentissage appartenant à $C_k \times E_j$. La probabilité empirique $\widehat{p}_{kj}$ est définie par :

$$\widehat{p}_{kj} = \widehat{P_{Y|X}}(E_j|C_k) = \frac{\widehat{P}_{XY}(E_j \cap C_k)}{\widehat{P}_X(C_k)} = \frac{N_{kj}}{\sum_j N_{kj}}$$

La loi forte des grands nombres justifie théoriquement l'estimation proposée :

$$\widehat{p}_{kj} \overset{n \to +\infty}{\longrightarrow} p_{kj}$$

presque sûrement.

Approche par les probabilité floues. Une autre méthode est envisageable, basée sur les probabilités floues, tenant compte des frontières floues des classes. Cette méthode consiste à estimer les probabilités d'appartenance aux classes floues selon la définition de Zadeh (cf. (équation 3.57)). La probabilité d'appartenir à l'ensemble flou $\tilde{\boldsymbol{c}}_k \times \tilde{e}_j$ est définie par

$$P_{XY}(\tilde{\boldsymbol{c}}_k \cap \tilde{e}_j) = \int_{\mathcal{X} \times \mathcal{Y}} (\tilde{\boldsymbol{c}}_k \cap \tilde{e}_j)(\boldsymbol{x}, y) dP_{XY}(\boldsymbol{x}, y) = \int_{\mathcal{X} \times \mathcal{Y}} [\tilde{\boldsymbol{c}}_k(\boldsymbol{x}) \wedge \tilde{e}_j(y)] \, dP_{XY}(\boldsymbol{x}, y)$$

où $\wedge$ est une t-norme (le produit). Là encore, nous pourrions appliquer la loi des grands nombres et calculer la probabilité empirique associée

$$\widehat{P_{XY}}(\tilde{\boldsymbol{c}}_k \cap \tilde{e}_j) = \frac{1}{N} \sum_{i=1}^{N} \tilde{\boldsymbol{c}}_k(\boldsymbol{x}_i) \wedge \tilde{e}_j(y_i).$$

Mais cette définition conduit à une propriété indésirable :

$$\sum_{k,j} \widehat{P}_{XY}(\tilde{\boldsymbol{c}}_k \cap \tilde{e}_j) \neq 1.$$

Dans ce cas, Yager [172] propose de modifier cette formule en utilisant les degrés d'appartenance *normalisés* μ^i_{kj} à $\tilde{\boldsymbol{c}}_k \times \tilde{e}_j$ pour chaque $(\boldsymbol{x}_i, y_i)$, définis par :

$$\mu^i_{kj} = \frac{\tilde{\boldsymbol{c}}_k(\boldsymbol{x}_i) \wedge \tilde{e}_j(y_i)}{\sum_{k,j} \tilde{\boldsymbol{c}}_k(\boldsymbol{x}_i) \wedge \tilde{e}_j(y_i)}$$

Il s'agit d'une extension de l'approche fréquentiste aux ensembles flous. Par analogie avec la méthode précédente, on peut alors déterminer les estimations des probabilités de façon suivante :

- $\widehat{P_{XY}}(\tilde{e}_j \cap \tilde{\boldsymbol{c}}_k) = \frac{1}{N} \sum_{k=1}^{N} \mu^i_{kj}$,
- $\widehat{P_X}(\tilde{\boldsymbol{c}}_k) = \frac{1}{N} \sum_j \sum_i \mu^i_{kj}$,
- $\widehat{p}_{kj} = \widehat{P_{Y|X}}(\tilde{e}_j|\tilde{\boldsymbol{c}}_k) = \dfrac{\sum_{i=1}^{N} \mu^i_{kj}}{\sum_j \sum_i \mu^i_{kj}}$

La preuve de la convergence des probabilités empiriques, calculées selon cette méthode, vers les probabilités théoriques n'a pas été faite. En effet, la loi des grands nombres ne s'applique pas immédiatement. Néanmoins, l'intérêt de cette approche est de tenir compte de l'imprécision des frontières entre les classes.

5.3 Procédures de simplification

5.3.1 Principe général

Motivations

L'inconvénient majeur des méthodes basées sur la théorie des croyances est qu'elles sont généralement coûteuses en temps de calcul pendant la phase de combinaison des structures, à cause de la multiplication des éléments focaux. Par exemple, dans le modèle Proto2 à K prototypes, nous avons vu qu'on pouvait obtenir jusqu'à 2^K éléments focaux. De plus, outre la lenteur des calculs, le résultat de la structure finale $\widehat{m}(.|\boldsymbol{x})$ pose également un problème de lisibilité. En effet, la plupart des éléments focaux de $\widehat{m}(.|\boldsymbol{x})$ sont difficiles à interpréter et certains sont quasi indiscernables.

Pour ces deux raisons, la rapidité des calculs et l'interprétabilité de la structure, il convient donc de contrôler sa complexité au cours de la phase de combinaison. Dans cet esprit, nous proposons une procédure de simplification d'une structure de croyance quelconque (classique ou floue).

Réduction du nombre d'éléments focaux

Le principe général est de transformer une structure de croyance m, floue ou non, en une autre structure m', telles que

$$|F(m')| < |F(m)|.$$

Ce principe est très différent des deux autres types d'approximations que nous avons vus jusqu'à présent, les prototypes et les plus proches voisins, qui consistent à réduire le nombre de structures à combiner. Ici, nous cherchons à réduire le nombre d'*éléments focaux* pendant la combinaison des structures. A chaque fois que nous combinons deux structures m_1 et m_2 ayant respectivement p et q éléments focaux, la structure résultante $m_{12} = m_1 \cap m_2$ possède pq éléments focaux. Avant de combiner m_{12} avec une troisième structure m_3, nous proposons de réduire d'abord le nombre d'éléments focaux de m_{12} en créant une structure approchée $m'_{12} = \text{App}(m_{12})$ dont les caractéristiques sont supposées voisines de m_{12}. L'étape suivante consiste à combiner m'_{12} et m_3 en une nouvelle structure m_{123}.

Le schéma d'obtention de la structure approchée finale est donc un algorithme itératif en deux étapes alternées :

Pour i=1 à N

- *fusion :* $m_{1\ldots i+1} = m_i \cap m'_{1\ldots i}$
- *approximation (si nécessaire):* $m'_{1\ldots i+1} = App(m_{1\ldots i+1})$

Fin

Si nous appliquons ce principe à notre modèle, nous obtenons donc la structure approchée $\widehat{m}'(.|\boldsymbol{x})$ de $\widehat{m}(.|\boldsymbol{x})$:

$$\widehat{m}'(.|\boldsymbol{x}) = \text{App}\left(\ldots(\text{App}\left(\widehat{m}_1(.|\boldsymbol{x}) \cap \widehat{m}_2(.|\boldsymbol{x})\right) \cap \widehat{m}_3(.|\boldsymbol{x}))\ldots \cap \widehat{m}_N(.|\boldsymbol{x})\right)$$

Nous pouvons remarquer que l'ordre de la combinaison a maintenant son importance. L'opération de combinaison n'est plus associative.

Nous allons maintenant nous concentrer sur les méthodes d'approximation, en décrivant d'abord les principales méthodes existantes, définies dans le cadre de structures à éléments focaux *classiques* et sur un référentiel *discret*. Puis, nous développerons de nouvelles méthodes, plus adaptées à la gestion d'éléments focaux flous et à l'existence d'un domaine continu.

5.3.2 Méthodes existantes

Soit une structure m d'éléments focaux $\{F_1, \ldots, F_p\}$. Le problème est donc de chercher une structure $m' = \text{App}(m)$, similaire à m, dans un sens à définir, et telle que $|F(m')| < |F(m)|$.

Plusieurs méthodes ont été proposées dans le cas des structures de croyances *classiques*, sur un référentiel $\mathcal{Y}$ discret [9, 153, 46, 161].

Approximation probabiliste [161]

La structure m' est bayésienne, les éléments focaux de m' sont des singletons. Voorbraak a proposé la transformation suivante :

$$\begin{cases} m'(\{y\}) &= \frac{\sum_{B\in F(m)} m(B)\delta_B(y)}{\sum_{B\in F(m)} m(B)|B|} \quad \forall y \in \mathcal{Y} \\ m'(A) &= 0 \quad \text{if } |A| > 1 \end{cases} \tag{5.4}$$

Nous considérons qu'il serait plus judicieux d'utiliser tout simplement la probabilité pignistique de m. En effet, dans la méthode de Voorbraak, les éléments focaux peu spécifiques sont sur-représentés.

Que l'on utilise l'une ou l'autre de ces transformations, la méthode bayésienne n'est pas représentative de la réalité dans les cas de grande nonspécificité des éléments focaux. De plus, elle nécessite des modifications si on veut l'appliquer dans le cas continu, car on obtiendrait dans ce cas une infinité d'éléments focaux. Une solution simple consiste cependant à discrétiser l'ensemble de référence $\mathcal{Y}$.

Approximation consonante [46]

Ce groupe de méthodes est défini en détail dans un article de Dubois et Prade [46]. Elles consistent à déterminer la structure consonante m', c'est-à-dire dont les éléments focaux sont emboîtés, la plus proche de m selon un certain critère. L'un de ces critères fait appel à la notion d'inclusion (dite *forte*) entre structures de croyance. Cependant, certains auteurs [153] ont montré que cette méthode est peu adaptée à la règle de combinaison des structures. L'ordre de présentation des structures peut modifier énormément le résultat final. De plus, la propriété de consonance n'est pas préservée par la règle de combinaison.

Nous rejetons donc l'emploi de cette méthode pour notre objectif.

Elimination d'éléments focaux

Ces méthodes ne sont pas basées sur une structure particulière de m', contrairement aux précédentes. Dans ce groupe de méthodes, les éléments focaux F_i peu significatifs, c'est-à-dire de poids $m(F_i)$ « faible », selon divers critères, sont éliminés. Les paramètres pouvant intervenir dans la sélection sont les suivants :

- le pourcentage total de poids des éléments focaux conservés ;
- un nombre minimal et/ou maximal d'éléments focaux.

Soit $K(m)$ l'ensemble des éléments focaux de m conservés. On a donc $F(m') = K(m)$.

Dans les méthodes les moins sophistiquées, comme celles définies par Lowrance et al. ou Tessem [98, 153], la masse de $F_i \in F(m)\backslash K(m)$ est répartie uniformément sur les éléments conservés. L'approximation m' est alors définie par :

$$\begin{cases} m'(A) &= \frac{m(A)}{\sum_{B\in K(m)} m(B)} \quad \forall A \in K(m) \\ m'(A) &= 0 \quad \forall A \notin K(m). \end{cases} \tag{5.5}$$

L'inconvénient de la méthode de Tessem est qu'elle ne tient pas compte de la proximité des F_i éliminés avec les éléments restants.

Bauer [9] a proposé un algorithme permettant de mieux répartir la masse $m(F_i)$ selon les F_j de $K(m)$ tels que $F_i \subset F_j, F_i \cap F_j \neq 0$ ou $\mathcal{Y}$.

Dans ces méthodes, il n'y a pas de transformation ni de création d'éléments focaux, mais seulement une absorption des éléments focaux éliminés. Ces méthodes se généralisent immédiatement aux structures de croyance floues.

5.3.3 Classification des éléments focaux

Les méthodes que nous proposons sont basées sur l'*agrégation d'ensembles flous semblables*. De façon générale, toute méthode de regroupement ou de classification automatique, étendue aux ensembles flous peut convenir : nous avons envisagé une classification de type hiérarchique [119]. La procédure est itérative : à chaque étape t, on remplace deux ensembles similaires par un seul. Ceci nécessite de définir :

- un *critère de proximité* I entre ensembles flous ;
- un *opérateur d'agrégation* d'ensembles flous ∇ ;
- et un *critère d'arrêt* δ de la procédure.

Soit $m^{(t)}$ la structure de croyance floue à l'étape t, dont les éléments focaux sont $\{F_1^{(t)}, \ldots, F_{p-t}^{(t)}\}$. Une nouvelle structure de croyance $m^{(t+1)}$ est obtenue en remplaçant les deux plus proches éléments focaux $F_i^{(t)}$ et $F_j^{(t)}$ de $m^{(t)}$ par un ensemble F' et en conservant les autres éléments focaux. Ainsi,

$$F(m^{(t+1)}) = (F(m^{(t)}) \cup \{F'\}) \backslash \{F_i^{(t)}, F_j^{(t)}\}$$
$$\text{avec } m^{(t+1)}(F') = m^{(t)}(F_i^{(t)}) + m^{(t)}(F_j^{(t)}).$$

Critères de proximité

Parmi les nombreuses mesures de similarité entre *ensembles flous* définies dans la littérature [129, 183], nous avons choisi la mesure ensembliste suivante :

$$S(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

On regroupe les deux ensembles A et B maximisant ce critère.

Ainsi, dans notre algorithme, à l'étape t, on choisit les ensembles $F_i^{(t)}$ et $F_j^{(t)}$ tels que

$$S(F_i^{(t)}, F_j^{(t)}) = \max_{k,l} S(F_k^{(t)}, F_l^{(t)}).$$

D'autres critères, tenant également compte de la masse des éléments focaux, sont envisageables. Ainsi, par analogie avec la méthode de Ward [165] en classification hiérarchique classique (cf. annexe **??**), on peut définir le critère suivant :

$$I(A, B) = \frac{m^{(t)}(A) + m^{(t)}(B)}{m^{(t)}(A)m^{(t)}(B)} \frac{S(A, B)}{1 - S(A, B)} \tag{5.6}$$

qui permet d'absorber les éléments focaux de faible importance.

Opérateurs d'agrégation

De nombreux opérateurs binaires d'agrégation ∇ sont *a priori* envisageables, produisant un nouvel ensemble F'. L'utilisation de l'opérateur d'union,

$$F' = F_i \cup F_j,$$

présente la propriété d'augmenter systématiquement la nonspécificité N. On vérifie en effet facilement que

$$N(m^{(t)}) \leq N(m^{(t+1)}).$$

Cette propriété est cohérente, car l'opération d'approximation suppose que l'on perde de l'information, ce qui se traduit ici par une plus grande nonspécificité. De plus, il est intéressant d'avoir une suite de structures « emboîtées », c'est-à-dire monotones selon un critère donné. Cependant, le danger de ce type de critères est de définir une structure très éloignée de la structure de départ après un grand nombre d'itérations. Il semble plus judicieux de faire intervenir là encore les masses des éléments focaux. En outre, l'objectif n'est pas forcément de chercher des structures emboîtées, mais plutôt des structures semblables. Nous proposons alors l'opérateur plus neutre de la *moyenne pondérée*, telle que :

$$\forall y \quad F'(y) = \frac{m^{(t)}(F_i)F_i(y) + m^{(t)}(F_j)F_j(y)}{m^{(t)}(F_i) + m^{(t)}(F_j)} \tag{5.7}$$

toujours par analogie avec la méthode de Ward.

Critère d'arrêt

Cette procédure d'agrégation s'arrête à l'itération t dès que, pour un seuil δ donné,

$$\max_{k,l} I(F_k^{(t)}, F_l^{(t)}) < \delta.$$

5.3.4 Méthodes d'optimisation

Les critères précédents sont définis *a priori*, sans optimisation d'une mesure globale d'erreur entre les structures $m = m^{(t)}$ et $m' = m^{(t+1)}$. On ne s'intéresse qu'à des critères entre éléments focaux sans tenir véritablement compte de l'impact global sur la structure. Le choix de A et B n'est peut-être pas celui qui minimise un critère d'erreur donné.

On peut alors de manière plus systématique définir un critère d'erreur C entre les structures $m^{(t)}$ et $m^{(t+1)}$, basé sur :

- des mesures d'information, comme la nonspécificité, la discordance ou l'information mutuelle,
- des mesures « probabilistes » , utilisant la probabilité pignistique, la crédibilité, la plausibilité.

Soit $\Phi_\nabla(.; A, B)$ la transformation, qui à la structure m, associe la structure m' telle que :

$$\begin{aligned} &F(m') = (F(m) \cup \{A \nabla B\}) \backslash \{A, B\}, \\ &m'(A \nabla B) = m(A) + m(B), \\ &m'(C) = m(C) \quad \forall C \in F(m') \backslash A \nabla B, \end{aligned} \tag{5.8}$$

où ∇ est un opérateur binaire flou.

Le problème peut ainsi se formuler de manière générale comme un problème d'optimisation :

$$\min_{A,B\in F(m)} C\left(m, \Phi_\nabla(m; A, B)\right) \tag{5.9}$$

Le critère d'arrêt de la procédure est défini comme précédemment. Ce problème de minimisation définit une fonction $f_{\nabla,C}$ de A et B et peut donc se réécrire de façon synthétique par le problème suivant :

$$\min_{(A,B)\in F(m)} f_{C,\nabla}(A, B). \tag{5.10}$$

Dans la suite, nous proposons quelques résultats théoriques, essentiellement pour l'opérateur d'union, dans le cas de structures de croyances *classiques*.

Critères « probabilistes »

On peut comparer les distributions de probabilité pignistiques et utiliser une distance quelconque entre deux distributions de probabilité, comme la distance de Kullback-Leibler :

$$C(m, m') = \int_{\mathcal{Y}} \ln\left(\frac{p_{bet}(y)}{p'_{bet}(y)}\right) p_{bet}(y)dy \tag{5.11}$$

ou la distance suivante :

$$C(m, m') = \int_{\mathcal{Y}} g[|p_{bet}(y) - p'_{bet}(y)|]dy. \tag{5.12}$$

où g est une fonction monotone croissante $\mathbb{R}^+$ dans $\mathbb{R}^+$. De manière générale, toute distance entre distributions de probabilité peut convenir.

On peut expliciter le critère défini par l'équation (5.10) en prenant l'opérateur d'agrégation $\cup$ et la mesure d'erreur (5.11). Dans le cas de structures de croyance classique m, le critère à minimiser en fonction de $(A, B) \in F(m)$ est obtenu par un calcul direct :

$$\begin{aligned} f_{\cup,C}(A, B) \;=\; & |A\backslash B|\, g\left(\left|\tfrac{m(A)}{|A|} - \tfrac{m(A)+m(B)}{|A\cup B|}\right|\right) + |B\backslash A|\, g\left(\left|\tfrac{m(B)}{|B|} - \tfrac{m(A)+m(B)}{|A\cup B|}\right|\right) \\ & + |A\cap B|\, g\left(\left|\tfrac{m(A)}{|A|} + \tfrac{m(B)}{|B|} - \tfrac{m(A)+m(B)}{|A\cup B|}\right|\right). \end{aligned} \tag{5.13}$$

Ce critère est valable dans le cas présent où $\mathcal{Y}$ est continu, mais également, dans le cas discret, en utilisant la définition adéquate de la cardinalité $|.|$.

On peut également se baser sur d'autres quantités comme la plausibilité ou la crédibilité. On définit ainsi le critère suivant :

$$C(m, m') = \sum_{F\in F(m)} g[|\mathrm{bel}(F) - \mathrm{bel}'(F)|]. \tag{5.14}$$

La fonction $f_{\cup,C}$ à minimiser devient alors :

$$\begin{aligned} f_{\cup,C}(A, B) \;=\; & g(m(A))\mathrm{Card}\{C|A \subseteq A \text{ et } B \not\subseteq C\} \\ & + g(m(B))\mathrm{Card}\{C|B \subseteq C \text{ et } A \not\subseteq C\} \end{aligned} \tag{5.15}$$

où Card(E) dénote le nombre d'éléments de l'ensemble E.

Mesures d'information

On peut également définir des critères d'erreur en se basant sur des mesures d'incertitude comme la nonspécificité N ou l'une des 5 mesures de conflit Confl définies dans la section 3.2.5.

$$C(m, m') = N(m') - N(m),$$

$$C(m, m') = \text{Confl}(m') - \text{Confl}(m).$$

Le choix de la structure m' correspond au *principe d'incertitude minimale* [83]. On cherche les éléments focaux A et B de façon à perdre le moins d'information possible.

Dans le cas de la nonspécificité, le critère à minimiser, pour un opérateur d'agrégation quelconque, est le suivant :

$$f_{\nabla,C}(A, B) = m(A) \log_2 \frac{|A \nabla B|}{|A|} + m(B) \log_2 \frac{|A \nabla B|}{|B|}. \tag{5.16}$$

Ces critères sont *a priori* à utiliser avec prudence car deux structures très différentes pourraient présenter des caractéristiques semblables tout en étant géométriquement éloignées. Cependant, cette situation est dans ce contexte improbable, car on ne compare ici la nonspécifivité que pour des structures emboîtées.

5.4 Résultats

5.4.1 Mesures de précision et d'information

Jeux de données et objectifs

Afin de montrer la validité et les capacités de prédiction de notre méthode, nous présentons les résultats obtenus par des simulations ou sur des données provenant d'applications réelles. Nous proposons d'étudier quatre problèmes de régression, illustrant chacun un ou plusieurs points des modèles présentés.

1. Données simulées d'une fonction de régression classique. On montre que notre méthode généralise les méthodes de régression classiques. Dans certaines zones, la fonction n'est pas connue ou imprécise,

2. Données réelles (broyage). Les données à estimer sont imprécises,

3. Données simulées multi-valuées (problème inverse),

4. Données réelles, multi-dimensionnelles (taux de mercure dans les tissus de poissons). Les données sont imprécises.

Mesures de précision et d'information

Soit $\{(\boldsymbol{x}_1, m_1), \ldots, (\boldsymbol{x}_n, m_n)\}$ un échantillon de données testées. Le critère de précision dépend de la nature des m_i. Si les m_i sont définis comme des nombres réels (modèle SCF1), la

précision de l'estimation est mesurée en terme d'erreur quadratique moyenne :

$$J_n = \frac{1}{n} \sum_{i=1}^{n} \left[(y_i - \widehat{y}_i(\boldsymbol{x}_i) \right]^2 ,$$

$\widehat{y}_i(\boldsymbol{x}_i)$ étant une des sorties ponctuelles définies en section 4.4.2. Par défaut, nous avons utilisé l'espérance pignistique (équation 3.53).

Si les m_i sont des intervalles, des ensembles flous, ou des structures de croyance, la précision est mesurée par le critère :

$$C(m_i, \widehat{m}(.|\boldsymbol{x}_i, \boldsymbol{\theta})) = \sum_{F \in F(m_i^*)} \sum_{F' \in F(m_i^*(.|x))} m_i^*(F) \widehat{m}_i^*(F'|\boldsymbol{x}) \tilde{d}(F, F'),$$

avec

$$\tilde{d}(F, F') = \int_0^1 h_2(F_\alpha, F'_\alpha) d\alpha,$$

où h_2 est la distance de Hausdorff entre deux ensembles classiques. Nous rappelons que ce critère généralise J_n.

Afin de décrire l'information que contiennent les sorties, plusieurs mesures d'incertitude classiques [84] ont été utilisées :

- la mesure de nonspécificité $N(m)$ (équation (3.22)), qui représente l'imprécision des ensembles,
- quatre mesures de conflit : la discordance $D(m)$ (équation (3.27)), la dissonance $E(m)$ (équation (3.25)), le *strife* $S(m)$ (équation (3.28)), et la mesure symétrique $\Delta(m)$ (équation (3.30)).

Enfin, le cas échéant, l'approximation des structures a été effectuée par la procédure de simplification des éléments focaux définie dans la section 5.3.

5.4.2 Exemple 1: simulation

Description des données

Dans ce problème de régression monodimensionnelle [116], les $x_i \in \mathbb{R}$ sont issus d'une variable aléatoire X dont la loi est un mélange de 2 distributions gaussiennes de mêmes proportions :

$$X \sim 0.5\mathcal{N}(-1.5, 0.5) + 0.5\mathcal{N}(1.75, 0.5)$$

et la sortie est définie par la fonction bruitée suivante :

$$y = \sin(3x) + x + \varepsilon(x),$$

où $\varepsilon(x) \sim \mathcal{N}(0, 0.01)$ si $x \leq 0$ et $\epsilon(x) \sim \mathcal{N}(0, 1)$ si $x > 0$. Ici, la taille de l'ensemble d'apprentissage étant relativement élevée, $N = 200$, nous avons utilisé la version Proto2 de notre méthode.

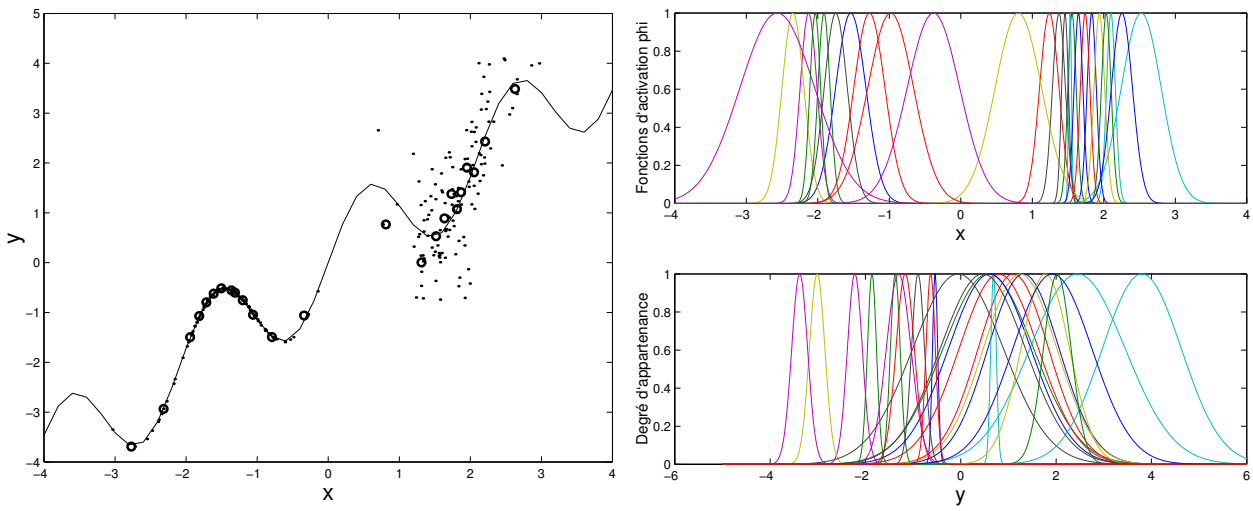


FIG. 5.1 – **Exemple 1. A gauche** *: données réelles y (-) ; ensemble d'apprentissage (.) ; prototypes (o).* **A droite** *: fonctions d'activation ϕ_k (en haut); éléments focaux $\tilde{e}_k$ des m_k (en bas)*

Estimation des paramètres du modèle

La phase d'identification (cf. section 1.3.2) nous a amené à retenir $K = 25$ prototypes c_k déterminés par classification de l'ensemble d'apprentissage. La technique d'identification est la même que celle utilisée dans le chapitre 2. La figure 5.1, à gauche, représente l'ensemble d'apprentissage et les prototypes (c_k, e_k), les e_k étant les évaluations du régresseur de Nadaraya-Watson en c_k. Les fonctions d'activation, de paramètres $c_k, q_x\sigma_k^x$ et les nombres flous gaussiens $\tilde{e}_k$, de paramètres $e_k, q_y\sigma_k^y$ sont représentés à droite, avec $q_x = 3$ et $q_y = 1$. Ces valeurs ont été obtenues par optimisation du critère classique de validation croisée, puisque les y_i sont des nombres réels.

Détermination de la structure $\widehat{m}$

Même dans cet exemple de complexité limitée, les procédures de simplification sont nécessaires. En effet, sans simplification, le nombre d'éléments focaux serait de 2^{25} ! Plusieurs options étaient envisageables. Les résultats des figures 5.2 et 5.4 ont été obtenus en utilisant uniquement les 3 plus proches prototypes dans la combinaison des structures. Les structures ont donc au maximum 8 éléments focaux.

La figure 5.2 donne un exemple de sortie $\widehat{m}$ et $\widehat{m}^*$, pour une valeur particulière $x = 0$. La figure 5.3 représente la probabilité pignistique correspondante.

La figure 5.4 montre l'intervalle interdécile de la distribution pignistique $p_{bet}(.|x)$ ainsi que des mesures d'imprécision. Les régions de faible densité sont caractérisées par des valeurs élevées pour la nonspécificité et l'ignorance $\widehat{m}(\mathcal{Y}|x)$ et par un intervalle interdécile très étendu. Les résultats confirment également une plus grande nonspécificité et un intervalle interdécile moins précis pour les régions ou la variance est plus grande.

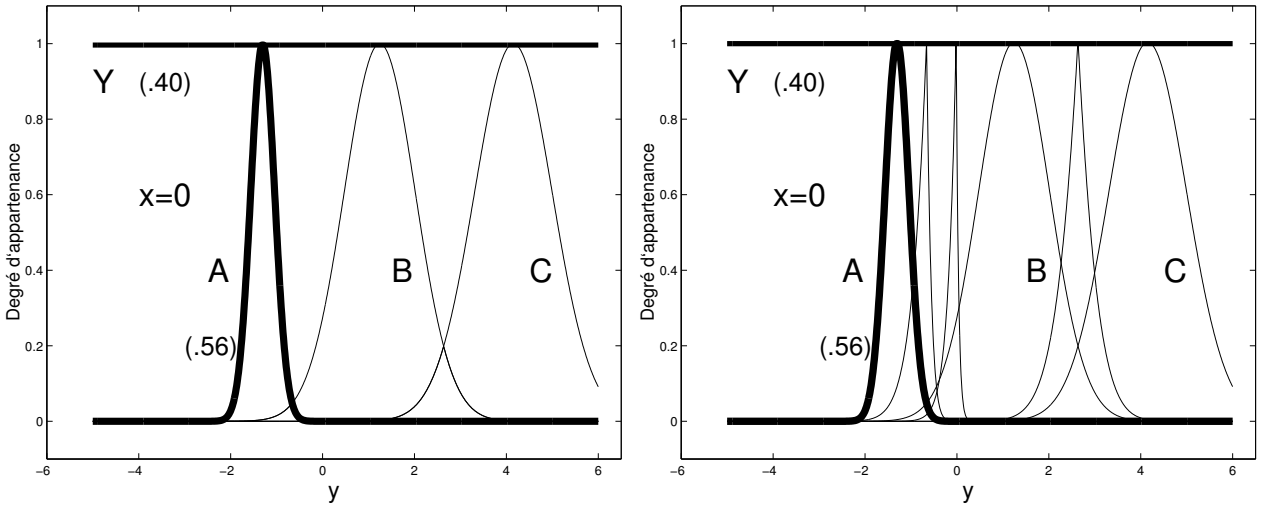


Fig. 5.2 – **Exemple 1.** *Sortie finale pour* $x = 0$. **A gauche** *:* $\widehat{m}$. **à droite***:* $\widehat{m}^*$. $A, B, C, \mathcal{Y}$ *sont des éléments focaux des deux structures. Le poids de* A *et* $\mathcal{Y}$*, les deux éléments prépondérants, est indiqué entre parenthèses.*

	Splines	NW	« Lazy »	Yager	Proto2
EQM 2	0.035	0.016	0.019	0.010	0.012
EQM 1	10^{-5}	0.003	0.013	0.008	0.011

Tab. 5.2 – **Exemple 1***: EQM pour différents modèles de régression: EQM 1 pour* $x \in [-2.5, -0.7]$*, EQM 2 pour* $x \in [-2.5, -0.7] \cup [0.8, 2.6]$.

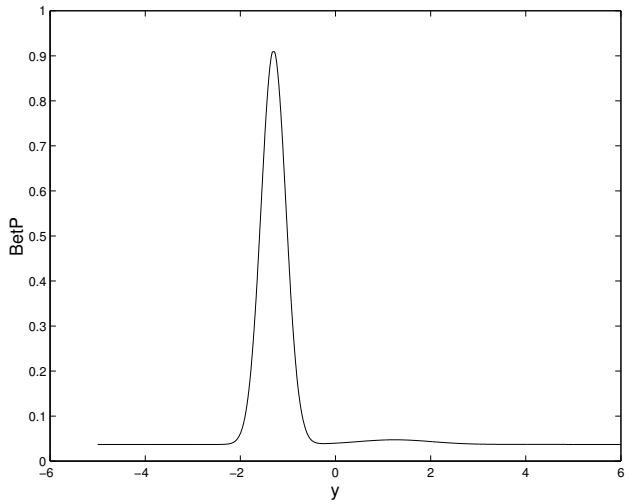


Fig. 5.3 – **Exemple 1.** *Sortie finale pour* $x = 0$ *:* p_{bet}

Estimation ponctuelle : comparaison avec des méthodes classiques

Afin de vérifier l'efficacité de notre méthode en termes de précision, nous l'avons comparée à plusieurs méthodes de régression classiques (cf. figure 5.5). Le tableau 5.2 indique l'erreur quadratique moyenne J obtenue pour deux méthodes de régression non paramétriques

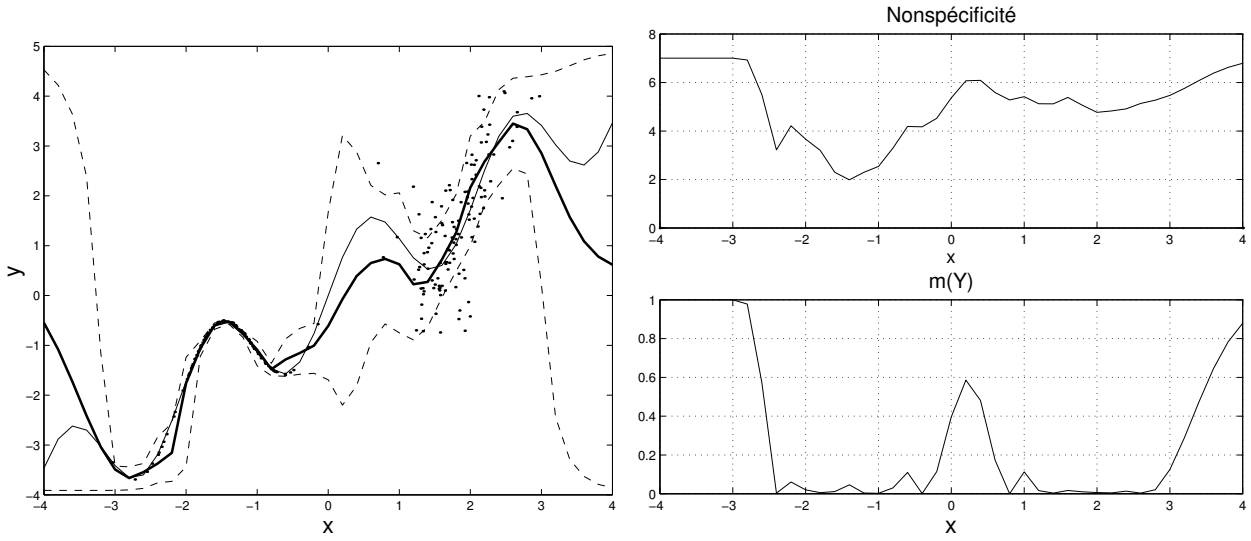


FIG. 5.4 – **Exemple 1. A gauche**. *données réelles y (-);); estimations* $\widehat{y}(\boldsymbol{x})$ *(-, en gras); 1er et 9eme decile de* $p_{bet}(.|\boldsymbol{x})$ *(- -); ensemble d'apprentissage (.)* **A droite**. *nonspécificité ; masse affectée au domaine entier* $\widehat{m}(\mathcal{Y}|\boldsymbol{x})$.

classiques, les splines cubiques et le régresseur de Nadaraya-Watson (NW), une méthode de mémorisation (« Lazy »)[10] (cf. annexe A), le modèle de Yager et notre méthode (Proto2). Les erreurs J ont été calculées uniquement dans les régions de fortes densités, dans deux zones : l'une contient les données bien spécifiées, ($x \in [-2.5, -0.7]$), l'autre inclut également la région de plus grande variance : ($x \in [-2.5, -0.7] \cup [0.8, 2.6]$). Comme le montre le tableau 5.2, les performances de notre méthode sont à peu près équivalentes aux autres, même si l'estimation des données de la zone 1 est particulièrement bonne pour splines. Cependant, cette méthode s'adapte moins aux données imprécises de la zone 2, ce qui ne semble pas le cas pour notre méthode.

Conclusion

Ce premier exemple, qui peut parfaitement être traité à l'aide de méthodes classiques, permet justement de confronter l'efficacité et la précision de notre méthode à celle des méthodes de régression standard. Nous ne prétendons pas réaliser de meilleures performances, mais nous obtenons des résultats corrects. De plus, notre méthode permet d'évaluer avec des critères simples la qualité globale de la sortie.

5.4.3 Exemple 2 : Broyage

Cet exemple provient d'une base de données réelle. Pour des raisons illustratives, nous avons choisi un modèle à une seule variable explicative. Un problème multidimensionnel sera étudié dans l'exemple 4. Dans la base d'apprentissage, pour une vitesse d'alimentation donnée, la variable à expliquer, la rugosité de la surface n'est connue qu'au travers d'un intervalle de valeurs. L'utilisation de cet intervalle de valeurs permet implicitement de tenir compte de la pauvreté du modèle en l'absence d'autres variables. Tanaka a traité ce problème à l'aide d'une méthode de régression par intervalles [151].

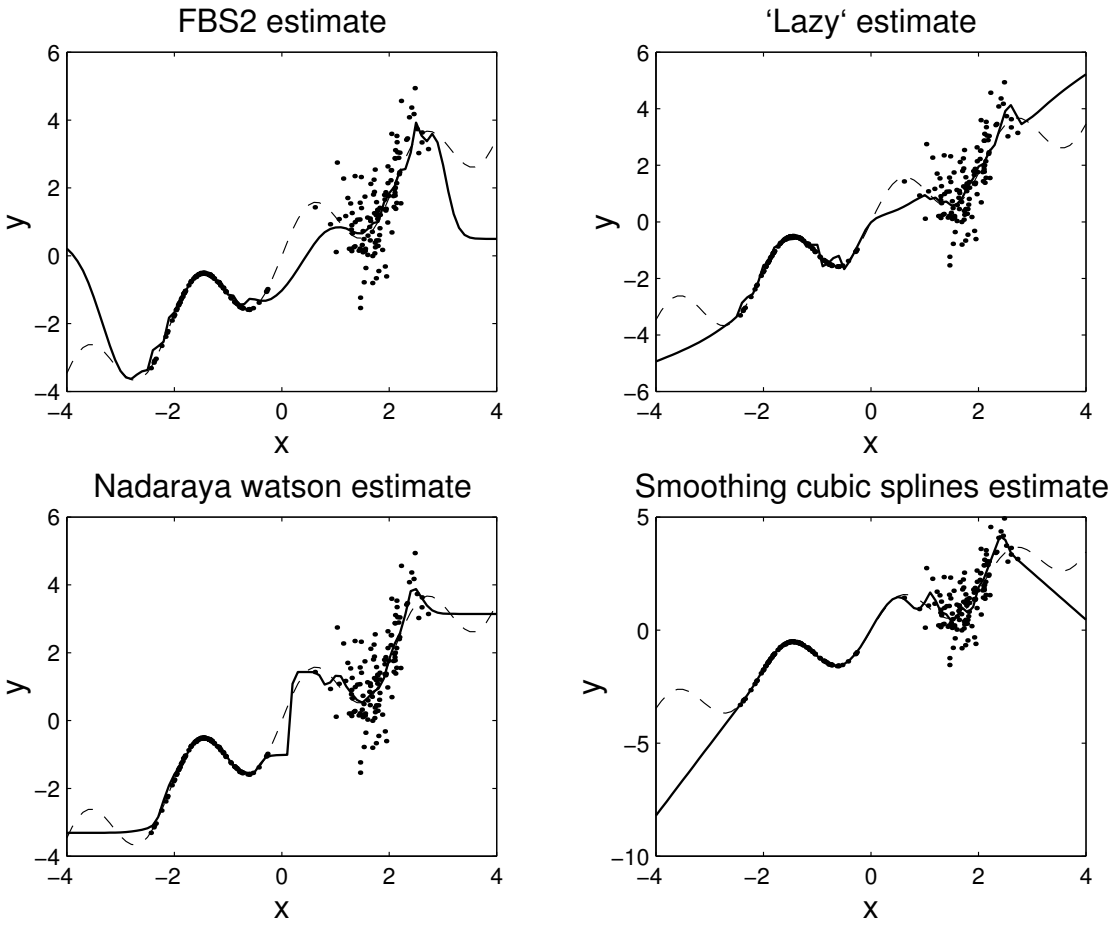


Fig. 5.5 – **Exemple 1.** *ensemble d'apprentissage (.); données réelles (–); estimation ponctuelle pour différents modèles (-)*

Description des données et solutions possibles

Dans cet exemple, le modèle est constitué d'une seule variable explicative x et on dispose d'un ensemble d'apprentissage très réduit de 8 triplets (x_i, y_i^-, y_i^+) où les valeurs y_i^-, y_i^+ représentent les valeurs minimale et maximale de la variable y observées pour une valeur fixée x_i de la variable d'entrée.

Il y a plusieurs façons de traiter ce type d'exemple. Si on opte pour une méthode probabiliste, bien que la configuration des données ne soit pas idéale, on peut par exemple construire deux modèles distincts utilisant respectivement les valeurs inférieure et supérieure de l'intervalle. On peut également prendre le milieu de l'intervalle et simuler un bruit uniforme autour de cette valeur.

Cependant, il est bien plus naturel de considérer que la sortie est définie directement par l'intervalle $Y_i = [y_i^-, y_i^+]$ et utiliser une méthode traitant ce type de données, comme la régression floue ou notre approche. Nous pouvons également construire la sortie comme un ensemble flou $\tilde{y}_i$.

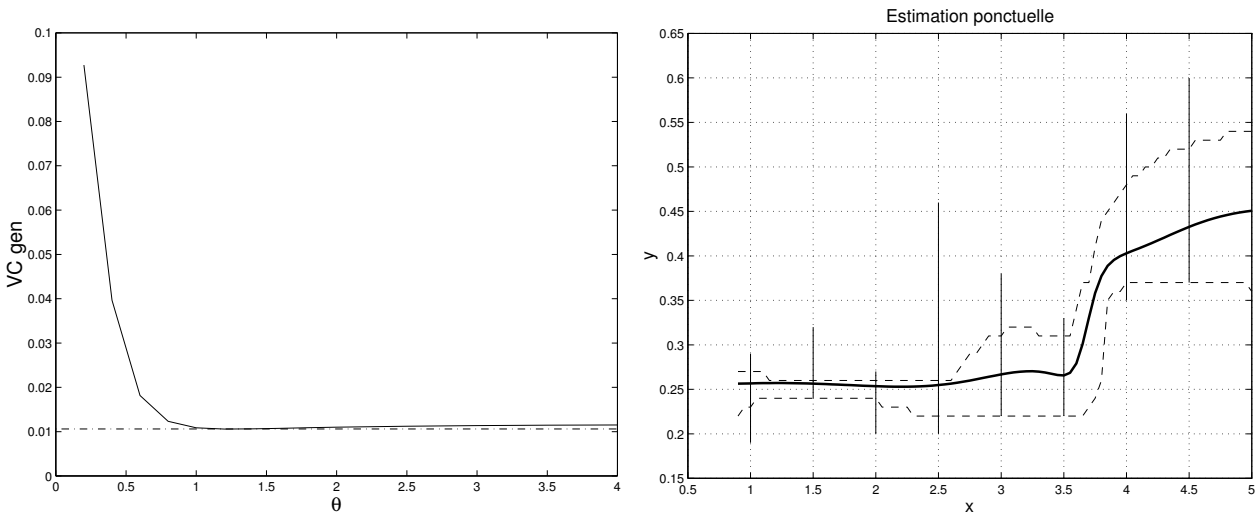


FIG. 5.6 – **Exemple 2 :** *ensemble d'apprentissage* (–) *et estimation ponctuelle* (–) ; distribution pignistique p_{bet} (- -).

Description des résultats

Les données étant peu nombreuses et ne présentant pas individuellement de conflit, nous avons utilisé la version SCF2 de notre méthode.

Identification

Chaque x_i est associé à la fonction d'activation exponentielle de paramètre θ:

$$\phi_i(\|x - x_i\|) = \text{Gauss}(x, x_i, \theta).$$

Dans cette méthode SCF2, l'identification se résume à l'estimation d'un seul paramètre, l'écart-type θ. La figure 5.6 présente l'évolution du critère de validation croisée généralisée CV (équation 4.28) pour différentes valeurs de θ. La valeur minimale est obtenue pour $\hat{\theta} = 1.2$ mais les valeurs comprises entre 1 et 4 semblent tout-à-fait correctes. Nous présentons également l'estimation ponctuelle obtenue avec le paramètre $\hat{\theta} = 1.2$.

Information

De manière générale, les mesures d'information rendent bien compte de la situation : la nonspécifité augmente pour les grandes valeurs de x et le conflit est faible, sauf pour les valeurs de x situées entre 3.5 et 4, ce qui est conforme aux données d'apprentissage (cf. figure 5.7). Nous avons représenté uniquement la mesure de « strife ». Les autres mesures de conflit donnent des résultats semblables.

Estimation par structure de croyance

La figure 5.8 présente la sortie sous la forme la plus générale, la structure de croyance non normalisée, pour une valeur d'entrée particulière : $x = 0.9$. Comme nous l'avons remarqué

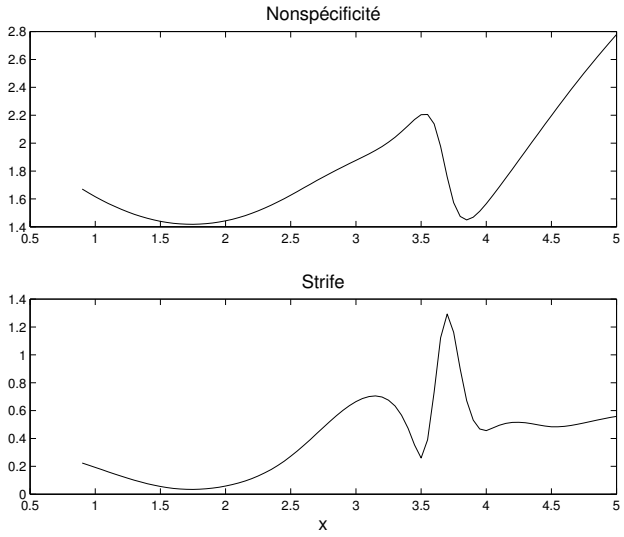


FIG. 5.7 – **Exemple 2 :** *Mesures d'information :* **en haut** *: nonspécificité ;* **en bas** *: strife.*

Critère d'arrêt	∞	10^8	10^6	10^5	10^3	100	10	0
$\lvert F(m(.\lvert x))\rvert$	20	18	15	12	8	4	2	1

TAB. 5.3 – *Nombre d'éléments focaux selon le critère d'arrêt δ pour $x = 0.9$*

dans le chapitre 4, la probabilité pignistique est constante par morceaux, discontinue en certaines extrémités des Y_i.

Sans aucune simplification, le nombre d'éléments focaux est de 20. C'est le même nombre quel que soit x. Il correspond à l'ensemble de toutes les intersections possibles entre les 9 éléments focaux initiaux : Y_i et $\mathcal{Y}$. Il est beaucoup plus faible que le maximum possible : 2^8 éléments.

Néanmoins, si les calculs sont rapides, la sortie est peu lisible. On peut donc simplifier le résultat, par exemple, par la probabilité pignistique, qui est maximale pour $y \in [0.2, 0.3]$. On peut aussi utiliser les plus proches voisins. Si on n'utilise que 3 voisins, la sortie est voisine, la qualité de la sortie ne s'altère pas (figure 5.9). Dans ce cas, on n'obtient plus que 6 éléments focaux. La figure 5.9 montre également le nombre moyen d'éléments focaux selon le nombre de voisins.

On peut enfin simplifier directement la structure finale en réalisant une classification des 20 éléments focaux. Nous avons utilisé le critère I défini par l'équation 5.6. Les figures 5.10 et 5.11 représentent la structure de croyance, toujours pour $x = 0.9$, pour différents critères d'arrêt : $\delta = 10^5, 1000, 100$ et 0. La sortie est de plus en plus lisible : le nombre d'éléments focaux correspondant à ces valeurs est respectivement de $12, 7, 3$ et 1. Le tableau 5.3 présente quelques détails sur l'influence de ce critère sur le nombre d'éléments focaux.

Représentation par nombres flous

Afin d'éviter d'obtenir une probabilité pignistique discontinue, phénomène dû à la représentation des données par intervalles, on peut également utiliser une représentation par des nombres flous. On peut par exemple définir les sorties comme des nombres flous triangulaires

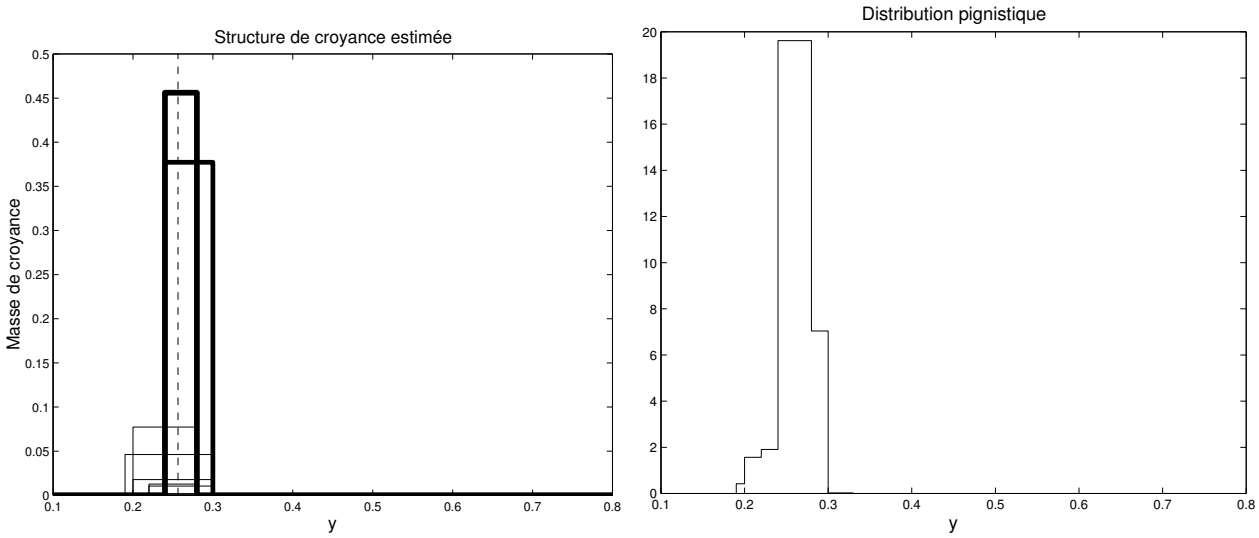


FIG. 5.8 – **Exemple 2 :** *structure de croyance* $\widehat{m}(.|x)$ *et probabilité pignistique* $p_{bet}(.|x)$ *pour* $x = 0.9$

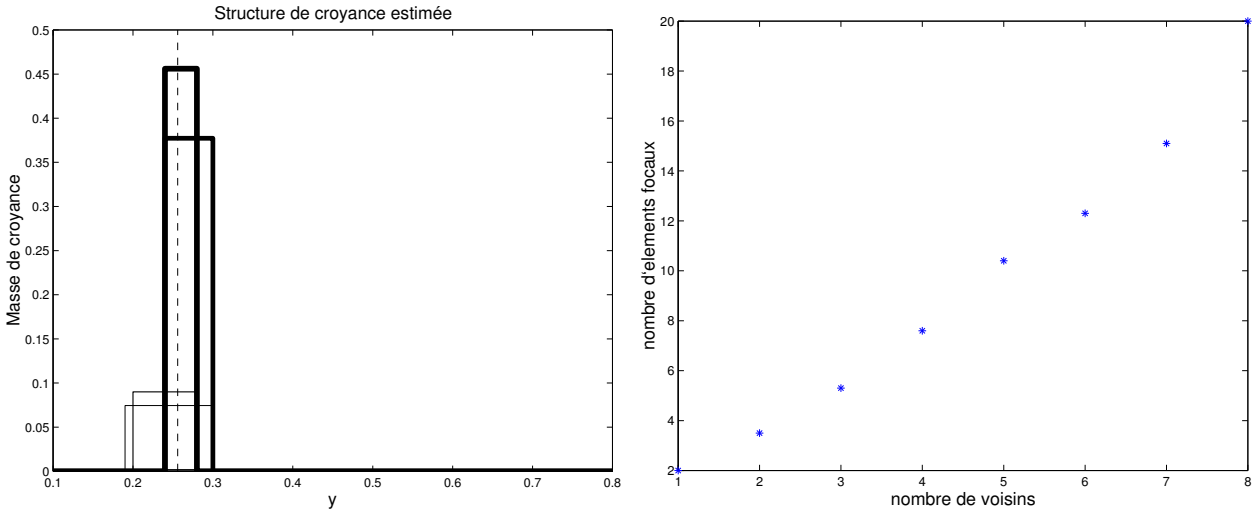


FIG. 5.9 – **Exemple 2 :** *structure de croyance avec* $k = 3$ *voisins ; nombre moyen d'éléments focaux en fonction du nombre de voisins utilisés pour la combinaison.*

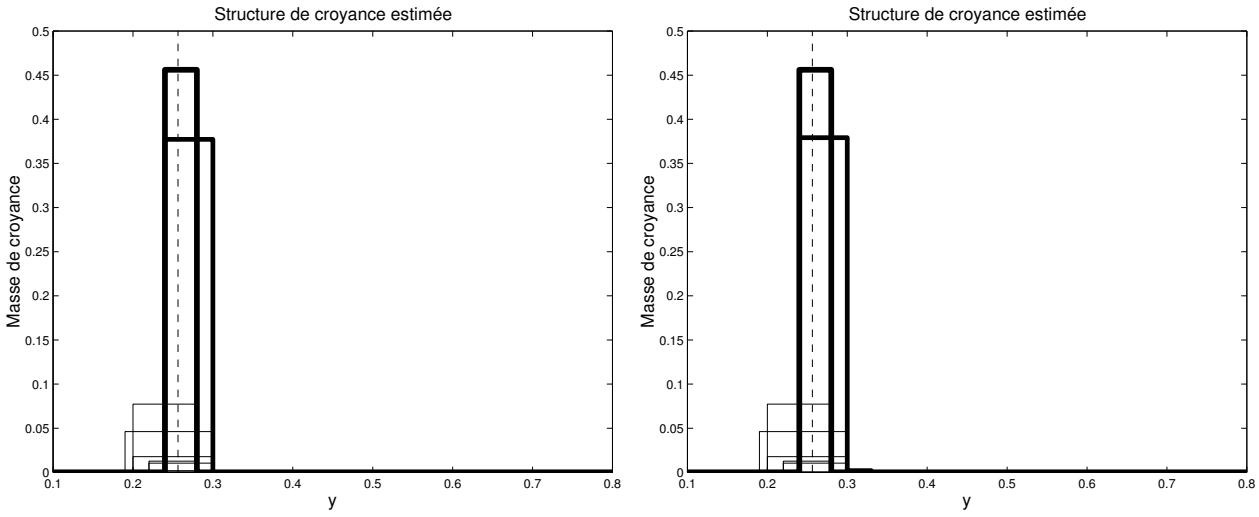


Fig. 5.10 – **Exemple 2 :** *structure de croyance simplifiée après la combinaison : selon le critère d'arrêt* $\delta = 10^5$ *et* 1000.

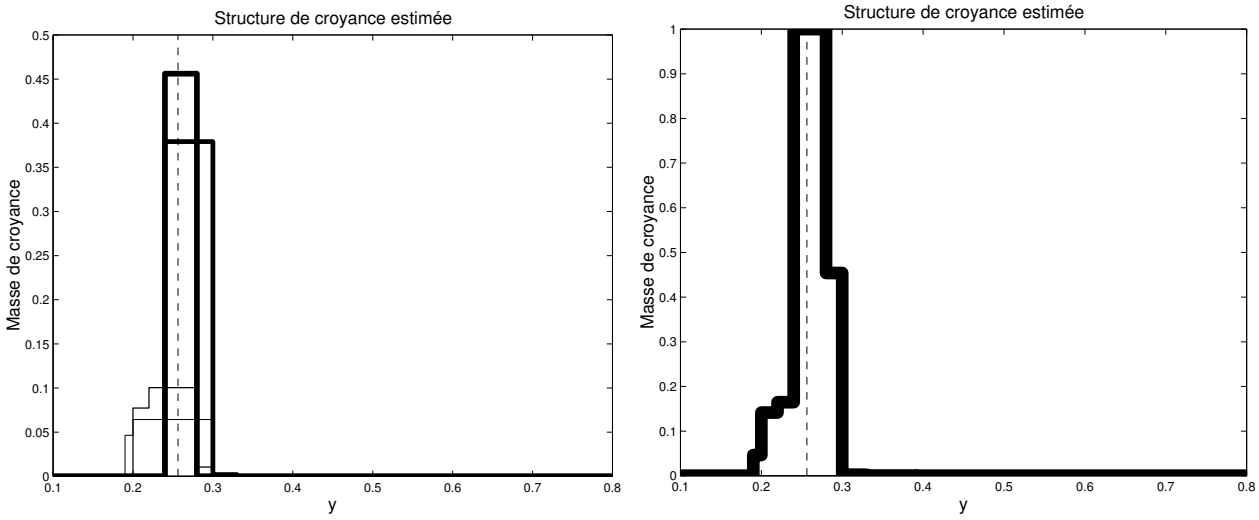


Fig. 5.11 – **Exemple 2 :** *structure de croyance simplifiée après la combinaison : selon le critère d'arrêt* $\delta = 100$ *et* 0.

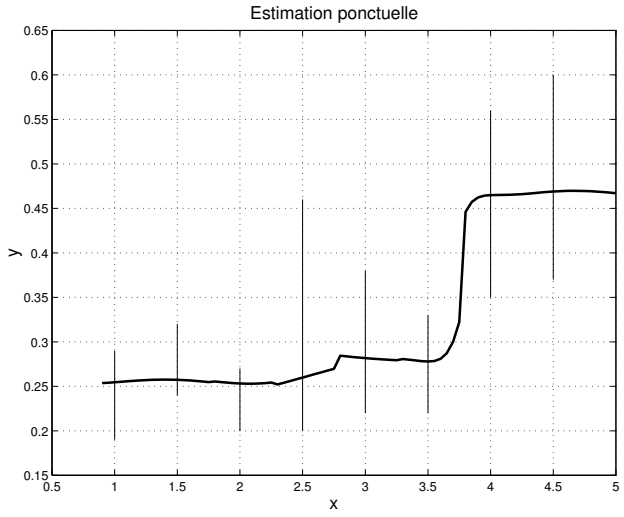


FIG. 5.12 – **Exemple 2 :** *Données d'apprentissage floues : estimation ponctuelle, SCF (3 voisins)*

T_i de support Y_i et centrés sur le milieu de l'intervalle : $T_i = \mathrm{Tri}(y_i^-, \frac{y_i^- + y_i^+}{2}, y_i^+)$.

Les résultats sont heureusement analogues aux précédents. Ils ont été obtenus à l'aide des 3 plus proches voisins. La figure 5.12 représente la sortie ponctuelle, très proche de l'exemple précédent. La figure 5.13 représente la structure et la probabilité pignistique point $x = 0.9$, comme précédemment.

Conclusion

Ce type d'exemple est particulièrement adapté à notre modéle. La connaissance imprécise de la variable à expliquer est traduite tout naturellement sous la forme d'un intervalle ou d'un ensemble flou. Notre méthode est conçue pour traiter ce type de données. De plus, la phase d'apprentissage est très rapide et ne nécessite que l'estimation d'un seul paramètre. Le petit nombre de données permet en effet d'éviter la phase préliminaire de classification. L'avantage de notre méthode est de pouvoir mettre en valeur le peu d'information disponible. Les méthodes probabilistes ne sont pas adaptées à ce type de données.

5.4.4 Exemple 3 : Problème inverse

Dans cet exemple, la variable x est générée par une fonction simple de la variable y :

$$x = y + 0.3\sin(2\pi y) + \varepsilon, \quad y \in [0, 1].$$

où ε est une variable aléatoire de loi uniforme sur $[-0.1, 0.1]$. Le problème qui nous intéresse est de reconstruire y à partir de x, ce qui est délicat, car y est une fonction multi-valuée pour certaines valeurs de x. La taille de l'ensemble d'apprentissage est de $N = 250$. Nous avons partitionné notre ensemble d'apprentissage en 20 classes (cf. figure 5.14).

Les résultats montrent que l'intervalle interdécile de la probabilité pignistique est beaucoup plus important pour les valeurs de x de l'intervalle $[0.3; 0.6]$ où la variance de y est très élevée. Ce résultat est la conséquence de la multimodalité des sorties correspondant aux points de

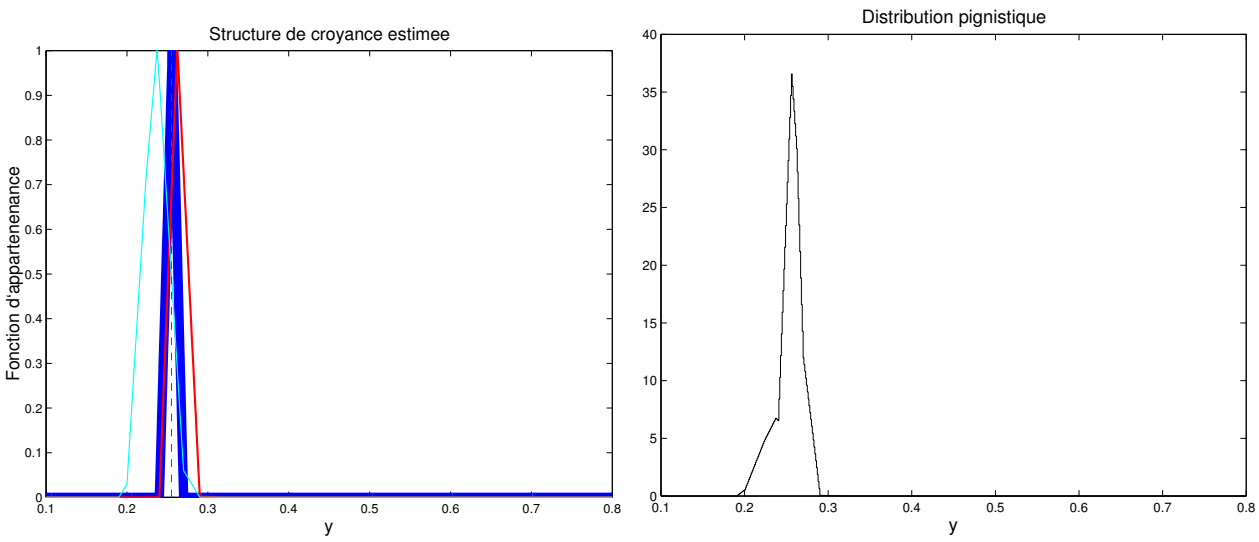


Fig. 5.13 – **Exemple 2 :** *Données d'apprentissage floues : estimation ponctuelle (3 voisins)*

cette région. En particulier, pour les points situés au voisinage de $x = 0.5$, la probabilité pignistique est trimodale (cf. figure 5.14, à droite). Les résultats que nous obtenons sont conformes aux prévisions. L'erreur d'estimation ponctuelle des données pour lesquelles il n'y a pas d'ambiguité est très faible.

Cet exemple a été traité par Bishop [12] à l'aide d'un modèle de mélange, construit à l'aide d'un réseau de neurones. Il est difficile de comparer quantitativement les deux méthodes : nous obtenons des résultats qualitatifs équivalents. Pour ce type de problème, la sortie ponctuelle n'a pas d'intérêt. Dans le cas probabiliste, l'information significative est contenue dans la probabilité conditionnelle. De même, dans notre méthode, l'information importante est contenue, soit directement dans la structure de croyance, soit au niveau de la probabilité pignistique.

5.4.5 Exemple 4 : Niveau de mercure dans des poissons

Description des données

Cet exemple provient d'une base de données réelles multi-dimensionnelles. On cherche à estimer la contamination en mercure dans 53 lacs de Floride. Pour cela, on a mesuré la concentration en mercure dans les tissus musculaires d'un échantillon de poissons de chaque lac. Pour chaque lac, on ne possède que la valeur minimale y_i^-, la moyenne $\bar{y}_i$ et la valeur maximale y_i^+ de cette concentration. On dispose de 4 variables explicatives $x_1, \ldots x_4$: le pH et les taux de calcium, d'alcalinité et de chlorophylle. Les relations entre les variables sont représentées sur la figure 5.15.

Comme dans l'exemple 2, la version SCF 2 de notre méthode semble appropriée pour traiter ce type d'exemple. La construction des intervalles ou des ensembles flous est faite de façon analogue à l'exemple 2. Nous avons choisi de représenter la connaissance m_i relative à la sortie de $\boldsymbol{x}_i$ par l'ensemble flou triangulaire $T_i = (y_i^-, \bar{y}_i, y_i^+)$.

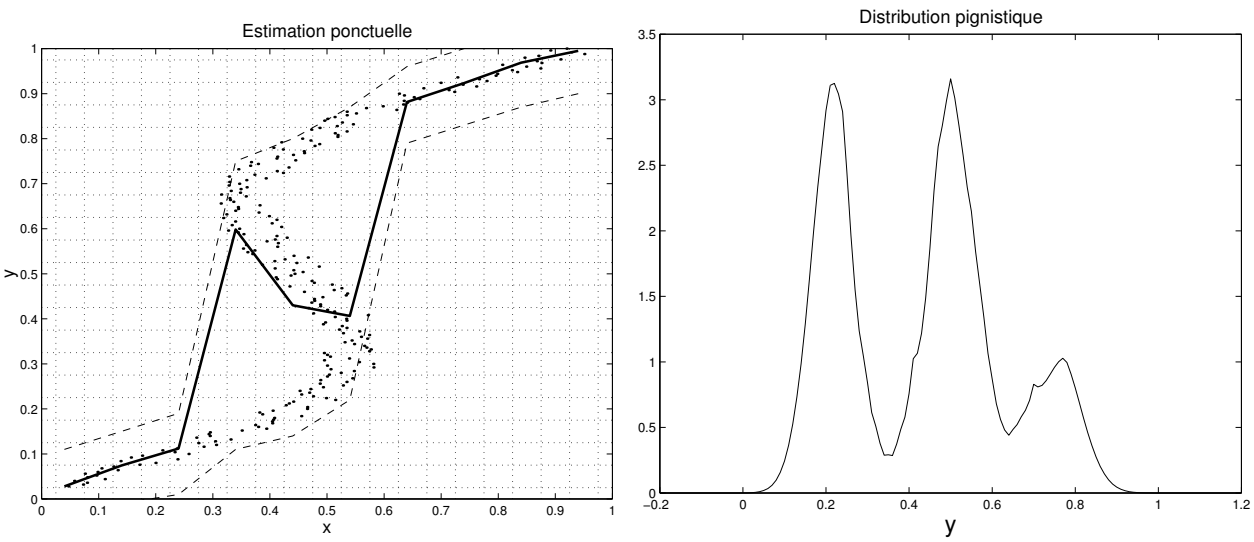


FIG. 5.14 – **Exemple 3 : A gauche**. *ensemble d'apprentissage (..) ; estimation ponctuelle (−, en gras); 1er et 9e décile de p_{bet} (- -).* **A droite**. *Distribution pignistique p_{bet} en $x = 0.5$. Cette distribution est trimodale.*

Résultats

Afin d'éviter la multiplication des paramètres, nous avons normalisé les données et nous avons pris le même écart-type pour toutes les fonctions d'activation. Il n'y a donc plus qu'un seul paramètre θ à estimer. En utilisant le critère CV, on obtient l'estimation $\widehat{\theta} = 5$.

Une fois le paramètre $\widehat{\theta}$ identifié, le traitement est le même que dans le cas monodimensionnel. Nous avons utilisé 33 lacs pour l'apprentissage et 20 pour le test. Nous avons employé une méthode de simplification par les s plus proches voisins pour la combinaison des masses individuelles. La figure 5.16, à gauche, présente les résultats ponctuels ainsi que l'intervalle interdécile des sorties, pour $s = 5$. L'estimation n'est pas très précise, car les données ne sont pas très bien corrélées. Nous avons comparé l'influence du nombre de voisins s dans la combinaison des masses. Les figures 5.17 et 5.18 montrent les résultats d'un lac particulier en utilisant respectivement 2 et 5 voisins. Sans faire une étude exhaustive, on peut remarquer sur cet exemple que ceux-ci sont assez différents. La sortie $m(.|\boldsymbol{x})$ utilisant 5 voisins comportant une vingtaine d'éléments focaux, nous avons appliqué notre algorithme de simplification afin de rendre cette sortie plus lisible, avec le critère d'agrégation $\delta = 1$. On obtient une structure $\widehat{m}'(.|\boldsymbol{x})$ très simplifiée, ne possédant que deux éléments focaux. Dans la figure 5.16 à droite, nous avons comparé la valeur du critère d'erreur, pour chacun des individus testés, avec ou sans simplification. On s'aperçoit que les 2 valeurs sont très proches. L'information perdue par la procédure de simplification est donc négligeable. En revanche, on observe que les critères d'erreur sont individuellement très différents dans le cas de 2 ou 5 voisins. Sur l'ensemble des 20 lacs testés, l'erreur moyenne est à peu près la même pour 2 ou 5 voisins : $C \simeq 0.07$. On peut conclure ici qu'un petit nombre de voisins est suffisant pour la détermination de la structure finale.

Mais il faudrait faire une étude plus systématique de l'influence de ce paramètre pour en déduire des remarques générales. Ceci est d'ailleurs valable pour l'ensemble des paramètres intervenant dans le modèle. Nous avons simplement voulu dans cette section illustrer notre méthode par quelques exemples simples, mais il est clair que des conclusions plus poussées

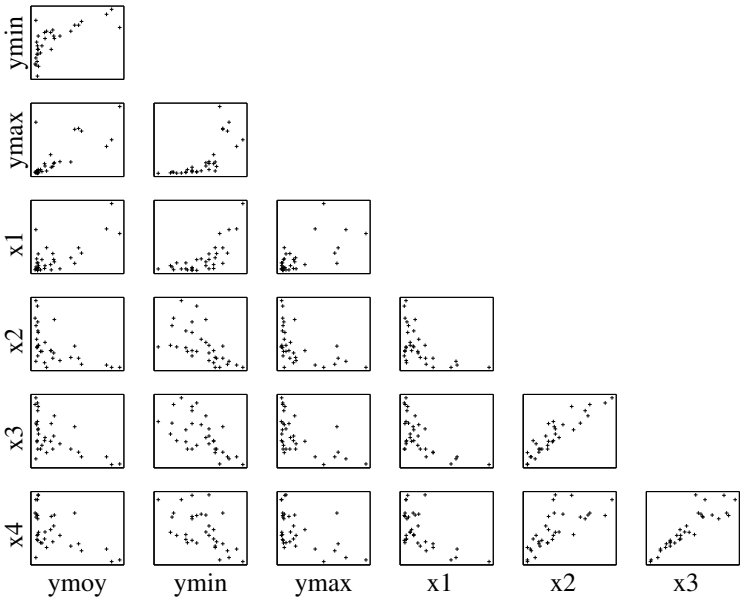


Fig. 5.15 – **Exemple 4**. *relations entre les variables prises 2 à 2.*

exigent l'étude de nouvelles applications.

5.5 Conclusion

Dans ce chapitre, nous avons présenté diverses méthodes de simplification de structures de croyance, afin de pallier l'augmentation indésirable du nombre d'éléments focaux pendant la phase de combinaison de notre modèle. Deux groupes de techniques ont été envisagées : les méthodes basées sur la réduction du nombre de structures à combiner et celles basées sur la diminution du nombre d'éléments focaux par structure.

Les premières méthodes ont pour but de sélectionner l'information pertinente fournie par les éléments de l'ensemble d'apprentissage, pour un vecteur $\boldsymbol{x}$ dont on cherche à estimer la sortie. Nous avons vu que notre méthode présentait des similitudes avec les méthodes de noyau. La distance aux éléments de l'ensemble d'apprentissage est primordiale. Bien souvent, la combinaison de structures issues des plus proches voisins de $\boldsymbol{x}$ suffit pour obtenir de bons résultats, même si nous ne l'avons pas montré formellement. Il est également possible de procéder à une classification globale de l'ensemble d'apprentissage et de construire des structures de croyance à partir de distances entre ces classes et le vecteur $\boldsymbol{x}$. Cette méthode suppose l'estimation des paramètres des classes.

Le deuxième groupe de méthodes repose sur l'agrégation d'éléments focaux similaires dans une même structure. La motivation de cette simplification est double : *pendant* la combinaison, le temps de calcul est plus réduit, *après* la combinaison, les résultats de la sortie sont plus facilement interprétables. L'algorithme que nous avons défini est basé sur la classification hiérarchique des éléments focaux d'une structure m. Le critère d'agrégation tient compte de la proximité des éléments focaux, ainsi que de leur contribution à la masse totale de la structure. Nous avons vu que cet algorithme est efficace, mais il reste une heuristique. En effet, il n'optimise pas de critère global entre la structure m et son approximation m'. Nous avons ainsi proposé une série de méthodes plus systématiques, basées sur des critères

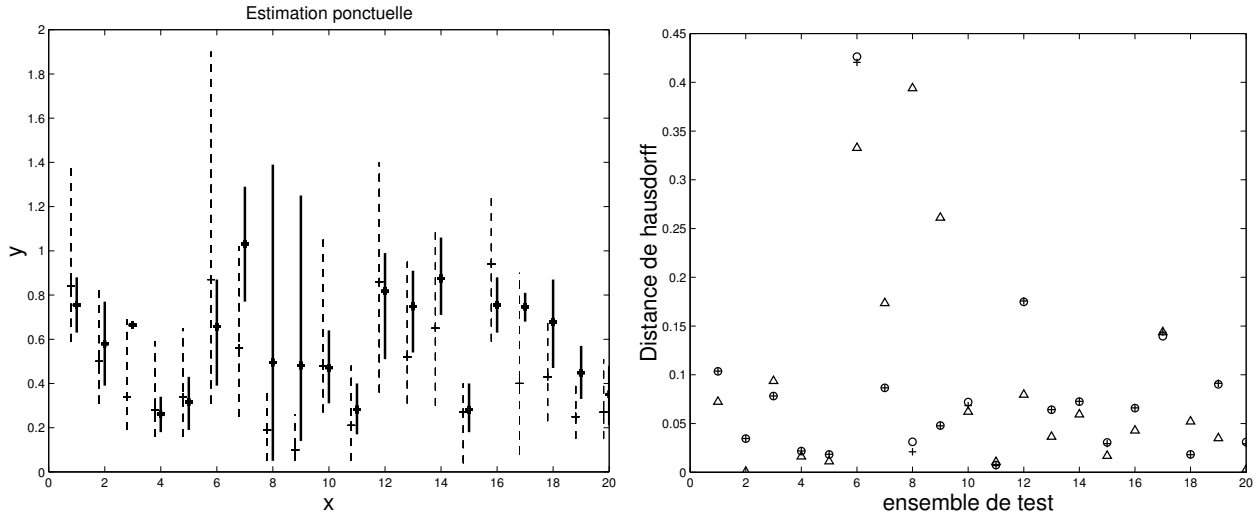


FIG. 5.16 – **Exemple 4** *: Estimation de la sortie à l'aide de 5 voisins.* **A gauche** *: Ensemble de test (- -) ; intervalle interdécile de* $p_{bet(.|\boldsymbol{x})}(-, engras)$ *; estimation ponctuelle (+) ;* **à droite** *: critère d'erreur* $C(m_i, \widehat{m}(.|\boldsymbol{x}_i))$ *pour* $s = 2$ *voisins* (Δ)*; pour* $s = 5$ *voisins, sans simplification du nombre d'éléments focaux (o); pour* $s = 5$ *voisins, avec simplification (+).*

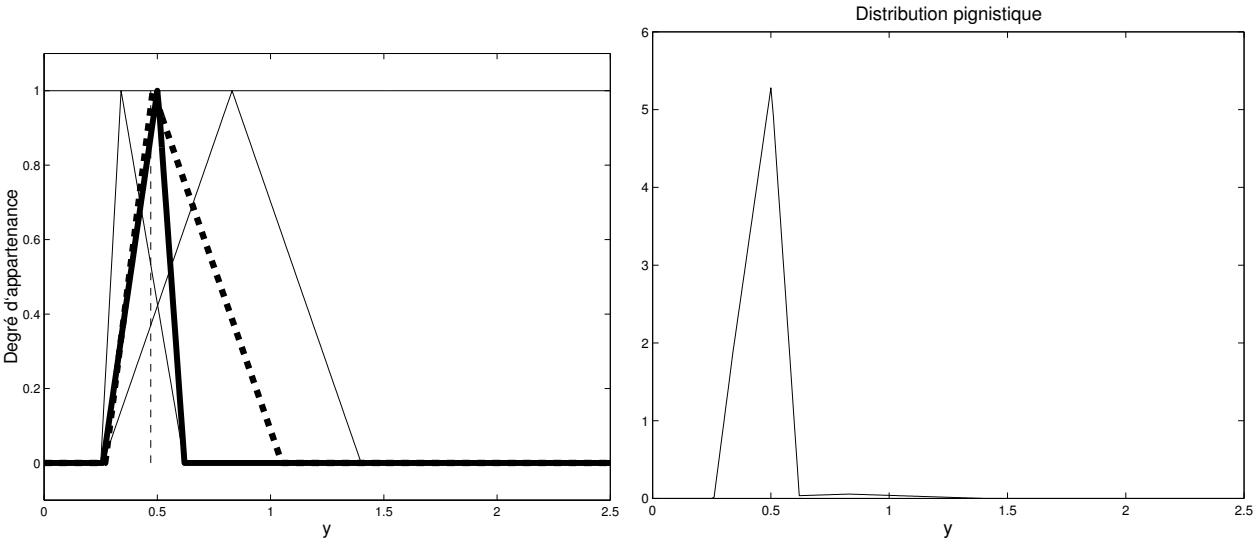


FIG. 5.17 – **Exemple 4** *: Estimation de la sortie à l'aide de 2 voisins.* **A gauche** *: Nombre flou triangulaire (–), sortie estimée* $\widehat{m}(.|\boldsymbol{x})$ *;* **à droite** *: distribution pignistique.*

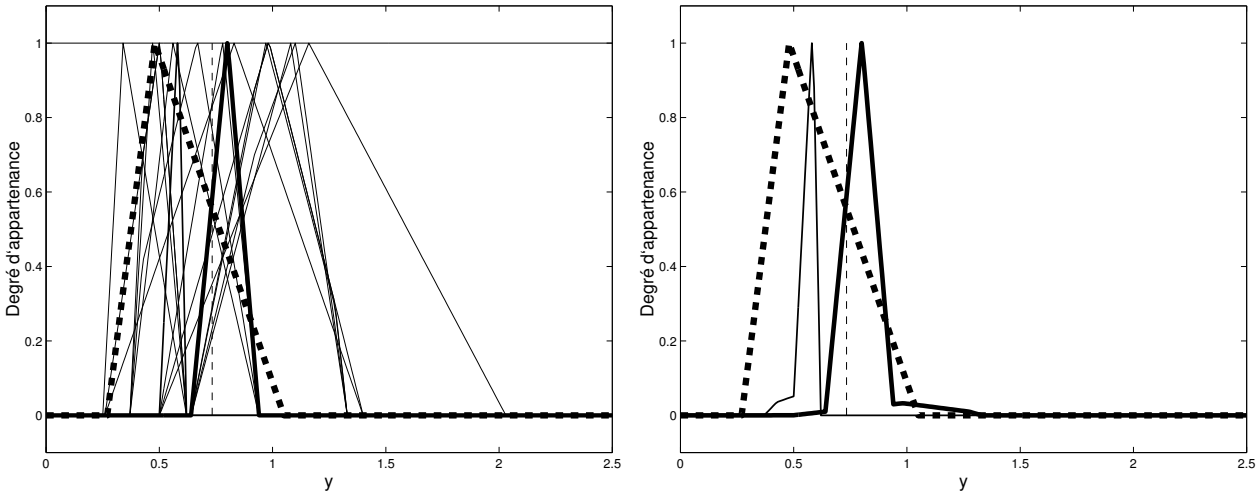


Fig. 5.18 – **Exemple 4 :** *Estimation de la sortie à l'aide de 5 voisins. Nombre flou triangulaire (–), sortie estimée* $\widehat{m}(.|\boldsymbol{x})$. **A gauche** *Sortie complète.* **à droite** *simplification de la structure* ($\delta = 1$)

d'information.

En résumé, les résultats de notre méthode sont les suivants. En termes de précision, nous obtenons des résultats comparables aux techniques statistiques dans le cas de données réelles. Les différentes méthodes de simplication permettent d'obtenir un temps de calcul acceptable. L'identification des différents modèles présentés n'est pas non plus un obstacle au développment de la méthode. Dans les versions SCF1, SCF2 et SCF3, l'identification se résume à l'estimation des paramètres des fonctions d'activation. Bien souvent, on peut se ramener à l'estimation d'un paramètre unique. Les méthodes Proto1 et Proto2 ne nécessitent qu'une phase préalable supplémentaire : la définition des classes, qui est en général assez rapide.

Un des avantages de notre méthode est de pouvoir évaluer avec des critères simples la qualité globale de la sortie, en fonction des informations disponibles. Nous avons montré que notre méthode était capable de faire face à des problèmes de régression très différents, que les données soient imprécises, multi-valuées, ou manquantes. Elle permet d'intégrer des informations *a priori* de nature très diverse.

Cependant, de nombreux points restent à développer ou explorer. Nous avons présenté plusieurs critères de simplification, mais nous ne les avons pas testés. L'identification est un problème crucial. Nous avons construit un critère essentiellement en cherchant une généralisation du critère quadratique classique. Mais nous n'avons pas défini concrètement ce qu'il mesurait. Enfin et surtout, notre méthode demande à être validée davantage par des applications réelles.

Chapitre 6

Conclusion

Synthèse des résultats

Le sujet général de cette thèse concerne l'estimation de variables d'un système dans un contexte où l'information disponible est incertaine et les données de l'ensemble d'apprentissage, sur lesquelles on s'appuie, sont imprécises.

Inadéquation des méthodes probabilistes

L'estimation d'une fonction de régression classique est un problème bien connu, pour lequel de nombreuses solutions ont été apportées, sous la forme de fonctions paramétriques ou non paramétriques à valeurs réelles, à partir d'un modèle probabiliste. Cependant, ces solutions sont généralement inadaptées à certaines situations où les données sont définies de façon imprécises. De plus, la prise en compte d'informations non numériques, comme le jugement d'un expert, ou l'expression de l'ignorance totale sur l'état d'une variable est mal aisée en termes probabilistes. La prise en compte directe d'une information parcellaire a priori n'est abordée que dans les méthodes bayésiennes. Les hypothèses restrictives sur la forme des données et sur la gestion de l'incertitude nous ont amenés très vite à recourir à d'autres théories. Nous avons donc écarté d'emblée la théorie des probabilités, en prenant soin de comparer malgré tout nos résultats, quand cela était possible, à la solution proposée par une ou plusieurs méthodes d'estimation statistique.

Estimation par les systèmes neuro-flous

La première « généralisation » de l'estimation fonctionnelle ou de l'identification de systèmes qui nous a semblé naturelle est celle de l'extension aux ensembles flous, puisque ceux-ci sont un excellent outil de représentation de données imprécises. Les systèmes flous résultant de cette extension sont maintenant bien connus. Parmi les différentes classes de systèmes flous, nous nous sommes particulièrement intéressés aux systèmes adaptatifs, capables d'ajuster finement les paramètres intervenant dans le système : les systèmes neuro-flous. L'engouement actuel pour les systèmes neuro-flous s'explique par la conjonction des avantages des réseaux de neurones, leur capacité d'apprentissage et de généralisation, et de ceux des systèmes flous, capables de traduire par des règles une connaissance issue du monde réel.

Nous avons choisi dans le chapitre 2 de nous placer dans ce cadre des systèmes neuro-flous pour traiter un problème particulier d'estimation : la reconstruction de données manquantes. L'avantage du modèle proposé par rapport aux autres méthodes de reconstruction classiques est de pouvoir estimer toutes les valeurs manquantes d'un vecteur dans un même modèle, quel que soit le nombre de variables disponibles. Le principe repose sur la traduction de la connaissance acquise à partir de données brutes en termes de règles d'inférence floues. Deux versions de la méthode ont été proposées selon la taille de l'échantillon. Si la taille est « réduite », chaque vecteur de l'ensemble d'apprentissage induit une règle d'inférence par la relation entre ses composantes. Si la taille est importante, le principe est le même, mais on partitionne l'espace de représentation des données en un certain nombre de classes. Cette fois, c'est à chaque classe qu'on associe une règle. La solution générale est donnée sous la forme d'une distribution de possibilité, dans son acception la plus large, c'est-à-dire non normalisée, définissant un degré de confiance dans les valeurs de l'espace des variables de sortie. On peut en déduire une estimation ponctuelle dans le cas où elle s'avère « nécessaire et raisonnable ».

Comme tout système neuro-flou, notre modèle présente des analogies avec certaines méthodes de régression. Sur le plan purement fonctionnel, l'expression de la valeur ponctuelle de chaque variable reconstruite, prise individuellement, est voisine de celle d'une méthode de noyau, si la totalité de l'ensemble d'apprentissage est utilisée. Si on passe par une classification préalable de l'ensemble d'apprentissage, son expression est celle d'une fonction de base généralisée, radiale, si on utilise des fonctions d'appartenance gaussiennes. Une analogie importante avec l'approche probabiliste des modèles de mélange a été observée. La définition des fonctions d'appartenance du modèle *a priori*, dont on estime les paramètres par la suite, correspond à la définition des distributions de probabilités *a priori* dans les modèles bayésiens. Le parallèle avec les méthodes probabilistes peut également se faire au niveau de la conception de notre méthode : notre modèle revient à estimer les paramètres de la distribution de possibilité *jointe* de l'ensemble des variables et à en déduire la distribution conditionnelle des valeurs *inconnues* par rapport aux valeurs *connues*.

Estimation « fonctionnelle » et théorie des croyances

Un apprentissage *partiellement* supervisé

L'étude des systèmes flous ne permet pas de traiter tous les types d'incertitude de façon satisfaisante. Les fonctions de croyance permettent de définir des degrés de confiance *a priori* sur les éléments de l'ensemble d'apprentissage. Ainsi, il est possible d'intégrer des informations externes de natures diverses sur une variable, comme la confrontation d'analyses d'experts ou les valeurs de plusieurs capteurs. La synthèse de ces informations définit une série de contraintes sur la valeur de la variable, qui représente l'état de nos connaissances, la *croyance* en cette variable. Nous nous sommes donc placés dans le cas d'un apprentissage partiellement supervisé : la base d'apprentissage est constituée de vecteurs $(\boldsymbol{x}_i, m_i)$ où m_i est la structure de croyance qui représente la connaissance de la variable à estimer y.

Trois types de raisons justifient selon nous l'utilisation de la théorie des croyances, pour l'estimation de variables.

- la gestion des données imprécises (les éléments focaux peuvent être des intervalles ou des ensembles flous) ;

- la représentation claire des différents types d'imperfection dans un objectif ultérieur d'aide à la décision ou de fusion d'informations ;

- l'exploration en soi, de la théorie des croyances en statistique.

L'application de la théorie des croyances nécessite quelques adaptations dans le cas de l'estimation fonctionnelle, car l'espace de référence, en général un intervalle de $\mathbb{R}$, devient continu. La difficulté essentielle provient du caractère non dénombrable de l'ensemble des éléments focaux potentiels. Mais la restriction à un nombre *fini* d'éléments focaux rend la généralisation immédiate, moyennant quelques adaptations simples.

Comme la plupart des méthodes de régression, notre méthode repose sur la proximité entre deux vecteurs observés $\boldsymbol{x}$ et $\boldsymbol{x}_i$. Mais ici, l'information apportée par chaque élément $\boldsymbol{x}_i$ de l'ensemble d'apprentissage est représentée par une structure de croyance, construite en fonction d'un indice de proximité et de l'information *a priori* délivrée par le vecteur $\boldsymbol{x}_i$. Cette information *a priori* m_i peut être un nombre réel, un intervalle, un ensemble flou, ou, plus généralement, une structure de croyance, éventuellement floue. La sortie proposée par notre méthode est exprimée sous la forme très générale d'une structure de croyance.

Notre méthode généralise les modèles probabilistes au sens où elle apporte des solutions correctes, aussi bien dans les cas traités par ces modèles que dans d'autres situations où l'information disponible ne permet pas de les utiliser. L'équivalent de l'intervalle de confiance peut être exprimé à l'aide de la distribution pignistique issue de la structure de sortie. Dans le cas où les données d'apprentissage dégènèrent en nombres réels, notre méthode peut être assimilée à une méthode de noyau, ce qui correspond en effet à l'esprit de la méthode.

En revanche, sur le plan des outils, notre méthode est conçue de manière radicalement différente des méthodes probabilistes. Elle est fondée sur la croyance, le jugement concernant une sortie possible. La définition de l'incertitude prend ici tout son sens. Ainsi, ce qui est important dans notre méthode, c'est que nous ne fournissons pas seulement des valeurs numériques ou intervalles de valeurs possibles, mais également des indications sur différents types d'incertitude, qui pourront être exploités ultérieurement.

Un aspect important de la méthode concerne l'identification du modèle. Nous avons proposé un critère généralisant le critère d'erreur quadratique classique, en utilisant trois extensions successives : une distance entre intervalles (la distance de Hausdorff), sa généralisation aux ensembles flous [183] et une extension naturelle aux structures de croyance, en fonction des poids des éléments focaux.

L'une des difficultés essentielles de notre méthode concerne l'aspect calculatoire. Le dernier chapitre est en partie consacré à la limitation du nombre d'opérations permettant d'obtenir la structure finale. Pour cela, deux types de méthodes ont été envisagés et peuvent être utilisés conjointement. L'un d'eux consiste simplement à résumer l'information par classification de l'ensemble d'apprentissage. L'autre, plus spécifique à notre méthode, consiste à faire une approximation des structures de croyance, basée sur la classification des éléments focaux. Cette deuxième solution nous a amené à définir une série de méthodes : une heuristique efficace, basée sur la similarité entre les ensembles flous et des méthodes systématiques basées sur l'optimisation de critères d'information.

Discussion et développements envisagés

Notre méthode de traitement de données manquantes a été en partie validée par l'application au projet européen EM^2S. Cependant, quelques points peuvent encore être développés. L'un d'entre eux concerne la généralisation de l'optimisation de l'apprentissage du réseau à des données floues. Dans ce cas, on obtient un réseau de neurones fuzzifié. L'apprentissage revient à optimiser un critère d'erreur entre deux ensembles flous. Une autre piste concerne l'étude de l'efficacité de la reconstruction des données pour un traitement ultérieur, par exemple, un problème de classification ou de régression. Ceci est un problème de fond commun à toutes les méthodes de reconstruction de données. Il faudrait, d'une part, étudier les conditions dans lesquelles il est utile de reconstruire un vecteur incomplet, d'autre part, comparer les résultats à d'autres méthodes.

En ce qui concerne les limites actuelles de notre méthode de régression, deux problèmes majeurs se posent : l'aspect calculatoire et le « fléau de la dimensionalité ».

Le problème du temps de calcul a fait l'objet d'une grande partie du dernier chapitre. Nous l'avons considérablement réduit grâce à bon nombre de méthodes d'approximation. Ces approximations visent à éliminer des informations effectivement négligeables : soit les éléments focaux agrégés ont un poids très faible, soit les points $\boldsymbol{x}_i$ sont très éloignés de $\boldsymbol{x}$. Ces heuristiques nous semblent donc raisonnables. Mais nous n'avons pas évalué de façon théorique l'erreur d'approximation commise.

Les méthodes de simplification de structures de croyance représentent en soi une ouverture intéressante. En particulier, il reste à exploiter les critères de simplification optimale. Tous les critères présentés devraient être confrontés à ceux définis dans la littérature [153, 161, 9]. Enfin, ces critères sont basés sur une classification hiérarchique, c'est à dire sur une simplification progressive. On pourrait également envisager d'utiliser le deuxième type de classification, à savoir, les méthodes de partitionnement.

En revanche, nous sommes conscients que la dimensionalité pose des problèmes d'estimation. Mais ce problème est commun à la plupart des méthodes de régression classique, en particulier à la méthode des noyaux. On pourrait par exemple étudier les limites du rapport acceptable entre la dimension et la taille de l'ensemble d'apprentissage.

D'autres aspects essentiels méritent d'être approfondis. Le point le plus important et le plus délicat, serait de définir un cadre axiomatique rigoureux qui manque actuellement, définissant la « consistence du modèle » [159], c'est-à-dire, permettant de décider ce qu'est une bonne approximation de la sortie. Le critère que nous avons défini pourrait être une solution possible, mais il doit être justifié ou modifié, selon les propriétés que l'on jugera nécessaires.

L'aspect illustratif est un point également crucial. Il est clair que le modèle demande à être validé davantage par des applications réelles. Nous nous sommes principalement concentrés sur les aspects théoriques du modèle, en vérifiant qu'il était effectivement applicable, mais nous sommes conscients qu'un exemple concret serait le bienvenu. A ce titre, une collaboration avec le laboratoire de recherche PSI de l'INSA de Rouen sur des données environnementales devrait débuter prochainement. Par ailleurs, une mise à disposition imminente sur l'Internet faciliterait la diffusion et les tests sur la méthode.

Ce travail n'est qu'un premier pas ouvrant des perspectives sur l'application de la théorie des croyances en statistique. Une présentation commune des problèmes de régression et de

classification commence à être mise en place. Un travail de fond mérite d'être fait pour harmoniser ces deux aspects du traitement de données et élargir le champ des applications à d'autres aspects.

La multiplication de toutes les théories de l'incertain est une bonne chose en soi, car elle permet l'élaboration de nouvelles techniques et induit une confrontation fructueuse de points de vue différents. Nous avons le sentiment qu'à terme, l'ensemble de ces théories se simplifiera. Nous sommes persuadés que des liens existent entre l'intelligence artificielle et les statistiques et que la frontière ne sera plus dans un avenir proche si rigide qu'elle l'est actuellement.

Annexes

Annexe A

Estimation fonctionnelle classique

Dans cette annexe, nous nous plaçons dans le contexte de la régression classique dont nous rappelons quelques résultats. Nous supposons que les variables de notre système sont divisées en deux catégories : l'entrée et la sortie. Sans perte de généralité, nous supposerons la sortie monodimensionnelle et réelle. L'entrée, notée $\boldsymbol{x}$, est un vecteur d'un espace $\mathcal{X} \subset \mathbb{R}^r, r \geq 1$ et la sortie, y, appartient à $\mathcal{Y} \subset \mathbb{R}$. Le but est ici de trouver une relation fonctionnelle entre l'entrée et la sortie, afin de prédire le résultat pour une nouvelle entrée. Nous disposons pour cela d'un échantillon fini d'observations couplées $(\boldsymbol{x}_i, y_i)_{i=1}^n$ formant un ensemble d'apprentissage.

A.1 Définition du modèle

On suppose que l'entrée et la sortie sont des réalisations respectives d'un vecteur aléatoire X de $\mathcal{X}$ et d'une variable aléatoire Y de $\mathcal{Y}$. Ces variables aléatoires sont supposées indépendantes et identiquement distribuées, de densités de probabilité respectives $P_X(\boldsymbol{x})$ et $P_Y(y)$. La source de cet aléa est multiple. Il peut par exemple provenir de la non-mesurabilité des entrées, du caractère stochastique du système lui-même ou d'un bruit dû à la sensibilité de l'instrument de mesure. On peut décomposer le système en deux parties distinctes, une partie fonctionnelle déterministe, qui ne dépend que des variables mesurables, et une partie aléatoire. Le modèle s'écrit alors :

$$y = g(\boldsymbol{x}) + \varepsilon \tag{A.1}$$

où g est une fonction de $\mathcal{X}$ dans $\mathcal{Y}$ et ε est une variable aléatoire centrée. Bien que la frontière soit un peu artificielle, on peut classer les méthodes d'estimation en deux catégories, selon que la forme générale de la fonction g est supposée connue ou non. Dans le premier cas, on cherche la solution dans une famille paramétrée de fonctions : il s'agit de méthodes paramétriques. Dans le deuxième cas, il n'y a aucun *a priori* sur la nature de la solution, les méthodes correspondantes sont dites non-paramétriques.

A.2 Méthodes paramétriques

Les méthodes paramétriques supposent une forme rigide de dépendance entre les entrées et la sortie. Dans ces modèles, la sortie estimée $\widehat{y}$ s'écrit sous la forme

$$\widehat{y} = h(\boldsymbol{x}, \boldsymbol{w}), \quad \boldsymbol{w} \in W$$

où h est une fonction déterminée et $\boldsymbol{w}$, un vecteur de paramètres à estimer parmi un ensemble W.

Quelque soit la forme du modèle choisi, celui n'est acceptable que dans la mesure où il « s'adapte bien » au jeu de données d'apprentissage. Ceci peut se mesurer par un critère d'erreur empirique, par exemple quadratique, entre les données réelles et estimées :

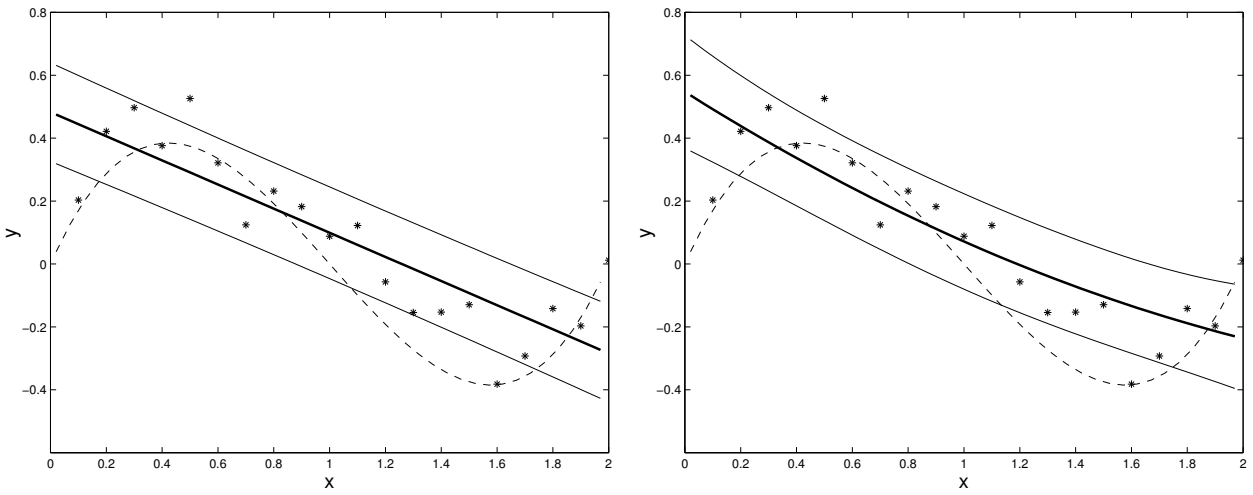
$$J_n(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h(\boldsymbol{x}_i, \boldsymbol{w}))^2. \tag{A.2}$$

La solution optimale $\widehat{\boldsymbol{w}}$ est celle qui minimise ce critère :

$$\widehat{\boldsymbol{w}} = \arg \min_{w \in W} J_n(\boldsymbol{w}).$$

Les paramètres peuvent également être déterminés à l'aide de techniques statistiques plus générales, comme la recherche du maximum de vraisemblance. Dans le cas où l'aléa ε suit une loi gaussienne, cette technique se ramène à la minimisation du critère précédent.

Les formes les plus courantes de modèles sont les modèles linéaires généralisés ou polynomiaux. Dans l'exemple de la figure A.1, les données semblent adaptées à une forme polynomiale d'ordre 3 ou 5, mais certainement pas à une forme linéaire, parabolique, ni à une forme polynomiale d'ordre «élevé». L'ordre du polynôme devient un hyperparamètre que l'on peut régler par une méthode de sélection de modèles (cf. section A.4).



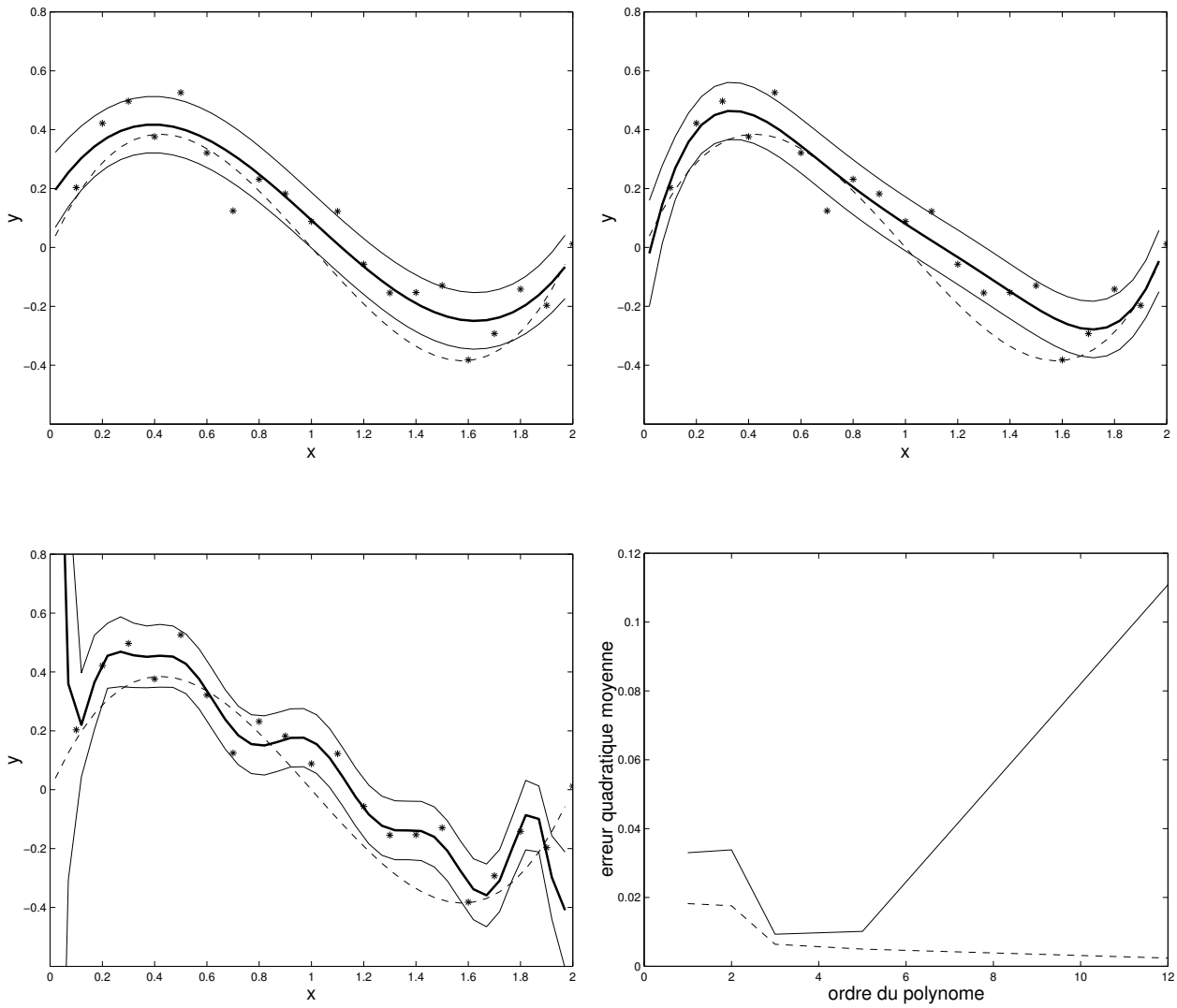


FIG. A.1 – **Régression polynomiale.** *Les cinq premiers graphiques présentent l'ajustement polynomial à un échantillon de 20 exemples. Dans chaque graphique, nous avons représenté la fonction cible g (–), l'échantillon (*) et la fonction d'estimation (-, en gras). Nous avons également représenté des barres d'erreur (-) qui contiennent 50% des prédictions, si ε suit une loi gaussienne. Nous avons utilisé successivement des polynômes d'ordre 1, 2, 3, 5 et 12. Dans chaque cas, le polynôme minimisant l'erreur quadratique moyenne a été sélectionné. Nous remarquons, dans le dernier graphique, que l'erreur définie sur un ensemble de test (-) atteint un minimum pour l'ordre 3 et augmente très vite pour des ordres élevés, pour lesquels il y a sur-apprentissage. Dans le même temps, l'erreur définie sur l'ensemble d'apprentissage (–) continue de décroître.*

On voit bien que dans le cas général, il est difficile de définir un type particulier de modèle adapté à un jeu de données. On peut pour cela avoir recours à d'autres méthodes, dites non-paramétriques ou flexibles. La non-flexibilité est l'inconvénient essentiel des méthodes paramétriques. Nous détaillerons ces méthodes dans les paragraphes suivants.

Certaines méthodes paramétriques permettent également d'éviter les solutions trop rigides, comme les méthodes de *régression par morceaux*. Contrairement aux méthodes classiques, qui ont une nature globale, ce sont des méthodes locales. L'espace des entrées est découpé en

différentes régions séparées par des nœuds. L'exemple le plus classique est celui des *polynômes par morceaux.*

A.3 Méthodes non paramétriques

A.3.1 Principe général

Lorsque la nature de la relation est inconnue, le principe consiste en général à déterminer, parmi toutes les fonctions mesurables, celle qui semble s'adapter le mieux aux données d'apprentissage. Dans cette optique, on peut chercher à évaluer l'erreur moyenne commise sur l'espace $\mathcal{X} \times \mathcal{Y}$. Il faut pour cela définir un critère d'erreur ponctuel entre la vraie valeur et son estimation :

$$C(y, g(\boldsymbol{x})).$$

L'erreur globale, appelée erreur de prédiction est alors :

$$R(g) = \mathbb{E}_{P_{XY}} \left(C(Y, g(X))\right] \tag{A.3}$$

La fonction f recherchée, dite fonction cible, est définie par :

$$f = \arg\min_{g \in \mathcal{M}} R(g) \tag{A.4}$$

où $\mathcal{M}$ est l'ensemble de fonctions mesurables de $\mathcal{X}$ dans $\mathcal{Y}$. Ce formalisme très général peut s'appliquer aussi bien au problème de régression qui nous concerne qu'à ceux de la classification ou de l'estimation d'une densité. Le choix du critère C doit être adapté à la situation. En régression, le critère C le plus couramment utilisé est le critère quadratique :

$$C(y, g(\boldsymbol{x})) = (y - g(\boldsymbol{x}))^2.$$

La solution de l'équation (A.4), appelée *fonction de régression*, est l'espérance conditionnelle de y sachant $\boldsymbol{x}$:

$$f(\boldsymbol{x}) = \mathbb{E}_{P_{Y|X}}(Y|X = \boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}. \tag{A.5}$$

D'autres critères peuvent être utilisés, parmi lesquels le critère des déviations absolues :

$$C(y, g(\boldsymbol{x})) = |y - g(\boldsymbol{x})|$$

Dans ce cas, la fonction cible, solution de (A.4) est la médiane conditionnelle :

$$f(\boldsymbol{x}) = mediane(Y|X = x).$$

Quelque soit le critère C, la connaissance de la densité jointe $P_{XY}(\boldsymbol{x}, y)$ est nécessaire à la détermination de la solution optimale f. Or, cette densité est très rarement disponible, difficile à estimer et la fonction f n'est donc pas accessible. Le problème est alors de déterminer une fonction $\widehat{f}$, estimant f, uniquement à l'aide des données. L'estimation de la densité est

elle-même un problème délicat, mais il est possible d'employer différentes heuristiques « raisonnables ». La plupart des estimateurs sont construits à partir d'une estimation simple de la densité, comme la densité empirique :

$$\hat{P}(\boldsymbol{x}, y) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\{x_i\}}(\boldsymbol{x}) \delta_{\{y_i\}}(y),$$

δ étant la fonction indicatrice.

Dans la suite, nous présentons les estimateurs les plus classiques de la fonction de régression définie par l'équation (A.5) [61, 64, 144, 57]. Bien que d'autres classifications soient possibles, nous regroupons les estimateurs en trois catégories, selon la façon dont on les construit : les estimateurs directs de la fonction de régression, les méthodes basées sur la minimisation du risque empirique, les méthodes de projection.

A.3.2 Estimation directe de la fonction de régression

Méthode des noyaux

Dans cette méthode, proposée simultanément par Nadaraya [110] et Watson [166], la proximité de $\boldsymbol{x}$ et $\boldsymbol{x}_i$ est représentée sous la forme d'un noyau K_σ, qui est une fonction de $\mathbb{R}^r$ dans $\mathbb{R}$, continue, bornée, et telle que

$$\int_{\mathcal{X}} K_\sigma(\boldsymbol{u}) d\boldsymbol{u} = 1,$$

σ étant un paramètre définissant la largeur de bande du noyau. Les noyaux les plus classiques sont la densité de probabilité gaussienne ou le noyau d'Epanechnikov [61], qui minimise asymptotiquement l'erreur quadratique moyenne.

On définit la densité jointe de Parzen et la densité marginale de Parzen respectivement de façon suivante :

$$\widehat{P}(\boldsymbol{x}, y) = \frac{1}{n} \sum_{i=1}^{n} K_\sigma(\boldsymbol{x} - \boldsymbol{x}_i) \delta_{y_i} \text{ et } \widehat{P}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_\sigma(\boldsymbol{x} - \boldsymbol{x}_i).$$

Et on définit naturellement la densité conditionnelle de Parzen de y sachant $\boldsymbol{x}$ par :

$$\widehat{P}(y|\boldsymbol{x}) = \frac{\widehat{P}(\boldsymbol{x}, y)}{\widehat{P}(\boldsymbol{x})}.$$

L'estimateur de Nadaraya-Watson est alors défini en remplaçant $P(y|\boldsymbol{x})$ par son estimation $\widehat{P}(y|\boldsymbol{x})$:

$$\widehat{f}(\boldsymbol{x}) = \int_y y \widehat{P}(y|\boldsymbol{x}) dy = \frac{\sum_{i=1}^{n} K_\sigma(\boldsymbol{x} - \boldsymbol{x}_i) y_i}{\sum_{i=1}^{n} K_\sigma(\boldsymbol{x} - \boldsymbol{x}_i)}. \tag{A.6}$$

Cette expression peut s'écrire sous la forme :

$$\widehat{f}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} H(\boldsymbol{x}, \boldsymbol{x}_i) y_i \tag{A.7}$$

où $H(\boldsymbol{x}, \boldsymbol{x}_i)$ correspond à l'influence relative de $\boldsymbol{x}_i$ sur $\boldsymbol{x}$. L'expression de l'estimateur apparaît comme une combinaison linéaire des y_i dont les poids dépendent de la proximité de $\boldsymbol{x}$ et $\boldsymbol{x}_i$. De nombreuses techniques de régression que nous allons voir dans les paragraphes suivants possèdent cette propriété de linéarité, comme les splines, les méthodes de projection, les fonctions de bases radiales, la méthode des plus proches voisins ou le régressogramme. La fonction $\boldsymbol{x} \to H(\boldsymbol{x}, \boldsymbol{x}_i)$ est appelée *noyau équivalent* de $\widehat{f}$ en $\boldsymbol{x}_i$.

Asymptotiquement, $\widehat{f}(\boldsymbol{x})$ converge en moyenne quadratique, et donc en probabilité vers $f(\boldsymbol{x})$, sous certaines conditions [61]. Le paramètre σ règle l'influence du biais et de la variance de l'estimateur. Une procédure de sélection de modèle est nécessaire pour déterminer une valeur raisonnable de ce paramètre. Si cette valeur est trop grande, la variance de l'estimateur est faible, mais l'estimateur est fortement biaisé. Inversement, si σ est trop petit, l'estimateur est moins biaisé mais la variance est grande.

Dans le cas multi-dimensionnel, le choix d'un paramètre pour chaque dimension donne des résultats plus performants mais nécessite une procédure de sélection plus lourde. La méthode des noyaux n'est pas conseillée pour les «grandes » dimensions ($r > 5$).

Dans la méthode classique, le paramètre σ est le même pour tous les noyaux. Des variantes de la méthode lèvent cette contrainte en permettant de construire des estimateurs ayant des largeurs de bande différentes. La difficulté de ces variantes est qu'elles nécessitent l'identification de nr paramètres.

Le régressogramme, l'un des plus anciens estimateurs non paramétriques, peut être vu comme un régresseur à noyau uniforme. Cet estimateur, obtenu en prenant la moyenne des valeurs de y pour les $\boldsymbol{x}_i$ appartenant à une région fixée de l'espace $\mathcal{X}$ contenant $\boldsymbol{x}$, est une fonction discontinue et n'est de ce fait que rarement utilisé.

Méthodes locales : les plus proches voisins

L'estimateur à noyaux est défini comme une moyenne pondérée des variables de sortie, en général dans un ensemble fixe, indépendant de $\boldsymbol{x}$. La méthode des k-plus proches voisins ne tient compte que d'un voisinage $V(\boldsymbol{x})$ de $\boldsymbol{x}$. Ce voisinage $V(\boldsymbol{x})$ est défini comme l'ensemble des k plus proches vecteurs $\boldsymbol{x}_i$ de $\boldsymbol{x}$ selon une certaine distance $\|.\|$ (distance Euclidienne ou distance de Mahalanobis en général). L'estimateur est alors défini simplement comme la moyenne des y_i sur $V(\boldsymbol{x})$:

$$\widehat{f}(\boldsymbol{x}) = \frac{1}{k} \sum_{\boldsymbol{x}_i \in V(\boldsymbol{x})} y_i \tag{A.8}$$

Cet estimateur, linéaire en y, peut se réécrire sous la forme de l'équation (A.7) avec

$$H(\boldsymbol{x}, \boldsymbol{x}_i) = \frac{n}{k} \text{ si } \boldsymbol{x}_i \in V(\boldsymbol{x}) \text{ et } 0 \text{ sinon.}$$

L'estimateur converge en moyenne quadratique vers $f(\boldsymbol{x})$, au taux optimal [145] de $n^{-4/5}$ si l'entrée est monodimensionnelle [88]. De nombreuses extensions ont été proposées dans la littérature, comme les techniques de régression locale pondérée [6, 49], parfois appelées méthodes à base de mémoire ou *lazy* [5, 10]. La plus largement répandue est celle de Cleveland [25], connue sous le nom de LOWESS. L'idée de ces méthodes est de combiner les avantages qu'offrent la méthode des noyaux et celle des plus proches voisins. Au lieu de

minimiser le critère quadratique empirique classique, comme par exemple dans les méthodes paramétriques, on tient compte de la proximité des vecteurs d'apprentissage en minimisant un critère quadratique pondéré par une fonction noyau K_σ:

$$J_n(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 K_\sigma(\boldsymbol{x} - \boldsymbol{x}_i). \tag{A.9}$$

Il s'agit d'une méthode locale, différente de la méthode des noyaux, dans la mesure où les poids $\widehat{\boldsymbol{w}}$ minimisant le critère précédent dépendent de l'entrée $\boldsymbol{x}$ à étudier. Cependant, Lejeune [94] et Müller [106] ont montré l'équivalence des méthodes de noyaux et de régression locale pondérée sous certaines conditions.

A.3.3 Minimisation du risque empirique

Au lieu de déterminer directement un estimateur de la fonction de régression, on revient à l'expression de l'erreur de prédiction (équation A.3) et l'on estime très simplement la densité jointe P_{XY} par la densité empirique. Le critère à minimiser, le coût empirique classique, est alors de nouveau l'erreur quadratique moyenne:

$$J_n(g) = \frac{1}{n} \sum_{i=1}^{n} (y_i - g(\boldsymbol{x}_i))^2.$$

Cependant, il existe une infinité de fonctions mesurables g, interpolant les données d'apprentissage $(\boldsymbol{x}_i, y_i)_{i=1}^n$, dont le coût empirique est nul. Ce problème de minimisation est alors *mal posé au sens de Hadamard* [160, 55]. Les techniques utilisées restreignent le champ des solutions en imposant des contraintes sur la forme de la solution, comme pour des perceptrons multi-couches, ou sur ses dérivées, comme pour les splines.

Perceptrons multi-couches

On se limite ici aux perceptrons à une couche cachée. La solution s'écrit sous la forme suivante:

$$\widehat{f}(\boldsymbol{x}) = \sum_{j=1}^{J} \theta_j \phi(\boldsymbol{w}_j^t \boldsymbol{x} - \boldsymbol{c}_j) + \theta_0, \tag{A.10}$$

où ϕ est une fonction dite d'activation, en général, la fonction logistique ou la fonction tangente hyperbolique, et $\boldsymbol{w}_j, \theta_j, \theta_0, \boldsymbol{c}_j$ sont des paramètres à ajuster en minimisant le coût empirique.

Contrairement aux techniques précédentes, les réseaux de neurones sont des estimateurs non linéaires par rapport aux y_i et nécessitent de ce fait des techniques d'optimisation non linéaires, comme la méthode de Newton ou ses variantes, comme la descente du gradient. L'algorithme itératif de rétroprogation du gradient, qui permet de calculer successivement les poids des différentes couches, a largement contribué à l'efficacité de ces méthodes.

Les réseaux de neurones sont des approximateurs universels [70], c'est-à-dire capables d'approcher indéfiniment toute fonction continue[1]. La principale difficulté de ces techniques est de définir la structure la mieux adaptée aux données, c'est-à-dire, la taille J du réseau. Ceci peut être réalisé à l'aide de méthodes de sélection et de validation de modèle (cf. section A.4).

Les estimateurs complexes présentent une grande variance, et donc une erreur de prédiction importante. Une des façons de contrôler la complexité du modèle est de pénaliser les structures faisant intervenir des poids élevés. Le critère à minimiser devient :

$$J_{pen}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(\boldsymbol{x}_i, \boldsymbol{w}))^2 + \lambda\|\boldsymbol{w}\|^2, \tag{A.11}$$

où λ est un paramètre de régularisation et $\|\boldsymbol{w}\|$ le carré de la norme de $\boldsymbol{w}$. Ces méthodes ont une interprétation bayésienne [12]. Le terme de pénalisation correspond à une distribution *a priori* sur les poids.

De nombreuses techniques similaires aux perceptrons multi-couches ont été développées dans la littérature. C'est en particulier le cas des projections révélatrices [51]. Il s'agit de réseaux à une couche cachée dont l'expression est la suivante :

$$\widehat{f}(\boldsymbol{x}) = \sum_{j=1}^{J}\theta_j\phi_j(\boldsymbol{w}_j^t\boldsymbol{x} - \boldsymbol{c}_j) + \theta_0. \tag{A.12}$$

Les fonctions d'activation non linéaires ϕ_j sont ici différentes les unes des autres et sont déterminées à partir des données pendant la procédure d'apprentissage. Les poids de chacune des couches sont optimisés de manière indépendante. Les projections révélatrices sont ainsi un moyen d'introduire une connaissance *a priori* sur la nature du problème.

Splines de lissage

Dans le cas des splines de lissage [58, 162], on évite le problème de l'interpolation en ajoutant au coût empirique un terme pénalisant les grandes variations locales.

On suppose ici que la variable x est mono-dimensionnelle. Une généralisation à deux ou plusieurs dimensions est possible mais requiert des coûts de calcul beaucoup plus importants (en $O(n^3)$ pour $r = 2$ contre $O(n)$ pour $r = 1$).

Parmi les différentes façons de quantifier les variations locales, une mesure courante pénalise les grandes valeurs de la dérivée seconde. Le problème est alors de trouver la fonction minimisant le critère suivant :

$$J_{reg}(g) = \frac{1}{n}\sum_{i=1}^{n}(y_i - g(\boldsymbol{x}_i))^2 + \lambda\int_{\mathcal{X}}(g''(\boldsymbol{x}))^2 \tag{A.13}$$

où λ, constante positive fixée, définie comme le paramètre de régularisation joue le rôle de paramètre de lissage. Le problème de minimisation de J_{reg} dans la classe des fonctions deux

1. Soit K un compact de $\mathcal{X}$ et $\mathcal{C} = \mathcal{C}(K, \mathcal{Y})$ l'ensemble des fonctions continues de K dans $\mathcal{X}$. L'ensemble de fonctions $A = \{f : K \longmapsto \mathcal{X}\}$ est un *approximateur universel* si et seulement si, $\forall f \in \mathcal{C}, \forall \epsilon > 0, \exists\, g \in A, \|f(x) - g(x)\| \leq \epsilon, \forall x \in K$.

fois différentiables, définies sur un intervalle de $\mathbb{R}$, a une solution unique, appelée *spline cubique*. Plus le paramètre λ est grand, plus la courbe est « lisse » et plus la variance de l'estimateur est faible.

Fonctions de base radiale

Dans cette approche, l'estimateur est une combinaison linéaire de J fonctions de base G_j non linéaires en $\boldsymbol{x}$, avec $J < n$, chaque G_j étant une fonction noyau particulière, fonction de la distance $\|\boldsymbol{x} - \boldsymbol{c}_j\|$, généralement euclidienne, entre deux vecteurs $\boldsymbol{c}_j$ et $\boldsymbol{x}$.

L'estimateur a donc la forme suivante [105]:

$$\widehat{f}(\boldsymbol{x}) = \sum_{j=1}^{J} \theta_j G_j(\|\boldsymbol{x} - \boldsymbol{c}_j\|) + \theta_0 \tag{A.14}$$

et peut être représenté par un réseau de neurones à une couche cachée (cf. chapitre 2, section 1.2.5). Cette famille de fonctions, comme les perceptrons, est aussi un approximateur universel.

Parmi les différentes fonctions de base considérées, la plus courante est la fonction locale Gaussienne, centrée en l'unité cachée $\boldsymbol{c}_j$:

$$G_j(\|\boldsymbol{x} - \boldsymbol{c}_j\|) = \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{c}_j)^t \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{c}_j)\right) \tag{A.15}$$

où $\boldsymbol{\Sigma}_j$ est une matrice symétrique définie positive.

Un des principaux avantages des fonctions de base radiales, par rapport aux perceptrons, est la possibilité d'optimiser rapidement et de manière indépendante, d'une part, les centres $\boldsymbol{c}_j$ et la matrice $\boldsymbol{\Sigma}_j$ et d'autre part les θ_j. En particulier, les centres peuvent par exemple être déterminés efficacement à l'aide de techniques de classification non supervisée de l'ensemble d'apprentissage. Les poids θ_j sont déterminés indépendemment,une fois les autres paramètres fixés, par minimisation du coût empirique J_n. Il est également possible d'introduire un terme de régularisation comme dans le cas des splines et de minimiser le coût J_{reg} défini par l'équation (A.13) (cf. paragraphe précédent) [55]. Quant à la structure du modèle, le nombre J d'unités cachées peut être déterminé à l'aide de techniques de sélection de modèles (cf. section A.4).

A.3.4 Méthodes de projection

Dans cette famille de méthodes, on suppose que la fonction de régression f peut être représentée par une série de Fourier :

$$f(\boldsymbol{x}) = \sum_{j=0}^{\infty} b_j \phi_j(\boldsymbol{x}) \tag{A.16}$$

où les $(\phi_j)_{j=0}^{\infty}$ forment une base orthonormée d'un espace $\mathcal{F}$ ($\mathcal{L}^2(\mathbb{R}^r)$, par exemple) de dimension infinie et les $(b_j)_{j=0}^{\infty}$ sont les coefficients de Fourier inconnus, définis par

$$b_j = \int_{\mathcal{X}} f(\boldsymbol{x})\phi_j(\boldsymbol{x})d\boldsymbol{x}. \tag{A.17}$$

On peut définir des conditions sous lesquelles une telle décomposition est possible [61]. Les polynômes de Laguerre, de Legendre, de Hermitte ou les transformées en ondelettes orthogonales [4] sont des exemples classiques de fonctions de base ϕ_j.

Puisque l'on ne possède qu'un nombre fini d'observations, il n'est possible d'estimer qu'un nombre fini de coefficients b_j et on ne conserve que J termes dans l'expression (A.16). L'estimateur, dit de série orthogonale, s'écrit alors de façon suivante :

$$\widehat{f}(\boldsymbol{x}) = \sum_{j=0}^{J} \widehat{b}_j \phi_j(\boldsymbol{x}), \tag{A.18}$$

où les $\widehat{b}_j$ sont des estimations des coefficients de Fourier b_j, linéaires en y. Ces estimateurs possèdent alors eux aussi la propriété de linéarité en y.

Le nombre de coefficients J est ici un hyperparamètre à estimer. Il joue le rôle de paramètre de lissage du modèle.

A.4 Sélection de modèles

A.4.1 Le compromis biais-variance

L'erreur de prédiction peut se décomposer en deux termes : l'erreur quadratique moyenne et un terme irréductible ne faisant intervenir que l'innovation du système ε. L'erreur quadratique moyenne de l'estimateur $\widehat{f}(\boldsymbol{x})$ de $f(\boldsymbol{x})$ peut se décomposer elle-même en deux termes antagonistes, la variance et le carré du biais :

$$\mathbb{E}_{P_X}[(\widehat{f}(\boldsymbol{x}) - f(\boldsymbol{x}))^2] = Var_{P_X}(\widehat{f}(\boldsymbol{x})) + [\mathbb{E}_{P_X}(\widehat{f}(\boldsymbol{x})) - f(\boldsymbol{x})]^2.$$

La plupart des méthodes que nous avons vues nécessitent de définir un ou plusieurs paramètres contrôlant ce compromis biais-variance.

A.4.2 Critères d'identification de modèles

En statistique, de nombreux critères ont été développés, souvent dans un contexte paramétrique, afin de mesurer la capacité de généralisation de modèles en utilisant les données d'apprentissage. Les plus usités sont le C_p de Mallows [102], le Critère d'Information d'Akaike (AIC) [2], ou le Critère d'Information Bayésien (BIC) [128]. Ces critères sont constitués de deux termes : le coût empirique et un terme pénalisant la complexité de l'estimateur.

Un deuxième type de critères est mieux adapté aux modèles non-paramétriques : ce sont les critères de rééchantillonnage, comme la validation-croisée et ses variantes, le jackknife ou le bootstrap [48]. Leur principe est basé sur la division de l'échantillon en plusieurs ensembles, qui servent alternativement à l'identification et à la validation de modèles. Nous présentons uniquement une des variantes de la validation croisée, connue sous le nom de « leave one out ».

Soit λ le paramètre de lissage, vectoriel ou momodimensionnel, contrôlant la complexité de la structure. Il peut s'agir aussi bien de la largeur de bande σ définie dans la méthode des noyaux, que du nombre k dans la méthode des plus proches voisins, du nombre J de

fonctions de base dans les méthodes de projection ou de la taille du modèle dans les réseaux de neurones. Pour un certain nombre de valeurs de λ, on effectue la procédure suivante. Le point $(\boldsymbol{x}_i, y_i)$ est retiré de l'échantillon et on estime la variable y en $\boldsymbol{x}_i$ à l'aide des $n-1$ exemples restants. L'estimateur de y_i obtenu étant noté $\widehat{f}_\lambda^{(-i)}(\boldsymbol{x}_i)$, on construit alors le critère de validation croisée suivant :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}_\lambda^{(-i)}(\boldsymbol{x}_i))^2 \tag{A.19}$$

La valeur choisie, $\hat{\lambda}$, est celle qui minimise ce critère.

A.5 Discussion

La diversité des problèmes soulevés dans cette partie rend difficile le choix d'une méthode idéale. Chaque méthode est à utiliser selon le contexte et l'information disponible. Si la taille de l'ensemble d'apprentissage est élevée et que les données sont bien réparties, on aura intérêt à utiliser une méthode statistique classique et peu coûteuse, comme les splines ou la méthode des noyaux. Si la dimension du vecteur d'entrée est importante, on utilisera plutôt une méthode de projection ou une technique connexioniste.

Cependant, ces techniques sont souvent utilisées bien que les données dont on dispose soient souvent issues d'un mélange de lois ou entachées de points aberrants. Dans le cas de données aberrantes, on utilisera de préférence une méthode robuste, comme la régression par la médiane, les estimateurs basés sur des statistiques d'ordre (L-estimateurs, R-estimateurs) ou la classe des M-estimateurs [71].

Différents types de méthodes traitent le problème de l'ambiguité entre plusieurs sorties possibles d'un système, connaissant l'entrée $\boldsymbol{x}$. Les méthodes statistiques visent en général à estimer la *probabilité conditionnelle* $P(y|\boldsymbol{x})$, qui contient une information plus riche qu'une simple valeur ponctuelle ou qu'un intervalle de confiance. Les *modèles de mélange* offrent un cadre approprié pour ce type de problème. Ils combinent les avantages des méthodes paramétriques et non paramétriques, en définissant une large classe de fonctions sans que la taille du modèle, c'est-à-dire le nombre de paramètres à estimer, soit trop importante. La probabilité conditionnelle est représentée comme une combinaison linéaire de fonctions :

$$P(y|\boldsymbol{x}) = \sum_{j=1}^{J}\theta_j(\boldsymbol{x})\phi_j(y|\boldsymbol{x}, \boldsymbol{w}_j)$$

où les $\phi_j(.|\boldsymbol{x}, \boldsymbol{w}_j)$, fonctions paramétrées par un vecteur $\boldsymbol{w}_j$, représentent la densité conditionnelle de y sachant $\boldsymbol{x}$ pour la j^{eme} composante du mélange. Les coefficients $\theta_j(\boldsymbol{x})$ peuvent être vus comme des probabilités *a priori*, connaissant $\boldsymbol{x}$, que la variable y soit générée par la j^{eme} composante du mélange. Ils doivent donc vérifier les contraintes

$$\sum_{j=1}^{J}\theta_j(\boldsymbol{x}) = 1 \text{ et } 0 < \theta_j(\boldsymbol{x}) < 1 \quad \forall j = 1 \ldots J.$$

Les paramètres $\boldsymbol{w}_j, \theta_j$ sont calculés par maximisation de la vraisemblance [12].

Lorsque la densité des points est faible dans une région de l'espace autour d'un point $\boldsymbol{x}$, le manque d'information est souvent indiqué, pour la plupart des méthodes statistiques par une variance importante $Var(\widehat{y}|\boldsymbol{x})$ de la variable expliquée, ce qui est simplement une forme d'incertitude sur la valeur inconnue. L'absence d'information, en tant que telle, en théorie des probabilités, peut se mesurer par des critères sur la probabilité conditionnelle $P(y|\boldsymbol{x})$. Pour ce type de problème, on peut mentionner les méthodes statistiques bayésiennes, qui proposent des solutions tenant compte de l'information globale disponible. Dans l'approche bayésienne, on définit une probabilité *a priori* sur le bruit ε et l'espace des fonctions possibles g définie dans le modèle générique (A.1). L'expression de la distribution conditionnelle $P(y|\boldsymbol{x})$ qui en découle est assez complexe, mais peut être estimée par des méthodes de Monte-Carlo ou d'estimation de la vraisemblance [54].

Annexe B

Estimation de données manquantes

B.1 Introduction

Dans un problème de classification ou de régression, se posent toujours en amont de la chaîne de traitement les phases préliminaires essentielles, comme le choix des variables ou la gestion des données manquantes. Dans cette annexe, nous proposons une synthèse bibliographique décrivant les principales méthodes d'estimation ou de gestion de données manquantes. Dans un tableau de données, l'existence de valeurs manquantes peut être due à des causes très diverses. La donnée peut être indisponible à cause d'un dysfonctionnement de l'appareil de mesure qui la délivre. Dans la collecte de données par sondages, il peut s'agir d'absence de réponses, de réponses contradictoires invalidées par l'analyste. Une absence de réponse peut elle-même refléter deux types de comportement : la donnée est complètement inconnue par le sondé ou au contraire elle provient d'un refus de répondre.

Soit $\{\boldsymbol{x}_1, \ldots \boldsymbol{x}_n\}$, un ensemble de n vecteurs r-dimensionnels, contenant des données observées et manquantes. On représentera par o ou $o(\boldsymbol{x})$, l'ensemble des indices de toutes les composantes observées d'un vecteur $\boldsymbol{x} = \{x_1, \ldots, x_r\}$ donné : $o \subset \{1, \ldots r\}$. De même, m ou $m(\boldsymbol{x})$ est l'ensemble des indices manquants, complémentaires à $\{1, \ldots, r\}$. Ainsi, $m \cup o = \{1, \ldots, r\}$. On note $\boldsymbol{x}^o$ la partie observée de $\boldsymbol{x}$ et $\boldsymbol{x}^m$, sa partie manquante. On suppose que toute observation contient au moins une composante observée et que toute variable est observée pour au moins un individu.

Hypothèses probabilistes

De manière générale, si l'on s'appuie sur des modèles probabilistes, lorsque l'on travaille avec des données partiellement manquantes, on peut décomposer la modélisation en deux parties [97, 53] :

- le mécanisme qui génère les données complètes ;
- le mécanisme qui définit la position des valeurs manquantes.

Soit $\boldsymbol{a} = \{a_1, \ldots, a_r\}$, le vecteur indicatrice de valeurs manquantes, où a_j vaut 1 si la valeur correspondante est manquante et 0 sinon. Les vecteurs $\boldsymbol{x}^o$, $\boldsymbol{x}^m$ et $\boldsymbol{a}$ sont supposés être des

réalisations des vecteurs aléatoires[1] respectifs X^o, X^m et A. La distribution de l'absence A des données, supposée paramétrée par un vecteur $\boldsymbol{\alpha}$, peut s'écrire de manière générale : $P(\boldsymbol{a}) = P(\boldsymbol{a}|\boldsymbol{x};\boldsymbol{\alpha})$. Selon Little et Rubin [97], on peut distinguer trois types de dépendance entre A et X :

- la donnée manquante est complètement aléatoire (*Missing Completely At Random*, MCAR), c'est-à-dire indépendante des données : $P(\boldsymbol{a}|\boldsymbol{x}^o, \boldsymbol{x}^m;\boldsymbol{\alpha}) = P(\boldsymbol{a};\boldsymbol{\alpha})$
- la donnée manquante est aléatoire (*Missing At Random*, MAR). Elle dépend seulement des données observées $\boldsymbol{x}^o$:

  $P(\boldsymbol{a}|\boldsymbol{x}^o, \boldsymbol{x}^m, \boldsymbol{\alpha}) = P(\boldsymbol{a}|\boldsymbol{x}^o;\boldsymbol{\alpha})$
- la donnée manquante est non aléatoire (*Not Missing At Random*, NMAR). Elle dépend également des valeurs manquantes. Les données sont dites *censurées*.

Face à un problème de données manquantes, il y a trois attitudes possibles. On peut éluder la question en ne retenant dans la base de données que les vecteurs complets pour des traitements ultérieurs ; on peut conserver les vecteurs incomplets sans chercher à estimer les valeurs manquantes ; au contraire, on peut estimer les valeurs manquantes à l'aide des informations disponibles. Ces attitudes donnent lieu à trois types de traitements : l'élimination des *vecteurs* incomplets, l'ignorance des *variables* inconnues et l'*estimation* des variables inconnues.

Elimination des vecteurs incomplets

Une technique courante consiste à supprimer tous les vecteurs d'observation $\boldsymbol{x}_i$ incomplets. On obtient alors un tableau de vecteurs complets auquel on peut appliquer directement des traitements et analyses statistiques classiques.

Cette technique simple ne donne des résultats satisfaisants que lorsque les données manquantes sont peu nombreuses et sont du type MCAR. En effet, si l'hypothèse MCAR n'est pas vérifiée, la distribution des données observées X^o est différente de celle de X. En particulier, la moyenne et la variance sont biaisées. De plus, l'un des inconvénients majeurs de cette méthode est la perte d'information qu'elle entraîne. Dans le cas de données multivariées où plusieurs caratéristiques d'un même vecteur sont manquantes, la suppression des données incomplètes, qui peuvent représenter une part importante des données, peut se révéler inefficace et entraîner une grande perte d'information, au sens de Fisher[2].

Ignorance des variables inconnues

Une méthode moins brutale consiste à conserver les vecteurs incomplets et à utiliser toute l'information délivrée par les variables disponibles pour des traitements comme la classification ou la régression.

1. Dans la suite, on notera en lettres capitales les variables ou vecteurs aléatoires et en gras, les réalisations des vecteurs aléatoires.

2. En effet, on peut montrer, d'après le *principe d'information manquante* [29], que l'information de Fisher complète est égale à la somme de l'information obtenue à partir des données observées et celle obtenue à partir des données manquantes.

Ainsi, pour un algorithme de classification non supervisée, l'adaptation aux données manquantes est immédiate. Si on prend l'exemple des cartes auto-organisatrices de Kohonen [86], on peut utiliser la distance aux prototypes sur le sous-espace des données observées $\mathcal{X}^o$ d'un exemple incomplet $\boldsymbol{x}$.

Soient $C = \{c_i(k)\}_{k=1}^K \in \mathcal{X}$ l'ensemble des prototypes (complets) de la carte à la présentation du i^{eme} vecteur de la base d'apprentissage $\boldsymbol{x}_i$ et $V_i(k) \subset C$ le voisinage associés à $c_i(k)$.

Après initialisation des unités $c(k)$ à $c_0(k)$, on itère la procédure suivante pour le vecteur d'apprentissage $\boldsymbol{x}_i$:

- recherche de l'unité gagnante $k^* = \arg\min_{k=1,\dots,K} \|\boldsymbol{x}_i^o - \boldsymbol{c}_i^o(k)\|$
- modification de l'unité gagnante et de ses voisines :

$$\forall k \in V_i(k^*), \quad \boldsymbol{c}_{i+1}^o(k) = \boldsymbol{c}_i^o(k) + \varepsilon_i(\boldsymbol{x}_i^o(k) - \boldsymbol{c}_i^o(k)),$$

où $\boldsymbol{c}_i^o(k)$ est la *projection* de $\boldsymbol{c}_i(k)$ sur l'espace $\mathcal{X}^o$, et ε_i, une suite décroissante bien choisie.

On procède de même pour tout algorithme de classification automatique, comme les centres mobiles, ou les centres mobiles flous.

Estimation

Les méthodes de substitution consistent à remplacer la valeur manquante de la caractéristique x_j d'un vecteur $\boldsymbol{x}$ en l'estimant à partir de ses valeurs connues ou de la valeur de la caractéristique sur l'ensemble des autres vecteurs. De nombreuses approches sont envisageables. On peut les classer en trois catégories :

- les traitements heuristiques : ce sont des règles pratiques, ne reposant sur aucun modèle particulier (substitution par la moyenne, la médiane...)
- les méthodes basées sur un modèle de régression ;
- les méthodes probabilistes, basés sur l'estimation de la densité de la variable à reconstruire ;

Nous allons détailler ces différentes approches dans les trois sections suivantes.

B.2 Traitements heuristiques

B.2.1 Traitement monovariable

De nombreuses techniques ont été développées par des praticiens sous forme de règles heuristiques [26]. L'expérience et certaines simulations ont montré qu'elles donnent d'assez bons résultats. Parmi les méthodes de remplacement les plus usuelles, les moins sophistiquées ne tiennent pas compte des relations entre les variables. Elles consistent à remplacer la valeur manquante x_j par la moyenne $\bar{x}_j$ (ou la médiane) de la j^{eme} variable définie sur les valeurs connues. Cette méthode n'est valable que pour les données MCAR. Elle préserve la moyenne observée, mais sous-estime la variance (d'un facteur de $\frac{n_j-1}{n-1}$, n_j étant le nombre d'exemples complets pour la variable j).

B.2.2 Méthodes connexionnistes

Il est possible de définir des architectures particulières de réseaux de neurones adaptées au traitement des données manquantes.

Certains réseaux reposent sur le codage de la variable d'entrée en *deux ou plusieurs* neurones selon les variantes [158]. Par exemple, la valeur des deux neurones est celle de la variable, x_j, si celle-ci est connue. Si la variable est inconnue, les deux neurones prennent respectivement une valeur « faible » et « élevée » de la variable, ces valeurs étant définies à partir de caractéristiques globales de la variable j, plus ou moins robustes : les valeurs minimale et maximale ; le premier et le troisième quartile ; les quantités $\bar{x}_j - s_j$ et $\bar{x}_j + s_j$, où s_j est l'écart-type estimé de la variable j.

D'autres formes de codages plus évolués ont été proposés, comme les codages flous [107, 104]. Cette fois, chaque donnée en entrée est codée par 3 neurones, mettant en œuvre 3 fonctions d'appartenance d'ensembles flous représentant les variables linguistisques $\{petit, moyen, grand\}$ de la caractéristique continue. Les coefficients de ces fonctions d'appartenance sont réglés de manière à ce qu'une valeur inconnue soit codée par le triplet $\{0.5; 0.5; 0.5\}$ correspondant au niveau d'activation respectif des cellules, ce choix étant justifié par le fait que la valeur 0.5 représente une ambiguïté maximale.

L'inconvénient de ces réseaux à codage est le doublement, voire le triplement des données d'entrée du réseau, et le caractère artificiel de certains codages qui peuvent perturber l'apprentissage.

B.3 Approches par estimation fonctionnelle

D'autres méthodes un peu plus sophistiquées sont basées sur la construction de modèles de régression dont les variables explicatives sont sélectionnées parmi les données connues, du type :

$$x_j = f_j(\boldsymbol{x}^o) + \varepsilon_j, \quad j \in m(\boldsymbol{x}), \qquad \text{avec } \mathbb{E}(\varepsilon_j) = 0. \tag{B.1}$$

Ainsi, pour tout $\boldsymbol{x}$, chaque variable inconnue x_j nécessite la construction et l'identification d'une fonction f_j.

L'estimateur de x_j est donc l'estimateur de la fonction de régression :

$$\widehat{x}_j = \widehat{\mathbb{E}(X_j | X^o = \boldsymbol{x}^o)} \tag{B.2}$$

L'approche la plus rudimentaire consiste à remplacer la valeur manquante par la valeur correspondante d'un individu ayant des caractéristiques semblables, le plus proche voisin au sens d'une distance $\|.\|$. On utilise souvent la notion de distance (de Mahalanobis, euclidienne ou autres) entre un individu et la base d'apprentissage, comme dans les méthodes de type « *hot deck* » [50], fréquente dans l'édition des questionnaires. Si $\boldsymbol{z} = (\boldsymbol{z}^o, \boldsymbol{z}^m)$ est l'individu pour lequel la distance au vecteur de données incomplètes $\boldsymbol{x}$ est minimale sur le sous -espace $\mathcal{X}^o$ des variables connues de $\boldsymbol{x}$, on obtient :

$$\widehat{\boldsymbol{x}^m} = \boldsymbol{z}^m, \qquad \text{avec } \boldsymbol{z} = \arg\max_{\boldsymbol{z}} \|\boldsymbol{x}^o - \boldsymbol{z}^o\| \tag{B.3}$$

On suppose ici pour simplifier que l'ensemble d'apprentissage constitué des vecteurs $\boldsymbol{z}$ est lui-même complet. Dans le cas contraire, on peut par exemple extraire de la base l'ensemble des vecteurs complets.

Dans un esprit de robustesse, on peut généraliser l'exemple précédent du plus proche voisin (B.3), en faisant une combinaison linéaire des valeurs des k plus proches voisins $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(p)}\}$. Si cette combinaison linéaire est pondérée par la distance $\|.\|$ aux voisins, on obtient un régresseur à noyau radial :

$$\widehat{x}_j = \sum_{k=1}^{p} K(\|\boldsymbol{z}^o, \boldsymbol{x}^o\|) z_j^{(k)} \quad \forall j \in m(\boldsymbol{x}).$$

Théoriquement, toute méthode d'approximation fonctionnelle convient (cf. annexe A). Quelle que soit la méthode de régression choisie, la procédure est en général la suivante : on extrait de l'ensemble d'apprentissage X le sous-ensemble des exemples complets $Z = (\boldsymbol{z}_k)$, qui sert à déterminer les paramètres des différentes fonctions de régression. Une variante consiste à remplacer les données manquantes par la moyenne et construire un ensemble de prototypes $C = (\boldsymbol{c}_k)$ représentatifs de l'ensemble d'apprentissage ainsi complété pour construire les fonctions de régression.

En pratique, on se limite à des méthodes simples comme l'imputation par le plus proche voisin (B.3) ou des modèles linéaires. Buck a été le premier à utiliser une méthode de régression linéaire [20] :

$$\widehat{x}_j = \widehat{\mathbb{E}(X_j|\boldsymbol{x}^o)} = \boldsymbol{w}^t \boldsymbol{x}^o$$

où $\boldsymbol{w}$ est l' estimateur des moindres carrés évalué sur la base complète Z ou C.

L'inconvénient majeur de ces méthodes est qu'elles nécessitent autant de modèles que de situations possibles définissant l'équation (B.1), c'est -à- dire :

$$r(2^{r-1} - 1)$$

modèles. En effet, pour chacune des r caractéristiques à reconstruire, il existe $(2^{r-1} - 1)$ combinaison de valeurs observées possibles. De plus, comme tout modèle de régression, elles supposent bien évidemment des corrélations entre ces variables, ce qui est illusoire quand peu de variables sont disponibles. La distribution jointe des données sera difficile à préserver. De plus, l'exploitation des résultats des valeurs estimées doit se faire avec prudence. Une mauvaise estimation peut par exemple conduire à des données manifestement aberrantes, hors du domaine de validité des variables. Il peut être utile de délivrer un indice de confiance sur la valeur proposée, qui peut être une fonction décroissante du nombre de valeurs observées, ou un intervalle de confiance si on obtient la distribution conditionnelle $P(\widehat{x}_j|\boldsymbol{x}^o)$ des valeurs possibles.

Crettaz et al. [26] font un intéressant parallèle entre l'étude des données manquantes et la détection de données aberrantes. Des méthodes robustes sont alors souvent recommandées afin de traiter les deux aspects en même temps.

B.4 Méthodes basées sur l'estimation de la densité conditionnelle

Une excellente description de ces méthodes est proposée dans le livre de Schafer [127].

B.4.1 Méthodes paramétriques basées sur la vraisemblance

La distribution du processus de génération des données, $P(\boldsymbol{x}|\theta)$, est supposée paramétrée par θ. Plusieurs méthodes d'estimation de données incomplètes peuvent se ramener à un problème classique de maximisation de la fonction de vraisemblance sur des données complètes [3, 97].

Sous la condition MAR, Little et Rubin [97, 127] ont montré que toute l'information statistique sur le paramètre θ était contenue dans la vraisemblance des données observées $L(\theta|\boldsymbol{x}^o)$ ou, dans un cadre bayésien, la probabilité *a posteriori* des données observées $P(\theta|\boldsymbol{x}^o)$. En général, ces expressions sont des fonctions complexes de θ, nécessitant l'utilisation de méthodes itératives, comme l'algorithme EM.

L'algorithme EM capitalise la dépendance entre $\boldsymbol{x}^m$ et θ. Le principe est le suivant. Pour un modèle paramétrique quelconque, la distribution des données complètes $\boldsymbol{x}$ peut s'écrire :

$$P(\boldsymbol{x}|\theta) = P(\boldsymbol{x}^o|\theta)P(\boldsymbol{x}^m|\boldsymbol{x}^o,\theta).$$

Si chaque terme est vu comme une fonction de θ, on obtient, en prenant le logarithme :

$$\ln L(\theta|\boldsymbol{x}) = \ln L(\theta|\boldsymbol{x}^o) + \ln P(\boldsymbol{x}^m|\boldsymbol{x}^o;\theta), \tag{B.4}$$

où $L(\theta|\boldsymbol{x})$ et $L(\theta|\boldsymbol{x}^o)$ sont respectivement la vraisemblance de l'ensemble des données et des données manquantes. Le terme $P(\boldsymbol{x}^m|\boldsymbol{x}^o;\theta)$ joue un rôle central dans l'algorithme EM, car il représente l'interdépendance entre $\boldsymbol{x}^o$ et θ. Puisque $\boldsymbol{x}^m$ est inconnu, le deuxième terme de l'équation (B.4) ne peut évidemment pas être calculé. Cependant, pour une valeur provisoire θ_k de θ, on peut calculer la valeur moyenne de l'expression (B.4) à partir des valeurs observées, selon la distribution conditionnelle $P(\boldsymbol{x}^m|\boldsymbol{x}^o;\theta_k)$:

$$Q(\theta|\theta_k) = \mathbb{E}\Big[\ln L(\theta|\boldsymbol{x})|\boldsymbol{x}^o;\theta_k\Big] = \int \ln\big(P(\boldsymbol{x}^m,\boldsymbol{x}^o,\theta)\big)\,P(\boldsymbol{x}^m|\boldsymbol{x}^o,\theta_k)\;d\boldsymbol{x}^m$$

qui est l'espérance conditionnelle du logarithme de la vraisemblance à partir des données observées $\boldsymbol{x}^o$, évalué pour le paramètre fixé θ_k. Ensuite, le paramètre θ est réestimé de manière à maximiser $Q(\theta|\theta_k)$. Cela revient à réestimer la densité conditionnelle $P(\boldsymbol{x}^m|\boldsymbol{x}^o;\theta)$. Après l'initialisation de θ à une valeur θ_0, l'algorithme EM alterne ainsi successivement deux étapes :

- Etape E (*Expectation*): calcul de $Q(\theta|\theta_k)$
- Etape M (*Maximisation*):

$$\theta_{k+1} = arg\max_{\theta} Q(\theta|\theta_k).$$

Dempster et al. [29] ont montré que $L(\theta_{k+1}|\boldsymbol{x}^o) \geq L(\theta_k|\boldsymbol{x}^o)$ et que l'algorithme converge vers un point stationnaire $\widehat{\theta}$, qui est *en général* le maximum de la vraisemblance $L(\theta|\boldsymbol{x}^o)$.

Les estimateurs des données manquantes s'écrivent alors:

$$\widehat{\boldsymbol{x}^m} = \mathbb{E}(\widehat{X^m|\boldsymbol{x}^o};\widehat{\theta}).$$

Un inconvénient des méthodes paramétriques est leur manque de flexibilité. Ce problème a été en partie résolu par l'utilisation de modèles de mélange [100], qui combinent la flexibilité

des modèles non paramétriques et les avantages analytiques des modèles paramétriques. Ghahramani et Jordan [53] ont appliqué la méthode précédente aux modèles de mélange, dont les paramètres sont estimés de manière analogue par l'algorithme EM. Notons que les applications initiales de EM étant à la fois la gestion de données manquantes et l'estimation de paramètres de modèles de mélange, qui peut lui-même être vu comme un problème de données manquantes, il est apparu naturel de combiner les deux problèmes.

Une approche robuste de cette modélisation [96], introduisant des poids dans l'algorithme EM, permet en outre de minimiser l'effet néfaste de données aberrantes. On peut aussi noter que si les données suivent une loi normale multivariée, la méthode de régression proposée par Buck [20] et l'algorithme classique EM donnent la même solution [97].

B.4.2 Méthodes de Monte-Carlo

Il s'agit d'un ensemble de méthodes utilisées de façon générale pour la *simulation* de distributions de probabilité. Dans un cadre bayésien, ces méthodes sont souvent vues de manière restrictive comme des méthodes de simulation de probabilités *a posteriori*. Les trois algorithmes les plus connus sont

- l'échantillonneur de Gibbs,
- la méthode de Métropolis-Hastings,
- l' « augmentation de données » (*data augmentation*).

Le principe général est de générer aléatoirement des valeurs d'une variable ou d'un vecteur aléatoire Z de loi P_Z, la distribution ciblée. Plutôt que de simuler cette loi directement, on génère une chaîne de Markov, c'est-à-dire, une suite de variables aléatoires $(Z_k)_{k=1,2...}$ censées être plus accessibles, dont la loi de chacune dépend de celles des précédentes et dont la distribution limite est la distribution ciblée P_Z.

Ainsi la loi de Z_k est définie à partir de celles de $Z_{k-1}, Z_{k-2}, \ldots$ et on a :

$$Z_k \xrightarrow{\mathcal{L}} Z.$$

A la différence d'une méthode d'optimisation comme l'algorithme EM, qui est déterministe et converge vers un point $\widehat{\theta}$ de l'espace des paramètres, les méthodes de Monte-Carlo sont stochastiques et convergent vers des distributions de probabilité.

L'une de ces méthodes, l'« augmentation de données » [152], se prête particulièrement bien au problème de données manquantes. Elle suppose que le vecteur aléatoire X, dont la loi P_X est difficile à simuler, est partitionné en deux sous vecteurs, $X = (X^o, X^m)$ et que les lois conditionnelles sont, elles, faciles à obtenir.

Le principe est le suivant. La probabilité *a posteriori* $P(\theta|\boldsymbol{x}^o)$ ne pouvant être facilement calculée, ni même simulée, $\boldsymbol{x}^o$ est « augmentée » d'une valeur déterminée de $\boldsymbol{x}^m$, la distribution de probabilité $P(\theta|\boldsymbol{x}^o, \boldsymbol{x}^m)$ étant censée être plus facile à simuler. Etant donnée une valeur provisoire θ_k du paramètre θ, on tire aléatoirement une valeur $\boldsymbol{x}^m_{k+1}$ à partir de la distribution conditionnelle de X^m_{k+1} :

$$X^m_{k+1} \sim P(\boldsymbol{x}^m|\boldsymbol{x}^o, \theta_k)$$

Ensuite, en conditionnant sur $\boldsymbol{x}_{k+1}^m$, on tire aléatoirement une nouvelle valeur θ_k de θ, à partir la loi de probabilité *a posteriori*:

$$\theta_{k+1} \sim P(\theta|\boldsymbol{x}^o, \boldsymbol{x}_{k+1}^m)$$

En partant d'une valeur initiale θ_0, on obtient un processus stochastique $(\theta_k, X_k^m)_{k=1,2,...}$ convergent en loi vers la distribution $P(\theta, \boldsymbol{x}^m|\boldsymbol{x}^o)$. La suite X_k^m converge en loi vers $P(\boldsymbol{x}^m|\boldsymbol{x}^o)$:

$$X_k^m \xrightarrow{\mathcal{L}} X^m,$$

et fournit donc une suite d'estimateurs convergent en loi vers $\boldsymbol{x}^m$.

Dans le cas où le vecteur ne possède qu'une seule seule donnée manquante, cette méthode correspond également à un cas particulier d'une autre méthode, l'échantillonneur de Gibbs [127].

Une méthode basée sur la simulation de versions plausibles de $\boldsymbol{x}^m$, la méthode de remplacement multiple (*multiple imputation*) a été proposée par Rubin [126]. La philosophie de cette approche est la même que celle de l'algorithme EM ou de l'augmentation de données : résoudre un problème de données incomplètes en répétant la résolution de problèmes avec des données complètes. Dans cette méthode, la donnée inconnue $\boldsymbol{x}^m$ est remplacée par p valeurs simulées $\boldsymbol{x}_{(j)}^m$, produisant p jeux de données complets qui seront analysés par des méthodes standard. Cette méthode est malheureusement coûteuse en temps de calcul.

D'autres méthodes plus complexes combinant ces différents types de méthodes de Monte-Carlo peuvent également être utilisées [127].

B.4.3 Autres méthodes probabilistes

Régulièrement, de nouvelles méthodes basées sur les développements récents des statistiques sont introduites. Par exemple, les techniques basées sur le rééchantillonnage ont donné lieu à l'utilisation du bootstrap [97] et du jackknife [125]. La modélisation se base alors souvent sur la fonction de répartition empirique. Cette approche peut être intégrée dans un contexte bayésien, par exemple dans le cas du bootstrap bayésien approximé [97]. Le bootstrap offre aussi une approche non-paramétrique pour vérifier la qualité d'un estimateur en présence de données manquantes.

Dans le contexte de la régression non paramétrique, quelques méthodes ont également été proposées, bien que la prise en compte de données incomplètes soit plus délicate dans ce contexte. Les études réalisées concernent l'estimation d'une densité sous l'hypothèse MCAR [154] et l'estimation d'une variable de sortie [24], sous l'hypothèse MAR, à l'aide d'une méthode basée sur les noyaux.

Les méthodes basées sur les réseaux de neurones présentent l'avantage d'intégrer les données manquantes dans une phase d'apprentissage des données. Ainsi, dans un contexte de classification [1] et de régression [157], Ahmad et Tresp montrent que des solutions approchées de techniques bayésiennes peuvent être obtenues dans le cas gaussien par des réseaux à fonction de base radiale. La méthode offre de grandes similitudes avec celle proposée par Ghahramani et Jordan [53] utilisant les modèles de mélange (voir plus haut).

Dans une méthode basée sur les réseaux bayésiens [113] proposée par Ramoni et al. [124], les auteurs estiment les probabilités conditionnelles du réseau et en déduisent des valeurs de

remplacement pour les données incomplètes.

B.5 Conclusion

Les méthodes d'estimation ou simplement de gestion de données manquantes font appel à des techniques très variées, qu'il est d'ailleurs difficile de classer, car elles relèvent de procédés différents mais aboutissent parfois au même résultat. C'est en particulier le cas pour certaines techniques statistiques que l'on peut voir sous l'angle connexionniste.

Les méthodes heuristiques, souvent utilisées par le praticien, comme le remplacement par la moyenne, la médiane ou une valeur quelconque de référence, permettent d'éluder le problème rapidement à l'aide de solutions peu coûteuses. L'estimation des données incomplètes n'est souvent pas un but en soit, mais un prétraitement de données. Néanmoins, les algorithmes ultérieurs de classification, d'estimation de sortie d'un système ou d'apprentissage peuvent être perturbés par une mauvaise reconstruction de données.

Si le nombre de caractéristiques r et le nombre de données à reconstruire sont élevées, les méthodes basées sur la régression et les méthodes neuronales équivalentes exigent un grand nombre de modèles ($r2^r - 1$ au maximum). D'autre part, le principe de construction de multiples modèles dénote un manque de souplesse dans ce type de méthodes.

Les algorithmes EM et de Monte-Carlo ont prouvé leur efficacité mais supposent la connaissance ou l'estimation des lois des variables.

Annexe C

Calcul du gradient de l'erreur J.

Erreur de reconstruction J

Soit $\boldsymbol{x} = (\boldsymbol{x}^o, \boldsymbol{x}^m)$, nous cherchons à minimiser le coût quadratique moyen:

$$J = \frac{1}{2|O(\boldsymbol{x})|} \sum_{l \in O(x)} (\widehat{x}_l - x_l)^2 \tag{C.1}$$

par rapport à σ_{kj} and c_{kj}, pour tout $k \in \{1, \ldots, n\}$ et tout $j \in \{1, \ldots, r\}$, avec :

$$\widehat{x}_l = \sum_{i=1}^{n} v_{il} c_{il} \quad \forall l \in O(\boldsymbol{x}), \tag{C.2}$$

$$v_{il} = \frac{\tau_{il}\sigma_{il}}{\sum_{p=1}^{n} \tau_{pl}\sigma_{pl}}, \tag{C.3}$$

$$\tau_{pl} = \exp\left\{ -\frac{1}{2} \sum_{q \in O(x)\backslash \boldsymbol{l}} \left(\frac{x_q - c_{pq}}{\sigma_{pq}} \right)^2 \right\} \tag{C.4}$$

L'expression des seuils τ_{pl} dépend ici de l, contrairement à la formule donnée en équation (2.9), car, dans la phase d'apprentissage, chaque x_l est reconstruit à l'aide des *autres* variables connues de $\boldsymbol{x}$, soit $x_q, q \in O(\boldsymbol{x})\backslash l$.

Calcul du gradient de J par rapport aux c_{kj}

On calcule donc :

$$\frac{\partial J}{\partial c_{kj}} = \frac{1}{|O(\boldsymbol{x})|} \sum_{l \in O(x)} (\widehat{x}_l - x_l) \frac{\partial \widehat{x}_l}{\partial c_{kj}}. \tag{C.5}$$

D'après l'équation (C.2),

$$\forall l \in O(\boldsymbol{x}), \quad \frac{\partial \widehat{x}_l}{\partial c_{kj}} = \sum_{i=1}^{n} \frac{\partial v_{il}}{\partial c_{kj}} c_{il} + v_{kl}\delta_{lj} \tag{C.6}$$

où $\delta_{lj} = 1$ si $l = j$ et 0 si $l \neq j$. D'après l'équation (C.3),

$$\frac{\partial v_{il}}{\partial c_{kj}} = \frac{1}{\sum_{p=1}^{n} \tau_{pl}\sigma_{pl}} \left\{ \frac{\partial \tau_{il}}{\partial c_{kj}} \sigma_{il} - v_{il} \sum_{p=1}^{n} \sigma_{pl} \frac{\partial \tau_{pl}}{\partial c_{kj}} \right\} \tag{C.7}$$

Or

$$\frac{\partial \tau_{il}}{\partial c_{kj}} = \begin{cases} \delta_{ik}\tau_{il}\dfrac{(x_j - c_{ij})}{\sigma_{ij}^2} & \text{si } j \in O(\boldsymbol{x})\backslash l \\ 0 & \text{sinon.} \end{cases} \tag{C.8}$$

$$\frac{\partial v_{il}}{\partial c_{kj}} = \frac{1}{\sum_{p=1}^{n} \tau_{pl}\sigma_{pl}} \left\{ \delta_{ik}\tau_{il}\sigma_{il}\frac{(x_j - c_{ij})}{\sigma_{ij}^2} - v_{il}\sigma_{kl}\tau_{kl}\frac{(x_j - c_{kj})}{\sigma_{kj}^2} \right\} \delta_{\{O(x)\backslash l\}}(j) \tag{C.9}$$

D'après l'équation (C.3), on obtient donc :

$$\frac{\partial v_{il}}{\partial c_{kj}} = \begin{cases} \dfrac{(x_j - c_{kj})}{\sigma_{kj}^2} v_{il}(\delta_{ik} - v_{kl}) & \text{si } j \in O(\boldsymbol{x})\backslash l \\ 0 & \text{sinon.} \end{cases} \tag{C.10}$$

D'après les équations (C.6) et (C.10),

$$\forall l \in O(\boldsymbol{x}), \quad \frac{\partial \widehat{x}_l}{\partial c_{kj}} = \begin{cases} \dfrac{(x_j - c_{kj})}{\sigma_{kj}^2} v_{kl} \left(c_{kl} - \sum_{i=1}^{n} c_{il} v_{il} \right) & \text{si } j \in O(\boldsymbol{x})\backslash l \\ v_{kl} & \text{si } j = l \end{cases} \tag{C.11}$$

Puisque $\widehat{x}_l = \sum_{i=1}^{n} c_{il} v_{il}$, on a donc :

$$\forall l \in O(\boldsymbol{x}), \quad \frac{\partial \widehat{x}_l}{\partial c_{kj}} = \begin{cases} \dfrac{(x_j - c_{kj})}{\sigma_{kj}^2} v_{kl} \left(c_{kl} - \widehat{x}_l \right) & \text{si } j \in O(\boldsymbol{x})\backslash l \\ v_{kl} & \text{si } j = l \end{cases} \tag{C.12}$$

On obtient finalement :

$$\frac{\partial J}{\partial c_{kj}} = \frac{1}{|O(\boldsymbol{x})|} \left\{ \frac{x_j - c_{kj}}{(\sigma_{kj})^2} \sum_{l \in O(x)\backslash\{j\}} (\widehat{x}_l - x_l)(c_{kl} - \widehat{x}_l) v_{kl} + v_{kj}(\widehat{x}_j - x_j) \right\} \tag{C.13}$$

si $j \in 0(\boldsymbol{x})$ et $\dfrac{\partial J}{\partial c_{kj}} = 0$ sinon.

Calcul du gradient de J par rapport aux σ_{kj}

Par un calcul analogue, on obtient :

$$\frac{\partial J}{\partial \sigma_{kj}} = \frac{1}{|O(\boldsymbol{x})|} \sum_{l \in O(x)} (\widehat{x}_l - x_l) \frac{\partial \widehat{x}_l}{\partial \sigma_{kj}}. \tag{C.14}$$

$$\forall l \in O(\boldsymbol{x}), \quad \frac{\partial \widehat{x}_l}{\partial \sigma_{kj}} = \sum_{i=1}^{n} \frac{\partial v_{il}}{\partial \sigma_{kj}} \sigma_{il}. \tag{C.15}$$

$$\frac{\partial v_{il}}{\partial \sigma_{kj}} = \frac{1}{\sum_{p=1}^{n} \tau_{pl}\sigma_{pl}} \left\{ \frac{\partial \tau_{il}}{\partial \sigma_{kj}} \sigma_{il} + \delta_{ki}\delta_{jl}\tau_{kl} - v_{il} \left(\sigma_{kl} \frac{\partial \tau_{kl}}{\partial \sigma_{kl}} + \tau_{kl}\delta_{jl} \right) \right\} \tag{C.16}$$

Or

$$\frac{\partial \tau_{il}}{\partial \sigma_{kj}} = \begin{cases} \delta_{ik}\tau_{il} \dfrac{(x_j - c_{ij})^2}{\sigma_{ij}^3} & \text{si } j \in O(\boldsymbol{x}) \backslash l \\ 0 & \text{sinon.} \end{cases} \tag{C.17}$$

On obtient alors :

$$\frac{\partial v_{il}}{\partial \sigma_{kj}} = \begin{cases} \dfrac{(x_j - c_{kj})^2}{\sigma_{kj}^3} v_{il}(\delta_{ik} - v_{kl}) & \text{si } j \in O(\boldsymbol{x}) \backslash l \\ \dfrac{v_{kl}}{\sigma_{kl}} (\delta_{ik} - v_{kl}) & \text{si } j = l \\ 0 & \text{sinon.} \end{cases} \tag{C.18}$$

D'après les équations (C.14), (C.15) et (C.18), on obtient finalement :

$$\frac{\partial J}{\partial \sigma_{kj}} = \frac{1}{|O(\boldsymbol{x})|} \left\{ \frac{(x_j - c_{kj})^2}{(\sigma_{kj})^3} \sum_{l \in O(x) \backslash \{j\}} (\widehat{x}_l - x_l)(c_{kl} - \widehat{x}_l) v_{kl} + \frac{v_{kj}}{\sigma_{kj}} (\widehat{x}_j - x_j)(c_{ij} - \widehat{x}_j) \right\}$$

si $j \in 0(\boldsymbol{x})$ et $\dfrac{\partial J}{\partial \sigma_{kj}} = 0$ sinon.

Annexe D

Algorithmes de classification non supervisés standards

Dans cette annexe, nous rappelons brièvement les algorithmes de classification non probabilistes les plus courants. Ces algorithmes appartiennent à deux groupes distincts : les méthodes de partionnement et les méthodes hiérarchiques. Nous renvoyons à [56, 27] pour la définition de la partition et de la hiérarchie.

Soit $\Omega = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ un ensemble de N vecteurs de $\mathbb{R}^r, r \in \mathbb{N}$. L'objectif est de regrouper ces vecteurs en un certain nombre de classes homogènes $C_1, \ldots, C_K$, de centres respectifs $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$.

D.1 Méthodes de partitionnement

D.1.1 Algorithme des centres mobiles

La méthode des centres mobiles [99], très fréquemment utilisée, repose sur l'alternance du calcul d'une partition $P = \{C_1, \ldots, C_K\}$ de Ω et des centres $\boldsymbol{c} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$ des classes de la partition P. Le nombre K de classes est supposé fixé.

Critère d'inertie

Le critère de choix de la partition et des centres se fait en minimisant le critère d'inertie :

$$I(P, \boldsymbol{c}) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2 \tag{D.1}$$

alternativement par rapport à P et $\boldsymbol{c}$. Soient $P^{(t)}$ et $\boldsymbol{c}^{(t)}$, la partition et les centres à l'itération t. On a donc

$$I(P^{(t+1)}, \boldsymbol{c}^{(t+1)}) \leq I(P^{(t+1)}, \boldsymbol{c}^{(t)}) \leq I(P^{(t)}, \boldsymbol{c}^{(t)}).$$

On peut montrer que l'algorithme converge vers une partition stable en un nombre fini d'itérations. Cette partition finale est un minimum local du critère ; elle dépend du choix des centres initiaux.

Algorithme

On aboutit alors à l'algorithme suivant :

- Initialiser les centres à des valeurs arbitraires $\boldsymbol{c}_k^{(0)}$
- **Répéter**
  1. Calcul de la partition $P^{(t)} = \{C_1^{(t)}, \ldots, C_K^{(t)}\}$
     Pour tout $i = 1, \ldots, N$: affecter $\boldsymbol{x}_i$ à la classe la plus proche $C_{k^*}^{(t)}$ telle que :
     $k^* = \arg\min_k \|\boldsymbol{x}_i - \boldsymbol{c}_k^{(t)}\|^2$
  2. Calcul des centres de gravité $\boldsymbol{c}_k^{(t+1)}$ des classes $C_k^{(t+1)}$
- **jusqu'à** $P^{(t+1)} = P^{(t)}$

On peut montrer que cet algorithme revient à minimiser l'inertie intra-classe de Ω :

$$I_W = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k\|^2,$$

où $\overline{\boldsymbol{x}}_k$ est le centre de gravité de C_k. Il existe de nombreuses variantes de l'algorithme des centres mobiles [56]. Cet algorithme est un cas particulier de la méthode des nuées dynamiques.

D.1.2 Méthodes connexionnistes

Dans le cadre des réseaux de neurones, de nombreux algorithmes relevant de la classification non supervisée ont été développés. Les algorithmes d'apprentissage par compétition [87] et des cartes de Kohonen [86] font partie des techniques les plus courantes. Ces deux méthodes suivent un principe similaire. Chaque classe C_k est représentée par un neurone ou prototype $\boldsymbol{c}_k$. Les coordonnées des prototypes s'adaptent au fur et à mesure qu'on leur présente les observations $\boldsymbol{x}_i$, l'objectif étant d'obtenir des vecteurs $\boldsymbol{c}_k$ représentatifs des données. Dans l'apprentissage par compétition, seul le prototype $\boldsymbol{c}_k$ le plus proche de $\boldsymbol{x}_i$, appelé prototype « gagnant », s'adapte en se rapprochant de $\boldsymbol{x}_i$. Dans l'algorithme des cartes auto-organisatrices de Kohonen, les prototypes sont supposés être reliés par une structure de *voisinage* définie à l'avance. Comme dans l'apprentissage par compétition, on cherche en premier lieu le prototype le plus proche du vecteur en présence $\boldsymbol{x}_i$. En revanche, on ne modifie pas uniquement ce prototype, mais également les unités situées dans un certain voisinage V_k de $\boldsymbol{c}_k$. Ces techniques donnent donc lieu aux deux algorithmes suivants :

Apprentissage par compétition

- Initialiser les prototypes à des valeurs arbitraires $\boldsymbol{c}_k^{(0)}$
- Pour une observation $\boldsymbol{x}^{(t)}$ (t représente l'itération courante.)
  1. Recherche du prototype gagnant $\boldsymbol{c}_{k^*}^{(t)}$ tel que :
     $k^* = \arg\min_k \|\boldsymbol{x}^{(t)} - \boldsymbol{c}_k^{(t)}\|^2$;

2. Mettre à jour les coordonnées du prototype gagnant :
   $\boldsymbol{c}_{k^*}^{(t+1)} = \boldsymbol{c}_{k^*}^{(t+1)} + \eta(t)(\boldsymbol{x}^{(t)} - \boldsymbol{c}_{k^*}^{(t)})$,

où $\eta(t)$ est une fonction décroissante bien choisie.

Cartes de Kohonen

- Initialiser les prototypes à des valeurs arbitraires $\boldsymbol{c}_k^{(0)}$
- Pour une observation $\boldsymbol{x}^{(t)}$

1. Recherche du prototype gagnant $\boldsymbol{c}_{k^*}^{(t)}$ tel que :
   $k^* = \arg\min_k \|\boldsymbol{x}^{(t)} - \boldsymbol{c}_k^{(t)}\|^2$;
2. Modifier le prototype gagnant et ses *voisins* ($\in V_{k^*}(t)$) :

$$\forall k \in V_{k^*}(t), \quad \boldsymbol{c}_{k^*}^{(t+1)} = \boldsymbol{c}_{k^*}^{(t+1)} + \varepsilon(t,k,k^*)(\boldsymbol{x}^{(t)} - \boldsymbol{c}_{k^*}^{(t)}),$$

où $\varepsilon(t,k,k^*)$ est une fonction décroissante par rapport à t et à la distance entre l'unité gagnante $\boldsymbol{c}_{k^*}$ et ses voisines.

D.1.3 Algorithme des centres mobiles flous

Partition floue

Dans les algorithmes précédents, chaque observation $\boldsymbol{x}_i$ est affectée à une et une seule classe de la partition finale. Afin d'autoriser une certaine souplesse dans la classification des vecteurs, Bezdek [11] a proposé d'introduire la notion de partition *floue*. Chaque vecteur $\boldsymbol{x}$ est supposé appartenir à toute classe C_k avec un certain degré d'appartenance μ_{ik}. La matrice $\boldsymbol{\mu} = (\mu_{ik})$, $i = 1, \ldots, N, k = 1, \ldots, K$, définit alors une *partition floue* sur Ω. La méthode de classification floue la plus connue est celle des *centres mobiles flous* ou *fuzzy-c-means* (FCM), nommée ainsi par analogie avec l'algorithme des centres mobiles classiques. Le principe de cette méthode est de chercher la partition floue $\boldsymbol{\mu}$ et les centres de classes $\boldsymbol{c}$ qui minimisent le critère suivant :

$$I_m(\boldsymbol{\mu}, \boldsymbol{c}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \mu_{ik}^m \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2, \tag{D.2}$$

sous les contraintes

$$\begin{cases} \mu_{ik} \in [0,1] \\ \sum_{i=1}^{K} \mu_{ik} = 1 \ \forall k = 1, \ldots K \\ 0 < \sum_{i=1}^{N} \mu_{ik} < N \ \forall i = 1, \ldots, N, \end{cases} \tag{D.3}$$

où l'exposant $m > 1$ est un paramètre fixé à l'avance. Les conditions (D.3) expriment que tout vecteur $\boldsymbol{x}_i$ appartient à la partition $\boldsymbol{\mu}$, et que dans toute classe, il existe au moins une observation de degré d'appartenance non nul. On peut noter que si la matrice de classification $\boldsymbol{\mu}$ est classique ($\mu_{ik} \in \{0,1\}$), on obtient le critère des centres mobiles classiques. L'algorithme FCM a une structure itérative similaire à celui des centres mobiles. A partir de centres initiaux, on optimise alternativement le critère I_m par rapport à la partition $\boldsymbol{\mu}$ et les centres $\boldsymbol{c}$:

Algorithme

On aboutit alors à l'algorithme suivant :

- Initialiser les centres à des valeurs arbitraires $\boldsymbol{c}_k^{(0)}$
- **Répéter**
  1. Calcul de la partition $\boldsymbol{\mu}^{(t+1)}$:

$$\forall i = 1, \ldots, N, k = 1 \ldots, K, \quad \mu_{ik}^{(t+1)} = \left(\sum_{j=1}^{K} \frac{\|(\boldsymbol{x}_i - \boldsymbol{c}_k^{(t)}\|}{\|\boldsymbol{x}_i - \boldsymbol{c}_j^{(t)}\|}^{\frac{2}{m-1}} \right)^{-1} \tag{D.4}$$

  2. Calcul des centres $\boldsymbol{c}_k^{(t+1)}$

$$\boldsymbol{c}_k^{(t+1)} = \frac{\sum_{j=1}^{K} (\mu_{ij}^{(t+1)})^m \boldsymbol{x}_i}{\sum_{i=1}^{N} (\mu_{ik}^{(t+1)})^m} \tag{D.5}$$

- **jusqu'à** $\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)}$ (ou $\|\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)}\| \leq \varepsilon$, avec $\varepsilon > 0$, une quantité fixée à l'avance).

D.2 Classification hiérarchique

D.2.1 Hiérarchie

Principe

Une hiérarchie ascendante sur Ω est un ensemble de partitions emboîtées, depuis l'ensemble des singletons $\{\{\boldsymbol{x}_i\}, \boldsymbol{x}_i \in \Omega\}$ jusqu'à l'ensemble Ω lui-même, en passant par des agrégations successives de ses sous-ensembles. La classification ascendante hiérarchique repose sur le calcul des *dissimilarités* entre les différentes parties de Ω. On définit un critère D, appelé *critère d'agrégation*, entre deux parties quelconques de Ω. La classification est un algorithme itératif, basé sur l'agrégation des deux classes A et B les plus proches au sens de D. On obtient alors l'algorithme suivant :

Algorithme

- Initialiser la hiérarchie : on part de N classes consituées des singletons de Ω. La partition $P^{(o)}\{\{\boldsymbol{x}_i\}, \boldsymbol{x}_i \in \Omega\}$
- **Répéter**
  1. Regroupement des 2 classes A et B les plus proches de la partition $P^{(t)}$ de Ω à l'itération t, en une nouvelle classe C :

$$C = A \cup B \text{ avec } (A, B) = \arg \min_{(F,G) \in P^{(t)}} D(F, G)$$

  2. Calcul des dissimilarités entre la nouvelle classe C et les autres classes.
- **jusqu'à** ce que la partition finale soit constituée d'un seul élément : $P^{(N)} = \Omega$ ou que l'on ait obtenu le nombre de classes désiré.

Critères d'agrégation

Parmi les critères d'agrégation les plus courants, on peut citer le critère du *lien minimum* :

$$D(A,B) = \min\{\|\boldsymbol{x}-\boldsymbol{y}\|, \boldsymbol{x} \in A \text{ et } \boldsymbol{y} \in B\},$$

le critère du *lien maximum*,

$$D(A,B) = \max\{\|\boldsymbol{x}-\boldsymbol{y}\|, \boldsymbol{x} \in A \text{ et } \boldsymbol{y} \in B\},$$

ainsi que la distance moyenne :

$$D(A,B) = \frac{1}{|A||B|}\sum_{x\in A}\sum_{y\in B}\|\boldsymbol{x}-\boldsymbol{y}\|^2.$$

Enfin, si les observations appartiennent à $\mathbb{R}^r$, on peut utiliser le critère d'agrégation de Ward [165].

D.2.2 Méthode de Ward

Critère de Ward

L'ensemble Ω est considéré comme un nuage de points de $\mathbb{R}^r$, muni de la distance euclidienne $\|.\|$. Le critère d'agrégation de Ward est défini par :

$$D_{ward}(A,B) = \frac{|A||B|}{|A|+|B|}\|\overline{\boldsymbol{x}}_A - \overline{\boldsymbol{x}}_B\|^2$$

où $\overline{\boldsymbol{x}}_A$ et $\overline{\boldsymbol{x}}_B$ sont les centres de gravité des ensembles A et B.

Minimisation de la perte d'inertie

Soit l'inertie intra-classes définie précédemment, pour une partition $P = \{C_1, \ldots, C_K\}$ donnée :

$$I_W(P) = \sum_{k=1}^{K}\sum_{x_i\in C_k}\|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k\|^2,$$

Si on fusionne les classes C_1 et C_2 de la partition P, on obtient une nouvelle partition

$$P' = P\backslash\{C_1, C_2\} \cup \{C_1 \cup C_2\}.$$

On peut alors montrer le résultat suivant :

$$I_W(P') - I_W(P) = D_{ward}(C_1, C_2).$$

On peut tirer deux conclusions de ce résultat :

- la fusion de deux classes augmente nécessairement le critère d'inertie intra-classe.
- le critère de Ward correspond à la *perte minimale* d'inertie intra-classe.

A chaque étape de l'algorithme de classification, le critère de Ward optimise localement le critère d'inertie intra-classes. Cependant, cet algorithme ne possède aucune propriété globale d'optimisation.

Bibliographie

[1] S. Ahmad and V. Tresp. Some solutions to the missing feature problem in vision. In Giles Hanson, Cowan, editor, *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, 1993.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[3] R. L. Anderson. Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52:200–203, 1957.

[4] A. Antoniadis, Grégoire G., and I. McKeague. Wavelet methods for curve estimation. *Journal of the American Statistical Society*, 89(428):1340–1353, 1994.

[5] C. G. Atkeson. *Memory-based approaches to approximating continuous functions*. Addison-Wesley, Harlow, UK, 1992.

[6] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

[7] R. Babuška and H. B. Verbruggen. Constructing fuzzy models by product space clustering. In H. Hellendoorn and D. Driankov, editors, *Fuzzy model identification : selected approaches*, pages 53–90. Sringer, Berlin, 1997.

[8] V. Barnett and T. Lewis. *Outliers in statistical data*. Wiley, 1994.

[9] M. Bauer. Approximations for decision making in the Dempster-Shafer theory. In *Uncertainty in Artificial Intelligence*, pages 339–344, Saarbrücken, 1996.

[10] H. Bersini, G. Bontempi, and M. Birattari. Is readibility compatible with accuracy? from neuro-fuzzy to lazy learning. In *Fuzzy-Neuro Systems '98*, pages 10–25, Munich, 1998.

[11] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, 1981.

[12] C.M. Bishop. *Neural networks for pattern recognition*. Clarendon Press, Oxford, 1996.

[13] D. Bosq and J.P. Lecoutre. *Théorie de l'estimation fonctionnelle*. Economica, Paris, 1987.

[14] D. H. Bossley. *Neurofuzzy modelling approaches in system identification.* PhD thesis, University of Southampton, 1997.

[15] S. Boucheron. *Théorie de l'apprentissage.Langue, raisonnement, calcul.* Hermès, Paris, 1992.

[16] B. Bouchon-Meunier. *La logique floue et ses applications.* Addison - Wesley, Paris, 1995.

[17] L. Breiman, J.H Friedman, R. Olshen, and C.J Stone. *Classification and regression trees.* Wadsworth, Belmont, CA, 1984.

[18] P. Brockwell and R. Davis. *Time series: theory and methods.* Springer, New York, 1991.

[19] M. Brown and C. Harris. *Neurofuzzy adaptive modelling and control.* Prentice Hall, Hemel-Hampstead, 1994.

[20] S. F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, series B*, 22:302–306, 1960.

[21] J. J. Buckley and Y. Hayashi. Fuzzy neural networks: a survey. *Fuzzy Sets and Systems*, 66:1–13, 1994.

[22] J. J. Buckley, Y. Hayashi, and E. Czogala. On the equivalence of neural nets and fuzzy expert systems. *Fuzzy Sets and Systems*, 53:129–134, 1993.

[23] W. F. Caselton and W. Luo. Decision making with imprecise probabilities: Dempster-Shafer theory and application. *Water Resources Research*, 28(12):3071–3081, 1992.

[24] C. K. Chu and P. E. Cheng. Nonparametric regression estimation with missing data. *Journal of Statistical Planning and Inference*, 48:85–99, 1995.

[25] W. S. Cleveland. Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[26] F. Crettaz de Roten and J.M. Helbling. Données manquantes et aberrantes : le quotidien du statisticien analyste de données. *Revue de Statistique Appliquée*, 44:105–115, 1996.

[27] V. M. Dang. *Classification de données spatiales : modèles probabilistes et critères de partitionnement.* PhD thesis, Université de Compiègne, 1998.

[28] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, AMS-38:325–339, 1967.

[29] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, 39:1–38, 1977.

[30] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[31] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

[32] T. Denœux. Function approximation in the framework of evidence theory: A connectionist approach. In *Proceedings of the 1997 International Conference on Neural Networks*, pages 199–203, Houston, 1997.

[33] T. Denœux. Application du modèle des croyances transférables en reconnaissance des formes. *Traitement du Signal*, 14(5):443–451, 1998.

[34] T. Denœux. Reasoning with imprecise belief structures. *International Journal of Approximate Reasoning*, 20:79–111, 1999.

[35] T. Denœux. Modeling vague belief structures. *Fuzzy Sets and Systems*, A paraître.

[36] T. Denœux, N. Boudaoud, S. Canu, G. Govaert, M. Masson, V. Mo Dang, S. Petit-Renaud, and S. Soltani. High level data fusion methods. Technical Report CNRS/EM2S/330/11-97 V 1.0, CNRS, 1997.

[37] T. Denœux and L.M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Soumis à Fuzzy Sets and Systems.*

[38] Ph. Diamond. Fuzzy least squares. *Information Sciences*, 46:141–157, 1988.

[39] Ph. 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, Norwell, 1998.

[40] D. Dubois and Prade H. *Possibility theory: an approach to computerized processing of uncertainty.* Plenum Press, New York, 1988.

[41] D. Dubois and H. Prade. Operations on fuzzy numbers. *International Journal System*, 9:613–626, 1978.

[42] D. Dubois and H. Prade. On several representations of an uncertainty body of evidence. In M. M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 167–181. North-holland, New-York, 1982.

[43] D. Dubois and H. Prade. A note on measures of specificity for fuzzy sets. *International Journal of General Systems*, 10(4):279–283, 1985.

[44] D. Dubois and H. Prade. On the unicity of Dempster rule of combination. *International Journal of Intelligent System*, 1:133–142, 1986.

[45] 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.

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

[47] B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(1):155–165, 1993.

[48] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, New York, 1993.

[49] J. Fan and I. Gijbels. *Local polynomial modeling and its applications*. Chapman and Hall, London, 1996.

[50] B. L. Ford. An overview of hot-deck procedures. In *Incomplete data in sample survey*. Academic Press, 1983.

[51] J. Friedman and Stuetzle. Projection pursuit algorithm. *Journal of American Statistical Association*, 76:817–823, 1981.

[52] T. George and N. Pal. Quantification of conflict in the Dempster-Shafer framework. *International Journal of General Systems*, 24(4):407–423, 1994.

[53] Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an EM approach. In Alspector Cowan, Tesauro, editor, *Advances in Neural Information Processing Systems 6*, pages 120–127. Morgan Kaufmann, 1994.

[54] M. N. Gibbs. *Bayesian gaussian processes for regression and classification*. PhD thesis, University of Cambridge, 1997.

[55] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7:219–269, 1995.

[56] G. Govaert. *Analyse de données*. Université de Compiègne, Polycopié de cours de DEA, 1996.

[57] Y. Grandvalet. *Injection de bruit dans les perceptrons multi-couches*. PhD thesis, Université de Compiègne, 1995.

[58] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall, New York, 1994.

[59] J.W. Guan and D.A. Bell. *Evidence theory and its applications*. Elsevier science, Amsterdam, 1991.

[60] J.W. Guan and D.A. Bell. A generalization of the Dempster-Shafer theory. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 10–25, Chambéry, France, 1993.

[61] W. Härdle. *Applied nonparametric regression*. Cambridge University Press, Cambridge, 1990.

[62] D. Harmanec and G. Klir. Measuring total uncertainty in Dempster-Shafer theory: a novel approach. *International Journal of General Systems*, 22(4):405–419, 1993.

[63] R.V.L. Hartley. Transmission of information. *The Bell Systems Technical Journal*, 7(3):535–563, 1928.

[64] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Chapman and Hall, New York, 1990.

[65] H. Hellendoorn and D. Driankov. *Fuzzy model identification : selected approaches.* Springer, Berlin, 1997.

[66] K. Hirota and W. Pedrycz. Knowledge-based networks in classification problems. *Fuzzy Sets and Systems*, 51:1–27, 1992.

[67] K Hirota and W. Pedrycz. A neural network architecture for classification of fuzzy inputs. *Fuzzy Sets and Systems*, 63:159–173, 1994.

[68] K. Hirota and W. Pedrycz. Or/and neuron in modeling fuzzy set connectives. *IEEE Transactions on Fuzzy Systems*, 2:151–161, 1994.

[69] U. Höhle. An evidence-theoretic neural network classifier. In *Proceedings of the IEEE International Symposium on the Multiple-Valued Logic*, pages 167–169, Vancouver, 1982.

[70] K Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universel approximators. *Neural networks*, 2:359–366, 1989.

[71] P. J. Huber. *Robust statistics.* Wiley, New York, 1981.

[72] K. J. Hunt, R. Haas, and R. Murray-Smith. Extending the functional equivalence of radial basis function networks and fuzzy inference systems. *IEEE Transactions on Neural networks*, 7:776–781, 1996.

[73] H. Ishibuchi, R. Fujioka, and H. Tanaka. Neural networks that learn from fuzzy if-then rules. *IEEE Transactions and Fuzzy Systems*, 1:85–97, 1993.

[74] H. Ishibuchi, K. Kwon, and H. Tanaka. A learning algorithm of fuzzy neural networks with triangular fuzzy weights. *Fuzzy Sets and Systems*, 71:277–293, 1995.

[75] F. Janez. *Fusion de sources d'information définies sur des référentiels non exhaustifs différents.* PhD thesis, Université d'Angers, 1996.

[76] J. S.R. Jang. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systemes, Man and Cybernetics*, 23:665–685, 1993.

[77] J. S.R. Jang and C. T. Sun. Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE Transactions on Neural Networks*, 4:156–159, 1993.

[78] J.S.R Jang, C.T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing.* Prentice-Hall, Upper Saddle River, N.J, 1997.

[79] A. Kandel. *Fuzzy expert systems.* CRC Press Inc., Boca Raton, Floride, 1992.

[80] H. Kim and P. H. Swain. Evidential reasoning approach to multisource-data classification in remote sensing. *IEEE Transactions on Systems, Man and Cybernetics*, 25(8):1257–1265, 1995.

[81] F. Klawonn and E. Schwecke. On the axiomatic justification of Dempster's rule combination. *International Journal of Intelligent Systems*, 7:469–478, 1992.

[82] G. Klir and A. Ramer. Uncertainty in the Dempster-Shafer theory: a critical re-examination. *International Journal of General Systems*, 18(2):155–166, 1990.

[83] G. Klir and M.J. Wierman. *Uncertainty-based-information : elements of generalized information theory.* Physica-Verlag, Heidelberg, NY, 1998.

[84] G. J. Klir and B. Yuan. *Fuzzy sets and fuzzy logic: Theory and applications.* Prentice Hall, Upper Sadddle River, N.J., 1995.

[85] J. Kohlas and P.A. Monney. *A mathematical theory of hints. An approach to Dempster-Shafer theory of evidence.* Springer-Verlag, 1995.

[86] T. Kohonen. Self organized formation of topological correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[87] B. Kosko. *Neural networks and fuzzy systems : a dynamical systems approach.* Prentice Hall, Upper Saddle River, NJ, 1991.

[88] S.L. Lai. *Large sample properties of k-nearest neighbours procedures.* PhD thesis, Department of Mathematics, UCLA, Los Angeles, 1977.

[89] M.T. Lamata and S. Moral. Measures of entropy in the theory of evidence. *International Journal of General Systems*, 14(4):297–305, 1988.

[90] M. Laviolette and J. Seaman. Efficacity of fuzzy représentations of uncertainty. *IEEE Transactions on Fuzzy Systems*, 2:4–15, 1994.

[91] J.P. Lecoutre and Ph. Tassi. *Statistique non paramétrique et robustesse.* Economica, Paris, 1987.

[92] C.C. Lee. Fuzzy logic in control systems: fuzzy logic controller, part 2. *IEEE Transactions on Systems, Man and Cybernetics*, 20:419–435, 1990.

[93] S. Lee and R. M. Kil. A gaussian potential function network with hierarchically self-organizing learning. *Neural networks*, 4:207–224, 1991.

[94] M. Lejeune. Estimation nonparamétrique par noyaux : régression polynômiale mobile. *Revue de Statistique Appliquée*, 23:43–67, 1985.

[95] D.V. Lindley. The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science*, 2:17–24, 1987.

[96] R. J. A. Little. Robust estimation of the mean and covariance matrix from data with missing values. *Applied statistics*, 37:23–38, 1988.

[97] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data.* Wiley, New York, 1987.

[98] J Lowrance, T. Garvey, and T. Strat. A framework for evidential-reasoning systems. In *Proceedings of the 5th National Conference on Artificial Intelligence*, pages 896–903, ., 1986. American Association for Artificial Intelligence.

[99] J. B. Mac Queen. Some methods for classification and analysis of multivariate observations. In *Statistics and Probability: 5th Berkeley Symposium*, pages 281–297, 1967.

[100] G. J. MacLachlan and K. E. Basford. *Mixture models, inference and application to clustering.* Marcel-Dekker, New York, 1988.

[101] Y. Maeda and H. Ichihashi. An uncertainty measure with monotonicity under the random set inclusion. *International Journal of General Systems*, 21(4):379–392, 1993.

[102] C. L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[103] E. H. Mamdani and S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Machine Studies*, 7:1–13, 1975.

[104] S. Mitra and K. Pal. Fuzzy multilayer perceptron, inferencing and generalisation. *IEEE Transactions on Neural Networks*, 6:51–63, 1995.

[105] J. Moody and C. J. Darken. Fast learning in networks of locally-tuned processing units. *Neural computation*, 1:281–294, 1989.

[106] H. G. Müller. Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, 82:231–238, 1987.

[107] S. Muller. *Un codage neuro-flou pour la classification de données incomplètes ou imprécises : application à la discrimination d'événements sismiques.* PhD thesis, Université Pierre et Marie Curie, 1998.

[108] R. Murray-Smith. *A local model network approach to nonlinear modelling.* PhD thesis, University of Strathclyde, 1994.

[109] R. Murray-Smith and T. A. Johansen. *Multiple model approaches to modelling and control.* Taylor and Francis, 1997.

[110] E. A. Nadaraya. On estimating regression. *Theory of probability and its applications*, 10:186–190, 1964.

[111] H. T. Nguyen and E. A. Walker. On decision making using belief functions. In *Advances in the Dempster-Shafer theory of evidence*, pages 311–330. Wiley, 1994.

[112] N.R Pal and J.C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, 3:370–379, 1995.

[113] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kauffman, San Mateo, CA, 1988.

[114] W. Pedrycz. Identification in fuzzy systems. *IEEE Transactions on Systems, Man and Cybernetics*, 14:361–366, 1989.

[115] S. Petit-Renaud and T. Denœux. Regression analysis based on the fuzzy evidence theory (best student paper award). In *Proceedings of the 8th International Conference on Fuzzy systems (FUZZ-IEEE'99)*, Séoul, Août 1999.

[116] S. Petit-Renaud and T. Denœux. Handling different forms of uncertainty in regression analysis: a fuzzy belief structure approach (ECSQARU' 99). In *Symbolic and quantitative approaches to reasoning and uncertainty*, pages 340–351, London, Juillet 1999.

[117] S. Petit-Renaud and T. Denœux. A neuro-fuzzy model for missing data reconstruction. In *Proceedings of the 1998 IEEE Workshop on Emerging Technologies, Intelligent Measurement and Virtual Systems for Instrumentation and Measurement ETIMVIS'98*, Saint-Paul, MN, Mai 1998.

[118] S. Petit-Renaud and T. Denœux. A fuzzy neuro system for reconstruction of multisensor information. In *Fuzzy Neuro Systems'98*, pages 322–329, Munich, Mars 1998.

[119] S. Petit-Renaud and T. Denœux. Application de la théorie des fonctions de croyance en régression. In *Rencontres francophones sur la logique floue et ses applications (LFA'99)*, Valenciennes, Octobre 1999.

[120] J. C. Platt. Learning by combining memorization and gradient descent. In R. P. Lipmann J. E. Moody, S.J. Hanson, editor, *Advances in Neural Information Processing Systems, 3*, pages 714–720. Morgan Kaufmann, 1991.

[121] H. Prade. A computational approach to approximate and plausible reasoning with applications to expert systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(3):260–283, 1985.

[122] A.L. Ralescu and D. A. Ralescu. Probability and fuzziness. *Information Sciences*, 34:85–92, 1984.

[123] A. Ramer. Uniqueness of information measure in the theory of evidence. *Fuzzy Sets and Systems*, 24(2):183–196, 1987.

[124] M. Ramoni and P. Sebastiani. Learning bayesian networks from incomplete databases. Technical Report 43, Knowledge Media Institute, Milton Keynes, UK, 1997.

[125] J. N. K. Rao and J. Shao. Jackknife variance estimation with survey data under hot dech imputation. *Biometrika*, 79:811–822, 1992.

[126] D. B. Rubin. *Multiple imputation for nonreponse in surveys.* Wiley, New York, 1987.

[127] J.L. Schafer. *Analysis of incomplete multivariate data.* Chapman and Hall, London, 1997.

[128] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[129] M. Setnes, R. Babuška, U. Kaymak, and H. R. van Nauta Lemke. Similarity measures in fuzzy rule base simplification. *IEEE Transactions on Systems, Man and Cybernetics - part B: Cybernetics*, 28:376–386, 1998.

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

[131] C.E. Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27(3):379–423,623–656, 1948.

[132] J. Sjöberg, Q. Zhang, L. Ljung, A. Benvéniste, B. Delyon, P. Y. Glonnerec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification : a unified overview. *Automatica*, 10:1691–1724, 95.

[133] Ph. Smets. *Un modèle mathématico-statistique simulant le processus du diagnostic médical.* PhD thesis, Université Libre de Bruxelles, 1978.

[134] Ph. Smets. The degree of belief in a fuzzy event. *Information Sciences*, 25:1–19, 1981.

[135] Ph. Smets. Information content of an evidence. *International Journal of Machine Studies*, 19:33–43, 1983.

[136] Ph. Smets. Constructing the pignistic probability function in a context of uncertainty. In M. Henrion, R.D. Shachter, L.N Kanal, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 5*, pages 29–40. North-Holland, 1989.

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

[138] Ph. Smets. The transferable belief model and possibility theory. Technical Report TR/IRIDIA/90-2, IRIDIA, Bruxelles, 1990.

[139] Ph. Smets. Belief functions: the disjonctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[140] Ph. Smets. What is Dempster-Shafer's model? In *Advances in the Dempster-Shafer theory of evidence*, pages 5–34. Wiley, 1994.

[141] Ph. Smets. The alpha-junctions: combination operators applicable to belief functions. In D.M Gabbay, R. Kruse, A. Nonnengart, and H.J. Ohlbach, editors, *Qualitative and quantitative practical reasoning*, pages 131–153. Springer, 1997.

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

[143] T. Söderström and P. Stoica. *System identification.* Prentice Hall, 1989.

[144] S. Soltani. *Application de la théorie des ondelettes en reconaissance des formes.* PhD thesis, Université de Compiègne, 1998.

[145] C. J. Stone. Optimal global rate of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

[146] T. Strat. Continuous belief functions for evidential reasoning. In *Proceedings of the 4th National Conference on Artificial Intelligence*, pages 308–313, Austin, Texas, 1984. American Association for Artificial Intelligence.

[147] T.M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4:391–418, 1994.

[148] M. Sugeno. *Theory of fuzzy integrals and its applications.* PhD thesis, Tokyo Institute of Technology, Japan, 1974.

[149] M. Sugeno and T. Yasukawa. A fuzzy logic based approach to qualitative modelling. *IEEE Transactions on Fuzzy Systems*, 1:7–31, 1993.

[150] T Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 15:116–132, 1985.

[151] H. Tanaka. *Possibilistic data analysis for operation research.* Physica-Verlag, Heidelberg, NY, 1999.

[152] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550, 1987.

[153] B. Tessem. Approximations for efficient computation in the theory of evidence. *Artificial Intelligence*, 61:315–329, 1993.

[154] D. M. Titterington and G. M. Mill. Kernel-based density estimates from incomplete data. *Journal of the Royal Statistical Society, series B*, 45:258–266, 1983.

[155] R. M. Tong. The construction and evaluation of fuzzy models. In R. R. Yager M.M. Gupta, R. K. Ragade, editor, *Advances in Fuzzy Set Theory and Applications*, pages 559–576. North Holland, Amsterdam, 1979.

[156] B. Tonn. An algorithmic approach to combining belief functions. *International Journal of Intelligent Systems*, 11:463–476, 1996.

[157] V. Tresp, S. Ahmad, and R. Neuneier. Training neural networks with deficient data. In Alspector Cowan, Tesauro, editor, *Advances in Neural Information Processing Systems 7*, pages 128–135. Morgan Kaufmann, 1994.

[158] P. Vamplew, D. Clark, A. Adams, and J. Muench. Techniques for dealing with missing data in feedforward networks. In *Proceedings of the 7th Australian Conference on Neural Networks*, pages 250–254, Canberra, 1996.

[159] V. N. Vapnik. *Estimation of dependences based on empirical data.* Springer-Verlag, New York, 1982.

[160] V. N. Vapnik. *The nature of statistical theory.* Springer-Verlag, New York, 1995.

[161] F. Voorbraak. A computationally efficient approximation of Dempster-Shafer theory. *International Journal of Machine Studies*, 30:525–536, 1989.

[162] G. Wahba. *Spline models for observational data.* SIAM, Philadelphia, 1990.

[163] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman and Hall, London, 1991.

[164] L. X Wang and J. M Mendel. Fuzzy basis functions, universal approximation and orthogonal least square learning. *IEEE Transactions on Neural Networks*, 3:807–814, 1992.

[165] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, 58:236–244, 1963.

[166] G. S. Watson. Smooth regression analysis. *Sankhya, Series A*, 26:359–372, 1964.

[167] R. R. Yager. Generalized probabilities of fuzzy events from fuzzy belief structures. *Information Sciences*, 28:45–62, 1982.

[168] R. R. Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18:183–190, 1988.

[169] R. R. Yager. Decision making under Dempster-Shafer uncertainties. Technical Report MII-915, Iona College, 1990.

[170] R. R. Yager, K. J. Engemann, and H. E. Miller. Fuzzy information engineering: a guided tour of applications. In *Risk management with imprecise information*, pages 531–542. Wiley, 1997.

[171] R. R. Yager, M. Fedrizzi, and J. Kacprzyk. *Advances in the Dempster-Shafer theory of evidence*. John Wiley and sons, 1994.

[172] R. R. Yager and D. P. Filev. *Essential of fuzzy modeling and control*. John Wiley and sons, 1994.

[173] R. R. Yager and D. P. Filev. Including probabilistic uncertainty in fuzzy logic controller modeling using Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(8):1221–1230, 1995.

[174] R.R. Yager. Entropy and specificity in a mathematical theory of evidence. *International Journal of General Systems*, 9(4):249–260, 1983.

[175] J. Yen. Generalizing the Dempster-Shafer theory to fuzzy sets. *IEEE Transactions on Systems, Man and Cybernetics*, 20(3):559–569, 1990.

[176] L. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[177] L. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man and Cybernetics*, 3:28–44, 1973.

[178] L. Zadeh. A theory of approximate reasoning. In J. Hayes, D. Michie, and L. I. Mikulich, editors, *Machine intelligence*, pages 149–194. Halstead Press, New York, 1979.

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

[180] L.M. Zouhal. *Contribution à l'application de la théorie des fonctions de croyance en reconnaissance des formes*. PhD thesis, Université de Compiègne, 1997.

[181] L.M Zouhal and T. Denœux. Reconnaissance des formes floues par la théorie de Dempster et Shafer. In *Actes des Rencontres francophones sur la logique floue et ses applications*, pages 3–8, Nancy, 1996.

[182] L.M Zouhal and T. Denœux. An evidence theoretic k-nn rule with parameter optimisation. *IEEE Transactions on Systems, Man and Cybernetics - Part C*, 28:263–271, 1998.

[183] R. Zwick, E. Carlstein, and D. V. Budescu. Measures of similarity among fuzzy concepts: a comparative analysis. *International Journal of Approximate Reasoning*, 1:221–242, 1987.